$$F = \frac{S_1^2}{S_2^2} = \frac{13 \cdot 5}{11 \cdot 3} = 1 \cdot 195$$

For $v_1 = 8 - 1 = 7$ and $v_2 = 10 - 1 = 9$, $F_{.05} = 3 \cdot 29$

The calculated value of F is less than the table value. Hence, we accept the null hypothesis and conclude that the difference in the varainces of two samples is not significant at 5% level.

Example 22. **Two random samples drawn from normal populations are :**

Sample I :	20	16	26	27	23	22	18	24	25	19		
Sample II :	27	33	42	35	32	34	38	28	41	43	30	37

Obtain estimates of the variances of the two populations and test whether two populations have the same variances.

Solution. Let us take the hypothesis that two populations have the same variance *i.e.*, $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_0^2$

Sample I X_1	$(X_1 - \bar{X}_1)$ $\bar{X}_1 = 22$	$(X_1 - \bar{X}_1)^2$	Sample II X_2	$(X_2 - \bar{X}_2)$ $\bar{X}_2 = 35$	$(X_2 - \bar{X}_2)^2$
20	− 2	4	27	− 8	64
16	− 6	36	33	− 2	4
26	+ 4	16	42	+ 7	49
27	+ 5	25	35	0	0
23	+ 1	1	32	− 3	9
22	0	0	34	− 1	1
18	− 4	16	38	+ 3	9
24	+ 2	4	28	− 7	49
25	+ 3	9	41	+ 6	36
19	− 3	9	43	+ 8	64
			30	− 5	25
			37	+ 2	4
$\Sigma X_1 = 220$ $n_1 = 10$		$\Sigma (X_1 - \bar{X}_1)^2 = 120$	$\Sigma X_2 = 420$ $n_2 = 12$		$\Sigma (X_2 - \bar{X}_2)^2 = 314$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{220}{10} = 22; \qquad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{420}{12} = 35$$

$$S_1^2 = \frac{\Sigma (X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{120}{10 - 1} = \frac{120}{9} = 13 \cdot 33$$

$$S_2^2 = \frac{\Sigma (X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{314}{12 - 1} = \frac{314}{11} = 28 \cdot 545$$

Applying F–test, we have

$$F = \frac{S_2^2}{S_1^2} = \frac{28 \cdot 545}{13 \cdot 33} = 2 \cdot 14 \qquad \text{where } S_2^2 > S_1^2$$

For $v_1 = 11$ and $v_2 = 9$, $F_{.05} = 3 \cdot 11$

Since, the calculated value of F is less than the table value, the null hypothesis is accepted and hence it may be concluded that the two populations have the same variance.

Example 23. **The following data relate to a random sample of Government employees in two states of Indian Union :**

	State I	State II
Sample size :	16	25
Mean monthly income (Rs.)	440	460
Sample variance	40	42

In the light of the data, test the hypothesis that the varainces of two populations are equal.

Solution. Let us take the null hypothesis that the variances of the two populations are equal *i.e.*, $H_0 : \sigma_1^2 = \sigma_2^2$

We are given : $n_1 = 16 \quad s_1^2 = 40$

$n_2 = 25 \quad s_2^2 = 42$

$$S_1^2 = \frac{n_1}{n_1 - 1} \cdot s_1^2 = \frac{16}{16 - 1} \times 40 = \frac{16}{15} \times 40 = \frac{640}{15} = 42 \cdot 67$$

$$S_2^2 = \frac{n_2}{n_2 - 1} \cdot s_2^2 = \frac{25}{25 - 1} \times 42 = \frac{25}{24} \times 42 = \frac{1050}{24} = 43 \cdot 75$$

$$F = \frac{43 \cdot 75}{42 \cdot 67} = 1 \cdot 025 \qquad \text{where, } S_2^2 > S_1^2$$

For $v_1 = 24$, and $v_2 = 15$, $F_{.05} = 2 \cdot 29$

Since, the calculated value of F is less than the table value of F, we accept the null hypothesis and hence it may be concluded that the variances of two populations are equal.

Example 24. **Two independent samples of 8 and 7 items respectively had the following values of variable (weight in grams) :**

Sample I :	9	11	13	11	15	9	12	14
Sample II :	10	12	10	14	9	8	10	

Do the two estimates of population variance differ significantly ?

Solution. Let us take the null hypothesis that the two populations have the same variance *i.e.*, $H_0 : \sigma_1^2 = \sigma_2^2$.

Sample I		Sample II	
X_1	X_1^2	X_2	X_2^2
9	81	10	100
11	121	12	144
13	169	10	100
11	121	14	196
15	225	9	81
9	81	8	64
12	144	10	100
14	196		
$\Sigma X_1 = 94$, $n_1 = 8$	$\Sigma X_1^2 = 1138$	$\Sigma X_2 = 73$, $n_2 = 7$	$\Sigma X_2^2 = 785$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{94}{8} = 11{\cdot}75; \qquad \bar{X}_2 = \frac{X_2}{n_2} = \frac{73}{7} = 10{\cdot}43$$

Since, the actual means are in fractions, we make use of original values. Thus,

$$S_1^2 = \frac{1}{n_1 - 1}\left[\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1}\right] = \frac{1}{8-1}\left[1138 - \frac{(94)^2}{8}\right] = \frac{33{\cdot}5}{7} = 4{\cdot}78$$

$$S_2^2 = \frac{1}{n_2 - 1}\left[\Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2}\right] = \frac{1}{7-1}\left[785 - \frac{(73)^2}{7}\right] = \frac{23{\cdot}7}{6} = 3{\cdot}95$$

Applying F–test, we have :

$$F = \frac{4{\cdot}78}{3{\cdot}95} = 1{\cdot}21$$

For $v_1 = 7$ and $v_2 = 6$, $F_{05} = 4{\cdot}21$

Since, the calculated value of F is less than the table value of F, we accept the null hypothesis and it may be concluded that the two estimates of population variances do not differ significantly.

AN IMPORTANT TYPICAL EXAMPLE

Example 25. **Can the following two samples be regarded as coming from the same normal population ?**

Sample	Size	Sample mean	Sum of squares of deviation from mean
1	10	12	120
2	12	15	314

Solution. To test if two independent samples have been drawn from the same normal population, we have to test (i) the equality of population means, and (ii) the equality of population variances.

Equality of means will be tested by applying t–test and equality of variance will be tested by F–test. Since, t–test assumes $\sigma_1^2 = \sigma_2^2$, we first apply F–test and then t–test.

***F*–test :** Set up $H_0 : \sigma_1^2 = \sigma_2^2$

We are given : $n_1 = 10$, $\Sigma(X_1 - \bar{X}_1)^2 = 314$

$$S_1^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{120}{10-1} = \frac{120}{9} = 13{\cdot}33$$

$$S_2^2 = \frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{314}{12-1} = \frac{314}{11} = 28{\cdot}55$$

Applying F–test,

$$F = \frac{S_2^2}{S_1^2} = \frac{28{\cdot}55}{13{\cdot}33} = 2{\cdot}14$$

For $v_1 = 11 - 1 = 10$ and $v_2 = 10 - 1 = 9$, $F_{05} = 3{\cdot}14$

The calculated values of F is less than the table value. Hence, we accept the null hypothesis and conclude that the difference in the varainces of two samples is not significant at 5% level.

Since, $\sigma_1^2 = \sigma_2^2$, we can now apply t–test for testing $H_0 : \mu_1 = \mu_2$

***t*–test :** Set up $H_0 : \mu_1 = \mu_2$

We are given : $n_1 = 10$ $\quad \bar{X}_1 = 12$ $\quad \Sigma(X_1 - \bar{X}_1)^2 = 120$

$n_2 = 12$ $\quad \bar{X}_2 = 15$ $\quad \Sigma(X_2 - \bar{X}_2)^2 = 314$

$$S = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}} = \sqrt{\frac{120 + 314}{10 + 12 - 2}} = \sqrt{\frac{434}{20}} = \sqrt{21{\cdot}7} = 4{\cdot}65$$

Aplying t–test, we have

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{12 - 15}{4{\cdot}65} \times \sqrt{\frac{10 \times 12}{10 + 12}} = 1{\cdot}506$$

Degrees of freedom $(v) = n_1 + n_2 - 2 = 10 + 12 - 2 = 20$

Table value of t for 20 d.f. at 5% level of signficance = 2·086.

Since, the calculated value of t is less the table value, we accept the null hypothesis and conclude that the difference in means is not significant.

Hence, we may regard that the given samples to have been drawn from same population.

EXERCISE – 6

1. Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 and 91 inches squares. Calculate the value of F and say whether it is significant or not at 5% level of significance ? (Given F_{05} for 8 and 7 d. f = 3·73)

OR

Two samples of sizes 9 and 8 give the sum of squares of deviations from their respective means equal to 160 and 91 inches squares. Can they be regarded as drawn from the same normal population at $\alpha = \cdot 05$? [Ans. $F = 1 \cdot 54$, the samples can be regarded as drawn from the same normal population]

2. Two random samples drawn from normal populations are :

Sample I :	66	67	75	76	82	84	88	90	92		
Sample II :	64	66	74	78	82	85	87	92	93	95	97

Obtain estimates of the variances of the two populations and test whether the two populations have the same variances. (Given $F = 3 \cdot 35$ at 5% level for $v_1 = 10$ and $v_2 = 8$)

[Ans. $F = 1 \cdot 414$, the two populations have the same variance)

3. For a random sample of 10 pigs fed on diet A, the increase in weight in pounds in certain periods were :

10, 6, 16, 17, 13, 12, 8, 14, 15, 9

For another random sample of 12 pigs fed on diet B, the increase in weight in the same period were :

7, 13, 22, 15, 12, 14, 18, 8, 21, 23, 10, 17

Test whether both the samples come from population having same variance.
(Given : F_{05} for $v_2 = 11$, $v_2 = 9$ is 3·112).

[Ans. $F = 2 \cdot 14$, samples come from population having same variance]

4. It is known that the mean diameters of rivets produced by two firms A and B are practically the same but standard deviations differ. For 22 rivets produced by firm A, the standard deviation is 2·9 mm, while for 16 rivets manufactured by firm B, the standard deviation is 3·8 mm. Compute the statistic you would use to test whether the product of firm A has the same variability as those of firm B.

[Ans. $F = 1 \cdot 748$, the two populations have the same variance]

5. In a test given to two groups of students drawn from two normal populations, the marks obtained were as follows :

Group A :	18	20	36	50	49	36	34	39	41
Group B :	29	28	26	35	30	44	46		

Examine at 5% level, whether the two populations have the same variance.

[Ans. $F = 2 \cdot 103$, populations have the same variances]

6. Two sets of random samples drawn from normal population are given below. Obtain the estimates of the variances of the two populations and test whether the two populations have the same variance. Use F–test.

Sample I :	20	16	26	27	23	22	18	24	25	19	30	37
Sample II :	27	33	42	35	32	34	38	28	41	43		

(Table value of F for $v_1 = 11$ and $v_2 = 9$ at 5% level = 3·112)

[Ans. $F = 2 \cdot 142$, population have the same variance]

7. The variability in the tensile strength of two types of steel wire is to compared. Given a sample of 10 observations of type A wire yielding a variance of 100·4 and a sample of 12 observations of type B wire yielding a variance of 115·5, test the hypothesis that the two populations have equal variances.

[Ans. $F = 1 \cdot 0625$, population have the same variance]

8. Two samples were drawn independently from two normal populations. The summary statistics are :

$n_1 = 8$, $\Sigma (X_1 - \bar{X}_1)^2 = 84 \cdot 4$ inches

$n_2 = 13$, $\Sigma (X_2 - \bar{X}_2)^2 = 102 \cdot 6$ inches

In the light of the data, test whether the two variances differ significantly.

[Ans. $F = 1 \cdot 410$, variances do not differ]

9. Can the following two samples be regarded as coming from the same normal population ?

Sample	Size	Sample mean	Sum of squares of deviation from mean
1	10	15	90
2	12	14	108

[**Hints** : Use F–test; $F = \frac{S_1^2}{S_2^2} = 1.019$ and t–test for $H_0 : \mu_1 = \mu_2$, $|t| = 0 \cdot 742$]

[Ans. The samples are drawn from the same normal population]

10. The means of two random samples of size 9 and 7 are 196·42 and 198·82 respectively. The sum of the squares of the deviations from the mean are 26·94 and 18·73 respectively. Can the samples by considered to have been drawn from the same normal population ?

[Ans. $F = 1 \cdot 078$, $|t| = 2 \cdot 634$, Reject H_0]

11. The following data relate to a random sample of Government employees in two states of Indian Union. First carry out a test of hypothesis that the variance of the two populations are equal. In the light of the result of the above test, carry out a test of hypothesis that the means of two populations are equal :

	State I	State II
Sample Size	16	25
Mean monthly income of sample employees (in days.)	440	460
Sample Variance	40	42

[Ans. $F = 1 \cdot 025$, Accept H_0 , $t = 9 \cdot 72$, Accept H_0]

MISCELLANEOUS SOLVED EXAMPLES

Example 26. **Prices of shares of a company on different days in a month were found to be :**

66, 65, 69, 70, 69, 71, 70, 63, 64 and 68

Discuss whether the mean price of the shares should be 65.

(The table value of t for 9 degree of freedom at 5% level is 2·262)

Solution. Let us take the hypothesis that the mean price of the share is 65, *i.e.*,

$H_0 : \mu = 65$ and $H_1 : \mu \neq 65$ (⇒ Two tailed test)

X	$A = 67$ $d = X - A$	d^2
66	−1	1
65	−2	4
69	+2	4

70	+ 3	9
69	+ 2	4
71	+ 4	16
70	+ 3	9
63	– 4	16
64	– 3	9
68	+ 1	1
$n = 10, \Sigma X = 675$	$\Sigma d = 5$	$\Sigma d^2 = 73$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{675}{10} \quad 67 \cdot 5$$

Since, the actual means of X is in fraction, we should take deviations from assumed mean to simplify the calculations.

$$\bar{d} = \frac{\Sigma d}{n} = \frac{5}{10} = 0 \cdot 5$$

$$S = \sqrt{\frac{\Sigma d^2 - n(\bar{d})^2}{n-1}} = \sqrt{\frac{73 - 10(0 \cdot 5)^2}{9}} = 2 \cdot 799$$

Applying t–test,

$$t = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n} = \frac{67 \cdot 5 - 65}{2 \cdot 799} \times \sqrt{10} = \frac{2 \cdot 5 \times 3 \cdot 162}{2 \cdot 799} = 2 \cdot 82$$

Degrees of freedom $(v) = n - 1 = 10 - 1 = 9$

For $v = 9$, $t_{0 \cdot 05}$ for two tailed test $= 2 \cdot 262$

Since, the calculated value of t is greater than the table vlaue, we reject the null hypothesis and therefore, conclude that mean price of the shares could not be equal to Rs. 65.

Example 27. **To compare the price of a certain commodity in two towns, ten shops were selected at random in each town. The following figures give the price found.**

Town A :	61	62	56	63	56	63	59	56	44	60
Town B :	55	54	47	59	51	61	57	54	64	58

Test whether the average price can be said to be the same in two towns.

Solution. Let us take the hypothesis that there is no difference in the average price of two towns : *i.e.*, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

X_1	$X_1 - \bar{X}_2$	$(X_1 - \bar{X}_1)^2$	X_2	$X^2 - \bar{X}_2$	$(X_2 - \bar{X}_2)^2$
61	3	9	55	– 1	1
62	4	16	54	– 2	4
56	– 2	4	47	– 9	81
63	5	25	59	3	9
56	– 2	4	51	– 5	25
63	5	25	61	5	25

59	1	1	57	1	1
56	– 2	4	54	– 2	4
44	– 14	196	64	8	64
60	2	4	58	2	2
$\Sigma X_1 = 580$ $n_1 = 10$	0	$\Sigma (X_1 - \bar{X}_1)^2 = 288$	$\Sigma X_2 = 560$ $n_2 = 10$	0	$\Sigma (X_2 - \bar{X}_2)^2 = 216$

$$\bar{X}_1 = \frac{\Sigma X_1}{n_1} = \frac{580}{10} = 58, \qquad \bar{X}_2 = \frac{\Sigma X_2}{n_2} = \frac{560}{10} = 56$$

$$S = \sqrt{\frac{\Sigma (X_1 - \bar{X}_1)^2 + \Sigma (X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{288 + 216}{10 + 10 - 2}} = \sqrt{\frac{504}{18}} = \sqrt{28} = 5 \cdot 29$$

Applying t–test,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{58 - 56}{5 \cdot 29} \sqrt{\frac{10 \times 10}{10 + 10}} = \frac{2 \times 2 \cdot 236}{5 \cdot 29} = 0 \cdot 845$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 10 + 10 - 2 = 18$

For $v = 18$, $t_{0 \cdot 05} = 2 \cdot 101$

Since, the calculated value of t is less than the table value, we accept the null hypothesis and conclude that there is no significant difference in the mean price.

Example 28. **An I.Q. test was administrated to 5 officers before and after they were trained. The results are given below :**

Candidates :	I	II	III	IV	V
I.Q. before training :	110	120	123	132	125
I.Q. after training :	120	118	125	136	121

Test whether there is any change in I.Q. after the training programme. [For $v = 4$, $t_{0 \cdot 01} = 4 \cdot 6$]

Solution. Let us take the hypothesis is that there is no change in I.Q. after the training programme. *i.e.*, $H_0 : \bar{d} = 0$ or $\mu_2 - \mu_1 = 0$ and $H_1 : \bar{d} > 0$ or $\mu_2 - \mu_1 > 0$

($\Rightarrow$ One tailed test)

I.Q. Before	I.Q. After II	d (II – I)	d^2
110	120	+ 10	100
120	118	– 2	4
123	125	+ 2	4
132	136	+ 4	16
125	121	– 4	16
$n = 5$		$\Sigma d = 10$	$\Sigma d^2 = 140$

$$\bar{d}=\frac{\Sigma d}{n}=\frac{10}{5}=2$$

$$S=\sqrt{\frac{\Sigma d^2-n(\bar{d})^2}{n-1}}$$

$$=\sqrt{\frac{140-5(2)^2}{5-1}}=\sqrt{\frac{140-20}{4}}=\sqrt{\frac{120}{4}}=\sqrt{30}=5\cdot477$$

Applying t–test,

$$t=\frac{\bar{d}}{S}\sqrt{n}=\frac{2\sqrt{5}}{5\cdot477}=0\cdot817$$

For $v=4$, $t_{0\cdot01}=4\cdot6$

The calculated value of t is less than the table value. We accept the null hypothesis and hence there is no change in I.Q. after the raining programme.

Example 29. **Two types of batteries are tested for their length of life and the following data are obtained :**

	No. of samples	Mean life in hours	Variance
Type A :	9	600	121
Type B :	8	640	144

Is there a significant difference in two mean ? Value of t for 15 degrees of freedom at 5% level is 2·131.

Solution. Let us take the hypothesis that there is no significant difference in mean life of two types of batteries *i.e.*, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test)

Given : $n_1 = 9$, $\bar{X}_1 = 600$, $s_1^2 = 121$

$n_2 = 8$, $\bar{X}_2 = 640$, $s_2^2 = 144$

$$S=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

$$=\sqrt{\frac{(9-1)\times121+(8-1)\times144}{9+8-2}}=\sqrt{\frac{968+1008}{15}}$$

$$=\sqrt{\frac{1976}{15}}=\sqrt{131\cdot73}=11\cdot47$$

Applying t–test,

$$t=\frac{|\bar{X}_1-\bar{X}_2|}{S}\cdot\sqrt{\frac{n_1n_2}{n_1+n_2}}$$

$$=\frac{|600-640|}{11\cdot47}\times\sqrt{\frac{9\times8}{9+8}}$$

$$=\frac{40}{11\cdot47}\times2\cdot057=\frac{82\cdot28}{11\cdot47}=7\cdot17$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 9+8-2 = 15$

For $v = 15$, $t_{0\cdot05} = 2\cdot131$

Since, the calculated value of t is greater than the table value, we reject H_0 and hence, the difference in the means is significant.

Example 30. **Two types of drugs were used on 5 and 7 patients for reducing their weights. Drug A was imported and drug B indigenous. The decreases in the weight after using drugs for six months are as follows :**

Drug A :	10	12	13	11	14		
Drug B :	8	9	12	14	15	10	9

Is there a significant difference in the efficacy of the two drugs ?
If not, which drug should you buy.
(For $v = 10$, $t_{0\cdot05} = 2\cdot223$)

Solution : Let us take the hypothesis that there is no significant difference in the efficacy of two drugs, *i.e.*, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$ ($\Rightarrow$ Two tailed test).

X_1	$\bar{X}_1 = 12$	$(X_1 - \bar{X}_1)^2$	X_2	$\bar{X}_2 - 11$	$(X_2 - \bar{X}_2)^2$
10	− 2	4	8	− 3	9
12	0	0	9	− 2	4
13	+ 1	1	12	+ 1	1
11	− 1	1	14	+ 3	9
14	+ 2	4	15	+ 4	16
			10	− 1	1
			9	+ 2	4
$\Sigma X_1 = 60$ $n_1 = 5$		$\Sigma (X_1 - \bar{X}_1)^2 = 10$	$\Sigma X_2 = 77$ $n_2 = 7$		$\Sigma(X_2 - \bar{X}_2)^2 = 44$

$$\bar{X}_1=\frac{\Sigma X_1}{n_1}=\frac{60}{5}=12, \qquad \bar{X}_2=\frac{\Sigma X_2}{n_2}=\frac{77}{7}=11$$

$$S=\sqrt{\frac{\Sigma(X_1-\bar{X}_1)^2+\Sigma(X_2-\bar{X}_2)^2}{n_1+n_2-2}}=\sqrt{\frac{10+44}{5+7-2}}=\sqrt{\frac{54}{10}}=\sqrt{5\cdot4}=2\cdot324$$

Applying t–test,

$$t=\frac{\bar{X}_1-\bar{X}_2}{S}\cdot\sqrt{\frac{n_1n_2}{n_1+n_2}}$$

$$=\frac{12-11}{2\cdot324}\cdot\sqrt{\frac{5\times7}{5+7}}$$

$$=\frac{1\times1\cdot708}{2\cdot324}=\frac{1\cdot708}{2\cdot324}=0\cdot735$$

Degrees of freedom $(v) = n_1 + n_2 - 2 = 5+7-2 = 10$

For $v = 10$, $t_{0\cdot05} = 2\cdot228$

Since, the calculated value of t is less than the table vlaue, we accept H_0 and conclude that there is no significant difference in the efficacy of two drugs. Since,

drug B is indigenous and there is no difference in the efficacy of imported and indigenous drug, we should buy indigenous drug B.

Example 31. **Below are given the gain in weights (lbs) of cows fed on two diets X and Y:**

Gain Weight (lbs)

Diet X:	25	32	30	32	24	14	32			
Diet Y:	24	34	22	30	42	31	40	30	32	35

Test at 5% level, whether the two diets differ as regards their effect on mean increase in weight (Table value of t for 15 degrees of freedom at 5% = 2·131).

Solution. Let us take the null hypothesis that diet X and Y do not differ significantly with regard to their effect on increase in weight, *i.e.*, $H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$
(⇒ Two tailed test)

X	$\bar{X} = 27$ $X - \bar{X}$	$(X - \bar{X})^2$	Y	$\bar{Y} = 32$ $Y - \bar{Y}$	$(Y - \bar{Y})^2$
25	− 2	4	24	− 8	64
32	+ 5	25	34	+ 2	4
30	+ 3	9	22	− 10	100
32	+ 5	25	30	− 2	4
24	− 3	9	42	+ 10	100
14	− 13	169	31	− 1	1
32	+ 5	25	40	+ 8	64
			30	− 2	4
			32	0	0
			35	+ 3	9
$\Sigma X = 189$ $n_1 = 7$	$\Sigma (X - \bar{X}) = 0$	$\Sigma (X - \bar{X})^2$ $= 266$	$\Sigma Y = 320$ $n_2 = 10$	$\Sigma (Y - \bar{Y}) = 0$	$\Sigma (Y - \bar{Y})^2$ $= 350$

$$\bar{X} = \frac{\Sigma X}{n_1} = \frac{189}{7} = 27, \qquad \bar{Y} = \frac{\Sigma Y}{n_2} = \frac{320}{10} = 32$$

$$S = \sqrt{\frac{\Sigma (X - \bar{X})^2 + \Sigma (Y - \bar{Y})^2}{n_1 + n_2 - 2}} = \sqrt{\frac{266 + 350}{7 + 10 - 2}} = \sqrt{\frac{616}{15}} = \sqrt{41 \cdot 066} = 6 \cdot 40$$

Applying t–test,

$$|t| = \frac{\bar{X} - \bar{Y}}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{|27 - 32|}{6 \cdot 40} \times \sqrt{\frac{7 \times 10}{7 + 10}}$$

$$= \frac{5}{6 \cdot 40} \times 2 \cdot 029 = \frac{10 \cdot 1459}{6 \cdot 40} = 1 \cdot 58$$

Degrees of freedom $= v = n_1 + n_2 - 2 = 7 + 10 - 2 = 15$
For $v = 15$, $t_{0 \cdot 05}$ for two tailed test $= 2 \cdot 131$
Since, the calculated value of t is less than the table value, we accept the null hypothesis and conclude that diets X and Y do not differ significantly as regards their effects on increase in weight is concerned.

Example 32. **A random sample of 18 pairs from a bivariate normal population showed a correlation coefficient of 0·4. Is this value significant of a correlation in the population ?**

Solution. Let us take the hypothesis that the variables are uncorrelated in the population.

$H_0 : \rho = 0$ and $H_1 = \rho \neq 0$ (⇒ two tailed test)

Applying t–test,

$$t = \frac{r \times \sqrt{n-2}}{\sqrt{1 - r^2}} = \frac{0 \cdot 4 \times \sqrt{18 - 2}}{\sqrt{1 - 0 \cdot 16}}$$

$$= \frac{0 \cdot 4 \times 4}{\sqrt{0 \cdot 84}} = \frac{1 \cdot 6}{0 \cdot 91} = 1 \cdot 76$$

Degrees of freedom $= v = n - 2 = 18 - 2 = 16$
For $v = 16$, $t_{0 \cdot 05}$ for two tailed test $= 2 \cdot 12$
Since, the claculated value of t is less than the table value, we accept H_0 and hence, the given value of r is not significant.

Example 33. **A correlation coefficient of 0·63 is obtained from a sample of 20 paired of observations. Is it significantly different from 0·5?**

Solution. We are given : $n = 20$, $r = 0 \cdot 72$, $\rho = 0 \cdot 8$
Let us take the null hypothesis that the correlation in the population is 0·5 *i.e.*,

$H_0 : \rho = 0 \cdot 5$ and $H_1 : \rho \neq 0 \cdot 5$ (⇒ two tailed test)

Z—transformation or r

$$Z_1 = 1 \cdot 1513 \log_{10} \frac{1 + r}{1 - r}$$

$$= 1 \cdot 1513 \log_{10} \frac{1 + \cdot 63}{1 - \cdot 63}$$

$$= 1 \cdot 1513 \log 4 \cdot 4$$

$$= 1 \cdot 1513 \times \cdot 6435 = 0 \cdot 741$$

Z—transformation of ρ

$$Z_2 = 1 \cdot 1513 \log_{10} \frac{1 + r}{1 - r}$$

$$= 1 \cdot 1513 \log_{10} \frac{1 + \cdot 5}{1 - \cdot 5}$$

$$= 1 \cdot 1513 \log 3$$

$$= 1 \cdot 1513 \times \cdot 4771 = 0 \cdot 549$$

$$SE_{Z_1 - Z_2} = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{20-3}} = \frac{1}{\sqrt{17}} = 0 \cdot 243$$

Applying Fisher's Z—test

$$|Z| = \frac{Z_r - Z_p}{SE_{Z_1 - Z_2}} = \frac{0 \cdot 741 - 0 \cdot 549}{0 \cdot 243} = \frac{0 \cdot 192}{0 \cdot 243} = 0 \cdot 79.$$

The critical value of Z at 5% for two tailed test = 1·96.
Since, the calculated value of $|Z|$ is less than 1·96 (5%), we accept null hypothesis and conclude that the given correlation coefficient is not significantly different from 0·5.

Example 34. **In a laboratory experiment, two samples gave the following results :**

Sample	Size	Sample Mean	Sum of squares of deviations from the mean
1	10	15	90
2	12	14	108

Test the equality of sample variances at 5% level of significance.

Solution. Let the null hypothesis be that the two population variances are equal *i.e.*,

$$H_0 : \sigma_1^2 = \sigma_2^2$$

We are given :

$n_1 = 10$, $\Sigma(X_1 - \bar{X}_1)^2 = 90$

$n_2 = 12$, $\Sigma(X_2 - \bar{X}_2)^2 = 108$

$$S_1^2 = \frac{\Sigma(X_1 - \bar{X}_1)^2}{n_1 - 1} = \frac{90}{10-1} = \frac{90}{9} = 10$$

$$S_2^2 = \frac{\Sigma(X_2 - \bar{X}_2)^2}{n_2 - 1} = \frac{108}{12-1} = \frac{108}{11} = 9 \cdot 82$$

Applying F–test,

$$F = \frac{S_1^2}{S_2^2} \qquad \text{where } S_1^2 > S_2^2$$

$$= \frac{10}{9 \cdot 82} = 1 \cdot 018$$

For $v_1 = 10 - 1 = 9$ and $v_2 = 12 - 1 = 11$, $F_{.05} = 2 \cdot 90$

Since, the calculated value of F is less than the table value, we accept the null hypothesis and conclude that the two populations have the same variance.

Example 35. **The profit of an automobile dealer varies from day to day. However, the dealer believes that the per day profit averages at least Rs. 3500. The profits per day during the past week were reported to be Rs. 2000, Rs. 3000, Rs. 5200, Rs. 3400, Rs. 2500 and 3700. Would you agree with the belief of the dealer. Use at 0·05 level of significance ?**

Solution. Let us take the hypothesis that the average profit of the dealer is at least Rs. 3500 *i.e.*, $H_0 : \mu \geq 3500$ and $H_1 : \mu < 3500$

[Since, the dealer belief would be false if the average rate is less than 3500]

It is one tailed test.

Sales X	$\bar{X} = 3300$ $(X - \bar{X})$	$d = (X - \bar{X})/100$	d^2
2000	– 1300	– 13	169
3000	– 300	– 3	9
5200	+ 1900	+ 19	361
3400	+ 1000	+ 1	1
2500	– 800	– 8	64
3700	+ 400	+ 4	16
$\Sigma X = 19800$		$\Sigma d = 0$	$\Sigma d^2 = 620$
$n = 6$			

$$\bar{X} = \frac{\Sigma X}{n} = \frac{19800}{6} = 3300$$

$$S = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}} \times c \qquad \text{[here, } (c = 100)\text{]}$$

$$= \sqrt{\frac{620}{5}} \times 100 = \sqrt{124} \times 100 = 11 \cdot 1355 \times 100 = 1113 \cdot 55$$

Applying t–test,

$$|t| = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

$$= \frac{|3300 - 3500|}{1113 \cdot 35} \sqrt{6} = \frac{200}{1113 \cdot 35} \times 2 \cdot 449 = \frac{489 \cdot 897}{1113 \cdot 35} = 0 \cdot 44$$

Degrees of freedom $= v = n - 1 = 6 - 1 = 5$

For $v = 5$, $t_{0 \cdot 05}$ for one tailed test $= 2 \cdot 015$

Since, the calculated value of t is less than the table value, we accept the null hypothesis and conclude that the claim of the dealer is justified.

IMPORTANT FORMULAE

t-test

1. Tests of Hypothesis about population mean :

$$t = \frac{\bar{X} - \mu}{S} \cdot \sqrt{n}$$

where, $S = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n-1}}$ or $\sqrt{\frac{\Sigma d^2 - (\bar{d})^2 \times n}{n-1}}$

d.f. $= n - 1$

2. Test of Hypothesis about the difference between two population means in case of independent samples :

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S} \cdot \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

Where,

$$S = \sqrt{\frac{\Sigma(X_1 - \bar{X}_1)^2 + \Sigma(X_2 - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

$$= \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

d.f. $= n_1 + n_2 - 2$

3. **Test of Hypothesis about the difference of the two population means with dependent samples**

$$t = \frac{\bar{d}}{S}\cdot\sqrt{n}$$

where,

$$S = \sqrt{\frac{\Sigma d^2 - (\bar{d})^2 \times n}{n-1}}$$

d.f. $= v = n-1$

4. **Test of Hypothesis about correlation coefficient :**

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

where,

d.f. $= v = n-2$

Fisher's Z-test

5. **Test of Hypothesis about correlation coefficient ($H_0 : \rho = \rho_0$) :**

$$|Z| = \frac{Z_r - Z_p}{SE_z}$$

6. **Test of Hypothesis about two correlation coefficients ($H_0 : \rho_1 = \rho_2$) :**

$$|Z| = \frac{Z_S - Z_p}{SE_{Z_1 - Z_2}}$$

F-test

8. **Test of Hypothesis about two population variances ($H_0 : \sigma_1^2 = \rho_2^2$) :**

$$F = \frac{S_1^2}{S_2^2} \qquad \text{where, } S_1^2 > S_2^2$$

QUESTIONS

1. Define student's t-test and explain some of its applications.
2. Explain how t-test is used to test the significance of the difference between the means of two samples.
3. Explain briefly various application of the t-test.
4. Explain how t-test is used to test the significance of the sample correlation coefficient in a sample drawn from a bivariate normal population.
5. Discuss Fisher's Z-test for testing the significance of correlation coefficient.
6. Discuss the F-test for testing the equality of two sample variances.
7. Discuss the usefulness of F-test.
8. Explain the procedure for testing hypothesis regarding equaltiy of two variances.
9. Explain how would test the significance of the correlation coefficient in case of small sample.

❀❀❀

5

Chi-Square Test

INTRODUCTION

The Chi-Square test (χ^2-test) is an important test amongst several tests of significance developed by the statisticians is. Chi-Square, symbolically written as χ^2 (Pronounced as Ki-square), is a statistical measure used in the context of sampling analysis for testing the significance of population variance. As a non-parametric test, it can be used as a test of goodness of fit and as a test of attributes. Thus, the Chi-Square test is applicable to a very large number of problems in practice which can be summed up under the following heads :

(1) **χ^2-test as a test for population variance.**

(2) **χ^2-test as a non-parametric test.**

Let us discuss them briefly

(1) **χ^2-test as a test for population variance :** χ^2-test is often used to test the significance of population variance i.e. we can use this test to judge if a random sample has been draw from a normal population with mean (μ) and with a specified variance (σ_0^2).

PROCEDURE :

(i) Set up the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$ and $H_1 : \sigma^2 > \sigma_0^2$

(ii) We compute χ^2 by using any one of the following formula :

$$\chi^2 = \frac{\Sigma(x-\bar{x})^2}{\sigma^2} \quad \text{or} \quad \frac{ns^2}{\sigma^2}$$

Where $\quad s^2 = \dfrac{\Sigma(x-\bar{x})^2}{n}$

$\Rightarrow \quad ns^2 = \Sigma(x-\bar{x})^2$

(iii) No. of degrees of freedom are worked out by using the following formula :
Degrees of freedom = v = n - 1

(iv) Obtain the table value of χ^2 with reference to the degrees of freedom for the given problem and the desired level of significance.

(v) If the calculated value of χ^2 > tabulated value of χ^2, we reject the null hypothesis H_0.
Otherwise, we accept H_0.

Note : In case H_1 is two tailed test, the procedure is slightly different. Accept H_0 if the calculated $\chi^2 < \chi^2_{\alpha/2}$ or calculated $\chi^2 < \chi^2_{1-\alpha/2}$ i.e. $\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{1-\alpha/2}$ and reject H_0 if calculated $\chi^2 > \chi^2_{\alpha/2}$ or calculated $\chi^2 < \chi^2_{1-\alpha/2}$.

Example 1 : **With variance of a normal population $\sigma^2 = 5.80$ and the sum of the squares of the deviations of 15 sample values from their mean is 150, compute the χ^2-value.**

Solution. We are given : $n = 15$, $\sigma^2 = 5.80$ and $\Sigma(X - \bar{X})^2 = 150$

The χ^2-value is calculated as :

$$\chi^2 = \frac{\Sigma(X - \bar{X})^2}{\sigma^2} = \frac{150}{5.8} = 25.86$$

Example 2 : **A sample of 10 units drawn from a normally distributed population shows a variance $s^2 = 25$. Test the hypothesis that the population variance $\sigma^2 = 36$ using $\alpha = 0.01$ level of significance.**

Solution. We are given : $n = 10$, $s^2 = 25$, $\sigma^2 = 36$, $\alpha = .01$

Let us take the null hypothesis that the population variance is 36 i.e.

$H_0 : \sigma^2 = 36$ and $H_1 : \sigma^2 > 36$ ($\Rightarrow$ Right tailed test)

The value of χ^2 is calculated as follows:

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

$$= \frac{10 \times 25}{36} = 6.94$$

Degrees of freedom $= v = n - 1 = 10 - 1 = 9$

For $v = 9$, $\chi^2_{.01} = 21.7$

Since, the calculated vlaue of χ^2 is less than the table value of $\chi^2_{0.01}$ we accept H_0 and conclude that the population variance $\sigma^2 = 36$.

Example 3 : **A random sample of size 20 from of normal population gives a sample mean of 42 and sample standard deviation of 6. Test the hypothesis that the population standard deviation is 9. Clearly state the alternative hypothesis you allow for and the level of significance adopted.**

Solution. We are given : $n = 20$, $\bar{X} = 42$, $s = 6 \Rightarrow s^2 = 36$, $\sigma = 9 \Rightarrow \sigma^2 = 81$

Let us take the null hypothesis that the population standard deviation is 6, i.e.

$H_0 : \sigma = 9$ and $H_1 : \sigma > 9$ ($\Rightarrow$ Right Tailed Test)

χ^2-value is calculated as :

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

$$= \frac{20 \times 36}{81} = 8.89$$

Degrees of freedom $= v = 20 - 1 = 19$

For $v = 19$, $\chi^2_{.05} = 30.1$

Since, the calculated vlaue of χ^2 is less than the table value of $\chi^2_{.05}$, we accept H_0 and conclude that the population standard deviation is 9.

Example 4 : **Weights in kg of 10 students are given below :**

38, 40, 45, 53, 47, 43, 55, 48, 52, 49

Can we say that variance of the distribution of weight of all students from which the above sample of 10 students was drawn, is equal to 20 kgs ? Test this at 5% and 1% level of significance. (At 9 d.f., $\chi^2_{.05} = 16.92$, $\chi^2_{.01} = 21.67$ at 10 d.f; $\chi^2_{.05} = 18.31$, $\chi^2_{.01} = 23.21$)

Solution. Let us take the null hypothesis that population variance is 20, *i.e.*

$H_0 : \sigma^2 = 20$ and $H_1 : \sigma^2 > 20$ ($\Rightarrow$ Right tailed test)

Applying χ^2- test

X	$\bar{X} = 47$ $X - \bar{X}$	$(X - \bar{X})^2$
38	-9	81
40	-7	49
45	-2	4
53	6	36
47	0	0
43	-4	16
55	8	64
48	1	1
52	5	25
49	2	4
$\Sigma X = 470$ $n = 10$		$\Sigma(X - \bar{X})^2 = 280$

$\therefore$ $\bar{X} = \frac{470}{10} = 47$

χ^2-value is calculated as :

$$\chi^2 = \frac{\Sigma(X - \bar{X})^2}{\sigma^2} = \frac{280}{20} = 14$$

Degrees freedom $= v = n - 1 = 10 - 1 = 9$

For $v=9$, $\chi^2_{.05}=16.92$

For $v=9$, $\chi^2_{.01}=21.67$

Since, the calculated value of χ^2 is less than the table value at 5% and 1% level of significance, we accept null hypothesis and conclude that the variance of the distribution of weights of all students in the population is equal to 20 kgs.

Example 5: **A random sample of 10 is drawn randomly from a certain population. The sum of the squared deviations from the mean of the given sample is 50. Test the hypothesis that the variance of the population is 5. Use $\alpha=0.05$ level of significance.**

Solution. We are given : $n=10$, $\Sigma(X-\bar{X})^2=50$, $\sigma^2=5$

Let us take the null hypothesis that the population variance is 5, *i.e.*

$H_0:\sigma^2=5$ and $H_1:\sigma^2>5$ ($\Rightarrow$ Right Tailed Test)

The χ^2- value is calculated as :

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

Putting the vlues, we have

$$\chi^2=\frac{50}{5}=10$$

Degrees of freedom $=v=n-1=10-1=9$

For $v=9$, $\chi^2_{.05}=16{\cdot}92$

Since, the calculated value of χ^2 is less than table value, we accept the null hypothesis and conclude that the variance of the population is 5.

EXERCISE – 1

1. A normal population has a standard deviation $\sigma=2.50$. A random sample of 12 values selected from this population yields sample variance $s^2=5.60$. Computer the X^2-value.
[**Ans.** $\chi^2=10{\cdot}75$]

2. A sample of size $n=17$ drawn from a normally distributed population shows a variance $s^2=25$. Test the hypothesis that the population variance $\sigma^2=35$ against the alternative $\sigma^2>35$ using $\alpha=0.05$ level of significance. [**Ans.** $\chi^2=12{\cdot}14$, Accept H_0]

3. A random sample of size 10 from a normal population gives the following values :
65, 72 68, 74, 77, 61, 63, 69, 73, 71
Test the hypothesis that the population variance is 32 [**Ans.** $\chi^2=7\cdot316$, H_0 is accepted]

4. A sample of 20 observations gave a variance of 0.009. Is this compatible with the hypothesis that the sample is from a normal population with variance 0.010 ?
[**Ans.:** $\chi^2=18$, Accept H_0]

5. A company producing TV tuners knows that on the average its product works satisfactorily for 7 years, with a standard deviation of 1.75 years. A sample of 5 tuners results in life times of 6, 8, 10, 7 and 9 years. Should the producer be satisfied that his product still continues to have a standard deviations of 1.75 years ?
[Ans. $\chi^2=3{\cdot}26$, Accept H_0]

6. A random sample of size 25 from a population gives the sample S.D. to be 8.5 Test the hypothesis that the population S.D. is 10. [Ans. $\chi^2=18{\cdot}06$, Accept H_0]

(2) χ^2-test as a non-paramtric test : χ^2-test is an important non-parametric test and as such no rigid assumptions are necessary in respect of the type of population. We require only the degrees of freedom for using this test. As a non-parametric test, χ^2-test can be used (i) as a test of goodnesss of fit (ii) as a test of independence of attributes.

(i) χ^2-test as a test of goodness of fit : Under the test of goodness of fit, we try to find out how for the observed values of a given phenomenon are significantly different from the expected values *i.e.* there is good compatibility between theory and experiment or the fit is good. The term goodness of fit is also used for comparison of observed sample distribution with the expected probability distributions (such as Binomial, Poisson, Normal). χ^2-test determines how well theoretical distributions (such as Binomial, Poisson), fit the empirical distributions (*i.e.* those obtained from sample data)

Procedure :

(1) Set up the null hypothesis that there is no significant difference between the observed and the expected (or theoretical) values *i.e.* there is good compatibiliy between theory and experiment or the fit is good.

(2) We compute the value of χ^2 by using the formula :

$$\chi^2=\Sigma\left[\frac{(O-E)^2}{E}\right]$$

Where, O = Observed frequency, E = Expected frequency

The above formula can also be written as :

$$\chi^2=\Sigma\left[\frac{O^2}{E}\right]-N$$

Where, N is the total expected frequency and $\Sigma O=\Sigma E=N$

Note : The second form of the formula is more convenient for computation in case the expected frequencies comes in fractions.

(3) Degrees of freedom are worked out by using the following formula :

Degrees of freedon, $v=n-1$

In case of Binomial, Poisson and Normal distributions, the degrees of freedom are obtained by subtracting the number of independent constraints from the total frequency (n). The number of independent constraints in a given data depends upon the number of parameters involved in the same data. This is indicated as under :

	Type of distribution	Constraints	No. of Constraints	Degrees of freedom
1.	Binomial distribution	Total frequency (n)	1	$n-1$
2.	Poisson distribution	Total frequency (n) and arithmetic mean (m)	2	$n-2$
3.	Normal distribution	n, $\bar{X}$ and σ	3	$n-3$

(4) The calculated value of χ^2 as such is than compared with the table value of χ^2 for given degrees of freedom at 5% and 1% level of significance. If the calculated χ^2 exceeds the table value of χ^2, we reject H_0 and conclude that the fit is not good. If the calculated value of χ^2 is less than the table value, we accept H_0 and conclude that the fit is good which means that the divergence between the observed and expected frequencies is attributable to fluctuations of sampling.

CONDITIONS FOR USING THE χ^2-TEST

χ^2-test as a goodness of fit can be used only when (i) n *i.e.* total frequency is large *i.e.* n > 50 (ii) The sample observations are independent (iii) The constraints on the cell frequencies, if any, are linear (iv) no theoretical (or expected) frequency should be small *i.e.* E < 5; if any E < 5, we use pooling technique *i.e.* we add the small frequencies with the preceding or succeeding frequency to obtain the requried sum > 5 and adjust degrees of freedom (d.f.) accordingly.

Example 6 : **A die is thrown 180 times with the following results :**

No. turned up :	1	2	3	4	5	6	Total
Frequency :	25	35	40	22	32	26	180

Test the hypothesis that die is unbiased.

Solution. Set up the null hypothesis that the die is unbiased. On the basis of the hypothesis, the expected frequency of each number turned up $= np = 180 \times \frac{1}{6} = 30$.

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
25	30	-5	25	0.833
35	30	+5	25	0.833
40	30	+10	100	3.333
22	30	-8	64	2.133
32	30	+2	4	0.133
26	30	-4	16	0.533
				$\Sigma(O-E)^2/E = 7.798$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 7.798$$

Degrees of freedom $= v = 6 - 1 = 5$

The tabulated value of χ^2 at 5% level of significance for 5 d.f. = 11.07

Since, the calculated value of χ^2 is less than the table value, we accept the null hypothesis and conclude that die is unbiased.

Example 7 : **The following figures show the distributions of digits in numbers chosen at random from a telephone directory :**

Digit :	0	1	2	3	4	5	6	7	8	9	Total
Frequency :	1026	1107	997	966	1075	933	1107	972	964	853	10,000

Test at 5% level whether the digits may be taken to occur equally frequently in the directory. (Given $\chi^2_{0.05}$ for 9 d.f. = 16.919).

Solution. Let us take the hypothesis that the digits may be taken to occur equally frequently in the directory. On the basis of this thypothesis, the expected frequencies are :

$$10{,}000 \times \frac{1}{10}, 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}$$

$$10{,}000 \times \frac{1}{10}, 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}, \quad 10{,}000 \times \frac{1}{10}$$

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
1,026	1,000	+ 26	676	0·676
1,107	1,000	+ 107	11,449	11·449
997	1,000	-3	9	0009
966	1,000	- 36	1,156	1·156
1,075	1,000	+ 75	5,625	5·625
933	1,000	- 67	4,489	4·489
1,107	1,000	+ 107	11,449	11·449
972	1,000	- 28	784	0·784
964	1,000	- 36	1,296	0·296
853	1,000	- 147	21,609	21·609
				$\Sigma(O-E)^2/E = 57·542$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 57·542$$

Degrees of freedom $(v) = 10 - 1 = 9$

The table vlaue of χ^2 for 9 d.f. at 5% level of significance = 16·919.

Since, the calculated value of χ^2 is greater than the table value, we reject the hypothesis and conclude that the digits may not be taken to occur equally frequently in the directory.

Example 8 : **A sample analysis of examination results of 500 students were made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in second class and 20 got a first class. Are these figures commensurate with the**

general examination result which is in the ratio of 4 : 3 : 2 : 1 for the various categories respectively ?

(The table value of χ^2 for 3 d.f. at 5% level of significance is 7.81).

Solution. Let us take the hypothesis that the observed results are commensurate with the general examination results which is in the ratio of 4 : 3 : 2 : 1 .

The expected no. of students who have failed $= \frac{4}{10} \times 500 = 200$

The expected no. of students who have obtained a III class $= \frac{3}{10} \times 500 = 150$

The expected no. of students who have obtained a II class $= \frac{2}{10} \times 500 = 100$

The expected no. of students who have obtained a I class $= \frac{1}{10} \times 500 = 50$

Applying χ^2-test :

Category	O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
Failed	220	200	+ 20	400	2·000
3rd class	170	150	+ 20	400	2·667
2nd class	90	100	- 10	100	1·00
Ist class	20	50	- 30	900	18·00
					23·667

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 23 \cdot 667$$

Degrees of freedom$(v) = 4 - 1 = 3$

The tabulated value of χ^2 at 5% level of significance for 3 d.f. = 781.

Since, the calculated value of χ^2 is greater than the table value of χ^2, we reject the null hypothesis and conclude that the observed results are not commensurate with the general examination result.

Example 9 : **In an experiment on peas breeding, Mendal obtained the following frequencies of seeds : 315 round and yellow, 101 wrinked and yellow, 108 round and green, 32 wrinkled and green. According to his theory of heredity the numbers should be in proportion 9 : 3 : 3 : 1. Is there any evidence to doubt his theory at 5% level of significance ?**

Solution. Let us take the hypothesis that there is no siginficant difference in the observed and expected values. On the basis of this assumpation, the expected frequencies should be :

$$556 \times \frac{9}{16} = 312 \cdot 75, \quad 556 \times \frac{3}{16} = 104 \cdot 25, \quad 556 \times \frac{3}{16} = 104 \cdot 25, \quad 556 \times \frac{1}{16} = 34 \cdot 75$$

Category	O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
Round and Yellow	315	312·75	2·25	5·0625	0·016
Wrinkled and Yellow	101	104·25	- 3·25	10·5625	0·101
Round and green	108	104·25	- 3·75	14·0625	0·135
Wrinkled and green	32	34·75	- 2·75	7·5625	0·218
	556				$\Sigma(O-E)^2/E$ = 0·47

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 0 \cdot 47$$

Degrees of freedom $= v = n - 1 = 4 - 1 = 3$

For $v = 3$, $\chi^2_{0.05} = 7.82$

Since, the calculated vlaue of χ^2 is less than the table value, we accept null hypothesis. Hence, there is no evidence to doubt the theory at 5% level of significance.

Example 10 : **The following table gives the number of aircraft accidents that occured during the various days of the week.**

Days of the week	Sun.	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Total
No. of accidents	14	16	8	12	11	9	14	84

Find whether the accidents are uniformily distributed over the week:

(Given the table value of $\chi^2_{0.05}$ for 6 d.f. is 12·59)

Solution. Let us take the hypothesis that the accidents are uniformily distributed over the week *i.e.* they are independent of the day of the week. On the basis of this hypothesis, we should expect 84/7 = 12 accidents on each day. Applying χ^2-test :

O	E	$(O-E)^2$	$(O-E)^2/E$
14	12	4	0·333
16	12	16	1·333
8	12	16	1·333
12	12	0	0·00
11	12	1	0·083
9	12	9	0·750
14	12	4	0·333
$\Sigma O = 84$			$\Sigma(O-E)^2/E$ = 4·165

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 4{\cdot}165$$

Degrees of freedom $= v = n - 1 = 7 - 1 = 6$

For $v = 6$, $\chi^2_{0.05} = 12.59$

Since, the calculated value of χ^2 is less than the table value, we accept the null hypothesis and conclude that the accidents are uniformily distributed over the week.

Example 11 : **The number of automobile accidents per week in a certain city were as follows :**

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Are these frequencies in agreement with the belief that accident's numbers were the same during these 10 week period.

Solution. Let us take the this hypothesis be that the number of accidents per week in a certain are equal during the 10 week period.

On the basis of this hypothesis, the expected number of accidents per week $= \frac{100}{10} = 10$.

Applying χ^2- test :

O	E	$(O-E)^2$	$(O-E)^2/E$
12	10	4	0·4
8	10	4	0·4
20	10	100	10·0
2	10	64	6·4
14	10	16	1·6
10	10	0	0·0
15	10	25	2·5
6	10	16	1·6
9	10	1	0·1
4	10	36	3·6
			$\Sigma(O-E)^2/E = 26{\cdot}6$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 26{\cdot}6$$

Degrees of freedom $= v = n - 1 = 10 - 1 = 9$

For $v = 9$, $\chi^2_{0.05} = 16{\cdot}92$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and conclude that the accident conditions were not the same (uniform) over the 10 week period.

Example 12 : **In a city 1000 children were born last week and out of these 600 were males and 400 females. Use chi-square test to assess the general hypothesis that the sex ratio for the newly born children is 1 : 1.**

Solution. Let us assume that the sex ratio for the newly born children is 1 : 1

Given : $N = 1000$ p = probability of a male child $= \frac{1}{2}$, q = probability of a female child $= \frac{1}{2}$. On the basis of the null hypothesis,

Expected no. of male child $= \frac{1}{2} \times 1000 = 500$

Expected no. of female child $= 1000 - 500 = 500$

	O	E	$(O-E)^2$	$(O-E)^2/E$
Males	600	500	10,000	20
Females	400	500	10,000	20
Total	1000	1000		$\chi^2 = 40$

Degrees of freedom $= v = n - 1 = 2 - 1 = 1$

For $x = 1$, $\chi^2_{.05} = 3.84$

Since, the calculated value of χ^2 is grater than the table value of χ^2, we reject the null hypothesis and conclude that the sex ratio for the newly bron children is not 1 : 1.

Test of Goodness of Fit of a Binomial Distribution :

Example 13 : **A set of 5 coins is tossed 3200 times and the number of heads appearing each time is noted. The result are given below :**

No. of heads	0	1	2	3	4	5
Frequency	80	570	1100	900	500	50

Solution. Let us take the null hypothesis that the coins are unbiased i.e. p (H) $= \frac{1}{2}$ On the basis of null hypothesis, the expected number of heads in a toss of 5 coins is calculated by the use of binomial distribution as follows :

$$P(x) = n_{C_x}\, q^{n-x}.p^x \qquad \text{where } x = 0, 1, 2, 3, 4, 5$$

Given : $n = 5$, $N = 3200$, $p = p(H) = \frac{1}{2}$, $q = \frac{1}{2}$

X	$fe(x) = N \times n_{C_x} \cdot P(X)$	E
0	$3200 \times {}^5C_0 \left(\frac{1}{2}\right)^5 \cdot \left(\frac{1}{2}\right)^0$	= 100
1	$3200 \times {}^5C_1 \left(\frac{1}{2}\right)^4 \cdot \left(\frac{1}{2}\right)^1$	= 500
2	$3200 \times {}^5C_2 \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^2$	= 1000

3	$3200 \times {}^5C_3 \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^3$	= 1000
4	$3200 \times {}^5C_4 \left(\frac{1}{2}\right)^1 \cdot \left(\frac{1}{2}\right)^4$	= 500
5	$3200 \times {}^5C_5 \left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^5$	= 100

Applying χ^2- test :

O	E	$(O-E)^2$	$(O-E)^2/E$
80	100	400	4·00
570	500	4900	9·80
1100	1000	10,000	10·00
900	1000	10,000	10·00
500	500	0	0·00
50	100	2500	25·00
			$\Sigma\left[\frac{(O-E)^2}{E}\right]$ = 58·8

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 58 \cdot 8$$

Degrees of freedom $= v = n - 1 = 6 - 1 = 5$

For $v = 5$, $\chi^2_{0.05} = 11 \cdot 07$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and conclude that the coins are biased.

Example 13. **A random sample of 100 families with four children each disclosed the following data :**

No. of Female Births	0	1	2	3	4
No. of families	5	25	40	20	10

Verify at $\alpha = 0 \cdot 05$ if these data are inconsistent with the hypothesis that male and female are equally likey.

Solution. Let us take the null hypothesis that the male and female births are equally probable *i.e.* $p = q = 1/2$ with 0, 1, 2, 3, 4 female basis.

On the basis of null hypothesis, the expected number of families can be calculated by the use of binomial distribution. The probability of x female birth in a family of 4 is given by :

$$P(x) = {}^nC_x \, q^{n-x} \, p^x \qquad \text{where, } x = 0, 1, 2, 3, 4$$

Given : $n = 4$, $N = 100$, $p = \frac{1}{2}$, $q = \frac{1}{2}$

x	$fe(x) = N \cdot P(x)$	E
0	$100 \times {}^4C_0 \left(\frac{1}{2}\right)^4 \cdot \left(\frac{1}{2}\right)^0$	= 6·25

1	$100 \times {}^4C_1 \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^1$	= 25
2	$100 \times {}^4C_2 \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^2$	= 37·5
3	$100 \times {}^5C_3 \left(\frac{1}{2}\right)^1 \cdot \left(\frac{1}{2}\right)^3$	= 25
4	$100 \times {}^4C_4 \left(\frac{1}{2}\right)^0 \cdot \left(\frac{1}{2}\right)^4$	= 6·25

Applying χ^2- test, we have

O	E	$(O-E)^2$	$(O-E)^2/E$
5	6·25	1·5625	0·25
25	25·0	0	0
40	37·5	6·25	0·17
20	25·0	25·0	1·00
10	6·25	14·0625	2·25
100			$\Sigma(O-E)^2/E = 3 \cdot 67$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 3 \cdot 67$$

Degrees of freedom $= v = n - 1 = 5 - 1 = 4$

For $v = 5$, $\chi^2_{0.05} = 9 \cdot 49$

Since, the calculated value of χ^2 is less than the table value of χ^2, we accept the null hypotheis and conclude that the data are consistent with the hypothesis that male and fermale births are equally probable.

Test of Goodness of Fit of a Poisson Distribution :

Example 15 : **The number of defects per unit in a sample of 330 units of a manfactured product was found as follow:**

No. of defect :	0	1	2	3	4
No of units :	214	92	20	3	1

Fit a Poisson distribution to the data and test goodness of fit.

(Given $e^{-\cdot 439} = \cdot 6447$)

Solution.

Fitting of Poisson Distribution

x	f	fx
0	214	0
1	92	92
2	20	40
3	3	9
4	1	4
	$N = 330$	$\Sigma fx = 145$

$$\bar{X} = m = \frac{\Sigma fx}{N} = \frac{145}{330} = 0 \cdot 439$$

$\therefore$ Mean of the distribution $= m = 0 \cdot 439$

$$P(0) = e^{-m} = e^{-439} = \cdot 6447 \quad \text{[From the table]}$$

By Poisson distribution, the expected frequencies are calculated as

$$fe(x) = N.P(x) = \frac{N.e^{-m} \cdot m^x}{x!}$$

Computation of Expected Frequencies

x	$fe(x) = N \times P(x)$	E
0	$f(0) = 330 \times e^{-439} = 330 \times 0 \cdot 6447$	$= 212 \cdot 75$
1	$f(1) = f(0) \cdot \frac{m}{1} = 212 \cdot 75 \times 0 \cdot 439$	$= 93 \cdot 4$
2	$f(2) = f(1) \cdot \frac{m}{2} = 93 \cdot 4 \times \frac{\cdot 439}{2}$	$= 20 \cdot 5$
3	$f(3) = f(2) \cdot \frac{m}{3} = 20 \cdot 5 \times \frac{\cdot 439}{3}$	$= 3 \cdot 00$
4	$f(4) = f(3) \cdot \frac{m}{4} = 3 \cdot 0 \times \frac{\cdot 439}{4}$	$= 0 \cdot 33$

After fitting Poisson distribution, we now apply χ^2 test of goodness of fit.

Let us take the null hypothesis that there is no significant difference between observed frequencies and the frequencies obtained by fitting Poisson distribution.

Appling χ^2-test, we have

Defects	O	E	$(O-E)^2$	$(O-E)^2/E$
0	214	212·75	1·5625	0·0073
1	92	93·4	1·96	0·0210
2	20	20·5		
3	3 (24)	3·0 (23·83)	0·0289	0·0012
4	1	0·33		
				$\Sigma(O-E)^2/E = 0 \cdot 0295$

In the above data, the frequencies of 3 and 4 defects are less than 5, so the frequencies for these defects have been pooled together with defects at 2 in order to make the sum total more than 5 or 5 and $E = 23 \cdot 83$.

Appling the formula.

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 0 \cdot 0295$$

Degrees of freedom $= v = n - 2 = 3 - 2 = 1$

[Since, after grouping only 3 classes are left, therefore $n = 3$]

For $v = 1$, $\chi^2_{.05} = 3 \cdot 84$

Since, the calculated value of χ^2 is much less than the table value, we accept the null hypothesis and conclude that the fit is good.

EXERCISE – 2

1. A die is thrown 100 times and the frequency of various faces are given as below :

Face	1	2	3	4	5	6
Frequency	17	14	20	17	17	15

Test the hypothesis that die is unbiased. Use 5% l.o.s. [Ans. $\chi^2 = 1 \cdot 2796$, Accept H_0]

2. 200 digits were chosen at random from a set of tables. The frequencies of the digits were as follows :

Digits	0	1	2	3	4	5	6	7	8	9	Total
Frequencies	18	19	23	21	16	25	22	20	21	15	200

Using χ^2-test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the table from which they were drawn (The 5% value of χ^2 for 9 d.f. is 16·92). [Ans. $\chi^2 = 4 \cdot 30$, Accept H_0]

3. A research investigator selected a random sample of 200 voters to find which political party they would vote in the municipal elections. The results were observed as under :

Political Party	A	B	C	D
No. of voters	40	90	50	20

Verify at $\alpha = 0 \cdot 05$ if the observed data provide sufficient evidence that the four political parties are equaly preferred. [Ans. $\chi^2 = 4 \cdot 30$, Accept H_0]

4. In an experiment on peas breeding, the following frequencies of feeds were obtained : 218 round and yellow; 72 wrinkled and yellow; 90 round and green, 20 wrinkled and green, Total 400. Theory products that the frequencies should be in the proportions 9 : 3 : 3 : 1. Examine the difference between theory and experiment (Value of χ^2 at 4 and 3 degrees of freedom are 9·448 and 7·815) at 5% level of significance. [Ans. $\chi^2 = 4 \cdot 338$, Accept H_0]

5. The theory predicts the proportion of beans in the four groups, A, B, C and D should be 9 : 3 : 3 : 1. In an experiment among 1600 beans, the number in the four groups were 882, 313, 287 and 118. Does the experiment result support the theory ? Apply χ^2 - test. [$\chi^2 = 4 \cdot 7226$, Accept H_0]

6. The following tables gives the number of books borrowed from a public library during a particular week :

Days of the week	Mon	Tue	Wed	Thu	Fri	Sat
No. of Books borrowed	140	132	160	148	134	150

Test the hypothesis that the number of books borrowed does not depend on the day of the week. Test at 5% level of significance. [Ans. $\chi^2 = 3{\cdot}941$, Accept H_0]

7. A survey of 320 families with 5 children each revealed the following distribution :

No. of boys :	5	4	3	2	1	0
No. of girls :	0	1	2	3	4	5
No. of families :	14	56	110	88	40	12

Is this result consistent with the hypothesis that male and female births are equally probable ? [Ans. $\chi^2 = 7{\cdot}16$. Accept H_0]

8. 4 coins were tossed 160 times and the following results were obtained :

No. of heads :	0	1	2	3	4
No. of obseved frequency :	17	52	54	31	6

Under the assumption that coins are unbiased, find the expected frequencies of getting 0, 1, 2, 3 or 4 heads and test the goodness of fit. [Ans. $\chi^2 = 12{\cdot}725$, Reject H_0]

9. A book has 700 pages. The number of pages with various number of misprints is recorded below. At 5% significance level, are the misprints distributed according to possion law :

No. of misprints :	0	1	2	3	4	5	Total
No. of pages with x misprints :	616	70	10	2	1	1	700

[Ans. $\chi^2 = 11{\cdot}04$, Reject H_0, Fit is not good]

10. The following mistakes per page were observed in a book :

No. of mistakes per page :	0	1	2	3	4	Total
No. of units :	211	90	19	5	0	325

Does this information verify that the mistakes are distributed according to Poisson distribution ? [$e^{-0{\cdot}44} = 0{\cdot}644$] [Ans. $\chi^2 = 0{\cdot}07$, Accept H_0]

11. The number of car accidents in a metropolitan city was found as 20, 17, 12, 6, 7, 15, 8, 5, 16 and 14 per month respectively. Use chi square test to check whether these frequencies are in agreement with the belief that occurrence of accidents was the same during the 10 months period. Test at 5% level of significance. (Take value at 5% level for $v = 9$ is 16.9) [Ans. $\chi^2 = 20{\cdot}331$ Reject H_0]

12. The following grades were given to a class of 100 students

Grade :	A	B	C	D	E
Frequency :	14	18	32	20	16

Test the hypothesis at the 0·05 level, that the distribution of grade is uniform.

Given :

Degree of Freedom :	3	4	5
χ^2-value :	7·81	9·49	11·07

[Ans. $\chi^2 = 10$, Raject H_0]

13. In the accounting department of a bank 100 accounts are selected at random and examined for errors. Suppose the following results have been obtained:

No. of errors :	0	1	2	3	4	5	6
No. of accounts :	36	40	19	2	0	2	1

On the basis of this information can it be concluded that the errors are distributed according to the poisson probability law ? [Ans. $\chi^2 = 1{\cdot}450$, Accept H_0]

14. A survey of 200 families with 3 children selected at random gave the following results.

Male Births	0	1	2	3
No. of families	40	58	62	40

Test the hypothesis that male and female are equally likely at 5% level of significance. [Ans. $\chi^2 = 24{\cdot}1$, Reject H_0]

15. In a city the percentage of smokers was 90. A random sample of 100 persons was taken and out of them universe 85 were found smokers. Use chi square test and tell whether sample ratio significantly differs from the universe ratio for the city [Ans. $\chi^2 = 2.778$, Accept H_0]

16. The manager of a theatre complex with four theatres wanted to see whether there was a difference in popularity of the four movies currently showing for saturday afternoon matinees. The number of custmers for each movie was recorded for one Satuday afternoon with the following results : 63, 55, 75 and 77 customers viewed movies, 1, 2, 3 and 4 respectively. Complete the test to see whether there is a difference at the 5% level of significance. [Ans. $\chi^2 = 4.78$, Reject H_0]

(ii) χ^2-test as a test of independence of attributes : χ^2-test enables us to examine whether or not two attributes are associated or independent of one another. For example, we may be interested in knowing whether a new medicine is effective in controlling fever or not. χ^2-test will help us in deciding this issue.

Procedure :

(i) Set up the null hypothesis that the two attributes (*viz.* new medicine and control of fever) are independent which means that new medicine is not effective in controlling fever.

(ii) On the basis of the null hypothesis, we calculate the expected frequencies by using the following formula :

$$\text{Expected frequency} = \frac{(R) \times (C)}{N}$$

Where, R = Row Total, C = Column total, N = Total number of observations.

(iii) We compute the χ^2-value by using the following formula :

$$\chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right]$$

(iv) For a contingency table which has 'r' rows and 'c' columns, degrees of freedom are worked out by using the formula.
Degrees of freedom $= v = (c-1)(r-1)$

(v) Obtain the critical value (or table value) of χ^2 with reference to the degrees of freedom for the given problem and the desired level of significance.

(vi) If the calculated value of χ^2 is less than the table value at a certain level of significance for given degrees of freedom, we accept the null hypothesis and conclude that the two attributes are independent (*i.e.* the new medicine is not effective in controlling the fever). But if the calculated value χ^2 is greater than its table value, we reject the null hypothesis and conclude that the two attributes are associated (*i.e.* new medicine is effective in controlling fever).

Example 16 : **A sample of 200 persons with a particular disease was selected. Out of them, 100 were given drug and others were not. The results were observed as follows:**

	No. of Persons Given		
	Drug	No Drug	Total
Cured	55	65	120
Not Cured	45	35	80
Total	100	100	200

Test whether the drug has been effective in curing the disease.

Solution. Let the null hypothesis be that drug has not been effective in curing the disease.
On the basis of this hypothesis, the expected frequencies are calculated as follow:

$$E_{11} = \frac{100 \times 120}{200} = 60$$

The remaining frequencies can be found by subtractions from the column and row totals.
The expected frequencies table would be as follows :

60	60	120
40	40	80
100	100	200

Applying χ^2- test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
55	60	− 5	25	0·416
45	40	+ 5	25	0·625
65	60	+ 5	25	0·416
35	40	− 5	25	0·625
				$\Sigma(O-E)^2/E = 2·082$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 2·082$$

Degrees of freedom $= v = (2-1)(2-1) = 1$
For $v = 1$, $\chi^2_{0.05} = 3·84$

Since, the calculated value of χ^2 is less than the table value of χ^2, we accept the null hypothesis and conclude that the drug has not been effective in curing the disease.

Example 17 : **The table given below shows the data during and epidemic of cholera :**

	Attacked	Not Attacked	Total
Inoculated	31	469	500
Not Inoculated	185	1,315	1,500
Total	216	1,784	2,000

Use χ^2 test to determine whether inoculation is effective in preventing the attack of cholera (Given as 5% level of significance, the value of $\chi^2_{0.05}$ for 1 d.f. = 3·84).

Solution. Let us take the hypothesis that inoculation is not effective in preventing the attack of cholera. On the basis of this hypothesis, the expected frequencies are :

$$E_{11} = \frac{216 \times 500}{2,000} = 54$$

The remaining frequencies can be found out by subtractions from the column and row totals.
The expected frequecies table would be as follows:

54	446	500
162	1338	1500
216	1784	2000

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
31	54	−23	529	9·797
185	162	+ 23	529	3·266
469	446	+ 23	529	1·187
1315	1338	− 23	529	0·396
				$\Sigma(O-E)^2/E = 14·646$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 14·464$$

Degrees of freedom $= v = (r-1)(c-1) = (2-1)(2-1) = 1$
For $v = 1$, $\chi^2_{0.05} = 3·84$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and conclude that inoculation is effective in preventing the attack of cholera.

Example 18 : Two investigators study the income of group of persons by the method of sampling. Following results were obtained by them :

Investigator	Poor	Middle class	Well-to-do	Total
A	160	30	10	200
B	140	120	40	300
Total	300	150	50	500

Show that the sampling technique of at least one of the investigator is suspected.

(Given the value of $\chi^2_{0.05}$ for 2 d.f. = 5·991)

Solution. Let us take the hypothesis that there is no suspicion about the sampling technique of the two investigators. On the basis of this hypothesis, the expected frequencies shall be :

$$E_{11} = \frac{300 \times 200}{500} = 120, \qquad E_{12} = \frac{150 \times 200}{500} = 60$$

The remaining frequencies can be found out by subtractions from the column and row totals.

The table of expected frequencies is given below :

120	60	20	200
180	90	30	300
300	150	50	500

applying χ^2-test

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
160	120	40	1600	13·333
140	180	− 40	1600	8·888
30	60	− 30	900	15·000
120	90	30	900	10·000
10	20	−10	100	5·000
40	30	+ 10	100	3·333
				$\Sigma (O-E)^2/E = 55·554$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 55·554$$

Degrees of freedom $= v = (r-1)(c-1) = (2-1)(3-1) = 2$

For $v = 2$, $\chi^2_{0.05} = 5·991$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and conclude that the sampling technique of at least one of the investigators is suspected.

Example 19 : **A milk producer's union wishes to test whether the preference pattern of consumers for its product is dependent on income levels. A random sample of 500 individuals gives the following data :**

Income	Product Preferred			
	Product A	Product B	Product C	Total
Low	170	30	80	280
Medium	50	25	60	135
High	20	10	55	85
Total	240	65	195	500

Can you conclude that the preference patterns are independent of income levels?

(For $v = 4$, $\chi^2_{0.05} = 9·49$)

Solution. Let us take the hypothesis that preference patterns are independent of income levels. On the basis of this hypothesis, the expected frequencies corresponding to different rows and columns shall be :

$$E_{11} = \frac{240 \times 280}{500} = 134·4 \qquad E_{12} = \frac{65 \times 280}{500} = 36·4$$

$$E_{21} = \frac{240 \times 135}{500} = 64·8 \qquad E_{22} = \frac{65 \times 135}{500} = 17·55$$

134·40	36·40	109·20	280
64·80	17·55	52·65	135
40·80	11·05	33·15	85
240	65·00	195·00	500

Applying χ^2- test :

O	E	$(O-E)^2$	$(O-E)^2/E$
170	134·40	1267·36	9·430
50	64·80	219·04	3·3802
20	40·80	432·64	10.603
30	36·40	40·96	1.125
25	17·55	55·50	3.162
10	11·05	1·10	0.099
80	109·20	852·64	7.808
60	52·65	54·02	1.026
55	33·15	477·42	14.402
			$\Sigma (O-E)^2/E = 51·036$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 51·036$$

Degrees of freedom $= v = (r-1)(c-1) = (3-1)(3-1) = 4$

For $v = 4$, $\chi^2_{0.05} = 14 \cdot 860$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and hence conclude that preference patterns are not independent of income levels.

Example 20 : **In a survey of 200 boys, of which 75 were intelligent, 40 had educated fathers; while 85 of the unintelligent boys had uneducated fathers. Do these figures support the hypothesis that educated fathers have intelligent boys. Use χ^2-test (The value of χ^2 for 1 degree of freedom at 5% level is 3·84).**

Solution. The given data can be tabulated as follows :

Boys/Fathers	Educated	Uneducated	Total
Intelligent	40	35	75
Unintelligent	40	85	125
Total	80	120	200

Let us take the hypothesis that there is no association between the education of fathers and intelligence of sons.

On the basis of this hypothesis, the expected frequencies shall be :

$$E_{11} = \frac{75 \times 80}{200} = 30$$

The remaining frequencies can be found be subtracting from the column and row totals.

The table of expected frequencies shall be as follows :

30	45	75
50	75	125
80	120	200

Applying χ^2-test :

O	E	$(O-E)^2$	$(O-E)^2/E$
40	30	100	3·333
40	50	100	2·000
35	45	100	2·222
85	75	100	1·333
			$\Sigma (O-E)^2/E = 8 \cdot 888$

$$\therefore \quad \chi^2 = \Sigma \left[\frac{(O-E)^2}{E} \right] = 8 \cdot 888$$

Degrees of freedom $= v = (r-1)(c-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{0.05} = 3 \cdot 84$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and hence that educated fathers have intelligent boys.

Alternative Formula for Finding the Value of χ^2 in a (2 × 2) table

There is an alternative formula of calculating the value of χ^2 in a (2 × 2) table. If we write the cell frequencies and marginal totals in case of a (2 × 2) table as :

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	N

then the formula for calculating the value of χ^2 will be written as follows :

$$\chi^2 = \frac{N.(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \qquad \text{where, } N = a+b+c+d$$

Note : The alternative formula is rarely used in finding the value of χ^2 as it is not applicable uniformaly in all cases but can be used only in a 2 × 2 contingency table.

Example 21. **In an anti-malaria campaign in a certain area, quinine was administrated to 812 persons out of a total population of 3248. The number of fever cases is shown below :**

Treatment	Fever	No Fever	Total
Quinine	20	792	812
No quinine	220	2216	2436
Total	240	3008	3248

Discuss the usefulness of quinine in checking malaria. (Given for $v = 1$, $\chi^2_{0.05} = 3 \cdot 84$)

Solution. Let us take the null hypothesis that the quinine is not effetive in checking malaria.

Arrange the given data in a designated form, we have

	Fever	No Fever	Total
Quinine	a 20	b 792	$a+b$ 812
No Quinine	c 220	d 2216	$c+d$ 2436
Total	$a+c$ 240	$b+d$ 3008	N 3248

For 2 × 2 table, using the direct formula of computing χ^2, we have

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Putting the values, we have

$$= \frac{3248\,(20 \times 2216 - 220 \times 792)^2}{(240)(3008)(812)(2436)} = 38 \cdot 48$$

Degrees of freedom $= v = (c-1)(r-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{0.05} = 3.84$

Since, the calculated value of χ^2 is greater than the table value, we reject the null hypothesis and conclude that the quinine is usefulness in checking malaria.

YATES CORRECTIONS IN A 2×2 TABLE

F. Yates has suggested corrections for continuity in χ^2 value calculated in a 2×2 table, particularly, when any cell frequency is less than 5. The correction suggested by Yates is popularly known as **Yates' Correction**. It involves the reduction of the deviation of observed from expected frequencies which of course reduces the value of χ^2. The rule for correction is to increase the observed frequencies which is less than 5 by 0·5 and than the remaining frequencies are adjusted by adding or subtracting 0·5 to them without disturbing the marginal totals. The observed values, thus corrected will be represented by O from which deviations of the corresponding expected values, E will be found.

Note : In a 2×2 table, the method of pooling cannot be applied.

Example 22: **The result of a certain survey of 50 ordinary shops of small size is given below :**

	Shops in		Total
	Towns	Villages	
Run by Men	17	18	35
Run by Women	3	12	15
Total	20	30	50

Can it be said that shops run by women are relatively more in villages than in towns. Use χ^2-test. (Table value of $\chi^2_{0.05}$ for one degree of freedom at 5% level of significance is 3·84).

Soluiton : Let us take the null hypothesis that shops run by women are equal in number in villages as well as in towns. On the basis of this hypothesis, the expected frequencies will be as follows :

$$E_{11} = \frac{20 \times 35}{50} = 14$$

The remaining frequencies can be found out by subtractions from the column and row totals.

The table of expected freqencies will be :

14	21	35
6	9	15
20	30	50

Since, one of the observed frequency is less 5, we increase the value of that observed frequency by 0·5 and adjust other frequencies using Yates' corrections. The adjusted observed frequencies after Yates' corrections will be as follow :

17 − 0·5 = 16·5	18 + 0·5 = 18·5	35
3 + 0·5 = 3·5	12 − 0·5 = 11·5	15
20	30	50

With the above expected and corrected observed values, the corrected value of χ^2 will be obtained as :

O	E	(O − E)	(O − E)²	(O − E)² / E
16·5	14	2·5	6·25	0·446
3·5	6	− 2·5	6·25	1·042
18·5	21	− 2·5	6·25	0·298
11·5	9	+ 2·5	6·25	0·694
				Σ (O − E)² / E = 2·48

$$\therefore \quad \chi^2 = \Sigma \frac{(O-E)^2}{E} = 2{\cdot}48$$

Degrees of freedom $= v = (c-1)(r-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{0.05} = 3{\cdot}84$

Since, the calculated value of χ^2 is less than the table value, we accept the null hypothesis and conclude that number of shops run by women are not relativiely more in villages than in towns.

Example 23 : **In an experiment on the immunization of goats from Anthrax, the following results were obtained. Derive your inference on the efficacy of the vaccine :**

	Diet of Anthrax	Survived	Total
Inoculation with vaccine	2	10	12
Not inoculated	6	6	12
	8	16	24

Solution. It is quite obvious from the above data that Yates' correction shall be applied here. Let us take the null hypothisis that there is no relationship between Inoculation with vaccine and death from anthrax.

Observed frequencies

2	10	12
6	6	12
8	16	24

Observed frequencies with Yates' corrections

2.5	9.5	12
5.5	6.5	12
8	16	24

Expection frequencies

$\frac{12 \times 8}{24} = 4$	12 − 4 = 8	12
8 − 4 = 4	12 − 4 = 8	12
8	16	24

$$\therefore \chi^2 = \frac{(2.5-4)^2}{4} + \frac{(9.5-8)^2}{8} + \frac{(5.5-4)^2}{4} + \frac{(6.5-8)^2}{8}$$

$$= 0.56250 + 0.28125 + 0.56250 + 0.28125 = 1.6875$$

Degrees of freedom $= v = (c-1)(r-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{.05} = 3.84$

Since the calculated value of χ^2 is less than the table value, we accept the null hypothesis and conclude that thre is no relationship between inoculation with vaccine and death from anthrax *i.e.* immunization is not effective.

Test of Equality of Several Population Proportions

Testing of equality of two or more population poportions is an extension of the χ^2-test of independence. χ^2-test is also used to examine the equality two or more population proportions. The following examples clarify the procedure.

Example 24 : **A social organisation claiming to be the promoters of sex education sought the views of parents from the states of Punjab, Bihar and Haryana introducing sex education at the school level. The views of 80 parents selected at random from each of the three states are as follows :**

	Punjab	Bihar	Haryana
In favour	50	20	45
Against	30	60	35

Do the sample provide enough evidence to the view that the proportion of parents in favour of introducing sex education in schools is the same in all three states ? Use $\alpha = \cdot 01$.

Solution. Let the null hypothesis be that the proportion of parents in favour of sex education in schools is the same in the three states.

One the basis of this hypothesis, the expected frequencies are calculated as follows :

$$E_{11} = \frac{80 \times 115}{240} = 38.3 \qquad E_{12} = \frac{80 \times 115}{240} = 38.3$$

The remaining frequencies can be found out by subtracting from the column and row totals.

The expected frequencies would be as follows :

38·3	38·3	38·4	115
41·7	41·7	41·6	125
80	80	80	240

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
50	38.3	11.7	136.89	3.57
20	38.3	–18.3	334.89	8.74
45	38.4	6.7	44.89	1.169
30	41.7	–11.7	136.89	3.28
60	41.7	18.3	334.89	8.03
35	41.6	– 6.6	43.56	1.04
				$\Sigma[(O-E)^2/E] = 25 \cdot 83$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 25 \cdot 83$$

Degrees of freedom $= v = (c-1)(r-1) = (3-1)(2-1) = 2$

For $v = 2$, $\chi^2_{0.01}\, 0 \cdot 1 = 9 \cdot 21$

Since, the calculated value of χ^2 is less than the table value of χ^2, we reject H_0 and conclude that the proportions of parents in favour of introducing education at the school level are not the same in the three states.

Example 25. **The following data gives the HDL-level in random samples of sizes 120, 200, 150 and 130 from the adult population of the four cities A, B, C and D.**

	A	B	C	D
High HDL	53	80	68	57
Not High DHL	67	120	82	73

Test the equality of proportions of adults with high HDL Cholesterol in these four cities. Use $\alpha = 0.025$.

Solution. Let P_1, P_2, P_3 and P_4 represents the true proportions of adults with high HDL cholesterol in the cities A, B, C and D respectively.

Set up the hypothesis :

Null hypothesis : $H_0 : P_1 = P_2 = P_3 = P_4$

Alternative hypothesis : $H_1 : P_1, P_2, P_3$ and P_4 are not all equal.

Compute the expected frequency for each observed frequency by the formula (under the hypothesis of independence) :

$$\text{Expected Frequency} = \frac{(\text{Row Total})\,(\text{Columm Total})}{\text{Grand Total}}$$

The observed and the expected frequencies are given in Table below. The bold figures in brackets ( ), represent the corresponding expected frequencies.

Observed and Expected Frequencies

	A	B	C	D	Total
High HDL	53	80	68	57	258
	(51.6)	(86)	(64.5)	(55.9)	
Not High HDL	67	120	82	73	342
	(68.4)	(144)	(85.5)	(74.1)	
Total	120	200	150	130	600

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

$$= \frac{(53-51.6)^2}{51.6} + \frac{(80-86)^2}{86} + \frac{(68-64.5)^2}{64.5} + \frac{(57-55.9)^2}{55.9}$$

$$+ \frac{(67-68.4)^2}{68.4} + \frac{(120-114)^2}{114} + \frac{(82-85.5)^2}{85.5} + \frac{(73-74.1)^2}{74.1}$$

$$= 0.0380 + 0.4186 + 0.1899 + 0.0216 + 0.0287 + 0.3158 + 0.1433 + 0.0163 = 1.1722$$

Degrees of freedom : $v = (r-1)(c-1) = (2-1)(4-1) = 3$

The critical value of chi square for 3 d.f. and level of significance 0.025 is $\chi^2_{3(0.025)} = 9.348$.

Since, the computed value of test statistic is less than the critical (tabulated) value, it is not significant. Hence, we fail to reject the null hypothesis at 0.025 level of significance.

Conclusion : H_0 may be accepted at level of significance $\alpha = 0.025$ and we may conclude that the proportion of adults with high HDL cholesterol level is most likely the same in all the four cities.

Example 26. **It is found that 35 of 250 housewives in Delhi, 22 of 220 housewives in Mumbai and 39 of 300 housewives in Chandigarh watch at least one talk show every day. At the 0.05 level of significance, test that there is no difference between the true proportions of housewives who watch talk shows in these cities.**

Solution. Let P_1, P_2 and P_3 represent the true proportion of housewives who watch talk shows in the cities of Delhi, Mumbai and Chandigarh, respectively.

Null hypothesis : $H_0 : P_1 = P_2 = P_3$

Alternative hypothesis : $H_1 : P_1$, P_2 and P_3 are not all equal.

Expected frequencies : Compute the expected frequency for each of the cell frequencies by the formula (under the hypothesis of independence) :

$$\text{Expected Frequency} = \frac{(\text{Row Total})\,(\text{Columm Total})}{\text{Grand Total}}$$

The observed frequencies, along with the expected frquencies [(in the bold in brackets ()] are given in the following Table.

Observed and Expected Frequencies

	Delhi	Mumbai	Chandigarh	Total
Watch Talk Show	35 **(31.2)**	22 **(27.4)**	39 **(37.4)**	96
Do not Watch Talk Show	215 **(218.8)**	198 **(192.6)**	261 **(262.6)**	674
Total	250	220	300	770

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

$$= \frac{(35-31.2)^2}{31.2} + \frac{(22-27.4)^2}{27.4} + \frac{(39-37.4)^2}{37.4} + \frac{(215-218.8)^2}{218.8} + \frac{(198-192.6)^2}{192.6} + \frac{(261-262.6)^2}{262.6}$$

$$= 0.4628 + 1.0642 + 0.0684 + 0.0660 + 0.1514 + 0.0097 = 1.8225.$$

Degrees of freedom : $v = (r-1)(c-1) = (2-1)(3-1) = 2$

The critical value of chi-square for 2 *d.f.* and 0.05 level of significance is 5.991.

Conclusion : Since, the calcluated value of test statistic ($\chi^2 = 1.8225$ is less than the tabulated value) $\chi^2 = 5.991$, it is not significant. Thus, the data do not provide enough evidence against the null hypothesis, which may be accepted at 5% level of significance. Hence, we conclude that the proportion of housewives who watch talk shows is same in all the three cities.

TEST OF HOMOGENITY

The test of homogenity is another extension of the χ^2-test of independence. Tests of homogenity are used to determine whether two or more independent random samples are drawn from the sample population. Instead of one sample as we use with independence problem, we shall now have two or more samples are from each population.

The following example clarify the procedure of the test :

Example 27 : **A insurance company has introduced a new scheme for employees. Independent random samples of 100 males and 120 females when examined to know their views about the new scheme yielded the following results :**

	For	Against	Indifferent	Total
Male	25	40	35	100
Female	35	55	30	120
Total	60	95	65	220

Test the hypothesis at $\alpha = \cdot 01$ that the two samples have come from a homogenous populations.

Solution. Let us take the null hypothesis that the two samples have come from a homogenous population. On the basis of the hypothesis, The expected frequencies are calculated as :

$$E_{11} = \frac{60 \times 100}{220} = 27.3 \qquad E_{12} = \frac{95 \times 100}{220} = 43.2$$

The remaining frequencies can be found out by subtracting from the column and row totals.

The expected frequencies worked be as follows :

27·3	43·2	29·5	100
32·7	51·8	35·5	120
60	95	65	220

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
25	27·3	− 2·3	5·29	0·1937
40	43·2	− 3·2	10·24	0·2370
35	29·5	5·5	30·25	1·025
35	32·8	2·3	5·29	0·1617
55	51·8	3·2	10·24	0·1976
30	35·5	5·5	30·25	0·8521
				$(O-E)^2/E = 2·6671$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 2 \cdot 6671$$

Degrees of freedom $= v = (r-1)(c-1) = (2-1)(2-1) = 1$

For $v = 1$, $\chi^2_{.05} = 3 \cdot 84$

Since, the calculated value of χ^2 is less than the table value, we accept the null hypothesis and conclude that the two samples have come from homogenous populations.

EXERCISE – 3

1. A survey amongst women was conducted to study the family life. The observations are as follows :

Education	Family Life		
	Happy	Not Happy	Total
Educated	70	30	100
Non-Educated	60	40	100
Total	130	70	200

Test whether there is any association between familiy life and education.
(The table value of $\chi^2_{0 \cdot 05}$ for 1 d.f. $= 3 \cdot 84$) [**Ans.** $\chi^2 = 2 \cdot 198$, Accept H_0]

2. Calculate the expected frquencies for the following data persuming the two attributes, *viz*, condition of home and condition of child as independent.

		Condition of Home	
		Clean	Dirty
Condition of Child	Clean	70	50
	Fairly Clean	80	20
	Dirty	35	45

Use chi-square test at 5% level of signifiance whether the two attributes are indendent.
(Table values of chi-square at 5% for 2 d.f. is 5.991 and for 3 d.f. is 7.815 and for 4 d.f. is 9.488.) [**Ans.** $\chi^2 = 25 \cdot 848$, Rejeet H_0]

3. Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence level. The results are as follows :

Researcher	No. of students in each level				Total
	Below Average	Average	Above Average	Genius	
X	137	164	152	147	600
Y	32	57	56	35	180
Total	169	221	208	182	780

Would you say that the sampling techniques adopted by the two researchers are significantly different? (For $v = 3$, $\chi^2_{0 \cdot 05} = 7 \cdot 82$ [**Ans.** $\chi^2 = 5 \cdot 801$, not different]

4. From the following data, find out whether there is any relationship between sex and prefenece for colour :

Colour	Males	Females	Total
Green	40	60	100
White	35	25	60
Yellow	25	15	40
Total	100	100	200

(Given for $v = 2$, $\chi^2_{0 \cdot 05} = 5 \cdot 991$) [**Ans.** $\chi^2 = 8 \cdot 166$, Reject H_0]

5. A sample of 400 students of under-graduate and 400 staudents of post-graduate classes was taken to know their opinion about autonomos colleges. 290 of the under-graduate and 310 of the post-graduate students favoured the autonomous status. Present these facts in the form of a table and test, at 5% level, that the opinion regarding autonomous status of colleges is independent of the level of classes of students. (Table value χ^2 at 5% level is 3.84 for 1 *d.f.*). [**Ans.** $\chi^2 = 2 \cdot 66$, Accept H_0]

6. Two treatments A and B were tried to control a certain type of plant disease. The following results were obtained.
A : 400 plants were examined and 80 were found infected.
B : 400 plants were examined and 70 were found infected.
Is the treatment B superior to treatment A ?
(Given that $\chi^2_{0 \cdot 05}(1) = 3 \cdot 84$; $\chi^2_{0 \cdot 05}(3) = 7 \cdot 82$) [**Ans.** $\chi^2 = 0 \cdot 82$, Accept H_0]

7. In an experiment on immunization of cattle from tuberculosis, the following were obtained :

	Affected	Not Affected	Total
Inoculated	4	20	24
No inoculated	6	50	56
Total	10	70	80

Calculate χ^2 and discuss the effect of vaccine in controlling suspectability to tuberculosis. [Applying Yates' correction] [**Ans.** $\chi^2 = 2.04$, Accept H_0]

8. The following table gives the frequencies of firms on automation and productivity :

	Productivity increased	Productivity not increased	Total
Automated	32	468	500
Not Automated	184	1316	1500
Total	216	1784	2000

Use χ^2 (Chi-Square) test to determine whether productivity is independent of the automation ($\chi^2_{0 \cdot 05}$ at 1 d.f. $= 3 \cdot 84$) [**Ans.** $\chi^2 = 13 \cdot 395$, Reject H_0]

9. A drug is said to be useful for the treatment of cold. In an experiment carried out on 160 persons suffering from cold, half of the persons were treated with the drug and rest half with sugar pills. The effect of treatment is described in the following table :

	Helped	Harmful	No Effect
Drug	52	10	18
Suger Pills	44	10	26

Test the hypothesis that in the treatment of cold the drug is not at all effective as compared to sugar pills.

[Given for $v=2$, $\therefore \chi^2_{.05} \cdot 05 = 5 \cdot 991$] [Ans. $\chi^2 = 2 \cdot 12$, Accept H_0]

10. Two sample polls of votes for two candidates A and B for public office are taken, one each from among residents of rural and urban areas. The results are given below. Examine whether the nature of area is related to voting preference in this election ?

Area/Candidate	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

Given $\chi^2_{0 \cdot 05} = 3 \cdot 841, 5 \cdot 991, 7 \cdot 82$ for 1, 2 and 3 d.f. respectively.

[Ans. $\chi^2 = 10 \cdot 32$, Accept H_0]

11. Two groups of 100 people each were selected for testing the use of a vaccine. 15 persons contracted the disease out of the inoculated persons in one group, while 25 contracted the disease in the other group. Test the efficacy of the vaccine using the χ^2 test.

(Given of $v=1, \chi^2_{0 \cdot 05} = 3 \cdot 84$) [Ans. $\chi^2 = 3 \cdot 124$, Accept H_0]

12. A company has head offices at three places : A, B and C, A random sample of 10 executives posted at A, of 12 posted at B, and of 15 posted at C were examined to find out the number of those suffering from hypertension. The sample results were found to be as given below:

	A	B	C
Hypertension cases	4	7	9
Non-hpertension cases	6	5	6

Verify at $\alpha = 0 \cdot 01$ if the proportion of executives suffering from the hypertension at three head offices are the same. [Ans. $\chi^2 = 1 \cdot 0975$, Accept H_0]

13. From the adult male population of four large cities, random samples of sizes given below were taken and the number of married and single men recorded. Can we say that the proportion of married men is same in all the four cities?

City →	A	B	C	D	Total
Married	137	164	152	147	600
Single	32	57	56	35	180
Total	169	221	208	182	780

[Ans. $\chi^2 = 5.801$, Accept H_0]

ADDITIVE PROPERTY OF χ^2

Chi-Square possesses the additive property. If a number of samples of similar data have been independently collected and a number of χ^2 values have been obtained there form, it is possible to combine them by the simple process of addition. This helps in getting a better idea about the significance or otherwise of the problem in hand as instead of on investigation (or one sample).

Example 28. **An investigation was made in eight big cities of a state with a view to test the effectiveness of inoculation during an epidmic of cholera. The following results were obtained :**

Cities	A	B	C	D	E	F	G	H
χ^2 value	2·32	3·64	3·15	4·54	2·24	3·66	4·87	6·72
d.f.	1	1	1	1	1	1	1	1

Find out the pooled χ^2 for all the eight cities of any state and test your result at 5% level of significance.

Solution. $H_0 : f_0 = f_e$ (Observed and expected distributions are the same)

$H_1 : f_0 \neq f_e$ (Difference between observed and expected distributions is significant)

$\alpha = \cdot 05$; d.f. $= 8, \chi^2 = 15 \cdot 507$

Cities	A	B	C	D	E	F	G	H	Total
χ^2 value	2·32	3·64	3·15	4·54	2·24	3·66	4·87	6·72	Pooled $\chi^2 = 31 \cdot 14$
d.f.	1	1	1	1	1	1	1	1	8

INTERPRETATION

The table value of χ^2 at 5% level of significance with 1 d.f. is 3·841 and 8 d.f. is 15·507. By the analysis of each city separately, it is clear that the difference is not significant in the cities A, B, C, E and F *i.e.* the null hypothesis is true where as the difference in cities D, G and H is significant and the null hypothesis is not ture.

But the combined (or pooled) calculated χ^2 is 31·14 which is greater than 15·507. Thus combined value is greater than the table value; hence the difference in the cities together is significant. That is the null hypothesis is not true by considering all the cities together.

MISCELLANCEOUS SOLVED EXAMPLES

Example 29 : **A controlled experiment was conducted to test the effectivencess of a new drug. Under this experiment 300 patients were treated with the new drug and 200 were not treated with the drug. The results of the experiments are presented below. Using the Chi-square test, comment on the effectiveness of drug.**

Details	Cured	Condition worsened	No effect	Total
Treated with the new drug	200	40	60	300
Not treated with the new drug	120	30	50	200
Total	320	70	110	500

Solution. Let us take the hypothesis that the new drug is not effective. On the basis of this hypothesis, the expected frequencies are calculated as follows :

$$E_{11} = \frac{320 \times 300}{500} = 192; \qquad E_{12} = \frac{70 \times 300}{500} = 42 \text{ and so on.}$$

The remaining frequencies can be found out by subtraction from the column and row totals.

The expected frequencies is given below :

192	42	66	300
128	28	44	200
320	70	110	500

Applying χ^2-test :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
200	192	8	64	0·333
120	128	− 8	64	0·500
40	42	− 2	4	0·095
30	28	2	4	0·143
60	66	− 6	36	0·545
50	44	+ 6	36	0·818
				$\Sigma(O-E)^2/E = 2·434$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 2·434$$

Degrees of freedom $= v = (c-1)(r-1) = (3-1)(2-1) = 2$

For $v = 2$, $\chi^2_{.05} = 5·99$

Since, the calculated value of χ^2 is less than the table value, we accept the hypothesis and conclude that the drug is effective.

Example 30 : **Fit a Poisson Distribution and test the goodness of fit from the following data :**

No. of mistakes per page	0	1	2	3	4	Total
No. of pages	211	90	19	5	0	325

(Given $e^{-0·44} = 0·6440$)

Solution. (i) Fitting of Poisson Distribution

Mistake (X)	Pages (f)	fX
0	211	0
1	90	90
2	19	38
3	5	15
4	0	0
	$\Sigma f = 325$	$\Sigma fX = 143$

$$\therefore \quad \bar{X} = m = \frac{\Sigma fx}{\Sigma f} = \frac{143}{325} = ·44$$

By Poisson distibution, the expected frequency (number) of pages containing x mistakes is given by :

$$f(X) = N \cdot P(X) = 325 \times \frac{e^{-44} \times (·44)^x}{x!}$$

Also $P(0) = e^{-44} = ·6440$ Computation of Expected Frequencies

Computation of Expected Frequencies

X	$fe(x) = N \times P(x)$	E_i
0	$f(0) = 325 \times e^{-44} = 325 \times ·64470$	= 209·30
1	$f(1) = f(o) \times \frac{m}{1} = 209·30 \times ·44$	= 92·09
2	$f(2) = f(1) \times \frac{m}{2} = 92·14 \times \frac{·44}{2}$	= 20·26
3	$f(3) = f(2) \times \frac{m}{3} = 20·26 \times \frac{·44}{3}$	= 2·97
4	$f(4) = f(3) \times \frac{m}{4} = 2·97 \times \frac{·44}{4}$	= ·3267

(b) Test of Goodness of Fit : Let us teke the hypothesis that there is no difference between the observed and expected frequencies. Since, the frequency at one corner are less than 5, they would be combined with the adjacent frequency :

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
211	209.3	+ 1·7	2·89	·0138
90	92.09	− 2·09	4·368	0·0474
19	20.26	− 1·26	1.587	0·078
5				
0	3.29	+ 1·71	2·924	0·88
				$\Sigma(O-E)^2/E = 1.0192$

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 1·0192$$

Degrees of freedom $(v) = n - 2 = 4 - 2 = 2$

$\chi^2_{.05}$ for 2 d.f. $= 5·99$.

Since, the calculated value of χ^2 is less than the table of χ^2, we accept the null hypothesis and conclude that the fit is good.

Example 31 : **Four coins are tossed 160 times and the following results were obtained :**

No. of heads	0	1	2	3	4
Observed frequency	17	52	54	31	6

Under the asumption that coins are unbiased, find the expected frequencies of getting 0, 1, 2, 3, or 4 heads and test the goodness of fit.

Solution. On the assumption that the coins are undiased, the expected frequencies of getting 0, 1, 2, 3, and 4 heads will be given by the formula of binomial distribution :

$$f(X) = N.P(X) = N.\,{}^{n}c_{x}\,.q^{n-x}\,.p$$

Here, $p = P(H) = 1/2$, $q = P(T) = 1 - \frac{1}{2} = \frac{1}{2}$, $n = 4$, $N = 160$

No. of Heads	$fe(X) = N \times P(X)$	E
0	$160 \times {}^{4}C_{0}\left(\frac{1}{2}\right)^{4}\cdot\left(\frac{1}{2}\right)^{0}$	= 10
1	$160 \times {}^{4}C_{1}\left(\frac{1}{2}\right)^{3}\cdot\left(\frac{1}{2}\right)^{1}$	= 40
2	$160 \times {}^{4}C_{2}\left(\frac{1}{2}\right)^{2}\cdot\left(\frac{1}{2}\right)^{2}$	= 60
3	$160 \times {}^{4}C_{3}\left(\frac{1}{2}\right)^{1}\cdot\left(\frac{1}{2}\right)^{3}$	= 40
4	$160 \times {}^{4}C_{4}\left(\frac{1}{2}\right)^{0}\cdot\left(\frac{1}{2}\right)^{4}$	= 10

(b) Test of Goodness of Fit : Let us take the hypothesis that there is no difference between the observed frequencies and expected freqeucies.

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
17	10	7	49	4·900
52	40	12	144	3·600
54	60	− 6	36	0·600
31	40	−9	81	2·025
6	10	− 4	16	1·600
				$\Sigma(O-E)^2/E = 12.725$

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 12.725$$

Degrees of freedom $(v) = n - 1 = 5 - 1 = 4$

Tabulated value of $\chi^2_{.05}$ for 4 d.f. $= 9.49$.

Since, the calculated value of χ^2 is greater than the table value, we reject the hypothesis and hence conclude that the fit is poor.

Example 32 : Given the following actual and theoretical frequencies, test the goodness of fit:

Actual Frequency	25	50	75	102
Theoretical Frequency	36	54	72	90

Solution. Let us take the hypothesis that there is no difference in the actual frequencies and theoretical frequencies.

Computation of χ^2

O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
25	36	− 11	121	3·361
50	54	− 4	16	0·296
75	72	+ 3	9	0·125
102	90	+ 12	144	1·600
				$\Sigma(O-E)^2/E = 5.382$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 5.382$$

Degrees of freedom $(v) = n - 1 = 4 - 1 = 3$

For $v = 3$, $\chi^2_{0.05} = 7.815$.

Since, the calculated value of χ^2 is less than the table value, we accept the hypothesis and therefore conclude that the there is no difference in the actual frequencies and theoretical frequencies *i.e.* the fit is good.

Example 33. In a certain sample of 2000 families, 1400 families are consumers of tea. Out of 1800 Hindu families, 1236 families consume tea. Use Chi-square test to test whether there is any significant difference between the consumption of tea among Hindu and Non-Hindu families. Use 5% level of significance.

Solution. The above data can be conveniently arranged in the following table as :

	Hindu	Non-Hindu	Total
No. of families consuming tea	1236	164	1400
No. of families not consuming tea	564	36	600
Total	1800	200	2000

Let the null hypothesis be that there is no significant difference between the consumers of tea among Hindu and Non-Hindu families or that the two attributes (consmption of tea and community) are independent.

On the basis of this hypothesis, the expected frequencies are.

$$E_{11} = \frac{1800 \times 1400}{2000} = 1260$$

Te remaining frequencies are found out by subtracting from the column and row totals.

The expected frequencies table would be as follows:

1260	140	1400
540	60	600
1800	200	2000

Applying χ^2-test :

O	E	$(O-E)^2$	$(O-E)^2/E$
1236	1260	576	0·457
564	540	576	1·067
164	140	576	4·114
36	60	576	9·600
			$\Sigma(O-E)^2/E = 15·238$

$$\therefore \quad \chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 15·238$$

Degrees of freedom $= v = (r-1)(c-1) = (2-1)(2-1) = 1$

For $v=1$, $\chi^2_{.05} = 3·84$

Since, the calculated value of χ^2 is much greater than the table value of χ^2, we reject the null hypothesis and conclude that the two communities differ significantly as regard to the consumption of tea among them.

IMPORTANT POINTS

Chi-square (χ^2) test is used for :

(i) Testing the Significance of Population Variance

χ^2-test is used to test the signficiance of the population variance. The significance is tested by using the formula :

$$\chi^2 = \frac{\Sigma(x-\bar{x})^2}{\sigma^2} \quad \text{or} \quad \frac{ns^2}{\sigma^2} \quad \text{or} \quad \frac{(n-1)\,\bar{s}_1^2}{\sigma^2}$$

Degrees of freedom $v = n-1$

Now the calculated value of χ^2 > tabulated value of χ^2, we reject the null hypothesis H_0.

Otherwise, we accept H_0.

(ii) Testing the independence of attributes in a contingency table of order $r \times c$.

In case of contingency table. we set up the hypothesis that the two attributes are independent and on the basis of this assumption, we calculate expected frequency of each cell with the following formula :

$$\text{Expected frequency (E)} = \frac{\text{Total of row in which it occurs} \times \text{Total of column in which it occurs}}{\text{Total no. of observations}}$$

and finally we calculate

$$\chi^2 = \frac{\Sigma(O-E)^2}{E}$$

Degrees of freedom $= (r-1)(c-1)$

Now, if calculated value of χ^2 < tabulated value of χ^2 at 5% level of significance for $(r-1)(c-1)$ d.f., we accept our hypothesis otherwise reject it and conclude accordingly.

(iii) Testing the goodness of fit.

χ^2-test is used in testing the hypothesis that the observed sample distribution agrees with the theoretical distribution *i.e.*, there is no difference between the observed and expected frequencies. The significance of the difference between observed and expected frequencies are tested as follows :

Given that :

O : $O_1, O_2, O_3, \ldots, O_n$

E : $E_1, E_2, E_3, \ldots, E_n$

We calculate :

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right]$$

d.f. $= k-1$

Now if the calculated value of χ^2 < table of χ^2 for $(k-1)$, d.f., then we accept the hypothesis and conclude accordingly.

QUESTIONS

1. Describe the χ^2-test of significance and state the various uses to which it can be put.
2. (a) What is χ^2-test of goodness of fit ? What precautions are necessary while using this test ?
   (b) What is Chi-square test of independence ? Under what conditions is it applicable ?
3. What is χ^2-test ? Give various uses of χ^2-test. What are the limiting values of χ^2 ? How will you determine the degrees of freedom for χ^2 -test ?
4. Discuss the uses of χ^2-test.
5. Discuss the precautions which should be kept in mind while using χ^2-test.

❀❀❀

6

F-Test and Analysis of Variance

INTRODUCTION

Analysis of Variance (abbreviated as ANOVA) is one of the most powerful techniques of statistical analysis. It was developed by **R.A. Fisher.** Initially, this technique was used in agricultural experiments but now a days it is widely used in natural, social and physical science. This technique is used to test whether the difference between the means of three or more populations is significant or not. By using the technique of analysis of variance, we can test whether the different varieties of seeds or fertilisers applied on different plots of land differ significantly or not as regard their average yields. A manager of a firm may use this technique to test whether there is significant difference in the average sale figures of different salesmen employed by the firm. Analysis of variance thus enables us to test on the basis of sample observations whether the means of three or more population is significantly different or not.

MEANING OF ANALYSIS OF VARIANCE

Analysis of variance is a statistical technique with the help of which the total variation of the data is split up into various components which may be attributed to various "sources" or "causes" of variation. There may be variation between the samples and also within the samples. By comparing the variance between the samples and variance within samples, analysis of variance helps in testing the homogenity of several population means. In the words of **Yule and Kendell, "The analysis of variance is essentially a procedure for testing the difference between different groups of data for homogenity" To quote R.A. Fisher,"Analysis of variance is the separation of the variance ascribible to one group of causes from the variance ascribile to other groups."** Thus, the analysis of variance obtains a measure of the variance within the samples and also variance between the samples and then test the significance of the difference between the means of two or more populations.

ASSUMPTIONS OF ANALYSIS OF VARIANCE

The underlying assumptions for the study of analysis of variance are :

(1) **Normal Population** : All the population from which samples have been drawn are normally distributed.

(2) **Independence of Samples** : The samples are randomly and independently drawn from the population. That is, each of the sample is independent of the other samples.

(3) **Same Population Variance** : The population from where the samples have been taken should have equal variance ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma^2$, say) where σ^2 is unknown.

(4) **Additivity** : The sum of variances of all the components should be equal to the total variance.

The analysis of variance technique is valid under the above assumptions. Otherwise the results will be illusory with less importance.

USES AND UTILITY OF ANALYSIS OF VARIANCE

The following are some of the uses of analysis of variance :

(1) **Test of the significance between the means of several samples :** The analysis of variance is used to test the hypothesis whether the means of several samples are significantly different or not.

(2) **Test of the significance between the variance of two samples :** F-ratio in the analysis of variance is used to test the significance of the difference between the variance of two samples.

(3) **Study of homogenity in case of two-way classification :** Homogenity of data can also be studied in analysis of variance of two-way classification because in this case the data are classified into different parts on two bases.

(4) **Test of correlation and regression :** The analysis of variance is used to test the significance of multiple correlation coefficient. The linearity of regression is also tested with its help.

TECHNIQUE OF ANALYSIS OF VARIANCE

The technique of analysis of variance is studied under the following two headings "

(A) **One way Classification, and**

(B) **Two way Classification.**

(A) **One way Classification :** In one-way classification, the data are classified on the basis of one factor or criterion only. For example, the yields of several plots of land may be classified according to different types of seeds, fertilisers, etc. In case of one-way classification, the analysis of variance can be done by the following methods :

(1) **Direct Method**

(2) **Short-cutn Method**

(3) **Coding Method**

(1) **Direct Method :** Under direct method, the following steps are followed :

(i) **Null Hypothesis** , $H_0 : \mu_1 = \mu_2 = \ldots . \mu_k$ *i.e.*, the means of the population from which the samples have been taken are equal and there is no difference among them.

(ii) **Variance between the samples :** Compute the mean ($\bar{x}$) of each sample. Find the combined mean ($\bar{\bar{x}}$) of all the sample means. Take deviations from $\bar{\bar{x}}$ *i.e.*, compute $\bar{x} - \bar{\bar{x}}$ and then square these deviations $(\bar{x} - \bar{\bar{x}})^2$. Find the sum of these squared deviations and divide it by the corresponding degrees of freedom $(k-1)$, where k is the number of samples. Thus, we find the variance between the samples. Symbolically,

Sum of squares of the deviations between samples (SSB)

$$= n_1 (\bar{x}_1 - \bar{\bar{x}}_2)^2 + n_2 (\bar{x}_1 - \bar{\bar{x}}_2)^2 \ldots + n_k (\bar{x}_k - \bar{\bar{x}})^2$$

Degrees of freedom, $v_1 = k - 1$

$\therefore$ Variance between samples = (MSB) = $\frac{SSB}{k-1}$

(iii) Variance within samples : Take the deviations in each sample from the respective sample means, $x_1-\bar{x}_1, x_2-\bar{x}_2 \ldots$ and find their squares, $(x_1-\bar{x}_1)^2, (x_2-\bar{x}_2)^2$. Divide the sum of these squares of deviations by relevant degrees of freedom $v_2 = N-k$, where N is the total number of observations. Thus, we find the variance with samples. Symbolically,

Sum of squares of the deviation within samples (SSW)

$$=(x_1-\bar{x}_1)^2+(x_2-\bar{x}_2)^2+\ldots+(x_k-\bar{x}_k)^2$$

Degrees of freedom, $v_2 = N-k$

Variance within the samples (MSW) $= \dfrac{SSW}{N-k}$

(iv) Analysis of Variance Table : The results of the above calculations is presented in a table, called **Analysis of variance** or ANOVA table as follows :

ANOVA Table

Sources of Variation	Sum of squares	Degrees of freedom (d.f.)	Mean sum of squares (MSS)	F-ratio
Between Samples	$\Sigma n_k\left(\bar{x}_k-\bar{\bar{x}}\right)^2$ (SSB)	$k-1$	$\dfrac{SSB}{k-1}=$ MSB	$F=\dfrac{MSB}{MSW}$
Within Samples	$\Sigma\left(x-\bar{x}_k\right)^2$ (SSW)	$N-k$	$\dfrac{SSW}{N-k}=MSW$	
Total	$\Sigma\left(x-\bar{\bar{x}}\right)^2$ (TSS)	$N-1$		

(v) The calculated value of F is compared with the table value of F for $(k-1, N-k)$ d.f. at a specified level of significance. If the calculated value of F is less than the table value of F, we accept the null hypothesis and conclude that all population means are equal, otherwise they may be taken to be unequal.

The following examples illustrate the procedure involved under direct method :

Example 1. **Three varieties *A*, *B* and *C* of wheat are sown in four plots each and the following yields per acre were obtained :**

Plots	Varieties		
	A	*B*	*C*
1	8	7	12
2	10	5	9
3	7	10	13
4	14	9	12
5	11	9	14

Set up a table of analysis of variance and find out where there is a significant difference between the mean yields of these varieties. (Given $F_{.05} = 4{\cdot}26, 3{\cdot}88$ and $3{\cdot}88$ at d.f. (9, 2) (3, 9) and (2, 12) respectively).

Solution. Null hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ *i.e.*, mean yields of three varieties are the same.

Computation of Arithmetic Mean

X_1	X_2	X_3
8	7	12
10	5	12
7	10	9
14	9	13
11	9	12
$\Sigma X_1 = 50\ n_1 = 5$	$\Sigma X_2 = 40\ n_2 = 5$	14
		$\Sigma X_3 = 60\ n_3 = 5$

$$\bar{x}_1 = \frac{50}{5} = 10, \quad \bar{x}_2 = \frac{40}{5} = 8, \quad \bar{x}_3 = \frac{60}{5} = 12$$

Grand Mean, or $\bar{\bar{x}} = \dfrac{10+8+12}{3} = 10$

Variance between Samples

Sum of squares of the deviations between samples (SSB)

$$SSB = n_1(\bar{x}_1-\bar{\bar{x}})^2 + n_2(\bar{x}_2-\bar{\bar{x}})^2 + n_3(\bar{x}_3-\bar{\bar{x}})^2$$
$$= 5(10-10)^2 + 5(8-10)^2 + 5(12-10)^2$$
$$= 5\times 0 + 5\times 4 + 5\times 4 = 40$$

Degrees of freedom, $v_1 = k-1 = 3-1 = 2$

Variance between samples (MSB) $= \dfrac{SSB}{k-1} = \dfrac{40}{2} = 20$

Variance within Samples

X_1	$(X_1-\bar{x}_1)$	$(X_1-\bar{x}_1)^2$	X_2	$(X_2-\bar{x}_2)$	$(X_2-\bar{x}_2)^2$	X_3	$(X_3-\bar{x}_3)$	$(X_3-\bar{x}_3)^2$
8	− 2	4	7	− 1	1	12	0	0
10	0	0	5	− 3	9	9	− 3	9
7	− 3	9	10	2	4	13	+ 3	9
14	+ 4	16	9	1	1	12	+ 0	0
11	+ 1	1	9	1	1	14	+ 2	4
$\bar{x}_1 = 10$		$\Sigma(X_1-\bar{x}_1)^2 = 30$	$\bar{x}_2 = 8$		$\Sigma(X_2-\bar{x}_2)^2 = 16$	$\bar{x}_3 = 12$		$\Sigma(X_3-\bar{x}_3)^2 = 14$

Sum of the squares of the deviations within samples (SSW)

$$SSW = \Sigma(x_1-\bar{x}_1)^2 + \Sigma(x_2-\bar{x}_2)^2 + \Sigma(x_3-\bar{x}_3)^2$$
$$= 30+16+14 = 60$$

Degrees of freedom, $v_2 = N-k = 15-3 = 12$.

Variance with samples (MSW) $= \dfrac{SSW}{N-k} = \dfrac{60}{12} = 5$

The results of the above calculation is presented in a table called ANOVA table as follows :

ANOVA Table

Sources of Variation	Sum of squares	Degrees of freedom (d.f.)	Mean sum of squares (MSS)	F-Ratio
Between Samples	40 (SSB)	2	$\frac{40}{2}=20$ (MSB)	$F=\frac{20}{5}=4$
Within Samples	60 (SSW)	12	$\frac{60}{12}=5$ (MSW)	
Total	100 (TSS)	14		

For $v_1=2, v_2=12$, the table value of F at 5% level os significance is 3·88. Since, the computed value of F is greater than the table value *i.e.*, $F>F_{0.05}$, we reject null hypothesis and conclude that the difference between the mean yields of 3 varieties is significant.

(2) Short cut Method : The direct method is much calculative and time consuming and moreover, the calculation becomes more complicated when the arithmetic mean is not in whole number. In such a case, short-cut method is used. It involves the following steps :

(i) **Find the sum of all sample observations and their squares :**
Sum of the sample values $=\Sigma X_1,\ \Sigma X_2, \Sigma X_3, \ldots \Sigma X_k$
Sum of the squares of sample values $=\Sigma X_1^2, \Sigma X_2^2, \Sigma X_3^2, \ldots \Sigma X_k^2$

(ii) **Find the correction factor :** To obtain the correction factor, divide the square of the total of all values by the number of values *i.e.*,

$$C.F.=\frac{T^2}{N}$$

where, $C.F.$ = Correction factor, T^2 = Square of the total units of the samples
N = Total no. of units of the samples

(iii) **Find the total sum of squares, TSS :** To find the total sum of squares subtract correction factor from the sum of the squares of all samples values *i.e.*,

$$TSS=(\Sigma X_1^2+\Sigma X_2^2+\ldots \Sigma X_k^2).-\frac{T^2}{N}$$

(iv) **Find the sum of squares between samples, SSB :** To find the SSB, divide the sum of the squares of each samples by their size and then find their sum. Subtract the correction factor from this sum *i.e.*,

$$SSB=\left[\frac{(\Sigma X_1)^2}{n_1}+\frac{(\Sigma X_2)^2}{n_2}+\ldots+\frac{(\Sigma X_k)^2}{n_k}\right]-\frac{T^2}{N}$$

(v) **Find the sum of squares within samples, SSW :** It is obtained by deducting the sum of squares between the samples from the total sum of squares *i.e.*,

$$SSW = TSS - SSB$$

(vi) **Analysis of Variance Table and Interpretation of Significance :** Analysis of variance table and interpretation are the same as in case of direct method.

Note : In case sample sizes are unequal, there is no change in the analysis of variance. Utmost care must be taken while calculating degrees of freedom in such cases.

Example 2. **Three varieties *A*, *B* and *C* of wheat are sown in four plots each and the following yields per acre were obtained.**

Plots	Varieties		
	A	*B*	*C*
1	8	7	12
2	10	5	9
3	7	10	13
4	14	9	12
5	11	9	14

Is there any significant difference in the production of three varieties ? Use short cut method.

Solution. Null hypothesis, $H_0 : \mu_1=\mu_2=\mu_3$ *i.e.*, there is no difference between the mean yield of three varieties.

A		*B*		*C*	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
8	64	7	49	12	144
10	100	5	25	9	81
7	49	10	100	13	169
14	196	9	81	12	144
11	121	9	81	14	196
$\Sigma X_1=50$ $n_1=5$	$\Sigma X_1^2=530$	$\Sigma X_2=40$ $n_2=5$	$\Sigma X_2^2=336$	$\Sigma X_3=60$ $n_3=5$	$\Sigma X_3^2=734$

$$T=\Sigma X_1+\Sigma X_2+\Sigma X_3=50+40+60=150$$

Correction Factor, $$C.F.=\frac{T^2}{N}=\frac{(150)^2}{15}=1500$$

TSS = Total sum of squares
$=(\Sigma X_1^2+\Sigma X_2^2+\Sigma X_3^2)-C.F.$
$=530+336+734-1500=100$

SSB = Sum of squares between samples
$$=\left[\frac{(\Sigma X_1)^2}{n_1}+\frac{(\Sigma X_2)^2}{n_2}+\frac{(\Sigma X_3)^2}{n_3}\right]-C.F.$$
$$=\left[\frac{(50)^2}{5}+\frac{(40)^2}{5}+\frac{(50)^2}{5}\right]-1500$$
$$=\frac{1}{5}\cdot[2500+1600+3600]-1500=1540-1500=40$$

$SSW = TSS - SSB = 100 - 40 = 60.$

ANOVA Table

Sources of Variation	Sum of squares	Degrees of freedom (d.f.)	Mean sum of square (MSS)	F-Ratio
Between samples	40 (SSB)	2	20	$F=\frac{20}{5}=4$
Within Samples	60 (SSW)	12	5	
Total	100 (TSS)	14		

For $v_1=2, v_2=12$, the table value of F at 5% level of significance is 3·88. Since the calculated value of F is greater than the table value *i.e.*, $F>F_{0.05}$, we reject the null hypothesis and hence, conclude that the difference between the mean yields of three varieties is significant.

(3) Coding Method : The short-cut method becomes tedious when the magnitude of the given values is large. Coding method simplified the calculations involved in the short-cut method and is popularly used in practice. Coding refers to the addition, subtraction, multiplication or division of data by a constant quantity. As the F-statistic in the analysis of variance is a ratio, its value does not change if all the given values are coded *i.e.*, either multiplied or divided by a common factor or if a common figure is either subtracted or added to each of the given values. By this method, big figures are reduced in magnitude by subtraction or division and the work in simplified without altering the value of F. Analysis of variance table and interpretation are the same as in case of short-cut method.

The following examples illustrate the coding method :

Example 3 **The following table gives the yields of four varieties of wheat grown in 2 plots :**

Plots	Varieties			
	A	*B*	*C*	*D*
1	200	230	250	300
2	190	270	300	270
3	240	150	145	180

Is there any significant difference in the production of these varieties ?

Solution. Null hypothesis, $H_0: \mu_1=\mu_2=\mu_3=\mu_4$ *i.e.*, there is no significant difference in the mean yield of four varieties.

In order to simplify the calculation, subtract 200 from each sample value and dividing the difference by 10.

Coded Data

A		*B*		*C*		*D*	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
0	0	3	9	5	25	10	100
−1	1	7	49	10	100	7	49
4	16	−5	25	−5·5	30·25	−2	4
$\Sigma X_1=3$, $n_1=3$	$\Sigma X_1^2=17$	$\Sigma X_2=5$, $n_2=3$	$\Sigma X_2^2=83$	$\Sigma X_3=9·5$, $n_3=3$	$\Sigma X_3^2=155·25$	$\Sigma X_4=15$, $n_4=3$	$\Sigma X_4^2=153$

$$T=\Sigma X_1+\Sigma X_2+\Sigma X_3+\Sigma X_4=3+5+9·5+15=32·5$$

Correction Factor, $$C.F.=\frac{T^2}{N}=\frac{(32·5)^2}{12}=88·02$$

TSS = Total sum of squares

$$=[\Sigma X_1^2+\Sigma X_2^2+\Sigma X_3^2+\Sigma X_4^2]-C.F.$$

$$=[17+83+155·25+153]-88·02=320·23$$

SSB = Sum of squares between the samples

$$=\left[\frac{(\Sigma X_1)^2}{n_1}+\frac{(\Sigma X_2)^2}{n_2}+\frac{(\Sigma X_3)^2}{n_3}+\frac{(\Sigma X_4)^2}{n_4}\right]-C.F.$$

$$=\left[\frac{(3)^2}{3}+\frac{(5)^2}{3}+\frac{(9·5)^2}{3}+\frac{(15)^2}{3}\right]-88·02=28·39$$

SSW = SST − SSB = 320·23 − 28·39 = 291·84

ANOVA Table

Sources of Variation	Sum of squares	Degrees of freedom (d.f.)	Mean sum of square (MSS)	F-Ratio
Between samples	28·39	3	9·46	$F=\frac{36·48}{9·46}=3·85$
Within samples	291·84	8	36·48	
Total	320·23	11		

For $v_1=8, v_2=3$, the table value of F at 5% level of significance is 4·07. Since the calculated value of F is less than the table value *i.e.*, $F<F_{0.05}$, we accept the null hypothesis and conclude that there is no significant difference in the mean yield of four varieties.

Example 4. **The following figures relate to producting in kg of three varieties *A*, *B* and *C* of wheat sown in 12 plots :**

A	14	16	18		
B	14	13	15	22	
C	18	16	19	19	20

Is there any significant difference in the production of these varieties ?

Solution. Null Hypothesis, $H_0: \mu_1=\mu_2=\mu_3$ *i.e.*, there is no difference in the production of three varieties.

In order to simplify the calculations, the given data are coded by subtracting 12 from each figure. The deviations and their squares are as follows :

Coded Data

A		*B*		*C*	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
2	4	2	4	6	36
4	16	1	1	4	16

6	36	3	9	7	49
—	—	10	100	7	49
—	—	—	—	8	64
$\Sigma X_1 = 12$, $n_1 = 3$	$\Sigma X_1^2 = 56$	$\Sigma X_2 = 16$, $n_2 = 4$	$\Sigma X_2^2 = 144$	$\Sigma X_3 = 32$, $n_3 = 5$	$\Sigma X_3^2 = 214$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 12 + 16 + 32 = 60$$

$$C.F. = \frac{T^2}{N} = \frac{(60)^2}{12} = 300$$

TSS = Total sum of squares

$$= [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - C.F.$$

$$= 56 + 144 + 214 - 300 = 84$$

SSB = Sum of squares between samples

$$= \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}\right] - C.F.$$

$$= \left[\frac{(12)^2}{3} + \frac{(16)^2}{4} + \frac{(32)^2}{5}\right] - 300$$

$$= 48 + 64 + 204 \cdot 8 - 300 = 16 \cdot 8$$

SSW = TSS − SSB

$$= 84 - 16 \cdot 8 = 67 \cdot 20$$

ANOVA Table

Source of Variation	Sume of squares	Degrees of freedom (d.f.)	Mean sum (MSS)	F-Statistic
Between Samples	16·8	2	8·4	$F = \frac{8 \cdot 4}{7 \cdot 467} = 1 \cdot 125$
Within samples	67·20	9	7·467	
Total	84	11	—	

For $v_1 = 2, v_2 = 9$, the table value of F at 5% level of significance is 4·261. Since, the calculated value of F is less than the table value of F. We accept the null hypothesis and conclude that there is no difference in the mean productivity of three varieties.

EXERCISE – 1

1. The following table gives the yields on 15 sample plots under three varieties of seed :

Seeds / Plots	P_1	P_2	P_3	P_4	P_5
S_1	20	21	23	16	20
S_2	18	20	17	15	25
S_3	25	28	22	28	32

Is the difference between varieties significant.
For (2, 12) degrees of freedom, $F_{0 \cdot 05} = 3 \cdot 88$. [Ans. $F = 8 \cdot 14$, Reject H_0]

2. The following data gives the retail prices of a commodity in some shops selected at random in four cities :

City	Prices (Rs./Kg)			
A	22	24	27	23
B	20	19	23	19
C	10	17	21	18
D	24	25	29	26

Carry out the analysis of variance to test the significance of the difference between the prices of the commodities in four cities
For (3, 12) degrees of freedom, $F_{0 \cdot 05} = 3 \cdot 49$
For (3, 9) degrees of freedom, $F_{0 \cdot 05} = 3 \cdot 86$ [Ans. $F = 2 \cdot 114$, Reject H_0]

3. Three varieties A, B and C of wheat were sown in four plots each and the following yields per acre were obtained :

Plots of Land	Varieties of Wheat		
	A	B	C
1	10	9	4
2	6	7	7
3	7	7	7
4	9	5	6

Set up a table of analysis of variance and find out whether there is a significant difference between the mean yield of three varieties (Given $F_{0 \cdot 05} = 4 \cdot 25$, 3·86 and 4·10 at d.f. (2, 9), (3, 9) and (2, 10) respectively). [Ans. $F = 1 \cdot 771 < F_{0 \cdot 05}$, Accept H_0]

4. A test was given to 5 students chosen at random from M. Com. class of each of the three universities in Haryana. Their scores were found as follows :

University	Scores				
A	90	70	60	50	80
B	70	40	50	40	50
C	60	50	60	70	60

Perform analysis of variance and show if there is any significant difference between the scores of students in the three universities. [Given F value at 5% = 3·89]
[Ans. $F = 3 \cdot 33$, Accept H_0]

5. The following figures relate to the number of units sold in five different areas by four salesmen :

Area	Number of units			
	A	B	C	D
1	80	100	95	70
2	82	110	90	75
3	88	105	100	82
4	85	115	105	88
5	75	90	80	65

Is there a significant difference in the efficiency of these salesmen ?

Hint : See Example 11.
Table values of $F_{0.05}$ for $v_1 = 3, v_2 = 16$ is 3·24.
[**Ans.** $F = 10{\cdot}61 > F_{0.05}$, Reject H_0 *i.e.*, there is a significant difference in the efficiency of the four salesmen]

6. The Amrit Vanaspati Company of Rajpura (Punjab) wishes to test whether its three salesmen *A*, *B* and *C* tend to make sales of the same size or whether they differ in their selling ability as measured by the average size of their sales. During the last week of October, 2004, There have been 14 sales calls – *A* made 5 calls, *B* made 4 calls and C made 5 calls.
Following are the weekly sales record of the three salesmen :

A (Rs.)	300	400	300	500	000
B (Rs.)	600	300	300	400	—
C (Rs.)	700	300	400	600	500

Perform the analysis of variance and draw your conclusions.
[Given $F_{0.05}(2, 11) = 3{\cdot}98$; $F_{0.05}(2, 13) = 3{\cdot}82$] [**Ans.** $F = 1{\cdot}83$, Accept H_0]

7. Yields of 3 varieties of wheat in 3 blocks are given below :

Blocks/Varieties	1	2	3
I	200	230	300
II	190	270	270
III	240	150	180

Is the difference between varieties significant ? [**Ans.** $F = 3{\cdot}84$, Accept H_0]

8. The following figures relate to the production in Kg of three varieties *A*, *B* and C on wheat sown in 12 plots :

A	122	128	124	126		
B	114	116	118	114	106	
C	130	128	124			

(Use Coding Method)
Is there any significant difference in the production of three varieties ?
[**Ans.** $F = 16{\cdot}90 > F_{0.05}$, Reject H_0]

9. Apply *F*-test on the following data :

X_1	X_2	X_3
25	31	24
30	39	30
36	38	28
38	42	25
31	35	28
160	185	135

(Hints : See Example 15)
Given ($F_{2, 12}$, 5 percent = 3·89) [**Ans.** $F = 7{\cdot}49$, Reject H_0]

10. To test the significance of the variation of the retail prices of a commodity in three principal cities : Bombay, Kolkata and Delhi, four shops were chosen at random in each city and prices observed in Rs. were as follows :

Bombay	16	8	12	14
Kolkata	14	10	10	6
Delhi	4	10	8	8

Do the data indicate that the prices in the three cities are significantly different ?
[**Ans.** $F = 2{\cdot}616$ Accept H_0]

11. To assess the significance of possible variation are in a performance in a certain test as between the grammar schools of a city, a common test was given to students take at random from the senior fifth form of each of the four schools. Carry out analysis of variance of the data and comment upon the results.

School	Marks Obtained by the students							
A	8	7	4	5	5	5	6	6
B	7	5	5	4	3	4	6	4
C	5	3	4	4	3	5	4	4
D	10	5	6	4	8	7	8	4

Given : $F_{0.05} = 2{\cdot}95$ [**Ans.** $F = 5{\cdot}10$, Reject H_0]

(B) ANALYSIS OF VARIANCE IN TWO-WAY CLASSIFICATION

In two-way classification, the data are classified according to two factors. For example, the production of a manufacturing concern can be studied on the basis of workers as well as machines. A company can analyse its sales according to salesmen and seasons. In two-way classification, the following procedure is adopted in the analysis of variance :

(i) Coding method can be used to simplify the calculation.

(ii) Find the correction factor by using the formula :

$$\text{Correction Factor } (C.F.) = \frac{T^2}{N}$$

Where, T = Grand total of all the values in all the samples, N = Total number of items.

(iii) Find total sum of squares (TSS) : It is obtained by subtracting the correction factor from the total of squared values of the sample *i.e.*,

$$TSS = [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 + \ldots \Sigma X_k^2] - \frac{T^2}{N}$$

(iv) Find the sum of squares between columns (SSC) : The total of each column is squared and divided by the number of items in the column. The correction factor is subtracted from it and SSC is obtained *i.e.*,

$$SSC = \Sigma\left(\frac{\Sigma X_c}{n_c}\right)^2 - \frac{T^2}{N}$$

Where, ΣX_c^2 = total of squared values in each column; n_c = number of items in each column.

(v) Find the sum of squares between rows (SSR) : The total of each row is squared and divided by the number of items in respective rows. The correction factor is subtracted from the total of, thus, arrived row and SSR is obtained, *i.e.*,

$$SSR = \Sigma\left(\frac{\Sigma X_r^2}{n_r}\right) - \frac{T^2}{N}$$

Where, ΣX_r^2 = Total of squared values in each row; n_r = number of items in each row.

(vi) Find the sum of the squares of the residual (SSE) : The sum of the squares of the residual is obtained by deducting the SSC and SSR from TSS. Thus

$$SSE = TSS - SSC - SSR$$

(vii) Find the number of degrees of freedom by using the formula :

No. of degrees of freedom between columns $= (c-1)$

No. of degrees of freedom between rows $= (r-1)$

No. of degrees of freedom for residual $= (c-1)(r-1)$

Total no. of degrees of freedom $= N-1$ or $cr-1$

(viii) ANOVA Table : In a two-way classification, the analysis of variance (ANOVA) table is prepared in the following way :

ANOVA Table (Two-way Classification)

Source of variation	sum of square	Degrees of freedom	Mean Sum of squares (MSS)	F-Ratio
Between columns	SSC	$(c-1)$	SSC $\div (c-1)$ = MSC	$F = \frac{MSC}{MSE}$
Between Rows	SSR	$(r-1)$	SSR $\div (r-1)$ = MSR	$F = \frac{MSR}{MSE}$
Residual	SSE	$(c-1)(r-1)$	SSE $\div (c-1)\times(r-1)$ = MSE	
Total	TSS	$(N-1)$ or $(cr-1)$		

(ix) Interpretation : The calculated value of F is compared with the table value of F and if the calculated value of F is greater than the table value at a specified level of significance, the null hypothesis is rejected and concluded that the difference is significant otherwise vice versa.

Example 5. **The following data represent the number of units of a commodity produced by 3 different workers using 3 different machines :**

Machine / Workers	A	B	C
X	16	64	40
Y	56	72	56
Z	12	56	28

Test (i) Whether the mean productivity is the same for the different machine types (ii) whether the three workers differ with regard to mean productivity.

Solution. Let us take the hypothesis that :

(i) The mean productivity for three different machines is the same.

(ii) Three workers do not differ with respect to their mean productivity.

Machine Workers	Data			Row Total	Squares of Data		
	X_1	X_2	X_3		X_1^2	X_2^2	X_3^2
X	16	64	40	120	256	4096	1600
Y	56	72	56	184	3136	5184	3136
Z	12	56	28	96	144	3136	784
Column Total	84	192	1245	$T = 400$	3536	12416	5520

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 84 + 192 + 14 = 400$$

$$C.F. = \frac{T^2}{N} = \frac{(400)^2}{9} = 17777 \cdot 78$$

$$TSS = [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - C.F.$$
$$= [3536 + 12416 + 5520] - 17777 \cdot 78$$
$$= 21472 - 17777 \cdot 78 = 3694 \cdot 22$$

$$SSC = \left[\frac{(84)^2}{3} + \frac{(192)^2}{3} + \frac{(124)^2}{3}\right] - 17777 \cdot 78 \text{ (C.F.)}$$
$$= \frac{1}{3}[7056 + 36864 + 15376] - 17777 \cdot 78$$
$$= 19765 \cdot 33 - 17777 \cdot 78 = 1987 \cdot 55$$

$$SSR = \left[\frac{(120)^2}{3} + \frac{(184)^2}{3} + \frac{(96)^2}{3}\right] - 17777 \cdot 78 \text{ (C.F.)}$$
$$= \frac{1}{3}[14400 + 33856 + 9216] - 17777 \cdot 78$$
$$= 19157 \cdot 33 - 17777 \cdot 78 = 1379 \cdot 55$$

$$SSE = TSS - SSC - SSR$$
$$= 3694 \cdot 22 - 1987 \cdot 55 - 1379 \cdot 55 = 327 \cdot 12$$

Degrees of freedom are

$TSS = N - 1 = 9 - 1 = 8$

$SSC = c - 1 = 3 - 1 = 2$

$SSR = r - 1 = 3 - 1 = 2$

$SSE = (c-1)(r-1) = 2 \times 2 = 4$

ANOVA Table

Source of variation	Sum of square	Degrees of freedom	Mean Sum of squares (MSS)	F-Ratio
Between columns (Machines)	1987·55	2	$\frac{1987 \cdot 55}{2} = 993 \cdot 77$	$F = \frac{993 \cdot 73}{81 \cdot 78} = 12 \cdot 15$
Between Rows (Workers)	1379·55	2	$\frac{1379 \cdot 55}{2} = 689 \cdot 77$	$F = \frac{689 \cdot 77}{81 \cdot 78} = 8 \cdot 43$
Residual/Error	327·12	4	$\frac{327 \cdot 12}{4} = 81 \cdot 78$	
Total	3694·22	8		

Interpretation

For Machines

(i) For $v_1 = 2, v_2 = 4$, the table value of $F_{0.05} = 6{\cdot}94$.

Since the calculated value of F is greater than the critical value of F, the null hypothesis is rejected. Hence the mean productivity is not the same for three different machines.

For Workers

(ii) For $v_1 = 2$ and $v_2 = 4$, the table value of $F_{0.05} = 6{\cdot}94$.

Since the calculated value of F is greater than the critical value of F, the null hypothesis is rejected. Hence the workers differ with regard to mean productivity.

Example 6. **The following table gives the number of refrigerators sold by 4 salesmen in three seasons-summer, winter and monsoon :**

Season	Salesmen			
	A	*B*	*C*	*D*
Summer	62	62	32	60
Winter	46	48	52	54
Monsoon	42	46	48	48

Is there a significant difference in the sales made by the four salesmen ?
Is there a significant difference in the sale made during different seasons ?

Solution. (i) Let us take the null hypothesis that the mean sales made by the four salesmen is the same.

(ii) The sales do not differ with regard to seasons.

In order to simplify calculations, the given data is coded by subtracting 50 from each observation. The data in the coded form are given below :

Season	Coded Data				Row Total	Squares of Coded Data			
	X_1	X_2	X_3	X_4		X_1^2	X_2^2	X_3^2	X_4^2
S	+ 12	+ 12	− 18	+ 10	+ 16	144	144	324	100
W	− 4	− 2	+ 2	+ 4	+ 0	16	4	4	16
M	− 8	− 4	− 2	− 2	− 16	64	16	4	4
Column Total	0	+ 6	− 18	+ 12	$T = 0$	224	164	332	120

$$T = 0 + 6 - 18 + 12 = 0$$

$$\text{C.F.} = \frac{T^2}{N} = \frac{(0)^2}{12} = 0$$

$$\text{TSS} = \Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2 - \text{C.F.}$$

$$= 224 + 164 + 332 + 120 - 0 = 840$$

$$\text{SSC} = \left[\frac{(0)^2}{3} + \frac{(6)^2}{3} + \frac{(-18)^2}{3} + \frac{(12)^2}{3}\right] - 0 \quad \text{(C.F.)}$$

$$= 0 + 12 + 108 + 48 - 0 = 168$$

$$\text{SSR} = \left[\frac{(16)^2}{4} + \frac{(0)^2}{4} + \frac{(-16)^2}{4}\right] - 0 \quad \text{(C.F.)}$$

$$= 64 + 0 + 64 - 0 = 128$$

$$\text{SSE} = \text{TSS} - \text{SSC} - \text{SSR} = 840 - 168 - 128 = 544$$

Degrees of freedom are :

$$\text{TSS} = N - 1 = 12 - 1 = 11$$
$$\text{SSC} = (c - 1) = (4 - 1) = 3$$
$$\text{SSR} = (r - 1) = (3 - 1) = 2$$
$$\text{SSE} = (c - 1)(r - 1) = 3 \times 2 = 6$$

The ANOVA table is shown as :

ANOVA Table

Source of variation	Sum of squares	Degrees of freedom	Mean Sum of squares (MSS)	F-Ratio
Between Columns (Salesmen)	168	3	$\text{MSC} = \frac{168}{3} = 56$	$F = \frac{\text{MSC}}{\text{MSE}} = \frac{90{\cdot}67}{56} = 1{\cdot}619$
Between Rows (Seasons)	128	2	$\text{MSR} = \frac{128}{2} = 64$	$F = \frac{\text{MSR}}{\text{MSE}} = \frac{90{\cdot}67}{64} = 1{\cdot}4167$
Residual/Error	544	6	$\text{MSE} = \frac{544}{6} = 90{\cdot}67$	
Total	840	11		

Interpretation

(i) For Salesmen : The calculated value of $F = 1{\cdot}619$

Table value of F for (6, 3) d.f. $= 8{\cdot}94$

Since the calculated value of F is less than the table value of F at 5% level of significance, the null hypothesis is accepted and it can be concluded that there is no difference in the sales of the four salesmen.

(ii) For seasons : The calculated value of $F = 1{\cdot}4167$

Table value for (2, 6) d.f. $F_{0.05} = 5{\cdot}14$

Since the calculated value of F is less than the table value of F at 5% level of significance, the null hypothesis is accepted and it can be concluded that all seasons are similar so far as sales is concerned.

Example 7. **Four observers determine the moisture content of samples of a powder, each man taking a sample from each of six consignments. Their assessments are given below :**

Observers	Consignments					
	1	2	3	4	5	6
1	9	10	9	10	11	11
2	12	11	9	11	10	10
3	11	10	10	12	11	10
4	12	13	11	14	12	10

Analysis the data and discuss whether there is any significant difference between consignments or between observers.
(Given that $F_{0.05}^{(3,15)} = 3.29$, $F_{0.05}^{(5,15)} = 2.90$)

Solution. Let us take the hypothesis that :
(a) There is no significant difference between consignments.
(b) There is no significant difference between observers.
In order to simplify the calculations, the given data are coded by subtracting 10 from each figure. The deviations and their squares are as follows :

Observers	Coded Data						Row Total	Squares of Coded Data					
	X_1	X_2	X_3	X_4	X_5	X_6		X_1^2	X_2^2	X_3^2	X_4^2	X_5^2	X_6^2
1	−1	0	−1	0	1	1	0	1	0	1	0	1	1
2	2	1	−1	1	0	0	3	4	1	1	1	0	0
3	1	0	0	2	1	0	4	1	0	0	4	1	0
4	2	3	1	3	2	0	12	4	9	1	16	4	0
Column Total	4	4	−1	7	4	1	$T = 19$	10	10	3	21	6	1

$$T = 4+4-1+7+4+1 = 19$$

$$\text{C.F.} = \frac{T^2}{N} = \frac{(19)^2}{24} = 15.04$$

$$\text{TSS} = [10+10+3+21+6+1] - 15.04 \quad (\text{C.F.}) = 35.96$$

$$\text{SSC} = \left[\frac{(4)^2}{4} + \frac{(4)^2}{4} + \frac{(-1)^2}{4} + \frac{(7)^2}{4} + \frac{(4)^2}{4} + \frac{(1)^2}{4}\right] - 15.04\,(\text{C.F.}) = 9.71$$

$$\text{SSR} = \left[\frac{(0)^2}{6} + \frac{(3)^2}{6} + \frac{(4)^2}{6} + \frac{(12)^2}{6}\right] - 15.04\,(\text{C.F.}) = 13.13$$

$$\text{SSE} = \text{TSS} - \text{SSC} - \text{SSR} = 35.96 - 9.71 - 13.13 = 13.12$$

The ANOVA table is shown as :

ANOVA Table

Source of variation	Sum of square	Degrees of freedom	Mean Sum of squares (MSS)	F-Ratio
Between columns (Between consignments)	9.71	6 − 1 = 5	$\frac{9.71}{5} = 1.94$	$F = \frac{1.94}{0.87} = 2.23$
Between Rows (Between Observers)	13.13	4 − 1 = 3	$\frac{13.13}{3} = 4.38$	$F = \frac{4.38}{0.87} = 5.03$
Residual/Error	13.12	23 − 8 = 15	$\frac{13.12}{15} = 0.87$	
Total	35.96	24 − 1 = 23		

Interpretation
(i) Between Consignments : The calculated value of $F = 2.23$
Table value of F of (5, 15) d.f. at 5% l.o.s. $= 2.90$
Since, the calculated value of F is less than the table value of F at 5% l.o.s., the null hypothesis is accepted and it can be concluded that there is no difference between consignments.
(ii) Between Observers : The calculated value of $F = 5.03$
Table value of F for (3, 15) d.f. at 5% l.o.s. $= 3.29$
Since, the calculated value of F is greater than the table value of F at 5% l.o.s., the null hypothesis is rejected and it can be concluded that there is significant difference between the observers.

Example 8. **Perform a two way ANOVA on the data given below :**

Plots of Land	Treatment			
	A	B	C	D
P	45	40	38	37
O	43	41	45	38
R	39	39	41	41

(Use coding method subtracting 40 from the given numbers)

Solution. Let us take the null hypothesis that there is no significant difference in the treatment and plots of land. By subtracting 40 from the given numbers, the deviations and their squares are given below :

Plots	Coded Data				Row Total	Squares of Coded Data			
	X_1	X_2	X_3	X_4		X_1^2	X_2^2	X_3^2	X_4^2
P	5	0	−2	−3	0	25	0	4	9
Q	3	1	5	−2	7	9	1	25	4
R	−1	−1	1	1	0	1	1	1	1
Column Total	7	0	4	−4	$T = 7$	35	2	30	14

$$T = 7+0+4-4 = 0$$

$$\text{C.F.} = \frac{T^2}{N} = \frac{7^2}{12} = \frac{49}{12} = 4.083$$

$$\text{TSS} = 35+2+30+14-4.083\,(\text{C.F.})$$
$$= 81 - 4.083 = 76.917$$

$$\text{SSC} = \left[\frac{(7)^2}{3} + \frac{(0)^2}{3} + \frac{(4)^2}{3} + \frac{(-4)^2}{3}\right] - 4.083\,(\text{C.F.})$$

$$= 22.917\,.$$

$$SSR = \frac{(0)^2}{4} + \frac{(7)^2}{4} + \frac{(0)^2}{4} - 4 \cdot 083 = 8 \cdot 167$$

$$SSE = TSS - SSC - SSR = 76 \cdot 917 - 22 \cdot 917 - 8 \cdot 167 = 45 \cdot 333$$

ANOVA Table

Source of variation	Sum of square	Degrees of freedom	Mans Sum of squares (MSS)	F-Ratio
Between Columns (Between Treatment)	22·917	3	$\frac{22 \cdot 913}{3} = 7 \cdot 639$	$F = \frac{7 \cdot 639}{7 \cdot 639} = 1$
Between Rows (Between Fields)	8·167	2	$\frac{8 \cdot 167}{2} = 4 \cdot 083$	$F = \frac{7 \cdot 639}{4 \cdot 083} = 1 \cdot 87$
Residual/Error	45·833	6	$\frac{45 \cdot 833}{6} = 7 \cdot 639$	
Total	76·917	11		

Interpretation

(i) Between Treatment : The calculated value of $F = 1$

Table value of F for (3, 6) d.f. at 5% l.o.s. = 4·76

Since, the calculated value of F is less than the table value of F, the null hypothesis is accepted. Hence, there is no significant difference between the treatments.

(ii) Between Plots of land : The calculated value of $F = 1 \cdot 87$

Table value of F for (6, 2) d.f. at 5% l.o.s. = 19·3

Since, the calculated value of F is less than the table value of F, the null hypothesis is accepted. Hence, there is no significant difference between plots of land.

EXERCISE – 2

1. A company appoints 4 salesmen A_1, A_2, A_3 and A_4 and observes their sales in three seasons : Summer, Winter and Monsoon. The figures (in lakhs) are given ahead :

Salesmen

Seasons	A_1	A_2	A_3	A_4	Total
Summer	5	4	4	7	20
Winter	7	8	5	4	24
Monsoon	9	6	6	7	28
Salesmen Total	21	18	15	18	72

Carry out Two-way Analysis of Variance.

[**Ans :** F Between Salesmen = 1·335, F Between Seasons = 1·498, In both the cases, H_0 is accepted]

2. Perform a two-way ANOVA on the data given below :

Plots of Land	Treatments			
	A	B	C	D
I	38	40	41	39
II	45	42	49	36
III	40	38	42	42

(Using coding method subtracting 40 from given numbers)

Given for (3, 6) d.f. F05 = 4·76

and for (2, 6) d.f. F·05 = 5·14 . [**Ans :** F Between the column = 1·312, H_0 is accepted; F Between the Rows = 1·218, H_0, is accepted]

3. The price of a certain commodity was ascertained in each of the four towns A, B, C and D in four quarters of a year. The prices are given below. Are the variations in prices between different towns and in different season significant ?

Quarters	Towns			
	A	B	C	D
I	60	50	60	50
II	50	40	65	50
III	45	35	45	50
IV	65	45	60	70

[**Ans :** F Between the column = 4·89, F between Season's = 5·00 In both the cases, H_0 is rejected]

4. The following table gives the number of units of production per day turned out by four different types of machines.

Employee	Type of Machines			
	M_1	M_2	M_3	M_4
E_1	40	36	45	30
E_2	38	42	50	41
E_3	36	30	48	38
E_4	46	47	52	44

Using analysis of variance (i) test the hypothesis that the mean production is the same for the four machines, and (ii) test the hypothesis that the employees do not differ with respect to mean production. [**Ans :** F Between Machines= 9·27, F Between Employees = 8·27, In both the cases, H_0 is rejected]

5. The following data represent the number of a commodity produced by 3 different workers using 3 different machines :

Workers	Machines		
	A	B	C
X	8	32	20
Y	28	36	38
Z	6	28	14

Test (i) whether the mean productivity is the same for different machines types, (ii) whether the three workers differ with respect to mean productivity.
[**Ans :** F Between Machines = 9·38, H_0 is rejected, F Between Workers = 10·31, in both cases, H_0 is rejected]

6. To study the performance of three detergents and three different water temperatures, the following whiteness readings were obtained with specially designed equipment :

Water temp.	Detergent A	Detergent B	Detergent C
Cold water	57	55	67
Warm water	49	52	68
Hot water	54	46	58

Perform a two way analysis of variance using 5% level of significance (Given $F_{05} = 6·94$)
[**Hint :** See Example 14]
[**Ans :** F Between Column = 9·845, H_0 is rejected. F Between Rows = 2·381, in both the cases, H_0 is accepted]

7. You are given the following data :

Workers	Machine Type			
	A	B	C	D
1	44	35	48	38
2	48	40	50	44
3	37	38	40	36
4	45	34	45	32
5	40	44	50	40

Discuss whether there is a significant difference in mean productivity between machine types or workers. [**Ans :** F (Between columns) = 7·85; F (between rows) = 3·74, in both cases H_0 is rejected]

8. The following table gives the number of refrigerators sold by 4 salesmen in three months, May, June and July :

Month	Salesmen			
	A	B	C	D
May	50	40	48	39
June	46	48	50	45
July	39	44	40	39

Is there a significant difference in the sales made by the four salesmen ?
Is there a significant difference in the sales made during different months ?
[**Ans :** F Between columns = 1·018; F Between Rows = 3·33. In both cases, H_0 is accepted]

9. The following data represent the sales (Rs. 1000) per month of three brands of a detergent related among three cities :

Cities	Detergent A	Detergent B	Detergent C
	A	B	C
I	12	48	30
II	42	54	57
III	9	42	21

Test whether the (i) mean sales of the three brands are equal and (ii) the mean sales of detergent in each city are equal.
[**Ans :** F Between brands = 9·4, F between Cities = 10·3. In both cases, H_0 is rejected]

MISCELLANEOUS SOLVED EXAMPLES

Example 9 : **To test the significance of the variations of the retail prices of a commodity in three principle cities : Bombay, Kolkata and Delhi, four shops were chosen at random in each city and prices observed in rupees were as follow :**

Bombay	16	8	12	14
Kolkata	14	10	10	6
Delhi	4	10	8	8

Do the data indicate the prices in the three cities are significantly different?

Solution : $H_0 : \mu_1 = \mu_2 = \mu_3$, *i.e.*, the mean prices in the three cities are the same.
In order to simipify the calculation, subtract 10 from each observation. The deviations and their squares are as follow :

Coded Data

Bombay		Kolkata		Delhi	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
6	36	4	16	− 6	36
− 2	4	0	0	0	10
2	4	0	0	− 2	4
4	16	− 4	16	− 2	4
$\Sigma X_1 = 10$	$\Sigma X_1^2 = 60$	$\Sigma X_2 = 0$	$\Sigma X_2^2 = 32$	$\Sigma X_3 = -10$	$\Sigma X_3^2 = 44$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 10 + 0 - 10 = 0$$

$$C.F. = \frac{T^2}{N} = \frac{(0^2)}{12} = 0$$

TSS = Total sum of squares
$$= [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - C.F.$$
$$= [60 + 32 + 44] - 0$$
$$= 136$$

$$SSB = \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}\right] - C.F.$$
$$= \left[\frac{(10)^2}{4} + \frac{(0)^2}{4} + \frac{(-10)^2}{4}\right] - 0 = 50$$

SSW = SST - SSB
= 136 - 50 = 86

The various sum of squares (S.S.) along with the degrees of freedom (d.f.) are shown in the following table

ANOVA Table

Source of variation	Sum of square (S.S.)	Degrees of freedom	Mean Sum of squares (MSS)	F-Radio
Between City	50	3 – 1 = 2	25	$F = \frac{25}{9 \cdot 556} = 2 \cdot 616$
Within city	86	9	9·556	
Total	136	12 – 1 = 11		

For $v_1 = 2$ and $v_2 = 9$, the table value of F at 5% l. o.s. = 4·261
Since the calculated value of F is less than the table value of F the null hypothesis is accepted. We thus conclude that the mean prices in the three cities is not significantly different.

Example 10 : **Three samples, each of size 5, were chosen from three uncorrelated normal population with equal variances. Test the hypothesis that the population means are equal at 5% level.**

Sample 1	Sample 2	Sample 3
10	9	14
12	7	11
9	12	15
16	11	14
13	11	16

Solution : Let us take the hypothesis that the population means are equal for three samples
i.e. $H_0 : \mu_1 = \mu_2 = \mu_3$
In order to simplify the calcuation, surtract 10 from each observation. The deviations and their squares are as follows :

Coded Data

Sample 1		Sample 2		Sample 3	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
0	0	– 1	1	4	16
2	4	– 3	9	1	1
– 1	1	2	4	5	25
6	36	1	1	4	16
3	9	1	1	6	36
$\Sigma X_1 = 10$	$\Sigma X_1^2 = 50$	$\Sigma X_2 = 0$	$\Sigma X_2^2 = 16$	$\Sigma X_3 = 20$	$\Sigma X_3^2 = 94$

$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 10 + 0 + 20 = 30$

$C.F. = \frac{T^2}{N} = \frac{(30^2)}{15} = 60$

TSS = Total sum of squares

$= [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - C.F.$

$= [50 + 16 + 94] - 60$

$= 160 - 60 = 100$

$SSB = \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3}\right] - C.F.$

$= \left[\frac{(10)^2}{5} + \frac{(0)^2}{5} + \frac{(20)^2}{5}\right] - 60$

$= [20 + 0 + 80] - 60$

$= 100 - 60 = 40$

SSW = TSS – SSB

$= 100 - 40 = 60$

The various sum of squares (S.S.) along with the degrees of freedom (d.f.) are shown in the following table:

ANOVA Table

Source of variation	Sum of square (S.S.)	Degrees of freedom	Mean Sum of squares (MSS)	F-Radio
Between City	40	3 – 1 = 2	20	$F = \frac{20}{5} = 4$
Within city	60	15 – 3 = 12	5	
Total	100	14		

For $v_1 = 2$ and $v_2 = 12$, the table value of F at 5% l. o.s. = 3·08.
Since the calculated value of F is more than the table value, the null hypothesis is rejected. Hence the population means of the three samples do not seem to be equal.

Example 11 : **The following figures related to the number of units of a product sold in five different areas by four salesmen.**

Area	Number of units			
	A	B	C	D
1	80	100	95	70
2	82	110	90	75
3	88	105	100	82
4	85	115	105	88
5	75	90	80	65

Is there a significant difference in the efficiency of these salesmen ?
(Given that Table value of $F_{.05}$ for $v_1 = 3$, $v_3 = 16$ is 3·24.)

Solution : Let us take the hypothesis that there is no significant difference in the performance of the four salesmen *i.e.* $\mu_1 = \mu_2 = \mu_3 = \mu_4$.
In order to simplify the calculation, subtract 80 from each observation. The deviations and their squares are as follows :

Coded Data

A		B		C		D	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2	X_4	X_4^2
0	0	20	400	15	225	− 10	100
2	4	30	900	10	100	− 5	25
8	64	25	625	20	400	2	4
5	25	35	1225	25	625	8	64
− 5	25	10	100	0	0	− 15	225
$\Sigma X_1 = 10$	$\Sigma X_1^2 = 118$	$\Sigma X_2 = 120$	$\Sigma X_2^2 = 3250$	$\Sigma X_3 = 70$	$\Sigma X_3^2 = 1350$	$\Sigma X_4 = -20$	$\Sigma X_4^2 = 418$

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4 = 10 + 120 + 70 - 20 = 180$$

$$C.F. = \frac{T^2}{N} = \frac{(180)^2}{20} = 1620$$

$$TSS = [\Sigma X_1^2 + \Sigma X_2^2 + + \Sigma X_3^2 + \Sigma X_4^2] - C.F.$$

$$= [118 + 3250 + 1350 + 418] - 1620$$

$$= 5136 - 1620 = 3516$$

$$SSB = \left[\frac{(\Sigma X_1)^2}{n_1} + \frac{(\Sigma X_2)^2}{n_2} + \frac{(\Sigma X_3)^2}{n_3} + \frac{(\Sigma X_4)^2}{n_4}\right] - C.F.$$

$$= \left[\frac{(10)^2}{5} + \frac{(120)^2}{5} + \frac{(70)^2}{5} + \frac{(-20)^2}{5}\right] - 1620$$

$$= [20 + 2880 + 980 + 80] - 1620$$

$$= 3960 - 1620 = 2340$$

$$SSW = TSS - SSB$$

$$= 3516 - 2340 = 1176$$

The various sum of squares (S.S.) along with the degrees of freedom (d.f.) are shown in the following table :

ANOVA Table

Source of variation	Sum of square	Degrees of freedom	Mean Sum of squares (MSS)	F-Radio
Between City	2340	4 − 1 = 3	780	$F = \frac{780}{73·5} = 10·61$
Within city	1176	20 − 4 = 16	73·5	
Total	3516	19		

For $v_1 = 3$ and $v_2 = 16$, the table value of F at 5% l. o.s. = 3·24.
Since, the calculated value of F is greater than the table value, the null hypothesis is rejected. Hence there is a significant difference in the efficiency of the four salesmen.

Example 12. **Complete the following incomplete ANOVA table :**

Source of Variation	Sum of square (S.S)	Degree of Freedom (d.f.)	Mean sum of squares (MSS)	F-test
Between	–	$v_1 = 2$	5	F = ?
Within	14	$v_2 = -$	–	
Total	–	$v = 9$		

Solution : We get $v_2 = v - v_1 = 9 - 2 = 7$

$$\frac{SSB}{v_1} = MSB \quad \Rightarrow \quad SSB = MSB \times v_1 = 5 \times 2 = 10$$

$$TSS = SSB + SSW = 10 + 14 = 24$$

$$SSW = 14,\ MSW = \frac{SSW}{v_2} = \frac{14}{7} = 2$$

$$F = \frac{MSB}{MSW} = \frac{5}{2} = 2·5$$

COMPLETE TABLE

Source of Variation	Sum of square (S.S)	Degree of Freedom	Mean sum of squares (MSS)	F-test
Between	10	2	5	$F = \frac{5}{2} = 2·5$
Within	14	7	2	
Total	24	9	–	

Example 13 : **A company appoints four salesman A, B, C and D and observes their sales in three seasons in summer, winter and monsoon. The figure (in lakhs) are given in the following tables :**

Seasons	Salesmen			
	A	B	C	D
Summer	36	36	21	35
Winter	28	29	31	32
Monsoon	26	28	29	29

Do the salesman differ significantly in their performance ? (or Carry out an Analysis of Variance).

Solution : The above data are classified according to two criteria : (i) salesmen, and (ii) seasons. It is a two way classification.

Let us take the null hypothesis that there is no difference in the performance of the salesmen. This hypothesis means that there is no difference between the sales of salesmen and off seasons.

In order to simplify the calculation, we subtract 30 from each observation the deveations and their squares are as follows :

Seasons	Coded Data X_1	X_2	X_3	X_4	Row Total	Squares of Coded Data X_1^2	X_2^2	X_3^2	X_4^2
S	6	6	– 9	5	8	36	36	81	25
W	– 2	–1	1	2	0	4	1	1	4
M	– 4	– 2	– 1	– 1	– 8	16	4	1	1
Column Total	0	3	– 9	6	T = 0	56	41	83	30

$$T = 0+3-9+6 = 0$$

$$\text{C.F.} = \frac{T^2}{N} = \frac{(O)^2}{12} = 0$$

$$\text{TSS} = [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2 + \Sigma X_4^2] - C.F.$$

$$= 56+41+83+30-0 = 210$$

$$\text{SSC} = \left[\frac{(0)^2}{3} + \frac{(3)^2}{3} + \frac{(-9)^2}{3} + \frac{(6)^2}{3}\right] - C.F.$$

$$= 0+3+27+12-0 = 42$$

$$\text{SSR} = \left[\frac{(8)^2}{4} + \frac{(0)^2}{4} + \frac{(8)^2}{4}\right] - C.F. = 16+0+16-0 = 32$$

$$\text{SSE} = \text{TSS} - \text{SSC} - \text{SSR} = 210 - 42 - 32 = 136$$

ANOVA Table

Source of Variation	Sum of square (S.S)	Degree of Freedom	Mean sum of squares (MSS)	F-Ratio
Between Columns (Salesmen)	42	3	$\frac{42}{3} = 14$	$F = \frac{22 \cdot 63}{14} = 1 \cdot 62$
Betwen Rows (Seasons)	32	2	$\frac{32}{2} = 16$	$F = \frac{22 \cdot 67}{16} = 1 \cdot 42$
Residual/Error	136	3 × 2 = 6	$\frac{136}{6} = 22 \cdot 67$	
Total	210	11		

Interpetation

(i) **For salemen :** The calculated value of F = 1·62

Table value of F for (6, 3) d.f. at 5% l.o.s. = 8·94

Since the calculated value of F is less than the table value, we accept the null hypothesis and conclude that the sales of different salesmen do not differ significantly.

(ii) **For salemen :** The calculated value of F = 1·42

Table value of F for (6, 2) d.f. at 5% l.o.s. = 5·15

Since the calculated value of F is less than the table value, we accept the null hypothesis and conclude that there is not significant difference in the seasons so far as sales ae concerned

Example 14 : **To study the performance of three detergents and three different water temperatures, the following 'whiteness' readings were obtained with specially designed equipment :**

Water Temperature	Detergent A	Detergent B	Detergent C
Coled Water	57	55	67
Warm Water	49	52	68
Hot Water	54	46	58

Perform a two-way analysis of variance, using 5 percent level of significance.

Solution : Let us take the null hpothesis that there is no significant difference in the performance of three detergents due to water temperature and vice versa.

In order to simplify calculatins, let us subtract 50 from each figure. The deviations and their squares ae as follows :

Water Temp.	Coded Data X_1	X_2	X_3	Row Total	Squares of Coded Data X_1^2	X_2^2	X_3^2
Cold.	7	5	17	29	49	25	289
Warm	– 1	2	18	19	1	4	324
Hot	4	– 4	8	8	16	16	64
Column Total	10	3	43	T = 56	66	45	677

$$T = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 = 10+3+43 = 56$$

$$\text{C.F.} = \frac{T^2}{N} = \frac{(56)^2}{9} = 348 \cdot 44$$

$$\text{TSS} = [\Sigma X_1^2 + \Sigma X_2^2 + \Sigma X_3^2] - C.F.$$

$$= [66+45+677] - 348 \cdot 44$$

$$= 788 - 348 \cdot 44 = 439 \cdot 44$$

$$\text{SSC} = \left[\frac{(10)^2}{3} + \frac{(3)^2}{3} + \frac{(43)^2}{3}\right] - C.F.$$

$$= 33.3+3+616 \cdot 33 - 348 \cdot 44 = 304 \cdot 22$$

$$SSR=\left[\frac{(29)^2}{3}+\frac{(19)^2}{3}+\frac{(8)^2}{3}\right]-C.F.$$

$$=280\cdot33+120\cdot33+21\cdot33-348\cdot44=73\cdot55$$

$$SSE=TSS-SSC-SSR$$

$$=439\cdot56-304\cdot22-73\cdot55=61\cdot79$$

ANOVA Table

Source of Variation	Sum of square (S.S)	Degree of Freedom	Mean sum of squares (MSS)	F-Ratio
Between Columns (Detergents)	304·22	2	152·110	$F=\frac{152\cdot110}{15\cdot445}=9\cdot85$
Between Rows (Temperatures)	73·55	2	36·775	$F=\frac{36\cdot775}{15\cdot445}=1\cdot42$
Residual/Error	61·79	4	15·445	
Total	439·56	8		

Interpetation

(i) **For Detergents :** The calculated value of F = 9·85

Table value of F for (2, 4) d.f. at 5% l.o.s. = 6·94

Since the calculated value of F is greater than the table value, we reject the null hypothesis and conclude that there is significant different in the three varieties of deregents.

(ii) **For Temperature :** The calculated value of F = 2·38

Table value of F for (2, 4) d.f. at 5% l.o.s. = 6·94

Since the calculated value of F is less than the table value, we accept the null hypothesis and conclude that temperature do not make a significant difference.

Example 15 : Apply F-test on the following data

X_1	25	30	36	38	31	$\Sigma X_1=160$
X_2	31	39	38	42	35	$\Sigma X_2=185$
X_3	24	30	28	25	28	$\Sigma X_3=135$

Given ($F_{2,12}$, 5 percent = 3·89)

Solution : Null hypothesis, $H_0:\mu_1=\mu_2=\mu_3$ *i.e.* there is no significant difference in the mean of three samples.

In order to simplify the calculation, subtact 30 from each sample values.

Coded Data

Sample 1		Sample 2		Sample 3	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
−5	25	1	1	−6	36
0	0	9	81	0	0
6	9	8	64	−2	4
8	12	12	144	−5	25
1	5	5	25	−2	4
$\Sigma X_1=10$, $n_1=5$	$\Sigma X_1^2=126$	$\Sigma X_2=35$, $n_2=5$	$\Sigma X_2^2=315$	$\Sigma X_3=-15$	$\Sigma X_3^2=69$

$$T=\Sigma X_1+\Sigma X_2+\Sigma X_3=10+35-15=30$$

Correction Factor, $C.F.=\frac{T^2}{N}=\frac{(30)^2}{15}=60$

TSS = Total sum of squares

$$=[\Sigma X_1^2+\Sigma X_2^2+\Sigma X_3^2]-C.F.$$

$$=[126+315+69]-60=450$$

SSB = Sum of squares between the samples

$$=\left[\frac{(\Sigma X_1)^2}{n_1}+\frac{(\Sigma X_2)^2}{n_2}+\frac{(\Sigma X_3)^2}{n_3}\right]-C.F.$$

$$=\left[\frac{(10)^2}{5}+\frac{(35)^2}{5}+\frac{(-15)^2}{5}\right]-60=[20+245+45]-60=250$$

$$SSW=TSS-SSB=450-250=200$$

ANOVA Table

Source of Variation	Sum of squares	Degree of Freedom (d.f.)	Mean sum of squares (MSS)	F-Ratio
Between Samples	250	2	125	$F=\frac{125}{16\cdot667}=7\cdot5$
Between Samples	200	12	16·667	
Total	**450**	**14**		

For $v_1=2$, $v_2=12$, the table value of F at 5% level of significantce is 3·89 Since the calculated value of F is greater than the table value i.e. $F<F_{0\cdot05}$, we reject the null hypothesis and conclude that there is significant difference in the mean of three samples.

QUESTIONS

1. What is analysis of variance technique ? Explain its basic assumptions and uses.
2. (a) Discuss the assumptions of Analysis of Variance test (or techniques)
   (b) Distinguish between one-way and two-way ANOVA technique.
3. Discuss the technique of analysis of variance with an illustration for one-way classifcation.
4. Describe the technique of ANOVA for two-way classification.
5. What do you understand by analysis of variance? Explain the assumptions in an analysis of variance.
6. What is analysis of variance problem? Comment on the variance between the samples and within the samples.
7. What are the objectives, assumptions and uses of analysis of variance?
8. What is analysis of variance? Mention its applications.
9. Explain the meaning and significance of ANOVA. How is an ANOVA table set up and how a test is performed.

❀❀❀

7

Statistical Estimation Theory

INTRODUCTION :

Very often, we will need to make an estimate of the population parameter from the sample statistic. For example, suppose we are to find out the average amount of Pepsi Cola drunk by the students in Kurukshetra University, Kurukshetra, per day. It is difficult to find out the average of all the students and hence what usually done is, a sample taken and the average amount of Pepsi Cola drunk is found out. This sample mean is then used to find the average of the population. In fact, we estimate the population average on the basis of sample average. The theory of estimation deals with the estimation of the unknown population parameters (such as population mean and variance) from the corresponding sample statistics (such as sample mean and variance). **Statistical estimation in a procedure of estimating the unknown population parameters from the correspondiong sample statistics.**

SOME IMPORTANT TERMS

The following terms are used in the study of statistical estimation :

(1) Estimators and Estimates : Generally for the purpose of estimating a population parameter we can use various sample statistics. Those sample statistics (such as sample mean $\bar{X}$, sample median M, sample variance σ^2, etc.) which are used to estimate the unknown population parameters (such as population mean μ, population variance σ^2, etc.) are called estimators and the actual value taken by the estimators are called estimates. If θ (read as theta) denotes the parameters to be estimated, then its estimator will be denoted by $\hat{\theta}$ (read as theta hat). Thus, $\hat{\theta}$ is an estimator of the population parameter θ.

(2) Point Estimate and Interval Estimate : An estimate of the population parameter can be done in two ways :

(i) **Point Estimate :** A single value of a statistic that is used to estimate the unknown population parameter is called a point estimate. For example, the sample mean $\bar{X}$ which we use for estimating the population mean μ is a point estimator of μ. Similarly, the statistic s^2 is a point estimator of σ^2, where the value of s^2 is computed from a random sample. The point estimate is a single point on the real number scale and hence the name point estimator.

(ii) **Interval Estimate :** An interval estimate refers to the probable range within which the real value of a parameter is expected to lie. The two extreme limits of such a range are called **fiducial or confidence limits** and the range is called a **confidence interval.** These are determined on the basis of sample studies of a population. Thus, on the basis of sample studies when we estimate that the average monthly expenditure of students staying in a

certain hostel is between Rs. 1000 and Rs. 2000, it will be a case of interval estimate and the figures of 1000 and 2000 will be the two estreme limits within which the actual expenditure of the students would lie.

PROPERTIES OF A GOOD ESTIMATOR

There can be more than one estimators of a population parameter. For example, the population mean (μ) may be estimated either by sample mean ($\overline{X}$) or by sample median (M) or by sample mode (Z), etc. Similarly, the population variance (σ^2) may be estimated either by the sample variance (s^2), sample S.D. (s), sample mean deviation, etc. Therefore, it becomes necessary to determine a good estimator out of a number of available estimators. **A good estimator is one which is as close to the true value of the parameter as possible.** A good estimator possess the following characteristics or properties:

(1) Unbiasedness
(2) Consistency
(3) Efficiency
(4) Sufficiency

Let us consider them in detail :

(1) Unbiased Estimator : An estimator $\hat{\theta}$ is said be unbiased estimator of the population parameter θ if the mean of the sampling distribution of the estimator $\hat{\theta}$ is equal to the corresponding population parameter θ. Symbolically,

$$\mu_{\hat{\theta}} = \theta$$

In terms of mathematical expectation, $\hat{\theta}$ is an unbiased estimator of θ if the expected value of the estimator is equal to the parameter being estimated. Symbolically,

$$E(\hat{\theta}) = \theta$$

Example 1. **Sample mean $\overline{X}$ is an unbiased estimate of the population mean μ because the mean of the sampling distribution of the means $\mu_{\overline{X}}$ or $E(\overline{X})$ is equal to the pouplation mean μ. Symbolically,**

$$\mu_{\overline{X}} = \mu \quad \text{or} \quad E(\overline{X}) = \mu$$

Example 2. **Sample variance s^2 is a biased estimate of the population variance σ^2 because the mean of the sampling distribution of variance is not equal to the population variance. Symbolically,**

$$\mu_{s^2} \neq \sigma^2 \text{ or } E(s^2) \neq \sigma^2$$

However, the modified sample variance ($\hat{s}^2$) is unbiased estimate of the population variance σ^2 because

$$E(\hat{s}^2) = \sigma^2 \qquad \text{where, } \hat{s}^2 = \frac{n}{n-1} \times s^2$$

Example 3. **Sample proportion p is an unbiased estimate of the population proportion P because the mean of the sampling distribution of proportion is equal to the population proportion. Symbolically,**

$$\mu_p = P \quad \text{or} \quad E(P) = P$$

(2) Consistent Estimator : An estimator is said to be consistent if the estimator approaches the population parameter as the sample size increases. In other words, an estimator $\hat{\theta}$ is said to be consistent estimator of the population parameter θ, if the probability that $\hat{\theta}$ approaches θ is 1 as n becomes larger and larger. Symbolically,

$$P(\hat{\theta} \to \theta) \to 1 \quad \text{as} \quad n \to \infty$$

Note : A consistent estimator need not to be unbiased
A sufficient condition for the consistency of an estimator is that
(i) $E(\hat{\theta}) \to \theta$
(ii) $\text{Var}(\hat{\theta}) \to 0$ as $n \to \infty$

Example 1 : **Sample mean $\overline{X}$ is a consistent estimator of the population mean μ because the expected value of the sample mean approaches the population mean and the variance of the sample mean approaches zero as the size of the sample is sufficiently increased. Symbolically,**
(i) $E(\overline{X}) \to \mu$
(ii) $\text{Var}(\overline{X}) = \dfrac{\sigma^2}{n} \to 0$ as $n \to \infty$

Example 2 : **Sample median is also consistent estimator of the population mean because :**
(i) $E(M) \to \mu$
(ii) $\text{Var}(M) \to 0$ as $n \to \infty$

(3) Efficient Estimator : Efficiency is a relative term. Efficiency of an estimator is generally defined by comparing it with another estimator. Let us to take two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$. The estimator $\hat{\theta}_1$ is called an efficient estimator of θ if the variance of $\hat{\theta}_1$ is less than the variance of $\hat{\theta}_2$. Symbolically,

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$$

Then, $\hat{\theta}_1$ is called an efficient estimator.

Example : **Sample mean $\overline{X}$ is an unbiased and efficient estimator of the population mean (or true mean) than the sample median M because the variance of the sampling distribution of the means is less than the variance of the sampling distribution of the medians.**

The relative efficiency of the two unbiased estimators is given below :

We know that, $\text{Var}(\overline{X}) = \dfrac{\sigma^2}{n}$, $\text{Var}(M) = \dfrac{\pi}{2} \cdot \dfrac{\sigma^2}{n}$

$$\text{Efficiency} = \frac{\text{Var}(\overline{X})}{\text{Var}(M)} = \frac{\frac{\sigma^2}{n}}{\frac{\pi\sigma^2}{2n}} = \frac{2}{\pi} = \frac{14}{22} = \frac{7}{11} = 0{\cdot}64 \quad \left[\because \pi = \frac{22}{7}\right]$$

$$\therefore \quad \text{Var}(\overline{X}) = 0{\cdot}64\,\text{Var}(M)$$

Therefore, sample mean $\bar{X}$ is 64% more efficiency than the sample median.
Hence, the sample mean is more efficient estimator of the population mean as compared to sample median.

(4) Sufficient Estimator : The last property that a good estimator should possess is sufficiency. An estimator $\hat{\theta}$ is said to be a 'sufficient estimator' of a parameter θ if it contains all the informations in the sample regarding the parameter. In other words, a sufficient estimator utilises all informations that the given sample can furnish about the population. Sample means $\bar{X}$ is said to be a sufficient estimator of the population mean.

4. APPLICATION OF POINT ESTIMATION

The applications relating to point estimation are studied under two headings :

(1) Point Estimation in case of Single Sampling

(2) Point Estimation in case of Repeated Sampling.

(1) Point Estimation in case of Single Sampling : When a single independent random sample is drawn from a unknown population, the point estimate of the population parameter can be illustrated by the following examples :

Example 1. **A sample of 10 measurements of the diameter of a sphere gave a mean $\bar{X} = 4{\cdot}38$ inches and a standard deviation $= {\cdot}06$ inches. Determine the unbiased and efficient estimates of (a) the true mean (*i.e.*, population mean) and (b) the true variance (*i.e.*, population variance).**

Solution. We are given : $n = 10, \bar{X} = 4{\cdot}38, s = {\cdot}06$

(a) The unbiased and efficient estimate of the true mean μ is given by :

$$\bar{X} = 4{\cdot}38$$

(b) The unbiased and efficient estimate of the true variance σ^2 is :

$$\hat{s}^2 = \frac{n}{n-1} \cdot s^2$$

Putting the values, we get

$$\hat{s}^2 = \frac{10}{10-1} \times {\cdot}06 = 1{\cdot}11 \times 0{\cdot}06 = {\cdot}066$$

Thus, $\mu = 4{\cdot}38, \sigma^2 = 0{\cdot}666$

Example 2. **The following five observations constitute a random sample from an unknown population :**

6·33, 6·37, 6·36, 6·32 and 6·37 centimeters.

Find out unbiased and efficient estimates of (a) true mean, and (b) true variance.

Solution. (a) The unbiased and efficient estimate of the true mean (*i.e.*, population mean) is given by the value of

$$\bar{X} = \frac{\Sigma X}{n} = \frac{6{\cdot}33 + 6{\cdot}37 + 6{\cdot}36 + 6{\cdot}32 + 6{\cdot}37}{5} = \frac{31{\cdot}75}{5} = 6{\cdot}35$$

(b) The unbiased and efficient estimate of the true variance (*i.e.*, population variance) is :

$$\hat{s}^2 = \frac{\Sigma (X - \bar{X})^2}{n-1}$$

where, $\hat{s}^2$ = modified sample variance.

$$= \frac{(6{\cdot}33 - 6{\cdot}35)^2 + (6{\cdot}37 - 6{\cdot}35)^2 + (6{\cdot}36 - 6{\cdot}35)^2 + (6{\cdot}32 - 6{\cdot}35)^2 + (6{\cdot}37 - 6{\cdot}35)^2}{5-1}$$

$$= \frac{{\cdot}0022}{4} = {\cdot}00055 \text{ cm}^2$$

Example 3. **The following data relate to a random sample of 100 students in Kurukshetra University classified by their weights (kg) :**

Weight (kg) :	60–62	63–65	66–68	69–71	72–74
No. of Students :	5	18	42	27	8

Determine unbiased and efficient estimates of (a) population mean and (b) population variance.

Solution.

Calculation of Mean and Variance

Weight	No. of Students (f)	M.V. (m)	$A = 67$ $d = m - A$	$d' = d/3$	fd'	fd'^2
60–62	5	61	− 6	− 2	−10	20
63–65	18	64	− 3	− 1	− 18	18
66–68	42	67	0	0	0	0
69–71	27	70	+ 3	+ 1	+ 27	27
72–74	8	73	+ 6	+ 2	+ 16	32
	$n = 100$				$\Sigma fd' = 15$	$\Sigma fd'^2 = 97$

(a) The unbiased and efficient estimate of the population mean is given by the value :

$$\bar{X} = A + \frac{\Sigma fd'}{n} \times i$$

$$= 67 + \frac{15}{100} \times 3 = 67 + (0{\cdot}45) = 67{\cdot}45$$

(b) The unbiased and efficient estimate of the population variance is :

$$\hat{s}^2 = \frac{n}{n-1} \cdot s^2$$

where,

$$s^2 = \frac{\Sigma fd'^2}{n} - \left(\frac{\Sigma fd'}{n}\right)^2 \times i^2$$

$$= \left[\frac{97}{100} - \left(\frac{15}{100}\right)^2\right] \times 3^2$$

$$= [0{\cdot}97 - {\cdot}0225] \times 9 = 8{\cdot}5275$$

Now, $\hat{s}^2 = \frac{n}{n-1} s^2 = \frac{100}{99} \times 8{\cdot}5275 = 8{\cdot}6136$

Thus, $\mu = 67{\cdot}45, \sigma^2 = 8{\cdot}6136$

(2) Point Estimation in Case of Repeated Sampling : When large number of random samples of same size are drawn from the population with or without replacement, then the point estimates of the population parameter can be illustrated by the following examples :

Example 4. **A population consists of five values : 3, 4, 5, 6 and 7. List all possible samples of size 3 without replacement from this population and calculate the mean $\overline{X}$ of each sample. Verify that sample mean $\overline{X}$ is an unbiased estimate of the population mean.**

Solution. The population consists of the five values : 3, 4, 5, 6, 7. The total number of possible samples of size 3 without replacement are ${}^5C_3 = 10$ which are shown in the following table :

Sample No. (1)	Sample Values (2)	Sample Mean ($\overline{X}$) (3)
1	(3, 4, 5)	$\frac{1}{3}(3+4+5) = \frac{12}{3} = 4$
2	(3, 4, 6)	$\frac{1}{3}(3+4+6) = \frac{13}{3} = 4{\cdot}33$
3	(3, 4, 7)	$\frac{1}{3}(3+4+7) = \frac{14}{3} = 4{\cdot}67$
4	(3, 5, 6)	$\frac{1}{3}(3+5+6) = \frac{14}{3} = 4{\cdot}67$
5	(3, 5, 7)	$\frac{1}{3}(3+5+7) = \frac{15}{3} = 5{\cdot}0$
6	(3, 6, 7)	$\frac{1}{3}(3+6+7) = \frac{16}{3} = 5{\cdot}33$
7	(4, 5, 6)	$\frac{1}{3}(4+5+6) = \frac{15}{3} = 5{\cdot}00$
8	(4, 5, 7)	$\frac{1}{3}(4+5+7) = \frac{16}{3} = 5{\cdot}33$
9	(4, 6, 7)	$\frac{1}{3}(4+6+7) = \frac{17}{3} = 5{\cdot}67$
10	(5, 6, 7)	$\frac{1}{3}(5+6+7) = \frac{18}{3} = 6{\cdot}00$
Total	$k = 10$	$\Sigma\overline{X} = 50$

Mean of Sampling Distribution of Means $= \mu_{\overline{X}} = \frac{\Sigma\overline{X}}{k} = \frac{50}{10} = 5.$

$$\text{Population Mean} = \mu = \frac{3+4+5+6+7}{5} = 5$$

Since, $\mu_{\overline{X}} = \mu$, sample mean $\overline{X}$ is an unbiased estimate of the population mean μ.

Example 5. **Consider a hypothetical population comprising three values : 1, 2, 3. Draw all possible samples of size 2 with replacement. Calculate the mean $\overline{X}$ and variance s^2 for each sample. Examine whether the two statistics ($\overline{X}$ and s^2) are unbiased and efficient for the corresponding parameters.**

Solution. The population consists of three values : 1, 2 and 3. The total number of possible samples of size 2 with replacement are $N^n = 3^2 = 9$ which are given by

Sample No.	Sample Values	Sample Mean ($\overline{X}$)	Sample Variance $s^2 = \frac{1}{2}[(x_1-\bar{x})^2 + (x_2-\bar{x})^2]$	Modified Sample Variance $\left(\hat{s}^2 = \frac{n}{n-1} s^2\right)$
1.	(1, 1)	$\frac{1}{2}(1+1) = 1{\cdot}0$	$\frac{1}{2}\cdot[(1-1)^2 + (1-1)^2] = 0{\cdot}00$	$0{\cdot}00$
2.	(1, 2)	$\frac{1}{2}(1+2) = 1{\cdot}5$	$\frac{1}{2}\cdot[(1-1{\cdot}5)^2 + (2-1{\cdot}5)^2] = 0{\cdot}25$	$0{\cdot}50$
3.	(1, 3)	$\frac{1}{2}(1+3) = 2{\cdot}0$	$\frac{1}{2}\cdot[(1-2)^2 + (3-2)^2] = 1{\cdot}00$	$2{\cdot}00$
4.	(2, 1)	$\frac{1}{2}(2+1) = 1{\cdot}5$	$\frac{1}{2}\cdot[(2-1{\cdot}5)^2 + (1-1{\cdot}5)^2] = 0{\cdot}25$	$0{\cdot}5$
5.	(2, 2)	$\frac{1}{2}(2+2) = 2{\cdot}0$	$\frac{1}{2}\cdot[(2-2)^2 + (2-2)^2] = 0{\cdot}00$	$0{\cdot}00$
6.	(2, 3)	$\frac{1}{2}(2+2) = 2{\cdot}5$	$\frac{1}{2}\cdot[(2-2{\cdot}5)^2 + (3-2{\cdot}5)^2] = 0{\cdot}25$	$0{\cdot}50$
7.	(3; 1)	$\frac{1}{2}(3+1) = 2{\cdot}0$	$\frac{1}{2}\cdot[(3-2)^2 + (1-2)^2] = 1{\cdot}00$	$2{\cdot}00$
8.	(3, 2)	$\frac{1}{2}(3+2) = 2{\cdot}5$	$\frac{1}{2}[(3-2{\cdot}5)^2 + (2-2{\cdot}5)^2] = 0{\cdot}25$	$0{\cdot}50$
9.	(3, 3)	$\frac{1}{2}(3+3) = 3{\cdot}0$	$\frac{1}{2}[(3-3)^2 + (3-3)^2] = 0{\cdot}00$	$0{\cdot}00$
Total	$k = 9$	$\Sigma\overline{X} = 18$		$\Sigma\hat{s}^2 = 6$

(a) Mean of Sampling Distribution of Means $= \mu_{\overline{x}} = \frac{\Sigma\overline{X}}{k} = \frac{18}{9} = 2$. Here, K = No. of samples.

$$\text{Population Mean } \mu = \frac{1+2+3}{3} = 2.$$

Since, $\mu_{\overline{X}} = \mu$, sample mean $\overline{X}$ is an unbiased estimate of the population mean μ.

(b) Mean of the Sampling Distribution of Variance $= \mu_{s^2} = \frac{\Sigma s^2}{k} = \frac{3}{9} = \frac{1}{3}$

$$\text{Population Variance } \sigma^2 = \frac{(1-2)^2 + (2-2)^2\ (3-2)^2}{3} = \frac{2}{3}$$

Since, $\mu_{s^2} \neq \sigma^2$, sample variance s^2 is not an unbiased estimate of the population variance (σ^2).

But the modified sample variance defined as $\hat{s}^2 = \frac{n}{n-1} s^2$ will be unbiased estimate of the population variance σ^2 because :

$$\mu_{\hat{s}^2} = \frac{\Sigma \hat{s}^2}{k} = \frac{6}{9} = \frac{2}{3}$$

$$\sigma^2 = \frac{2}{3}$$

$$\therefore \quad \mu_{\hat{s}^2} = \sigma^2$$

Since, $\mu_{\hat{s}^2} = \sigma^2$, the modified sample variation is an unbiased estimate of the population variance.

Example 6. **Show that the sample mean ($\overline{X}$) is an unbiased estimate of the population mean.**

or

An independent random sample $x_1, x_2, x_3, \ldots, x_n$ is drawn from a population with mean μ. Prove that the expected value of the sample mean $\overline{X}$ equals the population mean μ.

Solution. A random sampling is one where each sample has an equal chance of being selected. We draw a random sample of size 'n'.

Then,

$$E(\overline{x}) = E\left[\frac{x_1 + x_2 + \ldots + x_n}{n}\right] \text{ Where } x_1 \text{ is the sample observation.}$$

$$= \frac{1}{n} \cdot [E(x_1) + E(x_2) + \ldots E(x_n)]$$

Now the expected values of x_i (a member of the population) is population mean μ.

$$\therefore \quad E(\overline{x}) = \frac{1}{n}[\mu + \mu + \ldots + \mu] \qquad [\because E(x_1) = E(x_2) = \ldots E(x_n) = \mu]$$

$$= \frac{1}{n} \cdot [n\,\mu] = \mu \qquad [\because \Sigma C = C_1 + C_2 + \ldots C_n = nC]$$

Thus, sample mean $\overline{X}$ is an unbiased estimate of population mean.

EXERCISE – 1

1. Measurements of a sample of masses were determined to be 8·3, 10·6, 9·7, 8·8, 10·2 and 9·4 kilograms (kg) respectively. Determine unbiased and efficient estimates of (a) the population mean and (b) the population variance, and (c) compare the sample standard deviation and estimated population S.D. **[Ans.** (a) 9·5 (b) ·736 (c) $\hat{s} = \sigma = 0{\cdot}86$, $s = {\cdot}78$]
2. A random samples of 9 individuals has the following heights in inches : 45, 47, 50, 52, 48, 47, 49, 53 and 51. Find the unbiased and efficient estimate of (a) true mean. (b) true variance. **[Ans.** (a) 49·11 (b) 6·91]
3. A population consists of four numbers : 3, 4, 2, 5. List all possible distinct samples of size two which can be drawn without replacement and verify that the population mean is equal to the mean of sample means.
4. A population consists of three numbers : 2, 5 and 8. List all possible distinct samples of size two which can be drawn without replacement from this population. Calculate the mean $\overline{x}$ for each sample. Verify that sample mean $\overline{x}$ is an unbiased estimate of the population mean.

5. A population consists of five numbers : 2, 3, 6, 8, 11. List all possible samples of size 2 which can be drawn with replacement from this population. Calculate the mean $\overline{x}$ and variance $s^2 = \frac{1}{2}[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2]$ for each sample. Examine whether the two statistics are unbiased for the corresponding population parameters. What is the sampling variance of $\overline{X}$?
6. A sample of 10 television tubes produced by a company showed a mean life of 1200 hr. and a standard deviation of 10 hr. Find the unbiased and efficient estimates of the (a) population mean and (b) population variance. **[Ans.** $\mu = 1200$ h. $\hat{s}^2 = 111{\cdot}11$]

INTERVAL ESTIMATION (OR CONFIDENCE INTERVAL)

A point estimator, however, good it may be, cannot be expected to coincide with the true value of the parameter and in some cases may differ widely from it. In the theory of interval estimation, we find an interval or two numbers within which the value of unknown population parameter is expected to lie with a specified probability. The method of interval estimation consists in the determination of two constant t_1 and t_2 such that $P[t_1 < \theta < t_2$ for given value of $t] = 1 - \alpha$, where α is the level of significance. The interval $[t_1, t_2]$ within which the unknown value of parameter θ is expected to lie is known as confidence interval and the limits t_1 and t_2 so determined are known as confidence limits and $1 - \alpha$ is called the confidence coefficient, depending upon the desired precision of the estimate. For example, $\alpha = 0{\cdot}05$ (or $0{\cdot}01$) gives 95% (or 99%) confidence limits.

Procedure for Setting up Confidence Interval (or Interval Estimation) or Limits for a Population Parameter

The following steps enable us to compute the confidence interval or confidence limits for the population parameter θ in terms of the sample statistic t :

(1) Compute or take the appropriate sample statistic t.
(2) Obtain the S.E. (t), the standard error of the sample statistic t.
(3) Select the confidence level and corresponding to that specified level of confidence, we note down the critical value of the statistic t.

Applications of Interval Estimation (or Confidence Interval)

The applications relating to interval estimation (or confidence interval) are studied under the following heads :

(A) Interval Estimation for Large Samples ($n > 30$)
(B) Inteval Estimation for Small Samples ($n \leq 30$)

Let us discuss them :

(A) Interval Estimation (or Confidence Interval) for Large Samples ($n > 30$) : In large sample ($n > 30$), the interval estimation is further studied under the following heads :

(1) Confidence Interval or Limits for Population Mean
(2) Confidence Interval or Limits for Population Proportion
(3) Confidence Interval or Limits for Population Standard Deviation
(4) Determination of a Proper Sample Size for Estimating μ or P.

(i) Confidence Interval or Limits for Population Mean μ (when $n > 30$) : The determination of the confidence interval or limits for the population mean μ in case of large sample ($n > 30$) requires the use of normal distribution.

(i) $(1-\alpha)$ 100% Confidence limits for μ are given by :

$$\bar{X} \pm Z_{\alpha/2} \cdot S.E_{\bar{X}}$$

or $$\bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$ where, σ is known.

or $$\bar{X} \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$ when σ is not known. [For large sample, $\sigma = s$]

(ii) $(1-\alpha)$ 100% confidence interval for μ is given by :

$$\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

or $$\bar{X} - Z_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$ where, σ is not known.

In particular, 95% confidence limits for μ are :

$$\bar{X} \pm 1{\cdot}96 \frac{\sigma}{\sqrt{n}}$$ [For large sample, $\sigma = s$]

Similary, 99% confidence limits for μ are

$$\bar{X} \pm 2{\cdot}58 \frac{\sigma}{\sqrt{n}}$$

Procedure: The construction of confidence interval for population mean μ involves the following steps :

(i) Compute $\bar{X}$ or take $\bar{X}$

(ii) Compute the $S.E_{\bar{X}}$ by using the following formula :

(a) $S.E_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, when σ is known.

(b) $S.E_{\bar{X}} = \frac{s}{\sqrt{n}}$, when σ is not known.

(iii) Select the desired confidence level and corresponding to that level of confidence, we find that value of $Z_{\alpha/2}$.

(iv) Substituting the value of $\bar{X}$, $S.E_{\bar{x}}$ and $Z_{\alpha/2}$ in the above stated formula.

Note :

1. If the population S.D. is not known, the sample S.D. (s) is used for large samples.
2. The values of $Z_{\alpha/2}$ (for large samples) corresponding to various level of confidence are given below :

Confidence Level $(1-\alpha)$ 100%	90%	95%	96%	98%	99%	Without any reference to the confidence level
Z-Value	± 1·64	± 1·96	± 2·06	± 2·33	± 2·58	± 3

For other confidence level, the values of $Z_{\alpha/2}$ can be found from the tables of area under the normal curve given at the end of the book.

Note : Where no reference to the confidence interval is given, then we always $Z_{\alpha/2} = 3$. This value corresponds to 99·73% level of confidence.

The following examples illustrate the procedure for setting up confidence limits for μ :

Example 7. **A random sample of 100 observations yields sample mean $\bar{X} = 150$ and sample variance $s^2 = 400$. Compute 95% and 99% confidence interval for the population mean.**

Solution. We are given : $n = 100, \bar{X} = 150, s^2 = 400 \Rightarrow s = 20$

$$S.E_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{20}{\sqrt{100}} = 2$$ [For large sample, $\sigma = s$]

At 95% confidence level, the value of $Z_{\alpha/2} = 1{\cdot}96$

At 99% confidence level, the value of $Z_{\alpha/2} = 2{\cdot}58$

(a) 95% confidence Interval or Limits for μ are :

$$\bar{X} \pm 1{\cdot}96\, S.E_{\bar{X}}$$

Putting the values, we get

$150 \pm 1.96 \times 2 = 150 \pm 3.92 = 153.92$ or 146.08

Thus, $146.08 < \mu < 153.92$

(b) 99% confidence interval or Limits for μ are :

$$\bar{x} \pm 2{\cdot}58 \,.\, S.E_{\bar{X}}$$
$$= 150 \pm 2{\cdot}58 \times 2$$
$$= 150 \pm 5{\cdot}16$$
$$= 155{\cdot}16 \text{ or } 144{\cdot}84$$

Thus, $144{\cdot}84 < \mu < 155{\cdot}16$

Example 8. **A random sample of 900 workers in a steel plant showed an average height of 67 inches with a standard deviation of 5 inches.**

(a) Establish a 95% confidence interval estimate of the mean height of all the workers at the steel plant.

(b) Establish a 99% confidence interval estimate of the mean height of all the workers at the steel plant.

Solution. We are given : $n = 900, \bar{X} = 67, s = 5$

$$S.E.(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{5}{\sqrt{900}} = 0{\cdot}167$$ [For large sample, $s = \sigma$]

At 95% confidence level, the value of $Z_{\alpha/2} = 1{\cdot}96$

At 99% confidence level, the value of $Z_{\alpha/2} = 2{\cdot}58$

(a) 95% confidence interval for μ is :

$$\bar{X} \pm 1{\cdot}96, S.E_{\bar{X}}$$

Putting the values, we get

$$67 \pm 1{\cdot}96 \times (0{\cdot}167)$$
$$= 67 \pm 0{\cdot}327 = 67{\cdot}327 \text{ to } 66{\cdot}673$$

Thus, $66{\cdot}673 < \mu < 67{\cdot}327$

(b) 99% confidence interval for μ are :

$$\overline{X} \pm 2{\cdot}58.\text{S.E}_{\overline{X}}$$

Putting the values, we get

$$= 67 \pm 2{\cdot}58 \cdot (0{\cdot}167)$$
$$= 67 \pm 0{\cdot}43$$
$$= 67{\cdot}43 \text{ to } 66{\cdot}57$$

Thus, $66{\cdot}57 < \mu < 67{\cdot}43$

Example 9. **Upon collecting a sample of 100 from a population with known standard deviation of Rs. 50, the mean is found to be Rs. 500.**
(i) Find 90% confidence interval for the population mean.
(ii) Find 98% confidence interval for the population mean.

Solution. We are given : $n = 100, \overline{X} = 500, \sigma = 50$

$$\text{S.E}_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{50}{\sqrt{100}} = 5 \qquad \text{[Here, } \sigma \text{ is known]}$$

At 90% confidence level, the value of $Z_{\alpha/2} = 1{\cdot}64$
At 98% confidence level, the value of $Z_{\alpha/2} = 2{\cdot}33$.

(a) 90% confidence interval for μ is :

$$\overline{X} \pm 1{\cdot}64 \,.\, \text{S.E}_{\overline{X}}$$

Putting the values, we get

$$= 500 \pm 1{\cdot}64 \cdot (5)$$
$$= 500 \pm 8{\cdot}2$$
$$= 508{\cdot}2 \text{ to } 491{\cdot}8$$

Thus, $491{\cdot}8 < \mu < 508{\cdot}2$

(b) 98% confidence interval for μ is

$$\overline{X} \pm 2{\cdot}33 \,.\, \text{S.E}_{\overline{X}}$$

Putting the values, we get

$$= 500 \pm 2{\cdot}33 \cdot (5)$$
$$= 500 \pm 11{\cdot}65$$
$$= 511{\cdot}65 \text{ to } 488{\cdot}35$$

Thus, $488{\cdot}35 < \mu < 511{\cdot}65$

Confidence Interval or Limits for population mean when sample is drawn without replacement from a finite population. In this case $(1-\alpha)$ 100% confidence interval or limits are given by :

$$\overline{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \qquad \text{when } \sigma \text{ is known}$$

or

$$\overline{X} \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \qquad \text{when } \sigma \text{ is not known.}$$

Where, $\sqrt{\frac{N-n}{N-1}}$ = Finite Population Correction Factor

Example 10. **A random sample of 100 articles selected from a batch of 2000 articles shows that the average diameter of the articles = 0·354 with a standard deviation = 0·048. Find 95% confidence limits for the average of this batch of 2000 articles. [Given Z value with 95% confidence is 1·96]**

Solution. We are given : $N = 2000, n = 100, \overline{X} = 0{\cdot}354$, and $s = 0{\cdot}048$

$$\text{S.E}_{\overline{X}} = \frac{s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \qquad \text{[For Large Sample, } \sigma = s\text{]}$$
$$= \frac{0{\cdot}048}{\sqrt{100}} \times \sqrt{\frac{2000-100}{2000-1}}$$
$$= \frac{0{\cdot}048}{10} \times \sqrt{\frac{1900}{1999}}$$
$$= 0{\cdot}0048 \times \sqrt{0{\cdot}95048} = 0{\cdot}0048 \times 0{\cdot}97493$$
$$= 0{\cdot}00468$$

At 95% confidence level, the value of $Z_{\alpha/2} = 1{\cdot}96$.

95% confidence limits for μ are :

$$\overline{X} \pm 1{\cdot}96 \cdot \text{S.E}_{\overline{X}}$$

Putting the values, we get

$$= 0{\cdot}354 \pm 1{\cdot}96 \times (0{\cdot}00468)$$
$$= 0{\cdot}354 \pm 0{\cdot}009173$$
$$= 0{\cdot}3448 \text{ to } 0{\cdot}3632$$

Example 11. **A manager wants an estimate of average sales of salesmen in his company. A random sample of 121 out of 600 salesmen is selected and average sales is found to be Rs. 760. If the population standard deviation is Rs. 150, manager specifies a 99% level of confidence. What is the interval estimate for population mean μ ?**

Solution. We are given : $N = 600, n = 121, \overline{X} = 760, \sigma = 150$

$$\text{S.E}_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \qquad \text{[Here, } \sigma \text{ is given]}$$
$$= \frac{150}{\sqrt{121}} \times \sqrt{\frac{600-121}{600-1}}$$
$$= 12{\cdot}2$$

At 99% confidence level, $Z_{\alpha/2} = 2{\cdot}58$

99% confidence limits for μ are given by :

$$\overline{X} \pm 2{\cdot}58 \cdot \text{S.E}_{\overline{X}} \qquad \text{[Here, } \sigma \text{ is given]}$$

Putting the values, we get

$$= 760 \pm 2{\cdot}58 \times 12{\cdot}2$$
$$= 760 \pm 31{\cdot}48 = 728{\cdot}52 \text{ to } 791{\cdot}48$$

Thus, $728{\cdot}52 < \mu < 791{\cdot}48$.

EXERCISE – 2

1. A random sample of 144 observations yields sample mean $\overline{X}=160$ and sample variance $s^2=100$. Compute a 95% confidence interval for population mean.
[**Ans.** $158\cdot37<\mu<161\cdot03$]

2. From a random sample of 64 farms are found to have a mean area of 45 hectares with a standard deviation of 12. What are the 95% and 99% confidence limits for the mean area ?
[**Ans.** (a) 47·94, 42·06 (b) 48·87, 41·63]

3. From a random sample of 100 farms are found to have a mean area of 250 hectares with a standard deviation of 50. Compute 99% confidence interval of the mean area. How does the width of the confidence interval change if the size of the sample were increased to 400 ?
[**Ans.** (a) $237\cdot1<\mu<262\cdot9$ (b) reduced to half]
[**Hint :** The width of the confidence interval is inversely related with the size of the sample]

4. Upon collecting a sample of 200 from a population with known standard deviation of 5·23, the mean is found to be 76·3.
(i) Find 90% confidence interval for the mean.
(ii) Find 98% confidence interval for the mean.
[**Ans.** (i) $75\cdot695<\mu<76\cdot905$ (ii) $75\cdot441<\mu<77\cdot159$]

5. A simple random sample of size 100 has mean 15, the population variance being 25. Find an interval estimate of the population mean with confidence level of (i) 99% and (ii) 95%. If the population variance is not given, then what should be done to find out the required interval estimates.
[**Ans.** (i) $14\cdot02<\mu<15\cdot98$, (ii) $13\cdot71<\mu<16\cdot29$; $\sigma^2=s^2=\frac{n}{n-1}\cdot s^2$]

6. A sample of size 64 was drawn from a population consisting of 128 units. The sample mean of the measurements of a certain characteristics was found to be 28. Set up a 96% confidence limits for the population mean, if it is known that the population. S.D. for the characteristic is 4.
[**Hints :** $S.E._{\overline{X}}=\frac{\sigma}{\sqrt{n}}\cdot\sqrt{\frac{N-n}{N-1}}$ For 96% , $Z=2\cdot05$]
[**Ans.** 28·7267 and 27·272]

---

(2) Confidence Interval or Limits for Population Proportion P : Though the sampling distribution associated with proportions is the binomial distribution, the normal distribution can be used as an approximation provided the sample is large (*i.e.*, $n>30$) and both np and $nq\geq5$ (when n is the size of the sample, p is the proportion of success and $q=1-p$).

(1) $(1-\alpha)$ 100% Confidence limits for P are given by :

$$p\pm Z_{\alpha/2}\cdot S.E.(p)$$

or $$p\pm Z_{\alpha/2}\cdot\sqrt{\frac{PQ}{n}}$$ when P is known.

or $$p\pm Z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}$$ when P is not known.

(ii) $(1-\alpha)$ 100% confidence interval for P is given by

$$p-Z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}<P<p+Z_{\alpha/2}\cdot\sqrt{\frac{pq}{n}}$$

In particular, 95% confidence limits for P are :

$$p\pm1\cdot96\cdot\sqrt{\frac{pq}{n}}$$

Similarly, 99% confidence limits for P are :

$$p\pm2\cdot58\cdot\sqrt{\frac{pq}{n}}$$

Procedure : The construction of confidence limits or interval for population proportion involves the following steps :
(i) Compute p or take p.
(ii) Compute the S.E. (p) by using the following formula :

$$S.E.(p)=\sqrt{\frac{PQ}{n}}$$ when P is known.

$$S.E.(p)=\sqrt{\frac{pq}{n}}$$ when P is not known.

(iii) Select the desired confidence level and corresponding to that level, we find the value of $Z_{\alpha/2}$.
(iv) Substituting the values of p, S.E. (p) and $Z_{\alpha/2}$, in the above stated formula.

Note :
1. If the population proportion (P) is not known, then sample proportion (p) is used for large samples.
2. When no reference to the confidence level is given, then always take $Z_{\alpha/2}=3$ for 99·73% confidence level.

Example 12. **Out of 1,200 tosses of a coin, it gave 480 heads and 720 tails. Find the 95 percent confidence interval for the heads.**

Solution. We are given : $n=1200$, $X=$Total heads $(np)=480$

$$p=\text{Sample proportion heads}=\frac{480}{1200}=0\cdot4$$

Also, the population proportion of heads $=P=0\cdot50$

$$\therefore\quad Q=1-P=1-0\cdot50=0\cdot50$$

$$S.E._{(p)}=\sqrt{\frac{PQ}{n}}$$ [For large sample, $p=P$]

$$=\sqrt{\frac{0\cdot5\times0\cdot5}{1200}}=0\cdot0144$$

For 95% confidence level, the value of $Z_{\alpha/2}=1\cdot96$

95% confidence interval for P is given by :

$$p\pm1\cdot96\,S.E._p$$

Putting the values, we get

$$= 0{\cdot}4 \pm 1{\cdot}96 \times 0{\cdot}0144$$
$$= 0{\cdot}4 \pm 0{\cdot}028$$
$$= 0{\cdot}372 \text{ to } 0{\cdot}428$$

Thus, $0{\cdot}372 < P < 0{\cdot}428$

Example 13. **A random sample of 1000 households in a city revealed that 500 of these had Gita. Find 95% and 99% confidence limits for the proportion of households in the city with Gita.**

Solution. We are given : $n = 1000$, x = No. of households having Gita = 500

$$p = \frac{500}{1000} = 0{\cdot}50$$
$$q = 1 - p = 1 - 0{\cdot}50 = 0{\cdot}50$$
$$\text{S.E.}(p) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0{\cdot}50 \times 0{\cdot}50}{1000}} = 0{\cdot}0158$$

For 95% confidence level, the value of $Z_{\alpha/2} = 1{\cdot}96$
For 99% confidence level, the value of $Z_{\alpha/2} = 2{\cdot}58$

(a) 95% confidence limits for P are given by :

$$p \pm 1{\cdot}96\ \text{S.E.}_{\bar{x}}$$

Putting the values, we get

$$= 0{\cdot}50 \pm 1{\cdot}96 \times 0{\cdot}0158$$
$$= 0{\cdot}50 \pm 0{\cdot}031$$
$$= 0{\cdot}531 \text{ to } 0{\cdot}469$$

(b) 99% confidence limits for P are given by :

$$p \pm 2{\cdot}58 \times \text{S.E.}_{\bar{x}}$$

Putting the values, we get

$$= 0{\cdot}50 \pm 2{\cdot}58 \times 0{\cdot}0158$$
$$= 0{\cdot}50 \pm 0{\cdot}040$$
$$= 0{\cdot}054 \text{ to } 0{\cdot}46$$

Example 14. **A random sample of 600 pineapples was taken from a large consignment and 75 of them were found to be bad. Estimate the proportion of bad apples in the consignment and obtain the standard error of the estimate. Assign the limits within which the percentage of bad pineapples in the consignment lies.**

Solution. We are given : $n = 600$ x = No. of bad pineapples = 75

Sample proportion, $p = \frac{75}{600} = 0{\cdot}125 = 12{\cdot}5\%$

$$q = 1 - 0{\cdot}125 = 0{\cdot}875$$
$$\text{S.E.}(p) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0{\cdot}125 \times 0{\cdot}875}{600}} = 0{\cdot}013$$

Since, the level of confidence is not specified, we assume it as 99·73%.
For, 99·73% confidence level, the value of $Z_{\alpha/2} = 3$.
99·73% confidence limits for P are given by

$$p \pm 3 \times \text{S.E.}_{\bar{x}}$$

Putting the values, we get

$$= 0{\cdot}125 \pm 3 \times 0{\cdot}013$$
$$= 0{\cdot}125 \pm 0{\cdot}039$$
$$= 0{\cdot}164 \pm 0{\cdot}086$$

Hence, the required percentage lies between 16·4% and 8·6%.

Confidence Interval or Limits for Population Proportion P when the sample is drawn without replacement from a finite population : In this case, $(1-\alpha)$ 100% confidence interval or limits are given by :

$$p \pm Z_{\alpha/2} \cdot \sqrt{\frac{pq}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

where, $\sqrt{\frac{N-n}{N-1}}$ = Finite Population Correction Factor

Note : If N is sufficiently large as compared to the sample size n, the finite population correction factor may be ignored.

Example 15. **Out of 20,000 customers ledger accounts, a sample of 600 accounts was taken to test the accuracy of posting and balancing where in 45 mistakes were found. Assign limits within which the number of defective cases can be expected at 95% level.**

Solution. We are given : $n = 600$, $N = 20{,}000$, x = No. of mistakes in the sample ledger accounts = 45

Sample proportion $p = \frac{x}{n} = \frac{45}{600} = 0{\cdot}075$

$$q = 1 - p = 1 - 0{\cdot}075 - 0{\cdot}925$$

Since, N is sufficiently large as compared to the sample size n, the finite population correction factor $\sqrt{\frac{N-n}{N-1}}$ may be ignored. Hence, assuming it as a sample from finite (large) population, the standard error of p is given by

$$\text{S.E.}(p) = \sqrt{\frac{pq}{n}}$$
$$= \sqrt{\frac{0{\cdot}075 \times 0{\cdot}925}{600}} = \sqrt{0{\cdot}0001156}$$
$$= 0{\cdot}011 \text{ (approx)}$$

For 95% confidence level, the value of $Z_{\alpha/2} = 1{\cdot}96$.
95% confidence limits for population P are given by :

$$p \pm 1{\cdot}96\ \text{S.E.}_{\bar{x}}$$

Putting the values, we get

$$= 0{\cdot}075 \pm 1{\cdot}96 \times 0{\cdot}011$$
$$= 0{\cdot}075 \pm 0{\cdot}022 = (0{\cdot}053,\ 0{\cdot}097)$$

Hence, the number of defective cases in a lot of 20,000 are expected to lie between : $20,000 \times 0 \cdot 053$ and $20,000 \times 0 \cdot 097$ *i.e.*, 1060 and 1940.

Note : If the finite population correction factor is not ignored then;

95% confidence limits for P are :

$$p \pm 1 \cdot 96 \cdot \sqrt{\frac{pq}{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

$$= 0 \cdot 075 \pm 1 \cdot 96 \sqrt{\frac{0 \cdot 075 \times 0 \cdot 925}{600}} \times \sqrt{\frac{20,000-600}{20,000-1}}$$

$$= 0 \cdot 075 \pm 1 \cdot 96 \times 0 \cdot 0108$$

$$= 0 \cdot 075 \pm 0 \cdot 021168$$

$$= (0 \cdot 0538,\ 0.096168)$$

Hence, the required number of defective cases in the lot lies between 20,000 (0·0538, 0·096168) *i.e.*, 1076 and 1924 .

EXERCISE – 3

1. A random sample 300 households in a city revealed that 123 of these houses had Gita. Find a 95 percent confidence interval for the proportion of hoseholds in the city with Gita. **[Ans.** $0 \cdot 355 < P < 0 \cdot 465$**]**
2. A random sample of 500 houses in a city disclosed that 125 of these houses had colour T.V. sets. Find a 98 percent confidence interval for the proportion of houses in the city with colour T.V. sets. (Table value of Z for 98% confidence level is 2·33). **[Ans.** $0 \cdot 205 < P < 0 \cdot 295$**]**
3. In a market survey for the introduction of a new product given in a town, a sample of 400 persons was drawn. When they were approached for sale, 80 of them purchased the product. Find a 95% confidence limits for the purchase of persons who would buy the product in the town. **[Ans.** 0·1608, 0·2392**]**
4. In a large consignment of oranges, a random sample of 100 organs revealed that 5 oranges were bad. Set up 96% confidence limits for the proportion of defective oranges in the whole consignment. **[Ans.** $0 \cdot 005 < P < 0 \cdot 095$**]**
5. A sample of 500 screws is taken from a large consignment and 65 are found to be defective. Estimate the percentage of defectives in the consignment and assign limits within which the percentage lies. **[Ans.** (a) $0 \cdot 085 < P < 0 \cdot 175$ (b) 8·5% to 17·5%**]**
6. In a random sample of 1000 civil servants, the proportion of those having favourable reaction to the newly introduced income tax structure was observed to be 0·45. Construct a 95 percent confidence interval of the proportions of all civil servants having favourable section. **[Ans.** $0 \cdot 419 < P < 0 \cdot 48$**]**
7. A random sample of 700 units from a large consignment showed that 200 were damaged. Find (i) 95 percent and (ii) 99 percent confidence limits for the proportion of damaged units. **[Ans.** (i) $0 \cdot 253 < P < 0 \cdot 319$ (ii) $0 \cdot 242 < P < 0 \cdot 330$**]**
8. Out of 10,000 customers ledger accounts, a sample of 400 accounts was selected to judge and accuracy of posting and balancing. It contained 40 mistakes. Assign limits within which the number of defective cases could be expected at 95 percent level. **[Ans.** (a) $0 \cdot 071 < P < 0 \cdot 129$ (b) $710 < x < 1290$**]**

(3) Confidence Interval or Limits for Population Standard Deviation : The determination of the confidence interval or limits for population S.D. σ in case of large sample ($n > 30$) requires the use of normal distribution.

(i) $(1-\alpha)$ 100% confidence limits for σ are given by

$$s \pm Z_{\alpha/2} \cdot \text{S.E.}_s$$

or

$$s \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{2n}} \qquad \text{when } \sigma \text{ is known.}$$

or

$$s \pm Z_{\alpha/2} \cdot \frac{s}{\sqrt{2n}} \qquad \text{when } \sigma \text{ is not known.}$$

(ii) $(1-\alpha)$ 100% confidence interval for σ is given by :

$$s - Z_{\alpha/2} \cdot \frac{s}{\sqrt{2n}} < \sigma < s + Z_{\alpha/2} \cdot \frac{s}{\sqrt{2n}}$$

In particular, **95% confidence limits for σ are :**

$$s \pm 1 \cdot 96 \cdot \frac{s}{\sqrt{2n}} \qquad \text{[For large sample, } s = \sigma\text{]}$$

Similarly, 99% confidence limits for σ are :

$$s \pm 2 \cdot 5 \pm \frac{s}{\sqrt{2n}}$$

Procedure : The construction of confidence limits for σ involves the following steps :

(i) Compute s or take s

(ii) Compute S.E. (s) by using the following formula :

$$\text{S.E.}(s) = \frac{\sigma}{\sqrt{2n}}$$

or

$$\text{S.E.}(s) = \frac{s}{\sqrt{2n}}$$

(iii) Select the desired confidence level and corresponding to that confidence level, the value of $Z_{\alpha/2}$.

(iv) Substituting the values of s, $Z_{\alpha/2}$ and n in the above stated formula.

Example 16. **A random sample of 50 observations gave a value of its standard deviation equal to 24·5. Construct a 95% confidence interval for population standard deviation σ.**

Solution. We are given : $n = 50$, $s = 24 \cdot 5$

$$\text{S.E.}(s) = \frac{s}{\sqrt{2n}} = \frac{24 \cdot 5}{\sqrt{100}} = 2 \cdot 45$$

For 95% confidence level, the values of $Z_{\alpha/2} = 1 \cdot 96$.

95% confidence interval for σ is given by :

$$s \pm 1 \cdot 96 \cdot \text{S.E.}_s$$

Putting the values we get

$$= 24 \cdot 5 \pm 1 \cdot 96 \times 2 \cdot 45$$

$$= 24 \cdot 5 \pm 4 \cdot 802$$

$$= 29 \cdot 032 \text{ to } 19 \cdot 698$$

Thus, $19 \cdot 698 < \sigma < 29 \cdot 302$

EXERCISE – 4

1. A sample of 100 items gives a standard deviation of 25. Set up the limits for the population standard deviation at 95% level of confidence. [Ans. 21·55, 28·46]
2. A sample of 100 items gives a standard deviation of 4700. Set up the limits for the population standard deviation at 99% confidence level of confidence. [Ans. 5032·4, 4367·60]

(4) Determination of a Proper Sample Size for Estimating μ or P:

So far we have calculated the confidence intervals based on the assumption that the sample size n is known. In most of the practical situation, generally, sample size is not known. The method of determining a proper sample size is studied under two headings :

(a) Sample Size for Estimating a Population Mean

(b) Sample Size for Estimating a Population Proportion

(a) Sample Size for Estimating a Population Mean : In order to determine the sample size for estimating a population mean, the following three factors must be known :

(i) the desired confidence level and the corresponding values of Z.

(ii) the permissible sampling error E.

(iii) the standard deviation σ or an estimate of σ (*i.e.*, $\bar{s}$)

After having known the above mentioned factors, the sample size n is given by :

$$n=\left(\frac{Z.\sigma}{E}\right)^2$$

Note :

1. The values of Z and E are predetermined.
2. The population S.D. (σ) may be actual or estimated.

Example 17. **A cigarette manufacturer wishes to use a random sample to estimate the average nicotin content. The sampling error should not be more than one milligram above or below the true mean, with 99 percent confidence level. The population standard deviation is 4 milligram. What sample size should the company use in order to satisfy these requirements ?**

Solution. We are given : $E=1$, $Z_{\alpha/2}=2{\cdot}58$ for 99% confidence level and $\sigma=4$.

Sample size formula is :

$$n=\frac{Z^2.\sigma^2}{E^2}$$

Substituting the values, we get

$$n=\frac{(2{\cdot}58)^2\,(4)^2}{1^2}=106{\cdot}50 \text{ or } 107$$

Hence, the required sample size $n=107$ which the company should use for their requirements to be fulfilled.

(b) Sample size for Estimating a Population Proportion : In order to determine the sample size for estimating population proportion, the following three factors must be known :

(i) the desired level of confidence and the corresponding value of Z.

(ii) the permissible sampling error E.

(iii) the actual or estimated true proportion of success P.

The sample size n is given by :

$$n=\frac{Z^2\times PQ}{E^2}$$

where, $Q=1-P$

Note :

1. The values of Z and E are predetermined.
2. The value of the population proportion P may be actual or estimated.

Example 18. **A firm wishes to determine with a maximum allowable error of 0·05 and a 98 percent level of confidence the proportion of consumer who prefer its product. How large a sample will be required in order to make such an estimate if the preliminary sales reports indicate that 25 percent of all the consumers prefer the firm's product ?**

Solution. We are given : $E=0{\cdot}05$, $P=0{\cdot}25$, $Q=1-0{\cdot}25=0{\cdot}75$, $Z=2{\cdot}33$ for 98% confidence level.

Sample size formula is

$$n=\frac{Z^2\times PQ}{E^2}$$

Substituting the values, we get

$$n=\frac{(2{\cdot}33)^2}{(0{\cdot}05)^2}(0{\cdot}25)(0{\cdot}75)$$

$$=\frac{5{\cdot}4289}{0{\cdot}0025}(0{\cdot}1875)=\frac{1{\cdot}0179}{0{\cdot}0025}=407{\cdot}16 \text{ or } 408$$

Hence, the required sample size $n=408$.

EXERCISE – 5

1. A firm wishes to estimate with an error of not more than 0·03 and a level of confidence of 98%, the proportion of consumers that prefers its brand of household detergent. Sales reports indicate that about 0·20 of all consumers prefer the firm's brand. What is the requisite sample size ? [Ans. $n=965$]
2. Mr. X wants to determine the average time to complete a certain job. The past records show that population standard deviation is 10 days. Determine the sample size so that Mr. X may be 95% confident that the sample average remains ± 2 days of the average. [Ans. $n=96$]
3. In measuring reaction time, a psychologist estimates that the standard deviation is 0·05 seconds. How large a sample of measurements must be taken in order to be 95% confident that the error of his estimate will not be exceeded 0·01 seconds. [Ans. $n=96$]

B. INTERVAL ESTIMATION (OR CONFIDENCE INTERVAL) FOR SMALL SAMPLES ($n \leq 30$)

The determination of confidence intervals in case of small sized sample ($n \leq 30$) is studied under two headings :

(1) Confidence Interval or Limits for Population Mean μ

(2) Confidence Interval or Limits for Population Variance σ^2

(1) Confidence Interval or Limits for Population Mean ($n \leq 30$) : When the samples size is small (*i.e.*, $n \leq 30$) and σ (the population S.D.) is unknown the desired confidence interval or limits for population mean μ can be found by making use of t-distribution. In case of small samples, t-values are used in place of Z-values.

(i) $(1-\alpha)$ 100% confidence limits for population Mean μ are given by :

$$\bar{X} \pm t_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}} \quad \text{where, } \hat{s} = \text{modified sample S.D.} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} \text{ or } \hat{s} = \sqrt{\frac{n}{n-1}} \cdot s^2$$

(ii) $(1-\alpha)$ 100% confidence interval for μ is given by :

$$\bar{X} \pm t_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}} < \mu < \bar{X} \pm t_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}$$

In particular, 95% confidence limits for μ are given by

$$\bar{X} \pm t_{0.025} \cdot \frac{\hat{s}}{\sqrt{n}}$$

Similarly, 99% confidence limits for μ are given by

$$\bar{X} \pm t_{0.005} \cdot \frac{\hat{s}}{\sqrt{n}}$$

Procedure : The construction of the confidence interval or limits in case of small sample ($n \leq 30$) involves the following steps :

(i) Compute $\bar{X}$ or take $\bar{X}$.

(ii) Compute modified sample S.D. using the following formula.

$$\hat{s} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$$

or $$\hat{s} = \sqrt{\frac{n}{n-1} \cdot s^2} \quad \text{when } s \text{ is given.}$$

(iii) Compute the degree of freedom (*d.f.*) using the formula :

$$d.f. = v = n-1$$

(iv) Select the desired confidence level and corresponding to that specified level of confidence and for given degrees of freedom, we note the value of the $t_{\alpha/2}$ from the t-table.

(v) Substituting the values of $\bar{X}$, $\hat{s}$ and $t_{\alpha/2}$ in the above stated formula.

Example 19. **A random sample of size 16 has 50 as mean with standard deviation of 3. Obtain 98 percent confidence limits of the mean of the population.**

Solution. We are given : $n=16, \bar{X}=50, s=3 \Rightarrow s^2=9$

$$\hat{s} = \sqrt{\frac{n}{n-1} \cdot s^2} = \sqrt{\frac{16}{16-1} \times 9} = 3.098$$

Degrees of freedom $= v = n = 16-1 = 15$

For a 98% confidence level, $\alpha = 0.02$ so that $\frac{\alpha}{2} = \frac{0.02}{2} = 0.01$

Using t-table the value of $t_{.01}$ for 15 d.f. $= 2.602$.

98% confidence limits for μ are given by

$$\bar{X} \pm t_{.01} \cdot \frac{\hat{s}}{\sqrt{n}}$$

Putting the values, we get

$$= 50 \pm 2.602 \times \frac{3.098}{\sqrt{16}}$$

$$= 50 \pm 2.015$$

$$= 52.015 \text{ to } 47.985$$

Example 20. **A random sample of 16 items from a normal population showed a mean of 53 and the sum of squares of deviations from this mean is equal to 150. Obtain 95% and 99% confidence limits for the mean of the population.**

Solution. We are given : $n=16, \bar{X}=53, \Sigma(X-\bar{X})^2 = 150$

$$\hat{s} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$$

$$= \sqrt{\frac{150}{16-1}} = \sqrt{\frac{150}{15}} = \sqrt{10} = 3.162$$

Degrees of freedom $= v = n-1 = 16-1 = 15$

For a 95% confidence level, $\alpha = 0.05$ so that $\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$

For a 99% confidence level, $\alpha = 0.01$ so that $\frac{\alpha}{2} = \frac{0.01}{2} = 0.005$

The table value of $t_{.025}$ for 15 d.f. $= 2.131$

The table vale of $t_{.005}$ for 15 d.f. $= 2.947$

(a) 95% confidence limits for population mean μ are :

$$\bar{X} \pm t_{0.025} \cdot \frac{\hat{s}}{\sqrt{n}}$$

Putting the values, we get

$$= 53 \pm 2.131 \times \frac{3.162}{\sqrt{16}}$$

$$= 53 \pm 2.131 \times \frac{3.162}{4}$$

$$= 53 \pm 1.684$$

$$= 51.316 \text{ to } 54.684$$

$$51\cdot316 < \mu < 54\cdot684$$

Thus,

(b) 99% confidence limits for population mean μ are

$$\overline{X} \pm t_{0\cdot005} \cdot \frac{\hat{s}}{\sqrt{n}}$$

Putting the values, we get

$$= 53 \pm 2\cdot947 \times \frac{3\cdot162}{\sqrt{16}}$$

$$= 53 \pm 2\cdot947 \times \frac{3\cdot162}{4}$$

$$= 53 \pm 2\cdot947 \times 0\cdot7905$$

$$= 53 \pm 2\cdot33$$

$$= 50\cdot67 \text{ and } 55\cdot33$$

Thus, $50\cdot67 < \mu < 55\cdot33.$

Example 22. **A sample of 5 individuals had the following heights in centimeters: 6·33, 6·37, 6·36, 6·32 and 6·37. Find out the unbiased and efficient estimates of (a) true mean and (b) the variance. Also find 95% confidence interval for true mean (*i.e.*, population mean).**

Solution. (a) Unbiased and efficient estimate of the true mean (*i.e.*, population mean) is given by :

$$\overline{X} = \frac{\Sigma X}{n} = \frac{6\cdot33 + 6\cdot37 + 6\cdot36 + 6\cdot32 + 6\cdot37}{5} = 6\cdot35 \text{ cm}$$

(b) Unbiased and efficient estimate of the true variance (*i.e.*, population variance is given by :

$$\hat{s}^2 = \frac{n}{n-1} \cdot s^2 = \frac{\Sigma (X - \overline{X})^2}{n-1}$$

$$= \frac{(6\cdot33 - 6\cdot35)^2 + (6\cdot37 - 6\cdot35)^2 + (6\cdot36 - 6\cdot35)^2 + (6\cdot32 - 6\cdot35)^2 + (6\cdot37 - 6\cdot35)^2}{5-1}$$

$$= 0\cdot00055 \text{ cm}^2$$

$$\hat{s} = \sqrt{0\cdot00055} = 0\cdot023 \text{ cm.}$$

Part B : We are given : $n = 5, \overline{X} = 6\cdot35, \hat{s} = 0\cdot023$

Degrees of freedom $= v = n - 1 = 5 - 1 = 4$

For a 95% confidence level $\alpha = 0\cdot05$, so that $\frac{\alpha}{2} = \frac{0\cdot05}{2} = 0\cdot025.$

The table value of $0\cdot025$ for 4 d.f. $= 2\cdot776$.

95% confidence limits for population mean μ are given by

$$\overline{X} \pm t_{0\cdot05} \cdot \frac{\hat{s}}{\sqrt{n}}$$

Putting the values, we get

$$= 6\cdot35 \pm 2\cdot776 \times \frac{0\cdot023}{\sqrt{5}}$$

$$= 6\cdot35 \pm 2\cdot776 \times \frac{0\cdot023}{2\cdot236}$$

$$= 6\cdot35 \pm 0\cdot0285$$

$$= 6\cdot3785 \text{ to } 6\cdot3215$$

Thus, $6\cdot3215 < \mu < 6\cdot3785$

EXERCISE – 6

1. A sample of 9 cigarettes of a certain brand was observed for nicotine content. It showed an average nicotine of 25 milligrams and a standard deviation of 2·8 milligrams. Construct a 99 percent confidence interval for the true average nicotine content of this particular brand of cigarettes. [Ans. $21\cdot67 < \mu < 28\cdot33$]
2. A random sample of 15 ladies from a colony in Chandigarh shows that their monthly expenditure on cosemetics is Rs. 120 with a standard deviation of Rs. 40. Construct 95 percent confidence interval for the true monthly average expenditure on cosemetics by all the ladies in Chandigarh. [Ans. $97\cdot01 < \mu < 142\cdot99$]
3. A random sample of size 9 has 49 as mean. The sum of squares of deviations taken from mean is 52. Obtain 95% and 99% confidence limits for the mean.
[Ans. (a) $47\cdot04 < \mu < 50\cdot96$ (b) $46\cdot14 < \mu < 51\cdot96$]
4. A sample of 10 measurements of the diameter of a sphere gave a mean $\overline{X} = 4\cdot38$ and standard deviation $s = 0\cdot06$ inches. Find (a) 95% and (b) 99% confidence limits for the actual diameter. [Ans. (a) 4·425 to 4·335 (b) 4·444 to 4·316]
5. A random sample of 10 families had the following percentage expenses on food : 68, 60, 75, 70, 73, 69, 59, 60, 49 and 44. Obtain 95% and 99% confidence limits for the population mean.
[Ans. (a) 70·72 to 55·38 (b) 73·23 and 52·17]
6. A survey of 17 agricultural labourers reveals an income of Rs. 40 per week with a standard deviation of Rs. 8. Find out the limits of mean weekly wages in the population with a confidence of 95%. (Given $t = 2\cdot131$ for 16 df). [Ans. $35\cdot74 < \mu < 44\cdot26$]
7. A sample of 6 persons in an office revealed an average daily smoking of 10, 12, 8, 9, 16, 5 cigarettes. Determine unbiased and efficient estimates of (a) the true mean and (b) the true variance. Also find 95% confidence interval for the true mean.
[Ans. (a) $\overline{X} = 10$ (b) $\hat{s}^2 = 14$ (c) $6\cdot92 < \mu < 13\cdot08$]

(2) Confidence Interval or Limits for Population Variance (When $n < 30$). The determination of confidence interval or limits for population variance σ^2 requires the use of χ^2 (Chi-square) distribution. Here χ^2-values are used in place t-values.

$(1-\alpha)$ 100% confidence interval for population variance σ^2 is given by :

$$\frac{(n-1)\hat{s}^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)\hat{s}^2}{\chi^2_{1-\alpha/2}}$$

In particular, 95% confidence interval for the population variance σ^2 is

$$\frac{(n-1)\hat{s}^2}{\chi^2_{0\cdot025}} < \sigma^2 < \frac{(n-1)\hat{s}^2}{\chi^2_{0\cdot975}}$$

Similarly, **99% confidence interval for the population variance σ^2 is**

$$\frac{(n-1)\,\hat{s}^2}{\chi^2_{0.005}} < \sigma^2 < \frac{(n-1)\,\hat{s}^2}{\chi^2_{0.995}}$$

Procedure : The construction of the confidence interval for the variance σ^2 involves the following steps :

(i) Calculate modified sample variance $(\hat{s})^2$ by using the formula

$$\hat{s}^2 = \frac{n}{n-1}s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$$

(ii) Select the desired confidence level and corresponding to that specified level of confidence, we note the value of the confidence coefficient $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ from the χ^2-table for certain degrees of freedom

(iii) Construct the confidence interval for σ^2 by putting the values of $\hat{s}^2$, $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ in the above stated formula.

Example 22. **A random sample of size 15 selected from a normal population has a standard deviation $s = 2{\cdot}5$. Construct a 95 percent confidence interval for variance σ^2 and standard deviation σ.**

Solution. We are given : $n = 15$, $s = 2{\cdot}5 \Rightarrow s^2 = 6{\cdot}25$

$$\hat{s}^2 = \left(\frac{n}{n-1}\right)s^2$$

$$= \frac{15}{15-1} \times 6{\cdot}25 = 6{\cdot}696$$

For a 95% confidence level, $\alpha = 0{\cdot}05$ so that $\frac{\alpha}{2} = 0{\cdot}025$ and $1-\alpha = 1-0{\cdot}025 = 0{\cdot}975$.

Degrees of freedom $(v) = n-1 = 15-1 = 14$

The table value of $\chi^2_{.025}$ for 14 d.f. $= 26{\cdot}1$

The table value of $\chi^2_{0{\cdot}0975}$ for 14 d.f. $= 5{\cdot}63$

(a) 95% confidence interval for σ^2 is

$$\frac{(n-1)\,\hat{s}^2}{\chi^2_{.025}} < \sigma^2 < \frac{(n-1)\,\hat{s}^2}{\chi^2_{.975}}$$

Putting the values, we get

$$\frac{(15-1)\times 6{\cdot}696}{26{\cdot}1} < \sigma^2 < \frac{(15-1)\times 6{\cdot}696}{5{\cdot}63}$$

or $\quad 3{\cdot}59 < \sigma^2 < 16{\cdot}65$

(b) 95% confidence interval for σ is :

$$\sqrt{3{\cdot}59} < \sigma < \sqrt{16{\cdot}65}$$

or $\quad 1{\cdot}89 < \sigma < 4{\cdot}08$

Example 23. **A sample of 5 individuals had the following heights in inches 63·3, 63·7, 63·6, 63·2 and 3.8. Construct 95% confidence interval for population variance.**

Solution.

$$\bar{X} = \frac{63{\cdot}3 + 63{\cdot}7 + 63{\cdot}6 + 63{\cdot}2 + 63{\cdot}8}{5} = 63{\cdot}52$$

$$\hat{s}^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$$

$$= \frac{(63{\cdot}3-63{\cdot}52)^2 + (63{\cdot}7+63{\cdot}52)^2 + (63{\cdot}6-63{\cdot}52)^2 + (63{\cdot}2-63{\cdot}52)^2 + (63{\cdot}8-63{\cdot}52)^2}{5-1}$$

$$= \frac{{\cdot}0484 + 0{\cdot}0324 + 0{\cdot}0064 + 0{\cdot}1024 + 0{\cdot}0784}{4}$$

$$= \frac{0{\cdot}268}{4} = 0{\cdot}067$$

95% confidence interval for σ^2 is

$$\frac{(n-1)\,\hat{s}^2}{\chi^2_{0.025}} < \sigma^2 < \frac{(n-1)\,\hat{s}^2}{\chi^2_{0.975}}$$

Degrees of freedom $(v) = n-1 = 5-1 = 4$

The table value of $\chi^2_{0.025}$ for 4 d.f. $= 11{\cdot}14$

The table value of $\chi^2_{0.0975}$ for 4 d.f. $= {\cdot}484$

95% confidence interval for σ^2 is

$$\frac{(5-1)\times(\cdot067)}{11{\cdot}14} < \sigma^2 < \frac{(5-1)(\cdot067)}{{\cdot}484}$$

$$\Rightarrow \quad \frac{0{\cdot}268}{11{\cdot}14} < \sigma^2 < \frac{0{\cdot}268}{{\cdot}484}$$

$$\Rightarrow \quad 0{\cdot}0240 < \sigma^2 < 0{\cdot}5537$$

EXERCISE – 7

1. A random sample of size 12 selected from a normal population has a standard deviation $s = 2{\cdot}4$. Construct 95 percent confidence interval for (a) variance σ^2 and (b) standard deviation σ. **[Ans.** (a) $3{\cdot}15 < \sigma^2 < 18{\cdot}08$ (b) $1{\cdot}77 < \sigma < 4{\cdot}25$**]**
2. A random sample of size 25 selected from normal population has a standard deviation $s = 7$. Construct 95% confidence interval for (a) variance σ^2, (b) standard deviation σ. **[Ans.** (a) $31{\cdot}12 < \sigma^2 < 98{\cdot}79$ (b) $5{\cdot}578 < \sigma < 9{\cdot}939$ **]**
3. A random sample of 15 ladies of a posh locality shows that their monthly expenditure on cosmetics is Rs. 120 with a standard deviation of Rs. 40. Construct 99 percent confidence interval for (a) variance σ^2 and (b) standard deviation σ. **[Ans.** (a) $766{\cdot}8 < \sigma^2 < 5896{\cdot}8$ (b) $27{\cdot}7 < \sigma^2 < 76{\cdot}7$**]**

QUESTIONS

1. What is statistical estimation ? Distinguish between point estimation and interval estimation. Describe the desirable properties of a good estimator.
2. What is an estimator ? Discuss the important properties of a good estimator. Show that the sample mean is a good estimate of population mean.
3. Differentiate between :
   (a) Estimator and Estimate
   (b) Statistic and Parameter
   (c) Point Estimator and Interval Estimate.
4. Explain the (i) consistency (ii) Unbiasedness (iii) Efficiency and (iv) Sufficiency properties of an estimator.
5. Define unbiased and efficient estimates of (a) true mean and (b) true variance.

   OR

   Define efficiency and unbiasedness of an estimator.
6. Define unbiased and consistency properties of an estimator.
7. Explain the concept of confidence interval or interval estimation. Outline the procedure for setting up a confidence interval for the population parameter.
8. Explain the procedure for setting up a confidence interval for (a) population mean (b) population proportions and (c) population variance.
9. Define the following terms and given an example of each :
   (i) Unbiased statistic (ii) Consistent statistic
   (iii) Efficient statistic (iv) Sufficient statistic
10. Show that the sample mean ($\overline{X}$) is an unbiased estimate of the population mean (μ).
11. Explain why a random sample of size 25 is to be preferred to a random sample of 20 to estimate population mean.

❀❀❀

8 Non-Parametric Tests

INTRODUCTION

Sampling tests, discussed so far, are known as parametric tests because these are based on the assumptions that the concerned sample has been obtained from a population with known values of its one or more parameters. For example, the use of *t*-test to test the $H_0 : \mu_1 = \mu_2$ requires that the two samples are drawn from normal population with equal variance ($\sigma_1^2 = \sigma_2^2$). Similarly, the use of *F*-test requires to test $H_0 : \sigma_1^2 = \sigma_2^2$ assumes that various samples are obtained from normal population with equal variance etc. The validity of the results of a parametric test depends upon the appropriateness of these assumptions. Thus, when these assumptions are not met, the parametric tests are no longer applicable. In such cases, it is essential to study non-Parametric tests.

MEANING OF NON-PARAMETRIC TESTS

Non-parametric tests do not require any assumptions about the parameters or about the nature of population. By non-parametric tests we mean those statistial tests which do not depend either upon the shape of the distribution or upon the parameters of the population mean, standard deviation, variance, etc. The assumption as to the normality or symmetricity of the population distribution from which the samples have been drawn is not required for these non-parametric tests. Non-parametric tests are sometimes referred to as **distribution free tests**. In addition to this, these non-parametric test do not require measurements so strong as that requires by parametric tests.

DIFFERENCE BETWEEN PARAMETRIC AND NON-PARAMETRIC TESTS

(1) In parametric tests, assumptions of normal population is taken whereas in non-parametric tests, no such assumption is taken about the population. It is because of this that non-parametric tests are known as "**distribution free tests**".
(2) Parametric tests comprise of *t*-test, *Z*-test and *F*-test whereas non-parametric, tests comprise Chi-square test, Sign test, Median test, Wilcoxon signed rank test, etc.
(3) In case of a parametric test a normal hypothesis is set up and the variable is tested for drawing inferences. In case of non-parametric tests, inverse (opposite hypothesis) is set up.

ADVANTAGES OR USES OF NON-PARAMETRIC TESTS

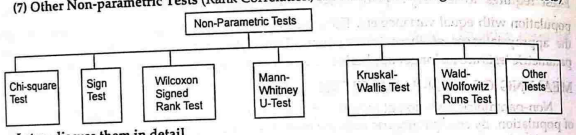
Some advantages (or uses) of non-parametric tests are mentioned below :
(1) Non-parametric tests are distribution free *i.e.*, they do not require any assumption to be made about population following normal or any other distribution.
(2) Generally, they are simple to understand and easy to apply when the sample sizes are small.

(3) Most non-parametric tests do not require lengthy and laboroius arithmetical computations and hence are less time-consuming.
(4) Non-parametric tests make fewer and less restative assumptions than do the parametric tests.
(5) There is no alternative to using a non-parametric test if the data are available in ordinal or nominal scale.
(6) Non-parametric tests are useful to handle data made up of samples from several populations without making assumptions.

TYPES OF NON-PARAMETRIC TESTS

There are many types of non-parametric tests. The important among them are :

(1) **Chi-square Test (discussed earlier in χ^2 Chapter)**
(2) **Sign Test**
(3) **Wilcoxon Signed Rank Test**
(4) **Mann-whitney U-Test**
(5) **Kruskal-Wallis Test**
(6) **Wald-Wolfowitz Runs Test**
(7) **Other Non-parametric Tests** (Rank Correlation, Median and Kolmogorov-Smirnov tests)



Let us discuss them in detail

(1) Sign Test : The sign test is the simplest type of all the non-parametric tests. Its name comes from the fact that it is based on the direction or the plus or minus signs of observations in a sample and not on their numerical magnitudes.

Types of Sign Test

The sign test can be of two types :
(a) One-sample sign test, and
(b) Paired-sample sign test

(a) One-sample sign test : In one-sample sign test, we set up the null hypothesis that + and − signs are the values of a random variables having the binomial distribution with $p=\frac{1}{2}$ *i.e.*,

$H_0 : p=\frac{1}{2}$ or that $\mu=\mu_0$

Procedure : This test involves the following steps :

(i) Find the + and − sign for the given distribution. Put a plus (+) sign for a value greater than the mean value (μ_0), a minus (−) sign for a value smaller than the mean value and a zero (0) for a value equal to the mean value.
(ii) Denote the total number of signs (ignoring zeros) by n and the number of less frequent signs by 'S'.
(iii) Obtain the critical value (K) of less frequent signs (S) preferably at 5% level of significance by using the following formula :

$$K=\frac{n-1}{2}-0\cdot98\sqrt{n}$$

(iv) Compare the value of 'S' with the critical value (K). If the value of S is greater than the value of K (*i.e.*, $S>K$) then the null hypothesis is accepted. If $S\le K$, the null hypothesis is rejected.

Alter : The problem relating to one sample test can also be solved by using Binomial Probability Distribution. When the sample size is fairly small (*i.e.*, $n\le25$), we find probability of the less frequent signs $p(S)$ by the sum of the probability of S of fewer S using the binomial distribution formula, ${}^nC_x q^{n-x}, p^x$ with $p=\frac{1}{2}$. Then, we compare the above calculated value of probability with the expected value at 5% level of significance *i.e.*, at $a=0\cdot05$ for one tailed/two tailed tests. If the calculated probability $P(S)$ is $\le0\cdot05$, null hypothesis is rejected and if $P(S)>0\cdot05$, the null hypothesis is accepted.

Example 1. **The production manager of a large underlaking randomly paid 10 visits to the worksite in a month. The number of workers who reported late for duty was found to be 2, 4, 5, 1, 6, 3, 2, 1, 7 and 8 respectively. Using the sign test verify the claim of the production supervisor that on an average, not more than 3 workers report late for duty. Use 5% level of significance.**

Solution: Let $H_0 : \mu\le3$ against $H_1 : \mu>3$ at 5% level of significance.

Determination of Signs w.r.t. $\mu=3$

X	Signs $(X-3)$	
2	−	
4	+	No. of Plus Signs = 5
5	+	
1	−	No. of Minus Signs = 4
6	+	No. of Zero = 1
3	0	
2	−	Total No. of Signs = n = 9 (ignoring 0's)
1	−	
7	+	
8	+	

From the above table, we get
Total no. of signs (ignoring zeros) = $n=9$
Number of less frequent signs = $S=4$
The critical value (K) of less frequent signs (S) at 5% level is given by :

$$K=\frac{n-1}{2}-0\cdot98\sqrt{n}=\frac{9-1}{2}-0\cdot98\sqrt{9}=4-2\cdot94=1\cdot06$$

Since number of less frequent signs $S(4)$ is more than the critical value of $K(1\cdot06)$ *i.e.*, $S>K$, the null hypothesis is accepted. It means that the sample data support the claim of the production supervisor.

Aliter : We can solve the above problem using Binomial Probability Distribution.

We are given, $n=9$, $p=\frac{1}{2}$, $\alpha=0\cdot05$, $S=4$

The probability of 4 or fewer success is given by :

$$P[S\le4]={}^9C_4\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^4+{}^9C_3\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^3+{}^9C_2\left(\frac{1}{2}\right)^7\left(\frac{1}{2}\right)^2+{}^9C_1\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^1+{}^9C_0\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^9$$

$$=126\left(\frac{1}{2}\right)^9+84\left(\frac{1}{2}\right)^9+36\left(\frac{1}{2}\right)^9+9\left(\frac{1}{2}\right)^9+\left(\frac{1}{2}\right)^9$$

$$=\frac{1}{512}\cdot[126+84+36+9+1]=\frac{256}{512}=\frac{1}{2}=0\cdot5$$

From the above, it is found that $P(S)>\alpha$ [$\alpha=\cdot05$]

This suggests that the null hypothesis is accepted. It means that the sample data support the claim of the production manager.

Example 2. **Suppose playing four rounds of golf at the city club 11 professionals totalled 280, 282, 290, 273, 283, 283, 275, 284, 282, 279 and 281. Use the sign test at 5% level of significance to test the null hypothesis that professional golfers average $\mu=284$ for four rounds against the alternative $\mu<284$.**

Solution: Let $H_0:\mu=284$ against $H_1:\mu<284$ at 5% level of significance.

Determination of signs w.r.t. $\mu=284$

X :	280	282	290	273	283	283	275	284	282	279	281
Signs $(X-284)$	–	–	+	–	–	–	–	0	–	–	–

From the above table, we get

No. of Plus signs = 1; No. of Minus signs = 9, No. of Zero = 1

Total no. of signs (ignoring zero) = $n=10$

Number of less frequent signs (+) = $S=1$

The critical value (K) of less frequent signs (S) at 5% level is given by

$$K=\frac{n-1}{2}-0\cdot98\sqrt{n}=\frac{10-1}{2}-0\cdot98\sqrt{10}=4\cdot5-3\cdot099=1\cdot40$$

Since S is less than critical value of K (1·40) i.e., $S<K$, the null hypothesis is rejected. It means that professional golfers average is less than 284 for four rounds of golf.

Aliter : We can solve the problem by using Binomial Probability Distribution.

We have, $n=10$, $p=\frac{1}{2}$, $\alpha=0\cdot05$, $S=1$

The probability of one or fewer successes with $n=10$ and $p=\frac{1}{2}$ is given by

$$P[S\le1]={}^{10}C_1\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right)^1+{}^{10}C_0\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^0$$

$$=10\times\frac{1}{1024}+\frac{1}{1024}=\frac{11}{1024}=0\cdot0107 \text{ app.}$$

From the above, it is found that $P(S)<\alpha$, [$\because\alpha=0\cdot05$]

The null hypothesis is rejected. It means that the professional golfers' average is less than 284 four for rounds of golf.

(b) Paired Sample Sign Test : The sign test has very important applications in problems involving paired data such as data relating to the collection of an account receivable before and after a new collection policy; responses of mother and daughter towards ideal family size etc. In such problems, each pair of sample values is replaced with a plus sign if the first value is greater than the second, a minus sign if the first value is smaller than the second or a zero if the two values are equal. Then we proceed in the same manner as in one-sample test.

Example 3. **Use the sign test of see if there is a difference between the number of days until the collection of an account receivable before and after a new collection policy. Use the 0·05 significance level.**

Before :	30	28	34	35	40	42	33	38	34	45	28	27	25	41	36
After :	32	29	33	32	37	43	40	41	37	44	27	33	30	38	36

Solution. Let us take the hypothesis that there is no significant difference before and after the new collection policy in the accounts receives.

Determination of the signs

Before (X)	30	28	34	35	40	42	33	38	34	45	28	27	25	41	36
After (Y)	32	29	33	32	37	43	40	41	37	44	27	33	30	38	36
Signs $(X-Y)$	–	–	+	+	+	–	–	–	+	+	–	–	+	0	

From the above table, we get

No. of Plus signs = 6; No. of Minus signs = 8, No. of Zeros = 1

Total no. of signs (ignoring zero) = $n=14$

The no. of less frequent signs (+) = $S=6$

The critical value (K) of less frequent signs (S) at 5% level is given by

$$K=\frac{n-1}{2}-0\cdot98\sqrt{n}=\frac{14-1}{2}-0\cdot98\sqrt{14}=6\cdot5-3\cdot666=2\cdot83$$

Since $S>K$, the null hypothesis is accepted. This means that there is no significant difference before and after the new collection policy in accounts receivable.

AN IMPORTANT TYPICAL EXAMPLE

Example 4. **A physical instructor claims that a particular exercise if done continuously for 7 days, reduced weight by 3·5 kg. Five overweight girls did the exercise for 7 days and their weights were observed as under :**

Girls :	1	2	3	4	5
Weight before exercise :	70	72	75	71	78
Weight after exercise :	66	70	72	66	72

Making use of the sign test, verify the claim at $\alpha=0\cdot05$ that the exercise reduces the weight by at least 3·5 kg.

Solution: Denoting the mean weight before and after exercise by μ_1 and μ_2 respectively, we have the following,

$$H_0 : \mu_1 - \mu_2 = 3{\cdot}5 \text{ against } H_1 : \mu_1 - \mu_2 < 3{\cdot}5, \alpha = 0{\cdot}05$$

Determination of Signs w.r.t. 3.5

Girls	X	Y	$D = (X - Y)$	Signs $(D - 3{\cdot}5)$
1	70	66	4	+
2	72	70	2	−
3	75	72	3	−
4	71	66	5	+
5	78	72	6	+

From the above table, we get
No. of Plus signs = 3; No. of Minus signs = 2
The total no. of signs or $n = 5$
Number of less frequent signs $= S = 2$ The critical value (K) of less frequent signs (S) is given by :

$$K = \frac{n-1}{2} - 0{\cdot}98\sqrt{n} = \frac{5-1}{2} - 0{\cdot}98\sqrt{5} = 2 - 2{\cdot}19 = -0{\cdot}19$$

Since $S > K$, H_0 is accepted. It means that the sample data support the claim that the exercise if continuously done for 7 days reduces the weight by at least 3·5 kg.

LARGE SAMPLE AND SIGN TEST

When the sample size is fairly large (i.e., $n > 25$), we use the normal approximation to the binomial distribution to carry out the sign test. The value of 'z' can be computed as :

$$Z = \frac{S - np}{\sqrt{np(1-p)}}$$

Then we get the critical value of Z at the desired level of significance.

If the calculated value of Z happens to be less than the critical value, then we accept the null hypothesis. If the case is reverse, then we reject the null hypothesis.

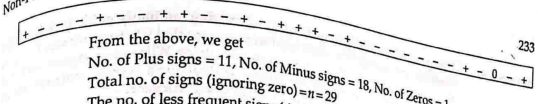
Example 5. **Given below are the data relating to the daily milking from a cow for 30 days :**

Milk in Litres	23	20·8	18·6	16·6	23·2	21	19·2	24·8	23·8	29·6
	19	23·2	20·4	22·8	24·6	22·8	20·8	22·6	21·4	18·6
	18·6	17·4	19·2	20·4	22·2	24·6	20	22·4	20	23

Using sign test to test the null hypothesis at 5% level of significance that the average daily milking from the cow is 22·4 litres as against the alternative hypothesis that it is less than 22·4 litres.

Solution: We have, $H_0 : \mu = 22{\cdot}4$, $H_1 : \mu < 22{\cdot}4$, $\alpha = 0{\cdot}05$ ($\Rightarrow$ left tailed test)

Determination of the signs w.r.t. 22·4



From the above, we get
No. of Plus signs = 11, No. of Minus signs = 18, No. of Zeros = 1
Total no. of signs (ignoring zero) $= n = 29$
The no. of less frequent signs (+) $= S = 11$
Since the size of samples is quite large (n i.e., $n > 25$), the test which is a close approximation to the Binomial distribution is used.
We have,

$$Z = \frac{S - np}{\sqrt{np(1-p)}}$$

Substituting the value in the above, we get

$$Z = \frac{11 - 29\left(\frac{1}{2}\right)}{\sqrt{29 \times \frac{1}{2}\left(1 - \frac{1}{2}\right)}} = \frac{11 - 14{\cdot}5}{\sqrt{14{\cdot}5 \times 0{\cdot}5}} = \frac{-3{\cdot}5}{2{\cdot}69} = -1{\cdot}3$$

The critical value of Z at 5% level for left tailed test $= -1{\cdot}645$
Since the calculated value of $Z <$ the critical value of Z, we accept H_0.
This suggests that the null hypothesis is accepted. It means that the average daily milking from cow is 22·4 litres.

(2) Wilcoxon's Signed-Rank Test

This is another non-parametric test which has been developed by Sir Wilcoxon. This test is based on the ranking of the sample of observations. Like sign test, Wilcoxon's signed-rank test can be of two types :

(a) **One-sample signed rank test.**

(b) **Paired-sample signed rank test.**

(a) Wilcoxon's One-Sample Signed-Rank Test : In a one sample signed-rank test, we test the null hypothesis that $\mu = \mu_0$ against an appropriate alternative hypothesis at a desired level of significance.

Procedure : This test involves the following steps :

(i) Calculate the difference $d = x - \mu$ with algebraic signs.

(ii) Assign ranks (ignoring the signs) to the difference in the increasing order of magnitude (i.e., from low to high) ignoring zero differences. In case of ties (i.e., when two or more values are the same), assign ranks to such pairs by averaging their rank positions.

(iii) Put all the ranks against the + ve difference in the + ve rank column (R^+) and all the ranks against the − ve differences in the − ve rank column (R^-)

(iv) Get the total number of ranks, n

(v) If $n \leq 25$, then calculate the value of the test statistic given by $T = \Sigma R^+$ or ΣR^- ranks which ever is less.

(vi) Then find the critical value of T from the Wilcoxon's T-table given at the end of the book with reference to the values of n and (i.e., significance level).

(vii) If the calculated value of T is less than or equal to its critical value, then reject the null hypothesis. In the reverse case, accept the null hypothesis.

Example 6. **The production manager of a large undertaking randomly paid 10 visits to the worksite in a month. The number of workers reported late for duty was found to be : 2, 4, 5, 1, 6, 3, 2, 1, 7 and 8 respectively. Using Wilcoxon's signed-rank test verify the claim of the production supervisor that on an average, not more than 3 workers report late for duty. Use 5% level of significance.**

Solution: Let us take the hypothesis that $H_0 : \mu = 3$ against $H_1 : \mu > 3$

X	$d = X - 3$	$\lvert d \rvert$	Ranks (± ignored)	Signed Ranks	
				R^+	R^-
2	– 1	1	2	–	2
4	1	1	2	2	–
5	2	2	5	5	–
1	– 2	2	5	–	5
6	3	3	7	7	–
3	0	0	–	–	–
2	– 1	1	2	–	2
1	– 2	2	5	–	5
7	4	4	8	8	–
8	5	5	9	9	–
Total			$n = 9$	$\Sigma R^+ = 31$	$\Sigma R^- = 14$

From the above table, it must be seen that total number of ranks $= n = 9$ (*i.e.*, < 25).
Since $n < 25$, the test statistic is given by :

T = smaller of the two sums of the signed-ranks = 14

Looking at Wilcoxon's T table at 5% level for one tailed test at $n = 9$, we get the critical value of T or $T_{0.05} = 8$.
Since the calculated value of T (14) is greater than its critical value (8), the null hypothesis is accepted. It means that the sample data support the claim made by the production manager.

(ii) Two Sample Signed Rank Test (or Paired–Sample Signed Rank Test) : The Wilcoxon's signed rank test has important applications in problem involving paired data. Such a test is widely used by the research scholars in their study of two related samples or matches pairs of ordinary data *viz.*, outputs of two similar machines, responses gathered before and after a treatment, etc. where we can find both the direction and magnitude of difference between the matched values. In these problems, we find the difference between each pair of values with algebraic signs. Then we proceed in the same manner as in the case of one sample signed rank test.

Example 7. **Use Wilcoxon's signed-rank test to see if there is a difference between the number of days until the collection of an account receivable before and after a new collection policy. Use the 0·05 level of significance.**

Before (X) :	30	28	34	35	40	42	33	38	34	45	28	27	25	41	36
After (Y) :	32	29	33	32	37	43	40	41	37	44	27	33	30	38	36

Solution: Let us take hypothesis H_0 : There is no difference between the number of days before and after a new collection policy in the accounts receivable.
And H_1 : There is a difference between the two.

Determination of Signed Ranks

X	Y	$d = X - Y$	$\lvert d \rvert$	Ranks (R) $\lvert R \rvert$ (± ignored)	Signed Ranks	
					R^+	R^-
30	32	– 2	2	6	–	6
28	29	– 1	1	3	–	3
34	33	1	1	3	3	–
35	32	3	3	9	9	–
40	37	3	3	9	9	–
42	43	– 1	1	3	–	3
33	40	– 7	7	14	–	14
38	41	– 3	3	9	–	9
34	37	– 3	3	9	–	9
45	44	1	1	3	3	–
28	27	1	1	3	3	–
27	33	– 6	6	13	–	13
25	30	– 5	5	12	–	12
41	38	3	3	9	9	–
36	36	–	–	–	–	–
Total				$n = 14$	$\Sigma R^+ = 36$	$\Sigma R^- = 69$

From the above table, it must be seen that the total number of ranks $= n = 14$ ($n \leq 25$).
Since $n < 25$, the test statistic is given by

T = smaller of two sums of the signed ranked = 36

Looking at Wilcoxon's T table at 5% level for a two tailed test at $n = 14$, we get the critical value of $T = 21$
Since the calculated value of T is greater than its critical value, null hypothesis is accepted. It means that there is no significance difference between the number of days before and after a new collection policy.

IMPORTANT TYPICAL EXAMPLE

Example 8. **A physical instructor claims that a particular exercise if done continuously for 7 days reduces weight by 15 kg. Five overweight girls did the exercise for 7 days and their weights were observed as under :**

Girls :	1	2	3	4	5
Weight before exercise :	70	72	75	71	78
Weight after exercise :	66	70	72	66	72

Making use of Wilcoxon's Signed-rank test verify the claim at $\alpha = 0{\cdot}05$ that the exercise reduces weight by at least 3·5 kg.

Solution: Denoting the mean weight before and after exercise by μ_1 and μ_2 respectively. We have the following

$H_0 : \mu_1 - \mu_2 = 3{\cdot}5$ against $H_1 : \mu_1 - \mu_2 < 3{\cdot}5, \alpha = 0{\cdot}05$

Girls	X	Y	$D = X - Y$	$d = D - 3{\cdot}5$	$\lvert d \rvert$	Ranks (R) (± ignored)	Signed Ranks	
							R^+	R^-
1	70	66	4	0·5	0·5	1·5	1·5	–
2	72	70	2	−1·5	1·5	3·5	–	3·5
3	75	72	3	−0·5	0·5	1·5	–	1·5
4	71	66	5	1·5	1·5	3·5	3·5	–
5	78	72	6	2·5	2·5	5	5	–
						$n = 5$	$\Sigma R^+ = 10$	$\Sigma R^- = 5$

From the above table, it must be seen that the total number of ranks $n = 5$ (*i.e.*, ≤ 25)

Since $n < 25$, the test statistic is given by :

T = smaller of the two sums of the signed ranks = 5

Looking at Wilcoxon's table at 5% level of significance for a one tailed test at $n = 5$ we get the critical value of T or $T_{0{\cdot}05} = 1$.

Since the calculated value of T (5) is greater than its critical value (1), the null hypothesis is accepted. It means that the sample data supports the claim that the exercise if continuously done for 7 days reduces weight by at least 3·5 kg.

EXERCISE – 1

1. A teacher claims that by imparting coaching for one month we can make the student worth securing at least 50 marks. A random sample of 10 students reveals the following scores in the examination. Test the hypothesis that the claim is accepted.

Students	1	2	3	4	5	6	7	8	9	10
Marks	52	54	60	65	45	42	58	64	50	48

[**Ans.** $K = 1{\cdot}06$, Hypothesis is accepted]

2. A new car model was put to test course for observing kilometer run per litre of petrol. The data record were :

20, 18, 22, 21, 15, 17, 14, 16, 22, 23

Use sign test and Wilcoxon signed rank test to test company claim that the new car model on an average runs, 20 km per litre modal.

[**Ans.** $K = 1{\cdot}06$ Hypothesis accepted; $T = 12{\cdot}5$ Hypothesis is accepted]

3. Use the sign test to see if there is a difference between the number of days until the collection of an account receivable before and after a new collection policy. Use at 5% level of significance.

Before	72	82	50	54	56	90	68	76	66	84	80	70	68	56	60
After	72	76	60	66	54	88	74	82	80	86	74	64	66	58	64

[**Ans.** $K = 2{\cdot}83$, Hypothesis is accepted]

4. The following are the number of patients treated by two doctors in a hospital during a fortnight :

By Doctor X	14	8	12	10	14	10	10	16	12	8	9	10	11	15	7
By Doctor Y	15	10	8	9	11	12	8	14	8	10	6	7	13	11	9

Use the sign test at 1% level of significance, test the null hypothesis that on an average both the doctors treat equal numbers of patients as against alternative hypothesis that Dr. X treats more patients than Doctor Y. [**Ans.** $K = 3{\cdot}2$]

5. A labour welfare officer visits 7 times in the product department where the frequent accidents are reported by trade union leaders. The number of accidents occured during his visits were reported by him as under :

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

Using Wilcoxon signed rank test to verify the claim of the labour that an average not more than 4 workers met the accident. Use $\alpha = 0.05$.

6. Nine adults agreed to test the efficiency of a new diet program. Their weights (lbs) were measured before and after the program and are given below :

Sr. No.	1	2	3	4	5	6	7	8	9
Before	132	139	126	114	122	132	142	119	126
After	134	141	118	116	114	132	145	123	121

Use Wilcoxon match paired signed-rank test to test the efficiency of the diet.

[**Ans.** $T = 15$]

(3) Mann-Whitney U-Test : This test was developed by Mann and Whitney and it is a rank sum test. This test is used to test whether two independent samples have come from the same (identical) population.

Procedure :

This test involves the following steps :

(i) Arrange the data of both the samples in one column in ascending order.

(ii) Assign ranks to them in increasing (from low to high) order of magnitude. In case of the repeated values, assign ranks to them by averaging their rank positions.

(iii) Then the ranks of the different samples are separated and summed up as R_1 and R_2.

(iv) If both n_1 and n_2 are sufficiently large (*i.e.*, > 8) then find the test statistic U by the following two models :

$$U_1 = n_1 n_2 + \frac{n_1 (n_1 + 1)}{2} - R_1$$

or

$$U_2 = n_1 n_2 + \frac{n_2 (n_2 + 1)}{2} - R_2$$

Then get the critical value of U from the U-table, if available, with reference n_1 and n_2 for comparsion with the above calculated value.

(v) If the U-table is not available, then get the transformed form of the U-statistic which is given by :

$$Z = \frac{U - (n_1 \cdot n_2)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

(vi) Then find the critical value of Z at 5% or 1% level of significance. The critical value of Z at 5% is 1·96 and at 1% is 2·58.

(vii) If the calculated value of Z is less than or equal to its critical value, then accept the null hypothesis or reject the same if the result appears to the reverse.

Note : The value of U that we use for the 'U' test is the smaller of U_1 and U_2.

Example 9. **Given below are the relating to production of rice in quintals per acre collected through two samples :**

Sample A	16	20	18	26	28	24	20	22	28	26	18	16
Sample B	22	20	18	22	24	12	16	08	18	22	14	–

Using the Mann-Whitney *U*-Test (a Rank Sum Test) at 5% level, verify the assertion that both the samples have come from the populations with the same means.

Solution:

Determination of the Ranks and their sums

Values of both the samples in ascending order	Rank in Increasing order		Ranks of the Sample A R_1	Ranks of the Sample B R_2
8–B		1	–	– 1
12–B		2	–	– 2
14–B		3	–	– 3
16–A	4	5	5	–
16–A	5	5	5	–
16–B	6	5	–	5
18–A	7	8·5	8·5	–
18–A	8	8·5	8·5	–
18–B	9	8·5	–	8·5
18–B	10	8·5	–	8·5
20–A	11	12	12	–
20–A	12	12	12	–
20–B	13	12	–	12
22–A	14	15·5	15·5	–
22–B	15	15·5	–	15·5
22–B	16	15·5	–	15·5
22–B	17	15·5	–	15·5
24–A	18	18·5	18·5	–
24–B	19	18·5	–	18·5
26–A	20	20·5	20·5	–
26–A	21	20·5	20·5	–
28–A	22	22·5	22·5	–
28–A	23	22·5	22·5	–
Total $N = 23$	**276**		$R_1 = 171$ $n_1 = 12$	$R_2 = 105$ $n_2 = 11$

We have, $R_1 + R_2 = \frac{N(N+1)}{2}$

$\Rightarrow$ $171 + 105 = \frac{23(23+1)}{2} = 276$ verified

Let us set up the Null Hypothesis, H_0 : Both the samples have come from the population with the same means.

And the Alternative Hypothesis, H_1 : The two samples are not from the same population.

Since, both n_1 and n_2 are > 8, the relevant test statistic U is given by

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$= 12 \times 11 + \frac{12(12+1)}{2} - 171$$

$$= 132 + 78 - 171 = 210 - 171 = 39$$

or $$U_2 = n_1 \cdot n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

$$= 132 + \frac{11(11+1)}{2} - 105 = 132 + 66 - 105 = 93$$

By transformation of U into Z we have,

$$Z = \frac{U - (n_1 . n_2)/2}{\sqrt{n_1 . n_2 (n_1 + n_2 + 1)/12}} = \frac{39 - 66}{16{\cdot}25} = -1{\cdot}66$$

Taking U at 93, we get $Z = \frac{93 - (12 \times 11)/2}{\sqrt{12 \times 11 (12 + 11 + 1)/12}} = \frac{93 - 66}{16{\cdot}25} = 1{\cdot}66$

The critical value of Z at 5% level as obtained from the Normal Curve Table is ±1·96.

Since the calculated value of Z (±1·66) is less than its critical value (±1·96), the null hypothesis is accepted.

Hence, we conclude that the two samples have come from the population with the same means.

(4) Kruskal-Wallis Test : This test was developed by Kruskal and Wallis jointly and it is an improvement over the sign test and wilcoxon's signed rank test of which ignored the actual magnitude of the paired observations. This test is applied to test whether two or three independent samples have come from the same (identical) population as against the alternative hypothesis that they are from population with different means.

Procedure :

This test involves the following steps :

(i) Arrange the data of both the sample in one column in ascending order.

(ii) Assign ranks to them in increasing (from low to high) order of magnitude. In case of the repeated values, assign ranks to them by averaging their rank positions.

(iii) Then the ranks of the different samples are separated and summed up as R_1, R_2, R_3, etc.

(iv) Then the test statistic H is calculated by using the following formula :

$$H=\frac{12}{n(n+1)}\left[\frac{R_1^2}{n_1}+\frac{R_2^2}{n_2}+.....+\frac{R_K^2}{n_k}\right]-3(n+1)$$

Note : The value of H calculated above should be increased slightly when there are very many ties in the rankings for that H is highly sensitive to ties.

(v) If each sample has at least 5 items, then get the value of χ^2 from the χ^2-table with $K-1$ degree of freedom at the desired significance. But if any of one sample it has less than 5 items, then χ^2 value should not be used.

(vi) If the calculated value of H happens to be less than the table value χ^2, the null hypothesis is accepted otherwise rejected.

Example 10. **Two sections of an elementary course consisting of 5 and 7 students respectively in Economics were taught by the teachers. The marks obtained on the final test were as under :**

	Marks						
Teacher I	**50**	**55**	**60**	**65**	**70**		
Teacher II	**60**	**63**	**58**	**70**	**55**	**68**	**73**

Using the Kruskal-Wallis test, verify at $\alpha=0{\cdot}05$ level the null hypothesis that the distribution of marks awarded by the two teachers are equal.

Solution: Let us take the hypothesis H_0 : There is no difference in the marks awarded by two teachers.

It is given that $n_1=5$ and $n_2=7$, $n=n_1+n_2=5+7=12$

Determination of Ranks and their Sums

Values of 2 samples arranged in ascending order	Rank (from low to high)		Ranks of different samples	
			R_1	R_2
50–I		1	1	–
55–I	2	2·5	2·5	–
55–II	3	2·5	–	2·5
58–II		4		4
60–I	5	5·5	5·5	–
60–II	6	5·5		5·5
63–II		7	–	7
65–I		8	8	–
68–II		9	–	9
70–I	10	10·5	10·5	–
70–II	11	10·5		10·5
73–II		12		12
	$n=12$		$R_1=27{\cdot}5$, $n_1=5$	$R_2=50{\cdot}5$, $n_2=7$

We now compute H-statistic which is given as :

$$H=\frac{12}{n(n+1)}\left[\frac{R_1^2}{n_1}+\frac{R_2^2}{n_2}\right]-3(n+1)$$

$$=\frac{12}{12(12+1)}\left[\frac{(27{\cdot}5)^2}{5}+\frac{(50{\cdot}5)^2}{7}\right]-3(12+1)$$

$$=\frac{12}{156}\cdot[151{\cdot}25+364{\cdot}32]-39$$

$$=\frac{12}{156}\cdot[515{\cdot}57]-39=39{\cdot}6592-39=0{\cdot}6592$$

Since there are many ties in the ranking, the value of H is slightly increased to 0·66

As none of the sample has less than 5 items, we get the critical value of χ^2 with 1 d.f. (*i.e.*, $K-1=2-1=1$) at 5% level from χ^2-table is 3·84.

Since the calculated value of H is less than the table value of χ^2, we accept H_0. It means that the distribution of marks awarded by the two teachers do not differ significantly.

Example 11. **Given below are the samples relating to number of minutes the patients has to wait in the clinics of three doctors :**

Doctor A	**44**	**39**	**38**	**33**	**47**	**45**	–	–
Doctor B	**34**	**45**	**43**	**39**	**42**	**40**	**46**	–
Doctor C	**46**	**34**	**43**	**36**	**30**	**42**	**41**	**44**

Using the Kruskal-Wallis test, verify at 5% level to verify the null hypothesis that all the three Doctors are equal in making the patients wait for the average time.

Solution: Let us take the hypothesis H_0 : There is no difference among the Doctors in making the patients wait.

It is given that $n_1=6$, $n_2=7$ and $n_3=8$, $n=6+7+8=21$.

Determination of Ranks and their Sums

Values of all the 3 samples in ascending order		Ranks (from low to high)		Ranks of the different samples		
				R_1	R_2	R_3
30–C	III	1		–	–	1
33–A	I	2		2	–	–
34–B	II	3	3·5	–	3·5	–
34–C	III	4	3·5	–	–	3·5
36–C	III	5		–	–	5
38–A	I	6		6	–	–
39–A	I	7	7·5	7·5	–	–
39–B	II	8	7·5	–	7·5	–

40–B	II		9	–	9	–
41–C	III		10	–	–	10
42–B	II	11	11·5	–	11·5	–
42–C	III	12	11·5	–	–	11·5
43–B	II	13	13·5	–	13·5	–
43–C	III	14	13·5	–	–	13·5
44–A	I	15	15·5	15·5	–	–
44–C	III	16	15·5	–	–	15·5
45–A	I	17	17·5	17·5	–	–
45–B	II	18	17·5	–	17·5	–
46–B	II	19	19·5	–	19·5	–
46–C	III	20	19·5	–	–	19·5
47–A	I		21	21	–	–
Total		$n = 21$		$R_1 = 69·5$, $n_1 = 6$	$R_2 = 82$, $n_2 = 7$	$R_3 = 79·5$, $n_3 = 8$

$$H = \frac{12}{n(n+1)} \cdot \left[\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \frac{R_3^2}{n_3}\right] - 3(n+1)$$

$$= \frac{12}{21(21+1)} \cdot \left[\frac{(69·5)^2}{6} + \frac{(82)^2}{7} + \frac{(79·5)^2}{8}\right] - 3(21+1)$$

$$= \frac{12}{462} \cdot \left[\frac{4830·25}{6} + \frac{6724}{7} + \frac{6320·25}{8}\right] - 66$$

$$= 0·025974\,[805·04167 + 960·57143 + 790·03125] - 66$$

$$= 0·025974\,(2555·6444) - 66 = 66·38 - 66 = 0·38$$

Since there are many ties in the rakings, the value of H is slightly increased to 0·50. As none of the samples has less than 5 items, we get the critical value of χ^2 with 2 d.f. (*i.e.*, $K-1=3-1=2$) at 5% level from the χ^2-table is 5·991.

Since the calculated value of H is less than the table value of χ^2, we accept the null hypothesis. It means that all the three doctors are equal in making the patients wait for the average time.

(5) Wald-Wolfowitz Runs Test : This test was developed by Wald and Wolfowitz. It is called **Runs Test.** This test is used to test the null hypothesis that the two populations from which the two independent samples are drawn form identical distributions.

Procedure :

This test involves the following steps :

(i) Arrange the data of both the samples as one and arrange them in ascending order.

(ii) Denote the values of the first sample and second sample by X (I) and Y (II) respectively.

(iii) Determine the number of uninterrupted runs of the sample in this sequence and denote it by R.

(iv) If the combined sample is greater than or equal to 20, the test statistic Z is calculated by using the following formula :

$$Z = \frac{R - \left(\frac{2n_1n_2}{n_1+n_2} + 1\right)}{\sqrt{\frac{2n_1n_2}{(n_1+n_2)^2} \times \frac{(2n_1n_2 - n_1 - n_2)}{(n_1+n_2+1)}}}$$

(v) Set up the null hypothesis H_0 against an appropriate alternative hypothesis at a desired level of significance.

(vi) Now that the calculated value of Z is compared with the critical value of Z at 5% or 1% for a two tailed test/one tailed test.

(vii) If the calculated value of Z is more than 1·96 at 5% for two tailed test, the null hypothesis is rejected otherwise accepted.

Example 11. **A manufacture uses two methods of production with the following results :**

Units Produced

Method I :	51	27	42	27	41	29	27	23	25	35	21	37	37
Method II :	36	19	31	19	25	16	39	25	21	17			

Carry out Wald-Wolfowitz test to find out whether the two examples are from the same population at 5% level of significance.

Solution: Let us take the hypothesis that the two samples have come from the same population.

It is given that $n_1 = 13, n_2 = 10, \alpha = ·05$

To determine runs, the observations of both samples are arranged in an ascending order.

16	17	19	19	21	21	23	25	25	25	27	27	27	29
II	II	II	II	II	I	I	I	II	II	I	I	I	I

31	35	36	37	37	39	41	42	51
II	I	II	I	I	II	I	I	I

R = The total number of runs in this sequence = 10.

We have,
$$Z = \frac{R - \left(\frac{2n_1n_2}{n_1+n_2} + 1\right)}{\sqrt{\frac{2n_1n_2}{(n_1+n_2)^2} \times \frac{(2n_1n_2 - n_1 - n_2)}{(n_1+n_2-1)}}}$$

$$= \frac{10 - \left(\frac{2 \times 13 \times 10}{13+10} + 1\right)}{\sqrt{\frac{2 \times 13 \times 10}{(13+10)^2} \times \frac{(2 \times 13 \times 10 - 13 - 10)}{(13+10-1)}}}$$

$$|Z| = \frac{|10-12\cdot3|}{\sqrt{\frac{260}{529}\times\frac{237}{22}}} = \frac{|-2\cdot3|}{\sqrt{0\cdot5\times10\cdot77}} = \frac{|-2\cdot3|}{\sqrt{5\cdot385}} = \frac{|2.3|}{2.320} = 0\cdot99$$

Since, the calculated value of Z is less than 1·96 at 5% level for two tailed test, the null hypothesis is accepted. It means that the two samples have come from the same population.

Example 12. **A random sample of 10 households selected from each of two, *A* and *B* communities revealed that their weekly medical expenses were as follows :**

Community A :	20	15	10	12	18	21	8	14	18	13
Community B :	12	14	20	17	21	22	15	13	14	18

Using the Wald-Wolfowitz test, verify at $\alpha = 0\cdot05$; if the two populations have the same underlying distributions.

Solution: Let us take up H_0 : The two populations have identical distributions against H_1 : The two populations follow different distributions.

$n_1 = 10, n_2 = 10$ is given

To determine runs (R), the observations of both samples are arranged in an ascending order :

8 10 12 12 13 14 14 14 15 15 17
18 18 18 20 20 21 21 22

R = the total number of runs in this sequence = 10

We have,

$$|Z| = \frac{R - \left(\frac{2n_1n_2}{n_1+n_2}+1\right)}{\sqrt{\frac{2n_1n_2}{(n_1+n_2)^2}\times\frac{(2n_1n_2-n_1-n_2)}{(n_1+n_2-1)}}}$$

$$= \frac{10-\left(\frac{2\times10\times10}{10+10}+1\right)}{\sqrt{\frac{2\times10\times10}{(10+10)^2}\times\frac{2\times10\times10-10-10}{(10+10-1)}}}$$

$$= \frac{10-11}{2\cdot18} = \frac{-1}{2\cdot18} = 0\cdot4587$$

Since, the calculated value of Z is less than 1·96 at 5% level of signficance for two tailed test, the null hypothesis H_0 is accepted. This means that the sample data support the hypothesis that the two populations have identical distributions.

EXERCISE – 2

1. Given below are the strengths of cables made from the different alloys I & II

Alloy I	18·3	16·4	22·7	17·8	18·9	25·3	16·1	24·2		
Alloy II	12·6	14·1	20·5	10·7	15·9	19·6	12·9	15·2	11·8	14·7

Using Mann-Whitney U test at 5% level of significance state whether or not there is a significant difference in the average of strength of the cable made from Alloy I and Alloy II.
[Ans. $U_1 = 10, U_2 = 70$, $|Z| = 2\cdot67$, Reject H_0]

2. A professor has two classes in Economics : a morning class of 9 students and an afternoon class of 12 students. On a final examination scheduled at the same time for all students, the classes received the following grades :

Morning Class	73	87	79	75	82	66	95	75	70	
Afternoon Class	86	81	84	88	90	85	84	92	83	91	53	84

Using Mann-Whitney U-test at 5% level of significance, state whether or not there is a significant difference in the average score of the morning and afternoon classes.
[Ans. $U = 80$ $Z = 1\cdot85$ Accept H_0]

3. To compare the effectiveness of three types of weight reducing diets, a homogenous group of 22 women was divided into there sub groups and each subgroup followed by one of those diet plans for a period of two months. The weight reduction in Kgs, were noted as given below :

Diet Plans	I	4·3	3·2	2·7	6·2	5·0	3·9			
	II	5·3	7·4	8·3	5·5	6·7	7·2	8·5		
	III	1·4	2·1	2·7	3·1	1·5	0·7	4·3	3·5	0·3

Use Kruskal-Wallis test to test the hypothesis that the effectiveness of three weight reducing diet plans are the same at 5% level of significance.
[Ans. $H = 15\cdot63$, H_0 is rejected $\chi^2 = 5\cdot99$]

4. Given below are the scores obtained by 16 pairs of boys and girls on a campus interview conducted by IBM (International Business Machines) Company

Boys	13	26	42	43	41	40	10	12	1	13	46	42	16	38	44	41
Girls	42	47	28	26	45	47	48	36	29	27	23	33	15	39	45	46

Using Kruskal Wallis test at 5% l.o.s., state whether or not there is a significant difference in the average I.Q. of the boys and girls. [Ans. $H = 2\cdot05$, H_0 is accepted]

5. Given below are the data relating to time in minutes the customer had to stand in a queue for encashing their cheques from the three banks in Haryana.:

SBI	10	25	42	21	28	36	10
PNB	35	30	25	22	28	23	15
UBI	34	32	36	23	25	10	18

Using Kruskal Wallis test at 5% l.o.s., state if the three banks are equally dilatory in payment of cheques. [Ans. $H = 0\cdot75$, $\chi^2 = 5\cdot99$]

6. A random sample of 10 students each from Boys and Girls revealed that their monthly expenditure on stationery was as follows :

Boys (B) :	10	12	15	20	17	14	11	13	25	18
Girls (G) :	12	14	15	8	18	22	20	18	10	15

Using the Wald-Wolfowitz runs test, verify at $\alpha = 0{\cdot}01$, if the two populations have identical distributions. **[Ans.** For $R = 12$, $Z = 0{\cdot}46$, H_0 is accepted]

7. The nicotine content of two brands of cigarettes measured in milligrams was found as follows:

Brand A	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3		
Brand B	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

Use rank sum test (Mann-Whitney Test), test the hypothesis at 0.05 l.o.s., state the average nicotine contents of two brands are equal against the alternative that they are not equal.
[Ans. $Z = 1.51$, Accept H_0.]

SOME OTHER NON-PARAMETERIC TESTS

(1) Rank Correlation Test : This test was developed by Charles Spearman and popularised by Hotelling in 1936. This test is used to test the null hypothesis that there is no correlation between the two populations against the alternative hypothesis that there is a correlation between the two.

Procedure : This test involves the following steps :

(1) We find the rank correlation coefficient between the two series by using the following formula :

(a) $R = 1 - \dfrac{6\Sigma D^2}{n^3 - n}$ (when ranks are not repeated)

(b) $R = 1 - \dfrac{6\left[\Sigma D^2 + \dfrac{1}{12}(m^3 - m) + \ldots\right]}{n^3 - n}$ (when ranks are repeated)

(2) Set up the null hypothesis that there is no correlation between the two populations *i.e.*, $H_0 = \rho = 0$.

(3) If the number of pairs is fairly small ($n \leq 30$), we find the critical value of R from the **Rank Correlation Table** (given at the end of the book) with reference to the values of n and α (level of significance).

(4) Compare the calculated value of R with the critical value of R. If the calculated value of R happens to be less than its critical value, then accept H_0. In the reverse case, it is rejected.

On the other hand, if $n > 30$, we compute test statistic Z based on Normal distribution by using the formula :

$$Z = \frac{R}{\frac{1}{\sqrt{n-1}}} = R \cdot \sqrt{n-1}$$

Then we get the critical value of Z from the Normal Curve Table with reference to significance level (α).

If the calculated value of Z happens to be less than its critical value, then we accept the null hypothesis otherwise we reject it.

The following example, illustrate the procedure of rank credit less.

Example 1. **Use the rank correlation test at 1% significance level, determine if there is any positive correlation between the prices of shares and prices of debentures given as below :**

Price of Shares (Rs.)	52	53	42	60	45	31	37	38	25	27
Price of Debentures (Rs.)	65	68	43	38	77	48	35	30	25	50

Solution. **Computation of Rank Correlation Coefficient**

Price of Shares (X)	R_1	Price of Debentures (Y)	R_2	$D = R_1 - R_2$	D_2
52	3	65	3	0	0
53	2	68	2	0	0
42	5	43	6	−1	1
60	1	38	7	−6	36
45	4	77	1	3	9
31	8	48	5	3	9
37	7	35	8	−1	1
38	6	30	9	−3	9
25	10	25	10	0	0
27	9	50	4	5	25
$n = 10$				$\Sigma D = 0$	$\Sigma D^2 = 90$

Applying the formula,

$$R = 1 - \frac{6\Sigma D^2}{n^3 - n}$$

$$R = 1 - \frac{6 \times 90}{10^3 - 10} = 1 - \frac{540}{990} = 1 - 0{\cdot}545 = 0{\cdot}455$$

Rank Correlation Test :

We have $n = 10, \alpha = 0{\cdot}01, R = 0{\cdot}455$

$H_0 : \rho = 0$ (*i.e.*, there is no correlation between the prices of shares and debentures)

and $H_1 : \rho > 0$ (there is positive correlation between the two)

Since $n < 30$, (*i.e.*, 10) the critical value of R as obtained from the Rank Correlation Table for $n = 10$ and $\alpha = 0{\cdot}01$ is $0{\cdot}7818$.

Since, the calculated value of R ($0{\cdot}455$) is less than its critical value ($0{\cdot}7818$), this suggests that the null hypothesis is accepted. Hence we conclude that there is no significant positive correlation between the price of the shares and price of debentures.

(2) Median Test : It is another important non-parametric test. It is used to test whether two or more samples are taken from the population with same median.

Procedure : This test involves the following steps :

(i) Set up the null hypothesis H_0 that there is no difference in the median of the two samples.

(ii) We find the median of the combined data. Both groups are combined and data are analysed in ascending order and find median (M).

(iii) We determine how many of the values in each sample fall above or below the median (*i.e.*, the two samples are classified in two groups above median and below median). The frequencies are counted for each group.

(iv) We present the data in the form of a (2×2) contingency table shown below :

	Sample I	Sample II	Total
Above Median	a	b	$a+b$
Below Median	c	d	$c+d$
Total	$a+c$	$b+d$	

(v) We now calculate the expected frequencies by using the formula :

$$\text{(Row} \times \text{Column)/Grand Total}$$

(vi) We compute the value of χ^2 by using the formula :

$$\chi^2 = \Sigma\left[\frac{(0-E)^2}{E}\right]$$

(vii) Determine the degree of freedom $= v = (r-1)(c-1)$.

(viii) The critical value of χ^2 at 0·05 level of significance for given degree of freedom is found.

(ix) If the calculated avlue of χ^2 exceeds the critical value of χ^2, H_0 is rejected. It implies that there is no evidence to suggest that the median is the same in case of two samples. In the reverse case, H_0 is accepted.

Example 2. **Two different fertilisers were used on a sample of eight plots :**

Plot No.	1	2	3	4	5	6	7	8
Fertiliser I	49	32	44	48	51	34	30	42
Fertiliser II	40	45	50	43	37	47	55	57

Use Median test to test the hypothesis that the two fertilisers have the same median.

Solution : H_0 : The two fertilisers have the same median.

Now, we calculate the combined median of the two series. Let us arrange the data of two series in an asceding order :

Sr. No.	X	Sr. No.	X
1	30	9	45
2	32	10	47
3	34	11	48
4	37	12	49
5	40	13	50
6	42	14	51
7	43	15	55
8	44	16	57

$$\text{Combined Median} = \text{Size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \frac{16+1}{2} = 8{\cdot}5\text{th item}$$

$$= \left[\frac{8\text{th} + 9\text{th}}{2}\right] \text{item} = = \frac{44+45}{2} = 44{\cdot}5$$

Form a 2×2 contingency table :

	Sample I	Sample II	Total
Above Median	3	5	8
Below Median	5	3	8
Total	8	8	16

Now, calculate the expected frequencies :

$$E(3) = \frac{8 \times 8}{16} = 4 \quad E(5) = 8-4 = 4$$

$$E(5) = 8-4 = 4 \quad E(3) = 8-4 = 4$$

O	E	$(O-E)$	$(O-0)^2$	$(O-E)^2/E$
3	4	-1	1	0·25
5	4	$+1$	1	0·25
5	4	$+1$	1	0·25
3	4	-1	1	0·25
				$\chi^2 = E(O-E)^2/E = 1$

$$\chi^2 = \Sigma\left[\frac{(O-E)^2}{E}\right] = 1$$

Degree of freedom $= v = (2-1)(2-1) = 1$

The critical value of χ^2 at 5% for d.f. = 3·84.

Since, the calculated value of χ^2 is less than the critical vale of χ^2 at 5% l.o.s., we accept null hypothesis and conclude that the two fertilisers have the same median.

(3) Kolmogorov – Smirnov Test

Kolmogorov – Smirnov test, named after statisticians A.N. Kolmogorov and N.V. Smirnov, is a simple non-parametric test for testing whether there is a significant difference between an observed frequency distribution and a theoretical or excepted frequency distribution. It is similar to the Chi-square test of goodness of fit. It is used when one is interested in comparing a set of values on an ordinal scale.

Procedure : This test involves the following steps :

(1) Set up the null hypothesis that there is no significant difference between the observed and the expected values (or theoretical values) *i.e.*, there is good compatibility between theory and experiment.

(2) On the basis of the null hypothes is, we calculate the expected frequencies.

(3) Compute the observed relative cumulative frequency (F_0) and expected relative cumulative frequencies (Fe).
(4) Determine the largest absolute deviations between F_0 and F_e *i.e.*, $D = |F_0 - F_e|$.
(5) Compute the critical value of D with reference to the values of n and α (l.o.s.) from Kolmogrov – Smirnov Test Table given at the end of book.
(6) Compare the calculated value of D with the critical value of D. If the calculated value of D happens to be less than the critical value, the accept H_0. In the reverse case, H_0 is rejected.

Example 3. **The following grades were given to a class of 100 students.**

Grade	A	B	C	D	E	
Frequency	50	60	20	40	30	$N = 200$

Test the hypothesis that the distribution of grade is uniform. Use Kolmogorov – Smirnov test.

Solution. Let us take the hypothesis that the distribution of grades are uniform *i.e.*, there is no difference in their distribution. No. of students given grade = 200. We should expect $\frac{200}{5} = 40$ to each student.

O	Observed Cumulative Frequency	Observed Relative Frequency F_0	E	Expected Cumulative Frequency	Expected Relative Frequency F_e	$D = \lvert F_0 - F_e \rvert$
50	50	0·25	40	40	0·20	0·05
60	110	0·55	40	80	0·40	0·15
20	130	0·65	40	120	0·60	0·05
40	170	0·85	40	160	0·80	0·05
30	200	1·00	40	200	1·00	0·00

From the table, we find that the largest absolute difference is 0·15 which is known as Kolmogorov - Smirnov D value.
Since the sample size is more than 35, the critical value of D with reference to value of n and α is $\frac{1{\cdot}36}{\sqrt{200}} = 0{\cdot}096$. As the calculated value of D exceeds the critical value of 0·096, we reject H_0 and conclude that the grades are not uniformly distributed.

EXERCISE-3

1. Use the rank correlation test at 1% level of significance, determine if there is any positive correlation between study time and scores :

Number of hours studied (X) :	8	5	11	13	10	5	18	15	2	8
Score (Y) :	56	44	79	72	70	54	94	85	33	65

[Ans. $R = 0{\cdot}9758$, $|Z| = 2{\cdot}94$, Reject H_0 and significant relationship.]

2. An I.Q. test was given to a random sample of 15 male and 20 female students of a university. Their scores were recorded as follows :

Male :	56	66	62	81	75	73	83	68	48	70	60	77	86	44	72					
Female :	63	77	65	71	74	60	76	61	67	72	64	65	55	89	45	53	68	73	50	81

Use median test to determine whether I.Q. of male and female students is same in the university. (Given the median of the combined sample = 68)

3. Below is the table of observed frequencies, along with the frequency to the observed under normal distribution : [Ans. $\chi^2 = 0{\cdot}252$, Accept H_0]
(a) Calculate the K-S statistic.
(b) Can we conclude that the distribution does in fact follow a normal distribution? Use 0·10 l.o.s. and Kolmogorov – Smirnov test.

Total Score :	51–60	61–70	71–80	81–90	91–100
Observed Frequency :	30	100	440	500	130
Expected Frequency :	40	170	500	390	100

[Ans. $D_5 = 0{\cdot}117$, Accept H_0]

QUESTIONS

1. What are non-parametric tests ? In what ways are they different from parametric tests.
2. Differentiate between parametric and non-parametric tests, thus, highlight the advantages of non-parametric tests.
3. Discuss the methods of using ordinary sign test and Wilcoxon's signed-rank test.
4. Explain Wilcoxon's signed rank test procedure.
5. Name various non-parametric tests. Describe by taking a suitable example Mann-Whitney *U*-Test.
6. Explain briefly the Chief features of Wald-Wolfwitz test and its uses in economic data analysis.
7. Write a short note on Kruskal-Wallis *H*-test.
8. What is sign test ? What is it used and what are its limitations ?
9. Explain the procedure of :
   (i) One sample sign test
   (ii) Wilcoxon's signed-rank test
   (iii) Kruskal-Wallis *H*-test
   (iv) Wald-Wolfwitz run test
10. What are non-parametric tests ? Briefly describe the process of Wilcoxon's Signed-Rank Test.
11. Write short notes on :
    (i) Rank correlation test (ii) Median test and
    (iii) Kolmogorov-Smirnov Test.

❀❀❀

9

Statistical Decision Theory

INTRODUCTION

In every field of life one has to make decisions in different alternative courses of action. Decision making is needed whenever an individual or an organisation (private or public) is faced with a situation of selecting an optimal (or best) course of action among several available alternatives. For example, an individual may have to decide whether to invest his money in stock, bonds or debentures; whether to build a house or to purchase a flat or live in a rented accommodation; whether to join a service or to start own business; which company's car/scooter should be purchased, etc. Similarly, a business firm may have to decide which product it should produce among various products; the type of technique to be used in production; what is the most appropriate method of advertising its product, etc. Decision making is a process of choosing an optimal course of action out of several alternative courses for the purpose of achieving a 'goal or goals'. Statistical decision theory consists of a large number of quantitative techniques which helps in analysing a decision situation and enable us to assure at a conclusion which is the best under given circumstances of the case.

ELEMENTS OF A DECISION PROBLEM

There are four elements of any decision problem. These are acts, states of nature and pay off matrix and regret matrix which are discussed below :

(1) Acts : The decision always involves a choice among several alternatives. These several alternatives are called acts. For example, a management is faced with the problem of choosing one of the three products X, Y and Z for manufacturing. It means that there are three acts out of one act is to be chosen. Thus, acts are the several alternative course of action or strategies, that are available to a decision maker. These are denoted by $A_1, A_2, A_3, \ldots A_n$.

(2) States of Nature (or Events) : In every act, there are events which are uncertain and beyond the control of a decision maker. These events are outside the firm and not under its control. In the above example where the management is faced with the problem of choosing one of three products for manufacutring, the potential demand for the product may turn out to be good, moderate or poor. The consumer's demand for the products are the events which are uncertain and beyond the control of the decision maker. Thus, events which are beyond the control of the decision maker are called states of nature. These are denoted by $S_1, S_2, S_3, \ldots S_n$.

(3) Pay off Table (or Pay off Matrix) : When the value of each event in the act are calculated directly in terms of gains or losses expressed in money, it is called a pay off. Each combination of a course of action in Acts and States of Nature is associated with pay off, which measures the net benefit to the decision maker. A pay off matrix (or a table) consists of the following two

things (i) alternative acts like $A_1, A_2, A_3, \ldots, A_n$ and (ii) various states of nature like $S_1, S_2, S_3, \ldots, S_n$ involved in every act.

The format of a pay off table/pay off matrix is given below :

States of Nature	Acts		
	A_1	A_2	A_3
S_1	X_{11}	X_{12}	X_{13}
S_2	X_{21}	X_{22}	X_{23}
S_3	X_{31}	X_{32}	X_{33}

In the above table column represents acts and row represents events (or States of Nature).

4. Opportunity Loss Table (or Regret Matrix) : Opportunity loss is the loss incurred as a consequence of the failure to take the best possible decisions. For any given state of nature (S_i), the opportunity loss for any given act (A_i) is defined as the difference between the maximum possible pay off over different acts for a given state of nature and the actual pay off for the act. (A_i) over that state of nature. A specimen of opportunity loss table (or regret matrix) is as follows :

Opportunity Loss Table

States of Nature	Acts		
	A_1	A_2	A_3
S_1	$M - X_{11}$	$M - X_{12}$	$M - X_{13}$
S_2	$M - X_{21}$	$M - X_{21}$	$M - X_{23}$
S_3	$M - X_{31}$	$M - X_{31}$	$M - X_{33}$

Here M = Maximum possible pay off.

DECISION MAKING ENVIRONMENTS

Decisions are made under three types of enviroments :

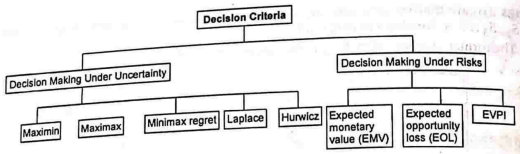
(1) Decision making under conditions of certainty : In this environment, only one state of nature exists *i.e.*, there is complete certainty about the future. It is easy to analyse the situation and make good decisions.

(2) Decision making under conditions of uncertainty : Here, more than one states of nature exist but the decision maker lacks sufficient knowledge to assign probabilities to the various states of nature.

(3) Decision making under conditions of risk : Here also, more than one states of nature exist but the decision maker has sufficient knowledge to assign probabilities to each of these states of nature.

DECISION CRITERIA

Every decision maker has to make choice among the best course of action (or act) under various states of nature. Different criteria are used for making decision under different decision making environments, which, from the view point of convenience of study have been presented by the following charts :



A. Decision Making Under Uncertainty

For decision making under uncertainty without the use of probability, the following different criteria are usually adopted :

(1) Maximin Criterion
(2) Maximax Criterion
(3) Minimax Regret Criterion
(4) Hurwicz Criterion
(5) Laplace Criterion

(1) Maximin Criterion : The maximin criterion was introduced by Wald. It is based on extreme pressimism. This decision criterion assumes that worse of the possible is going to happen. Therefore, it is designed to select the action alternative that maximises the minimum monetary pay off. This implies that the decision maker has to (i) determine the minimum pay off for each action, and (ii) select that act which maximises the minimum pay offs.

(2) Maximax Criterion : The maximax criterion is based on extreme optimism. It suggests that the decision maker should select that particular act under which it is possible for him to receive the most favourable pay off (the action that maximises the monetary pay offs). This implies that the decision maker has to (i) determine the maximum pay off for each action, and (ii) selects that act which maximises the maximum pay offs.

(3) Minimax Regret Criterion : This decision criterion was developed by **Savage.** He pointed out that the decision maker might experience regret after the decision has been made and the state of nature occurred. Thus, the decision maker should attempt to minimise the regret (minimax) before actually selecting a particular action. This implies that the decision maker has to (i) transform the pay off matrix in to a Regret Matrix. This can be done by subtracting each of the values of the act from the largest pay off of that act for a given state of nature (ii) identifies the maximum regret for each act, and (iii) selects that act which minimises the maximum regret.

(4) Hurwicz Alpha Criterion : Leonid Hurwicz has developed a criterion which is a combination of maximax (optimistic) and maximin (pessimistic) decision criterion. This criterion is based on the assumption that a decision maker has a degree of optimism, which is represented by the coeff. of optimism α. The maximum pay off of each act is multiplied by degree of optimus α and minimum pay off by the degree of pessimism $(1-\alpha)$. This implies that the decision maker has to (i) choose an appropriate degree of optimism, α so that $(1-\alpha)$ represents the degree of pessimism, (ii) determine the maximum as well as minimum of each alternative and obtain $P = \alpha \times$ maximum + $(1-\alpha) \times$ minimum for each act, and (iii) choose the act that yield the maximum value of weighted pay off, denoted by P.

(5) Laplace Criterion : Laplace criterion of decision making is applicable in those cases in which all the events (or states of nature) have equal opportunity. So equal probabilities are assigned to all the events. This implies that the decision maker has to (i) determine the average pay off for each act by using the formula $\frac{1}{n}(p_1 + p_2 + \ldots p_n)$ where n denotes the number of events and P denotes the pay offs and (ii) selects the act which results in maximum average pay off.

Applications of Decision Making Under Uncertainty

The applications relating to decision making under uncertainty are studied under the following heads :

(1) When the pay offs matrix with profit data is given
(2) When the pay offs matrix with cost data is given.

(1) When the pay offs matrix with profit data is given : When we are given pay off matrix with profit data, the uses of different decision criteria can be illustrated by following examples :

Example 1. **Given the following pay off matrix :**

States of Nature	Acts		
	A_1	A_2	A_3
S_1	700	500	300
S_2	300	450	300
S_3	150	100	300

Determine the best act to be chosen under :
(i) Maximin Criterion
(ii) Maximax Criterion and
(iii) Minimax Regret Criterion

Solution. **(i) Maximin Criterion :** When this criterion is adopted, we select that act which maximises the minimum pay off :

Acts	Minimum pay offs
A_1	150
A_2	100
A_3	300 ← Maximum

The maximum value of the minimum pay off is 300 which corresponds to act A_3. Hence, the decision maker selects A_3 as the best act by using Maximin criterion.

(ii) Maximax criterion : In this criterion, we select that act which gives the maximum pay offs.

Acts	Maximum pay offs
A_1	700 ← Maximum
A_2	500
A_3	300

The maximum value of maximum pay off is 700 which corresponds to the act A_1. Hence, the decision maker selects A_1 as the best act by using Maximax criterion.

(iii) Minimax Regret Criterion : In this criterion, the following steps are necessary to be followed :

(a) Determine the opportunity loss (or regret) for each act by subtracting from maximum pay off of each state of nature to the actual pay offs of all the acts under that state of nature.
(b) Determine the maximum of opportunity loss for each action.
(c) Select the act which minimises the maximum of the loss.

Opportunity Loss Table

States of Nature	Acts		
	A_1	A_2	A_3
S_1	700 – 700 = 0	700 – 500 = 200	700 – 300 = 400
S_2	450 – 300 = 150	450 – 450 = 0	450 – 300 = 150
S_3	300 – 150 = 150	300 – 100 = 200	300 – 300 = 0

Acts	Maximum Opportunity Loss
A_1	150 ← Minimum
A_2	200
A_3	400

The minimum value of the maximum opportunity loss is 150 which corresponds to act A_1. Hence, the decision maker selects A_1 as the best act by using minimax regret criterion.

Example 2. **Based on the following pay off (profit matrix) :**

Pay–Off Matrix

States of Nature	Acts			
	A	B	C	D
P	5	10	18	25
Q	8	7	8	23
R	21	18	12	21
S	30	22	19	20

Determine the alternative to be chosen under :
(i) Maximax criterion
(ii) Maximin criterion
(iii) Minimax regret criterion
(iv) Laplace criterion
(v) Hurwicz criterion (Use $\alpha = 0 \cdot 8$)

Solution. **(i) Maximax criterion :** In this criterion, the decision maker selects that alternative (act) which maximises the maximum profits :

Acts	Maximum pay offs
A	30 ← Maximum
B	22
C	19
D	25

The decision maker should choose act *A*.

(ii) Maximin criterion : In this criterion, the decision maker selects that alternative (act) which maximises the minimum pay off :

Acts	Minimum Pay off
A	5
B	7
C	8
D	20 ← Maximum

The decision maker should choose act *D*.

(iii) Minimax Regret Criterion : In this criterion, the decision maker select that alternative which minimises the maximum of the oppotunity losses.

Opportunity Loss Table

States of Nature	Acts			
	A	B	C	D
P	25 – 5 = 20	25 – 10 = 15	25 – 18 = 7	25 – 25 = 0
Q	23 – 8 = 15	23 – 7 = 16	23 – 8 = 15	23 – 23 = 0
R	21 – 21 = 0	21 – 18 = 3	21 – 12 = 9	21 – 21 = 0
S	30 – 30 = 0	30 – 22 = 8	30 – 19 = 11	30 – 20 = 10

Acts	Maximum Opportunity Loss
A	20
B	16
C	15
D	10 ← Minimum.

The decision maker should select act D for it minimises the maximum of the opportunity losses.

(iv) Laplace Criterion : In this criterion, we assign equal opportunity (or probability) to each state of nature. The decision maker selects that act which gives the maximum average pay offs.

Acts	States of Nature				Average Pay Off
	P	*Q*	*R*	*S*	
Probability	1 / 4	1 / 4	1 / 4	1 / 4	
A	5	8	21	30	$\frac{1}{4}[5 + 8 + 21 + 30] = 16$
B	10	7	18	22	$\frac{1}{4}[10 + 7 + 18 + 22] = 14 \cdot 25$
C	18	8	12	19	$\frac{1}{4}[18 + 8 + 12 + 19] = 14 \cdot 25$
D	25	23	21	15	$\frac{1}{4}[25 + 23 + 21 + 15] = 21$

The decision maker should select act D for it maximises the average pay off.

(v) Hurwicz Criterion : In this criterion we determine maximum and minimum of each action and obtain weighted pay off $(P) = \alpha \times$ maximum $+ (1 - \alpha)$ minimum for each act. The decision maker chooses that act which gives maximum weighted pay off

Here, $\alpha = 0{\cdot}8$

Acts	Maximum Pay off	Minimum Pay off	Weighted pay off $P = \alpha$ max $+ (1 - \alpha)$ min.
A	30	5	$30 \times {\cdot}8 + 5\,(1 - {\cdot}8) = 25$
B	22	7	$22 \times {\cdot}8 + 7\,(1 - {\cdot}8) = 19$
C	19	8	$19 \times {\cdot}8 + 8\,(1 - {\cdot}8) = 16{\cdot}8$
D	25	15	$25 \times {\cdot}8 + 15\,(1 - {\cdot}8) = 23$

The decision maker should select act A for it maximises the value of weighted payoff (p).

(2) When the pay off matrix with cost data is given : When we are given pay off matrix with cost data, then the reverse criterion **Minimax** and **Minimin** are used to decide the best course of action by a decision maker. The following example illustrate the uses of these criteria for solving such decision matrix with cost data :

Example 3. **A decision matrix with cost data is given below :**

Acts	States of Nature			
	S_1	S_2	S_3	S_4
A_1	1	3	8	5
A_2	2	5	4	7
A_3	4	6	6	3
A_4	6	8	3	5

Find the best act using (i) Minimax criterion; and (ii) Minimin criterion.

Solution. **(1) Minimax criterion :** In this criterion, the decision maker selects the act which minimise the maximum costs.

Acts	Row Maximum Values
A_1	8
A_2	7
A_3	6 ← Minimum
A_4	8

The decision maker should select act A_3 by using Minimax criterion.

(ii) Minimin Criterion : In this criterion, the decision maker selects the act which minimise the minimum values.

Act	Row Minimum Values
A_1	1 ← Minimum
A_2	2
A_3	3
A_4	3

The decision maker should select act A_1 by using Minimin criterion.

EXERCISE – 1

1. Suppose that a decision maker faced with three decision alternatives and four states of nature construct the following pay off table :

Acts/States of Nature	S_1	S_2	S_3	S_4
a_1	16	10	12	7
a_2	13	12	12	9
a_3	11	14	15	14

Assuming that the decision maker has no knowledge about the probabilities of occurrence of the four states of nature, find the decisions to be recommended under each of the following criteria : (i) Maximin; (ii) Maximax and (iii) Minimax Regret.
[Ans. (i) Act a_3 (ii) Act a_1 (iii) Act a_3]

2. A businessman has three alternatives open to him each of which can have four possible events. The conditional pay off for each event is given below :

Pay off

Alternative	States of Nature (Possible Event)			
	E_1	E_2	E_3	E_4
A_1	– 2	– 3	4	3
A_2	– 1	0	6	7
A_3	– 3	– 4	9	5

Determine which alternative should be chosen if he adopts
(i) Maximin criterion (or Wald's criterion)
(ii) Maximax criterion
(iii) Minimax regret criterion (Savage criterion)
(iv) Laplace criterion
[Ans. (i) Act A_2 (ii) Act A_3 (iii) Act A_2 (iv) Act A_2]

3. Construct the opportunity loss (or regret) table and find the best act using minimax regret criterion from the following pay off table :

Acts/States of Nature	S_1	S_2	S_3	S_4
A_1	2	0	8	– 2
A_2	– 3	9	– 1	4
A_3	– 7	6	12	10

[Ans. Act A_3 is the best]

4. A decision matrix with cost data is given below :

Pay off table

Acts/States of Nature	S_1	S_2	S_3	S_4
A_1	4	7	6	5
A_2	3	2	8	1
A_3	9	5	3	6

Find the best act using (i) Minimax criterion and (ii) Minimin criterion.
[Ans. (i) A_1 is the best act (ii) A_2 is the best act]

5. Suppose that a decision maker with three decision alternatives and four states of nature construct the following profit pay off table :

Acts/States of Nature	S_1	S_2	S_3	S_4
A_1	14	8	10	5
A_2	11	10	7	7
A_3	9	12	13	1

Assume that the decision maker has no knowledge about the probabilities of occurrence of the four states of nature, find the decision to be recommended under each of the following criterion :

(i) Maximin, (ii) Maximax and (iii) Minimax Regret

[**Ans.** (i) A_3 is the best act, (ii) A_1 is the best act (iii) A_3 is the best act]

6. Find the best act by Hurwicz criterion from the following pay off table.

Pay off Table

Acts/States of Nature	S_1	S_2	S_3	S_4
A_1	− 2	− 3	4	3
A_2	− 1	0	6	7
A_3	− 3	− 4	9	5

[Take the coefficient of optimism $\alpha = 0{\cdot}7$] [**Ans.** A_3 is the best act with $\alpha = 0{\cdot}7$]

(2) Decision Making Under Risks with Probability

In this case, the decision maker developed good probability estimates for the different states of nature. The following two criteria are usually adopted in decision making under risks with probability :

(1) Expected Monetary Value (EMV) Criterion

(2) Expected Opportunity Loss (EOL) Criterion

(3) Expected value of Perfect Informations (EVPI)

(1) Expected Monetary Value (EMV) Criterion : This criterion requires the calculation of expected monetary value (EMV) of each act which is obtained by multiplying the conditional pay offs for that act by the assigned probabilities of various states of nature. The decision maker selects that act that yields the highest EMV. The Expected Monetary Value (EMV) for a course of action X is given by :

$$\text{EMV}(X) = p_1X_{11} + p_2X_{21} + p_3X_{31}$$

where, X_{11}, X_{21} and X_{31} denote pay off of act X for S_1, S_2 and S_3 event or states of nature and p_1, p_2 and p_3 denote the probability of occurrence of S_1, S_2, S_3 event (or states of nature).

Procedure :

The calculation of EMV consists of following steps :

(i) Construct a pay off table listing the alternative course of action and the various states of nature, if not given. Enter the conditional profit for each decision act-event combination along with the associated probabilities.

(ii) Calculate the EMV for each decision act (or alternative) by multiplying the conditional profit by assigned probabilities and adding the resulting conditional values.

(iii) Select the act (or alternative) that yields the highest EMV.

(2) Expected Opportunity Loss (EOL) Criterion : An alternative criterion (to maximising EMV approach) is to minimise expected opportunity loss (EOL). Expected opportunity loss (or expected value of regrets) represents the amount by which maximum possible profit will be reduced under various possible actions. The course of action that minimises these losses or reductions is the optimal decision act (or alternative). The expected opportunity loss (EOL) of an act is obtained by multiplying the conditional opportunity loss for that act by the assigned probabilities of various states of nature. For a course of action X, the expected opportunity loss (EOL) is given by :

$$\text{EOL}(X) = p_1L_{11} + p_2L_{21} + p_3L_{31}$$

Where, L_{11}, L_{21} and L_{31} denote opportunity loss of act X for S_1, S_2 and S_3 event or states of nature and p_1, p_2 and p_3 denote the probability of occurrence of S_1, S_2 and S_3 event (or states of nature).

Procedure :

The calculation of EOL consists of the following steps :

(1) Construct a conditional pay off table for each act–event combination, if not given along with the associated probabilities.

(2) For each event (or states of nature) determine the conditional opportunity loss (EOL) by subtraction the pay off from the maximum pay off for that event (or states or nature).

(3) Calculate the expected opportunity loss (EOL) for each decision alternative (or act) by multiplying the conditional loss by the associated probability and then adding the values.

(4) Select the alternative (or act) that yields the lowest EOL.

Note : EOL criterion is similar to EMV criterion except the opportunity losses are considered instead of profits.

(3) Expected Value of Perfect Information (EVPI) : Under this ceriteion, it is assume that the decision maker has authentic and perfect information about the future. With perfect information, the retailer (decision maker) would know in advance the demand for each day and will store the exact number as per demand. The expected value of perfect information (EVPI) is the difference between expected pay off with perfect information (EPPI) and expected pay off with uncertainted (or EMV of Best Action). Symbolically :

EVPI = EPPI – EMV of Best Action

Where, EPPI = (Best pay off for 1st state of nature × probability of 1st state of nature) + (Best pay off for 2nd state of nature × probability of 2nd state of nature) +...+ (Best pay off last state of nature × probability of last state of nature)

APPLICATIONS OF DECISION-MAKING UNDER RISKS

The applications of decision making under risks with probability are studied under the following heads :

(1) When the conditional pay off matrix with profit data is given.

(2) When the conditional pay off matrix with profit data is not given.

(1) When the conditional pay off matrix with profit data is given : When we are given conditional profit table for different acts and the various states of nature along with the associated probabilities, the uses of EMV criterion and EOL criterion can be illustrated by the following examples :

Example 4. **Pay offs of three acts A,B and C and the states of nature *P*, *Q* and *R* are given below :**

States of Nature	Acts		
	A	B	C
P	– 35	120	– 100
Q	250	– 350	200
R	550	650	700

The probabilities of the states of nature are 0·5, 0·1 and 0·4 respectively. Tabulate the Expected Monetary Values for the above data and state which can be chosen as the best act.

Solution.

States of Nature	Probability	Pay off (Rs.)		
		A	B	C
P	0·5	– 35	120	– 100
Q	0·1	250	– 350	200
R	0·4	550	650	700

The expected monetary value (EMV) for the acts *A*, *B* and *C* are calculated below :

$$\text{EMV for } (A) = (0{\cdot}5) \times (-35) + (0{\cdot}1) \times (250) + (0{\cdot}4) \times (550)$$
$$= -17{\cdot}5 + 25 + 220 = 227{\cdot}5$$
$$\text{EMV for } (B) = (0{\cdot}5) \times (120) + (0{\cdot}1)(-350) + (0{\cdot}4)(650)$$
$$= 60 - 35 + 260 = \text{Rs. } 285$$
$$\text{EMV for } (C) = (0{\cdot}5) \times (-100) + (0{\cdot}1) \times (200) + (0{\cdot}4) \times (700)$$
$$= -50 + 20 + 280 = \text{Rs. } 250$$

Since, the Act B yields the highest EMV of Rs. 285, Act B can be chosen as the best act.

Example 5. **Given the following pay off matrix :**

States of Nature	Probability	Acts		
		X	*Y*	*Z*
P	0·3	– 120	– 80	100
Q	0·5	200	400	– 300
R	0·2	260	– 260	600

Using the Expected Monetary Values, decide which act can be chosen as the best.

Solution.

States of Nature	Probability	Acts		
		X	*Y*	*Z*
P	0·3	– 120	– 80	100
Q	0·5	200	400	– 300
R	0·2	260	– 260	600

The expected monetary value (EMV) for different acts *X*, *Y* and *Z* are calculated below :

$$\text{EMV for } (X) = 0{\cdot}3 \times (-120) + (0{\cdot}5)(200) + (0{\cdot}2)(260)$$
$$= -36 + 100 + 52 = 116$$
$$\text{EMV for } (Y) = (0{\cdot}3) \times (-80) + (0{\cdot}5) \times (400) + (0{\cdot}2)(-260)$$
$$= -24 + 200 - 52 = 124$$
$$\text{EMV for } (Z) = (0{\cdot}3) \times (100) + (0{\cdot}5) \times (-300) + (0{\cdot}2) \times (600)$$
$$= 30 - 150 + 120 = 0$$

Since, EMV for Act *Y* is maximum, Act *Y* may be selected to be the best act under EMV criterion.

Example 6. **A management is faced with the problem of choosing one of three products for manufacturing. The potential demand for each product may turn out to be good, moderate or poor. The probabilities for each of these of states of nature were estimated as follows :**

Nature of Demand

Product	Good	Moderate	Poor
X	0·70	0·20	0·10
Y	0·50	0·30	0·20
Z	0·40	0·50	0·10

The estimated profit or loss under the three states of demand may be taken as :

Product	Rs.	Rs.	Rs.
X	30,000	20,000	10,000
Y	60,000	30,000	20,000
Z	40,000	10,000	(–) 15,000 (Loss)

Calculate the expected monetary value and advise the management about the choice of the product to be manufactured.

Solution. The given data is rewritten in the form of the following table.

States of Nature	Expected pay off ('000 Rs.)					
	X		*Y*		*Z*	
	p	*X*	*p*	*Y*	*p*	*Z*
Good	0·70	30	0·50	60	0·40	40
Moderate	0·20	20	0·30	30	0·50	10
Poor	0·10	10	0·20	20	0·10	– 15

The expected monetary value (EMV) for different acts *X*, *Y* and *Z* are calculated as :

$$\text{EMV for } (X) = (0{\cdot}70) \times (30) + (0{\cdot}20) \times (20) + (0{\cdot}10) \times (10)$$
$$= 21 + 4 + 1 = \text{Rs. } 26$$
$$\text{EMV for } (Y) = (0{\cdot}50) \times (60) + (0{\cdot}30) \times (30) + (0{\cdot}20) \times (20)$$
$$= 30 + 9 + 4 = \text{Rs. } 43$$
$$\text{EMV for } (Z) = (0{\cdot}40) \times (40) + (0{\cdot}50) \times (10) + (0{\cdot}10) \times (0{\cdot}15)$$
$$= 16 + 5 - 1{\cdot}5 = \text{Rs. } 19{\cdot}5$$

Since, the expected value of product *Y* is highest, the management is advised to produce product *Y*.

Example 7. **Calculate the expected opportunity loss (EOL) from the following pay off table and hence decide which act is to be selected.**

States of Nature (Events)	Acts			
	A	*B*	*C*	*D*
S_1	50	20	− 10	− 20
S_2	120	50	200	300
S_3	200	240	400	350

The probabilities of the states of nature are 0·2, 0·5 and 0·3 respectively.

Solution. First of all, we construct the opportunity loss table from the given pay off table :

Opportunity loss (or Regret) Table

States of Nature	Acts			
	A	*B*	*C*	*D*
S_1	50 − 50 = 0	50 − 20 = 30	50 − (− 10) = 60	50 − (− 20) = 70
S_2	300 − 120 = 180	300 − 50 = 250	300 − 200 = 100	300 − 300 = 0
S_3	400 − 200 = 200	400 − 240 = 160	400 − 400 = 0	400 − 350 = 50

The above data is rewritten in the form of the following table :

States of Nature	Probability	Acts			
		A	*B*	*C*	*D*
S_1	0·2	0	30	60	70
S_2	0·5	180	250	100	0
S_3	0·3	200	160	0	50

Expected Opportunity Loss (EOL) for the four acts are calculated below :

EOL for $(A) = (0·2) \times (0) + (0·5) \times (180) + (0·3) \times (200) = 150$

EOL for $(B) = (0·2) \times (30) + (0·5) \times (250) + (0·3) \times (160) = 179$

EOL for $(C) = (0·2) \times (60) + (0·5) \times (100) + (0·3)\,(0) = 62$

EOL for $(D) = (0·2) \times (70) + (0·5) \times (0) + (0·3) \times (50) = 29$

Since, EOL is minimum for Act *D*, Act *D* may be selected to be the best act according to EOL criterion.

Example 8. **A group of students raises money each year by selling souvenirs outside the stadium after a cricket match betwen Teams A and B. They can buy any of the three different types of souvenirs from a supplier. Their sales are mostly dependent on which team win the match. A conditional pay off table is as under :**

Teams	Type of Souvenir		
	I	II	III
Team A wins	Rs. 1200	Rs. 800	Rs. 300
Team B wins	Rs. 250	Rs. 700	Rs. 1,100

(i) Construct the opportunity loss table, (ii) Which type of souvenir should the students buy if the probability of team A's winning is 0·6 ? (iii) Find out the cost of uncertainty.

Solution. (i) The required opportunity loss table is computed below :

Opportunity loss table (or Regret table)

Team Wins (State of Nature)	Type of Souvenir (Acts)		
	I	II	III
A	1200 − 1200 = 0	1200 − 800 = 400	1200 − 300 = 900
B	1100 − 250 = 850	1100 − 700 = 400	1100 − 1100 = 0

(ii) Since, the probability that Team A wins is 0·6, therefore, the probability that Team B wins = 1 − 0·6 = 0·4.

With the probabilities 0·6 and 0·4 for the teams *A* and *B* to win, the given data is written as :

Teams wins	Probability	I	II	III
A	0·6	0	400	900
B	0·4	850	400	0

The expected opportunity loss for the three acts I, II & III are calculated as below :

EOL (I) = 0·6 × 0 + 0·4 × 850 = Rs. 340

EOL (II) = 0·6 × 400 + 0·4 × 400 = Rs. 400

EOL (III) = 0·6 × 900 + 0·4 × 0 = Rs. 540

Since, the EOL for Type I is minimum, hence the students should buy Type I souvenir.

(iii) If there is certainty for a team to win, then there would be no opportunity loss and EOL would be zero. Hence, the cost of uncertainty is Rs. 340.

EXERCISE – 2

1. The pay offs (in Rs.) of three acts A_1, A_2 and A_3 and the possible states of nature S_1, S_2 and S_3 are given below :

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

The probabilities of the states of nature are 0·3, 0·4 and 0·3 respectively. Determine the optimal act using the expection principle. [Ans. Act A_1 is best]

2. Calculate the Expected Monetary Values for the data given below and state which act can be chosen as the best :

State of Nature	Probability	Pay Offs (in Rs.)		
		X	Y	Z
A	0·4	2,500	3500	5000
B	0·4	2500	3500	2500
C	0·2	2500	1500	1000

[Ans. Act Z is the best]

3. A manufacturing company is faced with the problem of choosing four products to manufacture. The potential demand for each product may turn out to be good, satisfactory and poor. The probabilities estimated for each of demand are given below :

Product	Probabilities of Types of Demand		
	Good	Statisfactory	Poor
A	0·60	0·20	0·20
B	0·75	0·15	0·10
C	0·60	0·25	0·15
D	0·50	0·20	0·30

The estimated profit or loss under different states of demand in respect of each product may be taken as :

Product	Rs.	Rs.	Rs.
A	40,000	10,000	1,100
B	40,000	20,000	(–) 7,000
C	50,000	15,000	(–) 8,000
D	40,000	18,000	15,000

Calculate the expected monetary value for different products and advice the company about the choice of the product to manufacture.

[**Ans.** The company should manufacture the product C]

4. Given the following pay off matrix :

State of Nature	Probability	Decision		
		Do not Expand (A_1)	Expand 200 units (A_2)	Expand 400 units (A_3)
High demand	0·4	2,500	3,500	5,000
Medium demand	0·4	2,500	3,500	2,000
Low demand	0·2	2,500	1,500	1,000

What should be the decision if we use (i) EMV criterion, (ii) The minimax criterion, (iii) The maximax crtierion, (iv) Minimax regret criterion.

[**Ans.** (i) Act A_3 (ii) Act A_1 (iii) Act A_3 (iv) Act A_2 or Act A_3]

5. A group of volunteers of a service organisation raises money each year by selling gift articles outside the stadium after a cricket match between teams *X* and *Y*. They can buy any of three types of gift articles from a dealer. Their sales are mostly dependent on which team wins the match. A conditional pay off table is as under :

Teams	Type of gift articles		
	I	II	III
Teams *X* wins	Rs. 1000	Rs. 900	Rs. 600
Teams *Y* wins	Rs. 400	Rs. 500	Rs. 800

(i) Construct the opportunity loss table.

(ii) Which type of gift articles should the volunteers buy if the probability of Team *X*'s winning is 0·8 ?

(iii) Find out the cost of uncertainty.

[**Ans.** (i) $\begin{bmatrix} 0 & 100 & 400 \\ 400 & 200 & 0 \end{bmatrix}$ (ii) The volunteers should buy type I gift articles

(iii) Cost of uncertainty = Rs. 80]

(2) When the conditional pay-off matrix is not given : When we are not given the conditional pay off matrix but the probability distribution of demand is known, we first construct pay off table listing the alternative actions (or acts) and the various states of nature and also write the associated probability of the states of nature. The use of EMV criterion in such problem can be illustrated by the following examples.

Example 9. **A baker produces a certain type of special pastry at a total average cost of Rs. 3 and sells it at a price of Rs. 5. This pastry is produced over the weekend and is sold during the following week : such pastry being produced but not sold during a week's time are totally spoiled and have to be thrown away. According to past experience, the weekly demand for the pastries is never less than 78 or greater than 80, you are required to formulate pay off table.**

Solution. It is clear from the problem given that the manufacture will not produce less than 78 or more than 80 pastries. Thus, there are three courses of action open to him *i.e.*, 78, 79 and 80 pastries. The states of nature is the weekly demand for pastries. There are three possible states of nature *i.e.*, demand is 78, 79 and 80 pastries. From the data given in the problem, we can calculate the conditional profit values for each action-event (demand) combination :

We are given : Profit = 5 – 3 = Rs. 2

Loss = Rs. 3

Conditional Pay off = 2 × units sold – 3 × units unsold

The resulting pay off is given as :

Conditional Pay off Table

Possible Demand (*D*)	Possible Purchase Action (*S*)		
	78 pastries	79 pastries	80 pastries
78	156	156 – 3 = 153	156 – 16 = 150
79	156	158	158 – 3 = 155
80	156	158	160

Example 10. **A newspaper boy has the following probabilities of selling a magazine :**

No. of Copies Sold :	10	11	12	13	14
Probability :	0·10	0·15	0·20	0·25	0·30

Cost of copy is 30 paise and sale price is 50 paise. He cannot return unsold copies. How many copies should he order ? Also calculate EVPI.

Solution. It is clear from the problem given that the newspaper boy would not purchase less than 10 copies and more than 14 copies. From the data given in the problem, we can calculate the conditional profit values for each purchase action–event (demand) combination. If CP denotes the conditional profit, S the quantity purchased and D the demands, then

We are given : Profit = 50 – 30 = 20 paise

Loss = 30 paise

Conditional Pay off = 20 × copies sold – 30 × copies unsold

The resulting pay off table is given below :

Conditional Pay Off Table

Possible demand (D) (No. of copies)	Probability	Possible Purchase Action (S)				
		10 copies	11 copies	12 copies	13 copies	14 copies
10	0·10	200	200 – 30 = 170	200 – 60 = 140	200 – 90 = 110	200 – 120 = 80
11	0·15	200	220	220 – 30 = 190	220 – 60 = 160	220 – 90 = 130
12	0·20	200	220	240	240 – 30 = 210	240 – 60 = 180
13	0·25	200	220	240	260	260 – 30 = 230
14	0·30	200	220	240	260	280

Expected Monetary Value (EMV) can now be computed by multiplying the probability of each state of nature with the conditional profit value and adding the resulting products.

EMV (10 copies)

$= 0{\cdot}10 \times 200 + 0{\cdot}15 \times 200 + 0{\cdot}20 \times 200 + 0{\cdot}25 \times 200 + 0{\cdot}30 \times 200 = \text{Rs. } 200$

EMV (11 copies)

$= 0{\cdot}10 \times 170 + 0{\cdot}15 \times 220 + 0{\cdot}20 \times 220 + 0{\cdot}25 \times 220 + 0{\cdot}30 \times 220 = \text{Rs. } 215$

EMV (12 copies)

$= 0{\cdot}10 \times 140 + 0{\cdot}15 \times 190 + 0{\cdot}20 \times 240 + 0{\cdot}25 \times 240 + 0{\cdot}30 \times 240 = \text{Rs. } 222{\cdot}5$

EMV (13 copies)

$= 0{\cdot}10 \times 110 + 0{\cdot}15 \times 160 + 0{\cdot}20 \times 210 + 0{\cdot}25 \times 260 + 0{\cdot}30 \times 260 = 220$

EMV (14 copies)

$= 0{\cdot}10 \times 80 + 0{\cdot}15 \times 130 + 0{\cdot}20 \times 180 + 0{\cdot}25 \times 230 + 0{\cdot}30 \times 280 = 205$

From the above calculations, we see that the highest value of EMV is Rs. 222·50 which corresponds to the purchase of 12 copies.

Hence, by EMV criterion, the newspaper boy should order for 12 copies of magazine as it gives maximum expected value.

Calculation of EVPI : From the above table, we notice the following :

Best pay off for the 1st state of nature $S_1 = 200$, $P(S_1) = 0{\cdot}10$

Best pay off for the 2nd state of nature $S_2 = 220$, $P(S_2) = 0{\cdot}15$

Best pay off for the 3rd state of nature $S_3 = 240$, $P(S_3) = 0{\cdot}20$

Best pay off for the 4th state of nature $S_4 = 260$, $P(S_4) = 0{\cdot}25$

Best pay off for the 5th state of nature $S_5 = 280$, $P(S_5) = 0{\cdot}30$

EPPI $= 200 \times 0{\cdot}10 + 220 \times 0{\cdot}15 + 240 \times 0{\cdot}20 + 260 \times 0{\cdot}25 + 280 \times 0{\cdot}30$

$= 20 + 33 + 48 + 65 + 84 = 250$

EVPI = EPPI – EMV of Best Act = 250 – 222·50 = Rs. 27·50

Example 11. **A physician purchases a particular vaccine on Monday for each week. The vaccine must be used within the week following, otherwise it becomes worthless. The vaccine costs Rs. 2 per dose and the physician charges Rs. 4 per dose. In the past 50 weeks, the physician has administered the vaccine in the following quantities :**

Doses per week :	20	25	40	60
No. of weeks :	5	15	25	5

Determine how many doses the physician should buy every week.

Solution. Here, number of doses of the vaccine purchased is an act and weekly demand of the vaccine is an event (or state of nature). As per given information, the physician must not purchase less than 20 or more than 60 doses per week. It is also given that each dose of vaccine administrated within a week yields a profit of Rs. (4 – 2) = Rs. 2, and otherwise, it is dead loss of Rs. 2.

We are given : Profit = 4 – 2 = Rs. 2

Loss = Rs. 2

Conditional Pay off = 2 × units sold – 2 × units unsold

The resulting conditonal pay off table is given below :

Event (demand per week) D	Probability	Act (purchase per week) – S			
		20	25	40	60
20	$\frac{5}{50} = 0{\cdot}1$	40	40 – 10 = 30	40 – 40 = 0	40 – 80 = – 40
25	$\frac{15}{50} = 0{\cdot}3$	40	50	50 – 30 = 20	50 – 70 = – 20
40	$\frac{25}{50} = 0{\cdot}5$	40	50	80	80 – 40 = 20
60	$\frac{5}{50} = 0{\cdot}1$	40	50	80	120

The expected monetary value (EMV) can now be computed as :

EMV (20) $= 0{\cdot}1 \times 40 + 0{\cdot}3 \times 40 + 0{\cdot}5 \times 40 + 0{\cdot}1 \times 40 = \text{Rs. } 40$

EMV (25) $= 0{\cdot}1 \times 30 + 0{\cdot}3 \times 50 + 0{\cdot}5 \times 50 + 0{\cdot}1 \times 50 = \text{Rs. } 48$

EMV (40) $= 0{\cdot}1 \times 0 + 0{\cdot}3 \times 20 + 0{\cdot}5 \times 80 + 0{\cdot}1 \times 80 = \text{Rs. } 54$

EMV (60) $= 0{\cdot}1 \times (-40) + 0{\cdot}3 \times (-20) + 0{\cdot}5 \times 40 + 0{\cdot}1 \times 120 = \text{Rs. } 22.$

Since, the purchase 40 doses yields the highest EMV of Rs. 54, the optimal act for the physician would be to purchase 40 doses of the vaccine per week.

EMV for Items that have a Salvage Value :

In the discussion so far it has been assumed that the product being stocked (or purchased) was completely worthless if not sold on the same day. This assumption that the product has no salvage, is not always realistic. If the product does a salvage value, then it must be considered in calculating profits for each stock action.

Example 12. **An ice-cream retailer buys ice-cream at a cost of Rs. 5 per cup and sells it for Rs. 8 per cup; any remaining unsold at the end of the day can be disposed off as a salvage price of Rs. 2 per cup. Past sales have ranged between 15 and 18 cups per day. The following is the record of sales :**

Cups Sold :	15	16	17	18
Probability :	0·10	0·20	0·40	0·30

Find how many cups, the retailer should purchase per day to maximise his profit. Also calculate EVPI

Solution. Here, number of cups of ice-cream purchased is an act and daily demand of the ice-cream cups is an event or state of nature.

We are given : Cost per cup = Rs. 5
Selling price = Rs. 8
Profit = Rs. 8 – Rs. 5 = Rs. 3 (if sold)
Disposal selling price = Rs. 2 (if unsold)
Loss = Rs. 5 – Rs. 2 = Rs. 3

Now, the various conditional profit (pay off) values for each act-event combination are given by :

Conditional Pay off = 2 × units sold – 3 × units unsold

The resulting conditional pay offs are given below :

Conditional Pay off (Rs.)

Event (demand per week) D	Probability	Act (purchase per week)			
		15	16	17	18
15	0·10	45	45 – 3 (1) = 42	45 – 3 (2) = 39	45 – 3 (2) = 36
16	0·20	45	48	48 – 3 (1) = 45	48 – 3 (2) = 42
17	0·40	45	48	51	51 – 3 (1) = 48
18	0·30	45	48	51	54

The expected monetary values for different acts are computed as :

EMV (15) = 0·10 × 45 + 0·20 × 45 + 0·40 × 45 + 0·30 × 45 = Rs. 45·00
EMV (16) = 0·10 × 42 + 0·20 × 48 + 0·40 × 48 + 0·30 × 48 = Rs. 47·40
EMV (17) = 0·10 × 39 + 0·20 × 45 + 0·40 × 51 + 0·30 × 51 = Rs. 48·60
EMV (18) = 0·10 × 36 + 0·20 × 42 + 0·40 × 48 + 0·30 × 54 = Rs. 47·40

Since, the act 'purchase 17 cups' yields the highest EMV of Rs. 48·60, the optimal act for the retailer would be to purchase 17 cups of ice-creams.

Calculation of EVPI : From the above table, we notice the following :

Best pay off for the 1st state of nature $S_1 = 45$, $P(S_1) = 0·10$
Best pay off for the 2nd state of nature $S_2 = 48$, $P(S_2) = 0·20$
Best pay off for the 3rd state of nature $S_3 = 51$, $P(S_3) = 0·40$
Best pay off for the 4th state of nature $S_4 = 54$, $P(S_4) = 0·30$

EPPI = 45 × 0·10 + 48 × 0·20 + 51 × 0·40 + 54 × 0·30
= 4·5 + 9·6 + 20·4 + 16·2 = Rs. 50·7

EVPI = EPPI – EMV of Best Act = 50·7 – 48·60 = Rs. 2·1

Example 13. **Each unit of a product and sold yields a profit of Rs. 50 if a unit produced but not sold results in a loss of Rs. 30. The probability distribution of the number of units demanded is as follows :**

No. of Units Demanded	Probability
0	0·20
1	0·20
2	0·25
3	0·30
4	0·50

How many units be produced to maximise the expected profits ? Also calculate EVPI.

Solution. Here the number of units produced is an act and the demand of the units is an event or state of nature.

We are given : Profit per unit = Rs. 50
Loss per unit = Rs. 30

The various conditional profit (pay off) values for each act – event combination are given by :

Conditional Pay off = 50 × units sold – 30 × units unsold

The resulting conditional pay off is

Event (demand) D	Proba-bility	Acts (Production)				
		0	1	2	3	4
0	0·20	0	0 – 30(1) = – 30	0 – 60 = – 60	0 – 90 = – 90	0 – 120 = – 120
1	0·20	0	50	50 – 30 (1) = 20	50 – 60 = – 10	50 – 90 = – 40
2	0·25	0	50	100	100 – 30 = 70	100 – 60 = 40
3	0·30	0	50	100	150	150 – 30 (1) = 120
4	0·50	0	50	100	150	200

The expected monetary value for different acts are computed as :

EMV (0) = 0·20 × 0 + 0·20 × 0 + 0·25 × 0 + 0·30 × 0 + 0·05 × 0 = 0
EMV (1) = 0·20 × (– 30) + 0·20 × 50 + 0·25 × 50 + ·30 × 50 + ·5 × 50 = Rs. 565

EMV (2) = 0·20 (– 60) + 0·20 (20) + 0·25 × 100 + 0·70 × 100 + ·5 × 100 + Rs. 137
EMV (3)= 0·20 (– 90) + 0·20 (– 10) + 0·25 × 70 + 0· 30 × 150 + 0·5 × 150 = Rs. 475
EMV (4) = 0·20 (– 120) + 0·20 (– 40) + 0·25 × 40 + 0· 30 × 120 + 0·5 × 200 = Rs. 114
Since, highest EMV is Rs. 137, it is optimal to produce 2 units.

(ii) The expected value of perfect information (EVPI) is the difference between the expected pay off with perfect information (EPPI) and the maximum expected pay off (Max. EMV) with no additional information *i.e.*,

EVPI = EPPI – EMV of the Best Action

EPPI is determined as below :

Event (1)	Prob. (2)	Best Pay off under perfect information (3)	Expected pay off Under PI (2 × 3)
0	0·20	0	0
1	0·20	50	10
2	0·25	100	25
3	0· 30	150	45
4	0·50	200	100
			Total = 180

Note : Best Pay off under PI is the max. pay off under each event (or state of nature).

The expected value of perfect information is

EVPI = EPPI – EMV of Best Actio = 180 – 137 = 43.

EXERCISE – 3

1. A newspaper vendor has to decide how many copies of a particular magazine he should buy for the month of Nov. Each magazine costs Rs. 5 and sells for Rs. 10. At the end of the month unused magazine has no value. The probability distribution to demand is given below :

No. of Copies Demanded :	10	11	12
Probability :	1/3	1/3	1/3

Construct a pay off table. According to EMV criterion, how many copies should be stock ?
[Ans. 10 copies]

2. A proprietor of a food stall has introduced a new item of food delicacy to which he calls Whim. He has calculated that the cost of manufacture is Rs. 1 per piece and sold at Rs. 3 per piece. It is however perishable goods and any goods unsold is a dead loss. The probability distribution to demand is given below :

No. of Pieces Demanded :	10	11	12	13	14	15
Probability :	·07	·10	·23	·88	·12	·10

How many pieces should be manufacture so that net profit expected is maximum.
[Ans. 13 pieces]

3. A retailer purchases berries every morning at Rs. 5 a case and sells for Rs. 8 a case. Any case remaining unsold as the end of the day can be disposed off the next day at a salvage value of Rs. 2 per case (there after they have no value). Past sales have ranged from 15 to 18 cases per day. The following is the record of sales for the past 120 days :

No. of cases sold :	15	16	17	18
No. of days :	12	24	48	36

Find how many cases the retailer should purchase per day to maximise his profit.
[Ans. 17 units]

4. A news paper distributor assigns probabilities to the demand for a magazine as follows :

Copies Demanded :	1	2	3	4
Probability :	0·4	0·3	0·2	0·1

A copy of magazine selles for Rs.7 and costs Rs 6. What can be maximum possible expected monetary value (EMV) if the distributor can return the unsold copies for Rs. 5 each ? Also find EVPI.
[Ans. EMV = Rs. 120, EVPI = 0·80]

5. Each unit of a product produced and sold yields a profit of a Rs. 50 if a unit produced but not sold results in a loss of Rs. 30. The probability distribution of the no. of units demanded is as follows :

No. of units demanded :	1	3	5
Probability :	0·6	0·3	0·1

How many units be produced to maximise the expected profits. Also calculate EVPI.
[Ans. (a) 3 units (b) EVPI = Rs. 46]

6. A fruit wholesaler buys cases of strawberries for Rs. 200 each and sells them for Rs. 500 each. Any case left would at the end of the day have a salvage value of only Rs. 50. In analysis of past sales record reveals the following probability distribution for the daily number of cases sold :

Daily Sales	Probability
10	0.15
11	0.20
12	0.40
13	0.20

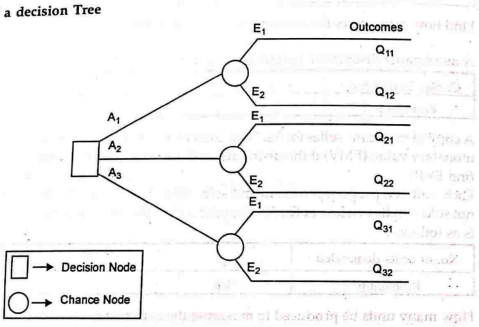
(i) What is the optimum stock action for the fruit seller ?
(ii) Also calculate EVPI for the same.
[Ans. (i) 11 units (ii) Rs. 42.50]

DESISION TREES

Decision making involves several stages and at each stage, each of the chocies open will result in a different payoff. For the sake of simplicity, these stages can be represented by an alternative method called tree formation listing out all events and the resultant outs comes. The tree diagram is also known as **decision tree.** A decision tree is a graphic device of a decision making process. It consists of nodes, branches, probabilities and the resultant pay-offs. Decision trees have standard

symbols. Square indicates a decision node. These are a number of branches leading from this square. These branches indicate various courses of action available to the decision maker. At the end of each branch there is a circle which represents chance node (or state of nature). Various outcomes emerge out of the chance node with their associate probability estimates. The net result of each outcome is indicated against each circle. The branches that are drawn from the decision node are named as decision branches while the branches drawn from chance node are named as chance branches. Following diagram gives the structure of the decision tree.

Sepecimen of a decision Tree



The following example illustrate the application of decision tree.

Example : **An organisation has two packaging machines : old and new. The new machine is more efficient if the materials are of good quality, on the other hand the old machine performs better if the materials are of poor quality. The following information are given :**

(i) 80% materials have been of good quality and 20% of poor quality.

(ii) The profit position is as under :

(a) Using old machine

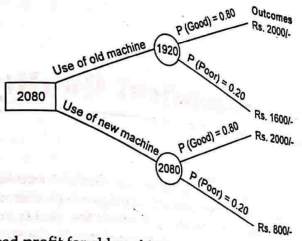
—If the materials are good	Rs. 2000
—If the materials are poor	Rs. 1600

(b) Using new machine

—If the materials are good	Rs. 2,400
—If the materials are poor	Rs. 800

Use a decision tree to decide which machine should be used under the condition that the quality of material is not known at this stage.

Solution:



Expected profit for old machine = $0{\cdot}8 \times 2000 + 0{\cdot}2 \times 1600$ = Rs. 1920

Expected profit for new machine = $0{\cdot}8 \times 2400 + 0{\cdot}2 \times 800$ = Rs. 2080

Since, the expected profit of new machine is high, hence select new machine.

QUESTIONS

1. Explain the concept of statistical decision theory and discuss its usefulness in business situations.
2. What are the elements that decision matrices usually contain ?
3. Explain briefly the following in the context of decision theory :
   (a) State of Nature (b) Act
   (c) Pay off Matrix (d) Minimax Regret Matrix.
4. Explain different methods for making decision under uncertainty.

   OR

   Briefly discuss three different criteria usually adopted for decision making under uncertainty without the use of probability.
5. Explain two different criterion usually adopted for decision making under risk with probability.

   OR

   Explain EMV and EOL criteria of decision making under risk.
6. Write short notes on :
   (i) Laplace strategy (ii) Savage Criterion
   (iii) Wald's Strategy (iv) Probability Distribution to demand
7. Write a brief note on statistical decision criteria.
8. What do you understand by 'Decision Theory' ? Describe some methods which are useful for decision making under uncertainty.
9. Discuss the criteria for taking decision under uncertainty with examples.
10. What is EVPI ? How is it calculated ?
11. Write a note on decision tree.

❀❀❀

10

Statistical Quality Control

INTRODUCTION

In this era of every-growing competition, it has become absolute necessary for a manufacturer/ producer to keep a continuous watch over the quality of the goods produced. But due to large scale production level, it is not possible for a producer to check the quality of each and every item produced. Therefore, to control quality of the manufactured goods, the study of statistical quality control, abbrviated as S.Q.C. is very important and useful.

MEANING OF STATISTICAL QUALITY CONTROL

Statistical Quality Control (S.Q.C.) refers to the use of statistical techniques in controlling the quality of manufactured goods. It is the means of establishing and achieving quality specification, which requires use of tools and techniques of statistics. It is an important application of the theory of probability and theory of sampling for the maintenance of uniform quality in a continuous flow of manufactured products. One of major tools of S.Q.C. is the control chart first introduced by **W.A. Shewhart** through the application of normal distribution.

DEFINITION OF STATISTICAL QUALITY CONTROL (S.Q.C.)

Some important definitions of statistical quality control are given below.

1. *"Statistical quality control can be simply defined as an economic and effective system of maintaining and improving the quality of outputs throughout the whole operating process of specification, production and inspection based on continuous testing with random samples."*
   —Ya Lun Chou
2. *Statistical quality control should be viewed as a kit of tools which may influence decisions to the functions of specification, production or inspection.* **—Eugene L. Grant**

From the above definitions, the essential characteristics of S.Q.C. may be brought about as under : (i) It is designed to control quality standard of goods produced for marketing, (ii) It is exercised by the producers during the production process to assess the quality of the goods, (iii) It is carried out with the help of certain statistical tools like Mean chart, Range chart, P-chart, C-chart, Sampling Inspection Plans, etc. and (iv) It is designed to determine the variations in quality of the goods and limits of tolerance.

ADVANTAGES (OR BENEFITS) ÓF STATISTICAL QUALITY CONTROL

The following are some of the advantages (or benefits) of statistical quality control :

(1) It provides an objective method of controlling the quality of product during the production process. It tells the production manager at a glance whether the quality of the product is under control or not.

(2) It provides a quick method to eliminate assignable causes of variation. By using the technique of statistical quality control, we can detect assignable causes of variation and necessary remedial action can be taken avoiding them.

(3) It provides better quality assurance at lower inspection cost. Sampling inspection is always cheaper vis-a-vis 100% inspection. Control charts are simple to construct and easy to interpret and economical.

(4) The acceptance sampling protects the interest of the consumers by helping him to reject a lot of bad quality. This is also helpful to the producers because they can know the probability of a good lot being rejected.

(5) The very presence of statistical quality control (S.Q.C.) in a manufacturing plant has a healthy influence on the psychology of workers and makes them quality conscious. They know that the quality is being checked.

(6) A quality conscious manufacturing unit is able to earn the goodwill from the consumers of its product which is of immense long run value.

(7) Past data on quality control may serve as a guide for the choice of a new plant and machinery as well as technical staff.

(8) It is possible to defend the qualtiy of output before any government agency on the basis of quality control records.

Limitations : Despite the great significance of statistical quality control, the technique of S.Q.C. suffers from certain limitations as under :

(i) It can not be applied indiscriminately as a panacea for all quality evils.

(ii) It cannot be used mechanically to all production process without studying their peculiar environment.

(iii) It involves mathematical and statistical problems in the process of analysis and interpretation of variations in quality.

(iv) It provides only an information services.

CAUSES OF VARIATION IN QUALITY CHARACTERISTICS

Every manufacturer / producer produces the product according to pre-determined standards. Though the product is carried out with the most sophisticated technology, some variations in the quality of products are bound to take place. For example, it is not possible that all pins, nuts or bolts produced in a factory would be exactly of the same quality. There must be some variation, however, minor it might be in the quality of the various items produced. There may be various causes of this variation. These causes are classified into the following two groups.

(i) **Assignable Causes :** These causes, as the name suggests, refer to those changes in the quality of the products which can be assigned or attributed to any particular cause like defective materials, defective labour, defective machine, etc. However, the effect of such variations can be eliminated with a better system of control like S.Q.C.

(ii) **Chance Causes :** These causes, as the name suggests, takes place as per chance or in a random fashion as a result of the cumulative effect of a multiplicity of several minor causes which cannot be identified. Such type of causes is inherent in every type of production and hence it is accepted as an allowable variation in any scheme of production.

Out of these two types of causes, nothing can be done about the chance causes. However, assignable varaitions can be detected and corrected.

METHODS OF STATISTICAL QUALITY CONTROL

Statistical quality control methods are applied to two distinct phases of plant operation. They are :

(1) Process control

(2) Product control

(1) Process control : Under the process control, the quality of the products is controlled while the products are in the process of production. The process control is secured with the technique of control charts. Control charts are used as a measure of quality control not only in the production process but also in the areas of advertising, packing, air line reservations, etc. Control charts ensures that whether the products confirm to the specified quality standard or not.

(2) Product Control : Under the product control, the quality of the products is controlled while the product is ready for sale and despatch to the customers. The product control is secured with the technique of acceptance sampling. In acceptance sampling, the manufactured articles are formed into lots, a few items are chosen randomly and lot is either accepted on the basis of certain set of rules, usually called sampling Inspection plans.

Thus, process control is concerned with controlling of quality of the goods during the process of manufacturing whereas product control is concerned with the inspection of finished goods, when they are ready for delivery.

CONTROL CHARTS

The control charts are the graphic devices developed by **Walter A. Shewhart** for detecting unnatural pattern of variation in the production process and determining the permissible limits of variaton. Control charts are the core of statistical quality control. These are based on the theory of probability and sampling. Control charts are simple to construct and easy to interpret and they tell the production manager at a glance whether or not the process is in control *i.e.*, within the tolerance limits. A control chart consists of three horizontal lines :

(1) Central Line (CL)

(2) Upper Control Limit (UCL), and

(3) Lower Control Limit (LCL)

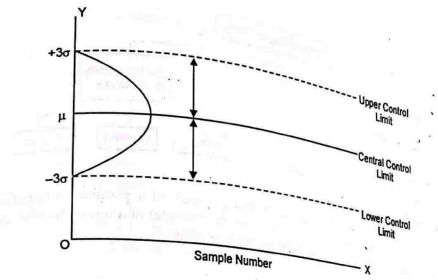
(1) Central Line (CL) : The central line is the middle line of the chart. It indicates the grand average of the measurements of the samples. It shows the desired standard or level of the process. The central line is generally drawn as bold line.

(2) Upper Control Limit (UCL) : The upper control limit is usually obtained by adding 3 sigma (3σ) to the process average. It is denoted by Mean + 3σ. The upper control limit is generally drawn as dotted line.

(3) Lower Control Limit (LCL) : The lower control limit is usually obtained by subtracting 3 sigma (3σ) to the process average. It is denoted by Mean – 3σ. The lower control limit is generally drawn as dotted line.

On the basis of these three lines, a control chart is constructed. The general format of a control chart is given in the diagram below :

In the control chart, the mean values of the statistics T (*i.e.*, Mean, Range, S.D., etc.) for successive samples are plotted and often joined by broken lines to provide a visual clarity. So long as the sample points fall within the upper and lower control limits, there is nothing to worry and in such a case the variation between the samples is attributed to chance causes.



Logic of Setting of Control Limits at $\pm 3\sigma$

Dr. Shewhart has proposed the 3σ limits for the control charts. From the probability, if a variable X is normally distributed, the probability that a random variable will be between $\bar{X} \pm 3\sigma$, where, $\bar{X}$ is the mean and σ is the standard deviation is 0·9973 which is extremely high. Thus, the probability of a random variable fall outside these limits is 0·0027, which is very low. In other words, occurence of events beyond the limits of $\bar{X} \pm 3\sigma$, provided the events lie on a normal curve, is on the whole nearly 3 out of 1000 events are extremely remote chance under normal circumstances. Thus, if $\pm 3\sigma$ limits are employed and the variable quality a characteristic is assumed to be normally distributed, then the probability of sample points falling outside these limits when the process is in control is very small.

Purpose and Uses of Control Charts

The control charts are useful in the following situations :

(1) It helps in determining the quality standard of the products while in process.

(2) It helps in detecting the chance and assignable variations in the quality standards of the produts by setting two control limits lines.

(3) It reveals varaitions in the quality standards of the products from the desired level.

(4) It indicates whether the production process is in control or not so as to take necessary steps for its correction.

(5) Control charts are simple to construct and easy to interpret.

(6) It ensures less inspection cost and time in the process control.

(7) Control charts tell the production manager at a glance whether or not the process is in control.

TYPES OF CONTROL CHARTS

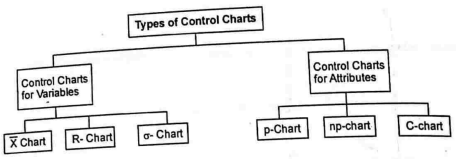
Control charts are of two types depending on whether a given quality or characteristics of a product is measurable or not. These are :

(A) Control Charts for Variables

(1) $\bar{X}$-Chart (2) R-Chart (3) σ-Chart

(B) Control Charts for Attributes

(1) p-chart (2) np-chart (3) C-chart



A. Control Charts for Variables

These charts are used when the quality or characteristics of a prdouct is capable of being measured quantitatively such as gauge of a steel alimirah, diameter of a screw, tensile strength of a steel pipe, resistance of a wire etc. Such charts are of three types :

(1) $\bar{X}$**-Chart (or Mean Chart)**

(2) *R*-**Chart (or Range Chart)**

(3) σ-**Chart (or Standard Deviation Chart)**

(1) $\bar{X}$**-Chart :** This chart is constructed for controlling the variations in the average quality standard of the products in a production process.

Procedure : The construction of $\bar{X}$-chart involves the following steps :

(i) Compute the mean of each sample *i.e.*,

$$\bar{X}_1, \bar{X}_2, \bar{X}_3 \ldots . \bar{X}_k$$

(ii) Compute the mean of the samples means by dividing the sum of the sample means by the number of samples *i.e.*,

$$\bar{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 \ldots . + \bar{X}_k}{\text{No. of samples}} = \frac{\Sigma \bar{X}}{k} \quad \text{where, } k = \text{No. of samples}$$

This grand mean ($\bar{\bar{X}}$) represents the Central Line (CL)

(iii) Determine the control limits by using the following formula :

(a) On the basis of standard deviation of the population (σ)

$$\text{Control Limits} = \bar{\bar{X}} \pm \frac{3\sigma}{\sqrt{n}}$$

$$\therefore \quad \text{UCL} = \bar{\bar{X}} + \frac{3\sigma}{\sqrt{n}} \quad \text{and}$$

$$\text{LCL} = \bar{\bar{X}} - \frac{3\sigma}{\sqrt{n}}$$

These control limits represents the upper control line and lower control line.

(b) On the basis of the Quality Control Factors A_2 and $\bar{R}$

Control Limits $= \bar{\bar{X}} \pm A_2 \bar{R}$, where, $\bar{R}$ = Mean of the ranges

$$\text{UCL} = \bar{\bar{X}} + A_2 \bar{R}$$

$$\text{LCL} = \bar{\bar{X}} - A_2 \bar{R}$$

Where, A_2 is a quality control factor whose value is obtained from the control chart table with reference to the size of the sample.

$$\bar{R} = \text{Mean of Ranges} = \frac{\Sigma R}{N}$$

(iv) Construct the mean chart ($\bar{X}$-chart) by plotting the sample number on x-axis and sample mean, UCL, LCL and Central Line on the y-axis.

(v) Interpret the $\bar{X}$-Chart. If all the sample means ($\bar{X}$) fall within the control limits, the production process is in a state of control otherwise it is beyond control.

(2) *R*-**Chart**

The Range Chart (*R*-chart) is constructed for controlling the variation in the dispersion or variability of the quality standard of the products in a production process.

Procedure :

The construction of *R*-chart involves the following steps :

(i) Compute the range (*R*) of each sample using the formula :

$$R = L - S \qquad L = \text{Largest value } S = \text{Smallest value}$$

(ii) Compute the mean of ranges by dividing the sum of the samples ranges (ΣR) by the number of samples *i.e.*,

$$\bar{R} = \frac{R_1 + R_2 + \ldots . + R_k}{k} = \frac{\Sigma R}{k} \quad \text{where, } k = \text{No. of samples}$$

The mean of ranges ($\bar{R}$) represents the Central line (CL) for the *R*-Chart

(iii) Determine the control limits by using the following formula :

(a) On the basis of Quality Control Factors D_3 and D_4 and $\bar{R}$:

$$\text{Upper control limit (UCL)} = D_4 \bar{R}$$

$$\text{Lower control limit (LCL)} = D_3 \bar{R}$$

where D_3 and D_4 are the quality control factors and their values are obtained from the control chart table with reference to the size of the sample.

$\bar{R}$ = Mean of Range

(b) On the basis of Quality Control Factors D_1, D_2 and population standard deviation (σ)

$$\text{UCL} = D_2 \sigma$$

$$\text{LCL} = D_1 \sigma$$

where, D_1 and D_2 are the quality control factors.

The value of *LCL* cannot be negative and in such case it would be reduced to zero.

(iv) Construct the *R*-Chart (Range Chart) by plotting the sample number on the x-axis and sample ranges (R), UCL, LCL and Central Line (CL) on the y-axis.

(v) Interpret the *R*-chart. If all the sample ranges (R) fall within the control limits, the production process is in a state of control otherwise it is beyond control.

Example 1. **Construct $\overline{X}$-Chart and Range Chart for the following data of 5 samples with each set of 5 items :**

Sample No.	Weights				
1	20	15	10	11	14
2	12	18	10	8	22
3	21	19	17	10	13
4	15	12	19	14	20
5	20	19	26	12	23

(Conversion factors for $n = 5$, $A_2 = 0{\cdot}577$, $D_3 = 0$, $D_4 = 2{\cdot}115$)

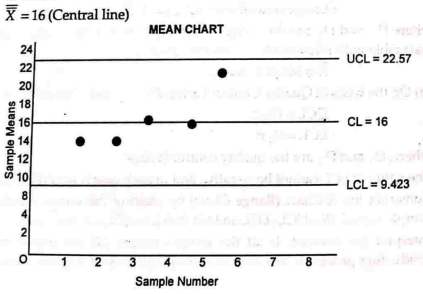
Solution. **Construction of $\overline{X}$ and R Charts**

Sample No.	Weights of Items in each sample (X)					Total Weights (ΣX)	$\overline{X} = (\Sigma X \div 5)$	Range $R = (L - S)$
1	20	15	10	11	14	70	14	10
2	12	18	10	8	22	70	14	14
3	21	19	17	10	13	80	16	11
4	15	12	19	14	20	80	16	8
5	20	19	26	12	23	100	20	14
$K = 5$							$\Sigma\overline{X} = 80$	$\Sigma R = 57$

$$\overline{\overline{X}} = \frac{\Sigma\overline{X}}{k} = \frac{80}{5} = 16 \qquad \overline{R} = \frac{\Sigma R}{k} = \frac{57}{5} = 11{\cdot}4$$

$\overline{X}$ Chart

$\overline{\overline{X}} = 16$ (Central line)



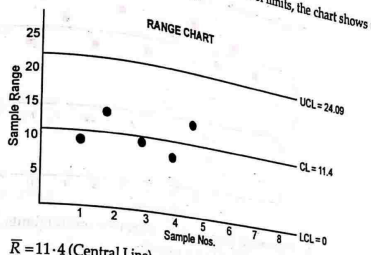
Control Limits

$$UCL = \overline{\overline{X}} + A_2\overline{R} = 16 + 0{\cdot}577 \times 11{\cdot}4 = 16 + 6{\cdot}577 = 22{\cdot}577$$

$$LCL = \overline{\overline{X}} - A_2\overline{R} = 16 - 0{\cdot}577 \times 11{\cdot}4 = 16 - 6{\cdot}577 = 9{\cdot}423$$

As all the sample mean values fall within the control limits, the chart shows that the given process is in staistical control.

Range Chart :



$\overline{R} = 11{\cdot}4$ (Central Line)

Control Limits :

$UCL = D_4\overline{R} = 2{\cdot}115 \times 11{\cdot}4 = 24{\cdot}09$ $\quad LCL = D_3\overline{R} = 0 \times 11{\cdot}4 = 0$

As all the range points fall within the control limits, so R-chart shows that the given process is in statistical control.

Example 2. **A machine is set to deliver packet of a given weight. 10 samples of size 5 each were recorded in the data given below :**

Sample No. :	1	2	3	4	5	6	7	8	9	10
Mean $\overline{X}$:	15	17	15	18	17	14	18	15	17	16
Range :	7	7	4	9	8	7	12	4	11	5

Construct the Mean Chart and Range chart and comment on state of control. (Conversion Factors for $n = 5$ are $A_2 = \cdot577$, $D_3 = 0$, $D_4 = 2{\cdot}115$)

Solution :

Sample No. :	1	2	3	4	5	6	7	8	9	10	Total
Mean $\overline{X}$:	15	17	15	18	17	14	18	15	17	16	162
Range :	7	7	4	9	8	7	12	4	11	5	74

$$\overline{\overline{X}} = \frac{\Sigma\overline{X}}{N} = \frac{162}{10} = 16{\cdot}2 \qquad \overline{R} = \frac{\Sigma R}{N} = \frac{74}{10} = 7{\cdot}4$$

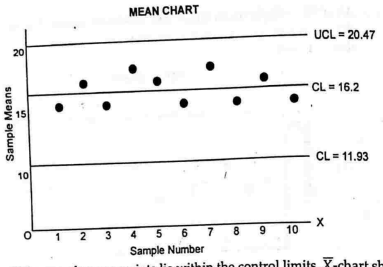
Mean Chart ($\overline{X}$ Chart)

$\overline{\overline{X}} = 16{\cdot}2$ (Central Line)

Control Limits

$$UCL = \overline{\overline{X}} + A_2\overline{R} = 16{\cdot}2 + \cdot577 \times 7{\cdot}4 = 20{\cdot}47$$

$$LCL = \overline{\overline{X}} - A_2\overline{R} = 16{\cdot}2 - \cdot577 \times 7{\cdot}4 = 11{\cdot}93$$



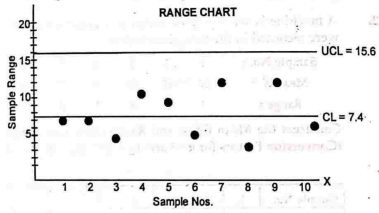
As all the sample mean points lie within the control limits, $\overline{X}$-chart shows that the given process is in statistical control.

Range Chart (*R*-Chart)

$\overline{R} = 7 \cdot 4$ (Central Line)

Control Limits

$UCL = D_4 \cdot \overline{R} = 2 \cdot 115 \times 7 \cdot 4 = 15 \cdot 65$ $\quad LCL = D_3 \overline{R} = 0 \times 7 \cdot 45 = 0$



As all the sample range points lie within the control limits, the *R*-Chart shows that the given process is in statistical control.

Example 3. **The following are the mean lengths and ranges of lengths of a finished product from 10 samples each of size 5. The specification limits for length are 200 ± 5 cm. Construct $\overline{X}$ and *R* charts and examine whether the process is under control and state your recommendations.**

Sample No. :	1	2	3	4	5	6	7	8	9	10
Mean $\overline{X}$:	201	198	202	200	203	204	199	196	199	201
Range :	5	0	7	3	4	7	2	8	5	6

Assume for $n = 5$, $A_2 = 0.577$, $D_3 = 0$ and $D_4 = 2.115$.

Solution. The specification limits for length are given to be 200 ± 5 cm. Hence, mean is known where as standard deviation is unknown.

Control limits for $\overline{x}$ chart

Central limit, $\quad CL = \overline{\overline{X}} = 200.$

$$UCL = \overline{\overline{X}} + A_2\overline{R}, \text{ where } \overline{R} = \frac{\Sigma R_i}{10} = \frac{47}{10} = 47.$$

$$UCL = 200 + 0.577 \times 4.7 = 202.712$$

$$LCL = 200 - 0.577 \times 4.7 = 197.29.$$

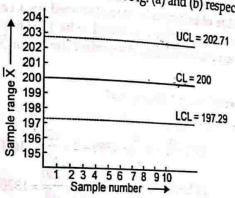
Control limits for R chart

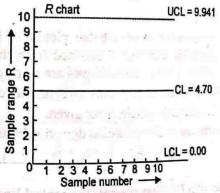
$$CL = \overline{R} = 4.7.$$

$$UCL = D_4\overline{R} = 2.115 \times 4.7 = 9.941$$

$$LCL = D_3\overline{R} = 0 \times 4.7 = 0.$$

The $\overline{X}$ and *R* charts are drawn in Fig. (a) and (b) respectively.





It can be seen that all points lie within the control limits of *R* chart. The process variability is, therefore, under control. However, three points corresponding to sample no. 5, 6 and 8 lie outside the control limits $\overline{X}$ chart. The process is, therefore, not in statistical control. The process, therefore, should be halted to check whether there are any assignable causes. If they are found, the process should be readjusted to remove them, otherwise fluctuations are going to be there.

Exmaple 4. **Twenty five samples of six items each were related from the assembly line of a machine having mean of 25 samples is 0·81 inches and range ·0025 inches.**

Compute the upper control limits and lower control limits of mean chart and range chart.
(For $n=6$, $A_2=0\cdot483$, $D_3=0$, $D_4=2\cdot004$)

Solution. Given : $\overline{\overline{X}}=0\cdot81$, $\overline{R}=0\cdot0025$, $n=6$

$\overline{X}$ **–Chart.**

Control Limits

$$UCL=\overline{\overline{X}}+A_2\overline{R}=\cdot81+(\cdot483)(\cdot0025)=\cdot8112$$

$$LCL=\overline{\overline{X}}-A_2\overline{R}=\cdot81-(\cdot483)(\cdot0025)=\cdot8088$$

Range Chart

Control Limits

$$UCL=D_4\overline{R}=(2\cdot004)\times(\cdot0025)=0\cdot0050.$$

$$LCL=D_3\overline{R}=0\times(\cdot0025)=0.$$

Example 5. **The mean life of battery cells manufactured by a certain plant as estimated on the basis of a large sample was found to be 1500 hrs with standard deviation of 180 hrs. Compute the 3-sigma (3σ) control limits for $\overline{X}$-Chart for a sample of size $n=9$.**

Solution. Given : $\overline{\overline{X}}=1500$ hrs, $\sigma=180$ hrs., $n=9$

Control Limits for $\overline{X}$ chart.

$$UCL=\overline{\overline{X}}+3\frac{\sigma}{\sqrt{n}}=1500+3\times\frac{180}{\sqrt{9}}=1680$$

$$LCL=\overline{\overline{X}}-3\frac{\sigma}{\sqrt{n}}=1500-3\times\frac{180}{\sqrt{9}}=1320$$

(3) σ-Chart

This chart is constructed to get a better picture of the variations in the quality standard in a process than that is obtained from the Range chart provided the standard deviation of the various samples are readily available.

Procedure : The constuction of σ-chart involves the following steps :

(i) Find the S.D. of each sample, if not given.

(ii) Compute the mean of the standad deviation by using the formula :

$$\overline{S}=\frac{\Sigma S}{k}=\frac{S_1+S_2+S_3+\ldots..S_k}{k}$$

The mean of S.D.s ($\overline{S}$) represents the central line (CL).

(iii) Find the upper and lower control limits by using the formula :

(a) **On the basis of quality control factors B_1, B_2 and population standard deviation (σ)**

$$UCL=B_2.\sigma \qquad LCL=B_1.\sigma$$

(b) **On the basis of quality control factors B_3, B_4 and estimated population standard deviation ($\overline{S}$)**

$$UCL=B_4.\overline{S} \qquad LCL=B_3.\overline{S}$$

Where, B_1, B_2, B_3 and B_4 are the quality control factors.

(iv) Construct σ-Chart by plotting the sample number on the x-axis and sample S.D. (σ), UCL, LCL and CL on the y-axis.

(v) Interpret the chart thus drawn.

Example 6. **Quality control is maintained in a factory with the help of mean and standard deviation charts. Ten items are chosen in every sample. Eighteen samples in all were chosen whose $\Sigma\overline{X}$ was 595·8 and ΣS was 8·28. Determine the three sigma limits of $\overline{X}$ and σ charts. You may use the following factors for finding 3σ limits**

n	A_1	B_3	B_4
10	0·949	0·28	1·72

Solution. Given : $\Sigma\overline{X}=595\cdot8$, $\Sigma S=8\cdot28$, $n=18$

$$\overline{\overline{X}}=\frac{\Sigma\overline{X}}{k}=\frac{595\cdot8}{18}=33\cdot1$$

$$\overline{S}=\frac{\Sigma S}{k}=\frac{8\cdot28}{18}=0\cdot46$$

$\overline{X}$**-Chart**

$\overline{\overline{X}}$ 33·1 (Central Line)

Control Limits

$$UCL=\overline{\overline{X}}+A_1\overline{\sigma}=33\cdot1+(0\cdot949)(\cdot46)=33\cdot1+\cdot43675=33\cdot53$$

$$LCL=\overline{\overline{X}}-A_1\overline{\sigma}=33\cdot1-(0\cdot949)(\cdot46)=32\cdot66$$

σ Chart

Control Limits

$$\overline{S}=0\cdot46$$

$$UCL=B_4\overline{S}=1\cdot72\times0\cdot46=0\cdot7912$$

$$LCL=B_3\overline{S}=0\cdot28\times0\cdot46=0\cdot1288$$

EXERCISE – 1

1. Construct $\overline{X}$-Chart and R-Chart for the following data of 12 samples with each set of 5 items :

42	42	19	36	42	51	60	18	15	69	64	61
65	45	24	54	51	74	60	20	30	109	90	78
75	68	80	69	57	75	72	27	39	113	93	94
78	72	81	77	59	78	95	42	62	118	109	109
87	90	81	84	78	132	138	60	84	153	112	136

(Given : $n=5$, $A_2=\cdot58$, $D_3=0$, $D_4=2\cdot115$)

[**Ans.** $UCL_{\overline{x}}=106\cdot2$, $LCL_{\overline{x}}=37\cdot0$, $UCL_{\overline{R}}=125\cdot9$, $LCL_{\overline{R}}=0$]

2. Construct Mean Chart ($\bar{X}$) and Range chart (R) for the following data :

Sample No.	Observations I	II	III
1	20	19	25
2	25	22	28
3	32	23	30
4	18	20	15
5	10	12	18
6	22	25	17
7	28	39	24
8	30	29	30

(Given : $n=3$, $A_2=1{\cdot}023$, $D_3=0$, $D_4=2{\cdot}575$)
[**Ans.** $UCL_{\bar{X}}=29{\cdot}266$, $LCL_{\bar{X}}=16{\cdot}734$, $UCL_R=15{\cdot}77$, $LCL_R=0$]

3. A machine is set to deliver packet of a given weights. 10 samples of size 5 each were recorded in the data given below :

Sample No.	1	2	3	4	5	6	7	8	9	10
Sample Mean ($\bar{X}$)	20	34	45	39	26	29	13	34	37	23
Sample Range (R)	23	29	15	5	29	17	21	11	90	10

Construct $\bar{X}$ chart and range chart and point out whether the process is within control (Conversion factors for $n=5$, $A_2=\cdot58$, $D_3=0$, D_4 2·115)
[**Ans.** $UCL_{\bar{X}}=41{\cdot}658$; $LCL_{\bar{X}}=18{\cdot}342$, $UCL_R=42512$, $LCL_R=0$]

4. The following data provide the mean ($\bar{X}$) and range (R) of 10 samples having 5 items each, construct mean chart and range chart and comment on the process of quality :

Sample No. :	1	2	3	4	5	6	7	8	9	10
Sample Mean ($\bar{X}$) :	11·2	11·8	10·8	11·6	11·0	9·6	10·4	9·6	10·6	10·0
Sample Range (R) :	7	4	8	5	7	4	8	4	7	9

(Conversion factor for $n=5$ are $A_2=0{\cdot}577$,, $D_3=0$ and $D_4=2{\cdot}115$)
[**Ans.** $UCL_{\bar{X}}=14{\cdot}2951$, $LCL_{\bar{X}}=7{\cdot}0249$, $UCL_R=13{\cdot}3245$, $LCL_R=0$]

5. Thirty samples of 5 items each were taken from the output of a machine and a critical dimension measured. The mean of 30 samples was 0·6550 inches and Range mean 0·0036 inch. Compute the control limits for $\bar{X}$ and R charts.
(Conversion factors for $n=5$, $A_2=0{\cdot}58$, $D_3=0$, $D_4=2{\cdot}115$)
[**Ans.** (i) ·6570, ·6529 (ii) ·0076, 0]

6. A drilling machine bores holes with a mean diameter of 0·5230 cm. and a standard deviation of 0·0032 cm. Calculate the 2-sigma and 3-sigma upper and lower control limits for means of samples of size 4, and prepare a control chart.
[**Ans.** 2–sigma limits : UCL = 0·5262; LCL = 0·5198
2–sigma limits : UCL = 0·5278; LCL = 0·5182]

7. Construct a control chart for mean ($\bar{X}$) and range (R) for the following data from 10 independent samples of 5 observations each from a production process :

Sample No. :	1	2	3	4	5	6	7	8	9	10
Sample Mean ($\bar{X}$) :	43	49	37	44	45	37	51	46	43	47
Sample Range (R) :	5	6	5	7	7	4	8	6	4	6

[**Ans.** $UCL_{\bar{X}}=47{\cdot}56$, $LCL_{\bar{X}}=40{\cdot}836$, $UCL_R=12{\cdot}267$, $LCL_R=0$]

8. A machine is set to deliver packets of a given weights. Ten samples of size 5 each were recorded. Below are given the relevant data :

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean :	45	51	39	47	57	39	53	48	45	49
Range :	9	7	9	9	7	8	8	2	9	9

Calculate the value of the central line and control limits for mean and range chart.
(Conversion factors for $n=5$ are $A_2=0{\cdot}58$, $D_3=0$, $D_4=2{\cdot}115$)
[**Ans.** $UCL_{\bar{X}}=51{\cdot}766$, $LCL_{\bar{X}}=42{\cdot}834$; $ULC_R=16{\cdot}28$, $LCL_R=0$]

B. CONTROL CHARTS FOR ATTRIBUTES

These charts are used when the quality or characteritics of a product cannot be measured in quantitative form and the data is studied on the basis of totality of attributes like defective and non-defectives. Such charts are of three types :

(1) ***p*-chart (or Fraction Defective Chart)**
(2) ***np*-chart (or Number of Defective Chart)**
(3) ***c*-chart (or Number of Defects per unit Chart)**

(1) ***p*-chart (Fraction Defective Chart) :** This chart is constructed for controlling the quality standard in the average fraction defective of the products in a process when the observed sample items are classified into defectives and non-defectives.

Procedure : The construction of p-chart involves the following steps :

(i) Find the fraction defective or proportion of defective in each sample *i.e.*,

$$p_1, p_2, p_3 \ldots\ldots p_k.$$

(ii) Find the mean of the fraction defectives by using the formula :

$$\bar{p}=\frac{\text{Total No. of Defectives}}{\text{Total No. of Units Inspected}}, \quad \bar{q}=1-\bar{p}$$

Aliter : The value of $\bar{p}$ can also be calculated as :

$$\bar{p}=\frac{p_1+p_2+\ldots.+p_k}{k} \quad \text{where, } k=\text{No. of samples}$$

The value of $\bar{p}$ represents the central line of the p-chart

(iii) Determine the control limits by using the formula :

$$\text{Control Limits}=\bar{p}\pm 3\sqrt{\frac{\bar{p}\,.\,\bar{q}}{n}}$$

$$\therefore \quad \text{LCL}=\bar{p}-3\sqrt{\frac{\bar{p}\,.\,\bar{q}}{n}}$$

$$\text{UCL}=\bar{p}+3\sqrt{\frac{\bar{p}\,.\,\bar{q}}{n}}$$

The value of LCL cannot be negative and in such a case it would be reduced to zero.

(iv) Construct the *p*-chart by plotting the sample number on *x*-axis and sample fraction defectives, UCL, LCL and central line on the *y*-axis.

(v) Interpret the *p*-chart. If all the sample fraction defective (*p*) fall within the control limits, the process is in a state of control otherwise it is beyond the control.

Note :

1. If the number of defectives is small, then *p*-chart should be constructed by finding the percentage defective.
2. This chart is specially useful when the size of the sample (*n*) is un equal. In such a case the value of *n* can be obtained by dividing the defective units in all the samples by the number of samples.

CASE I : EQUAL SAMPLE SIZE

Example 7. **The following data refers to visual defects found during the inspection of the first 10 samples of size 100 each from a lot of two-wheelers manufactured by an automobile company :**

Sample Number :	1	2	3	4	5	6	7	8	9	10
No. of defectives	5	3	3	6	5	6	8	10	10	4

Construct a control chart for fraction defective. What conclusions you draw from the control chart ?

Solution. We are given : n = size of sample = 100, k = No. of samples = 10

Computation of Fraction Defectives

Sample No. (k)	Size of Sample (n)	No. of defectives (d)	Fraction defectives d/n
1	100	5	5 / 100 = 0·05
2	100	3	3 / 100 = 0·03
3	100	3	0·03
4	100	6	0·06
5	100	5	0·05
6	100	6	0·06
7	100	8	0·08
8	100	10	0·10
9	100	10	0·10
10	100	4	0·04
k = 10	1,000	Σd = 60	0.60

$$\bar{p} = \frac{\text{Total No. of Defectives}}{\text{Total No. of Units}} = \frac{60}{1000} = 0{\cdot}06 \quad \Rightarrow \quad \bar{q} = 1 - \bar{p} = 1 - 0{\cdot}06 = 0{\cdot}94$$

The value of $\bar{p}$ represents the central line

Control Limits for *p*-chart

$$UCL = \bar{p} + 3\sqrt{\frac{\bar{p}\cdot\bar{q}}{n}}$$

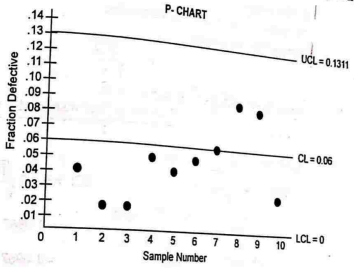
$$= 0{\cdot}06 + 3\sqrt{\frac{06 \times 94}{100}}$$

$$= 0{\cdot}06 + 3(0{\cdot}0237) = 0{\cdot}06 + {\cdot}0711 = 0{\cdot}1311$$

$$LCL = \bar{p} - 3\sqrt{\frac{\bar{p}\cdot\bar{q}}{n}} = 0{\cdot}06 - 3\sqrt{\frac{06 \times 94}{100}}$$

$$= {\cdot}06 - 3({\cdot}0237) = {\cdot}06 - {\cdot}0711 = -0{\cdot}0111 = 0.$$

Since, the fraction defective cannot be negative, LCL is taken as zero. The fraction defective chart (*p*-chart) is shown below :



The above chart shows that all the points lie within the control limits. This suggests that the process is in control.

CASE II : VARYING SAMPLE SIZE

Example 8. **The number of defective needles of sewing machine has been given in the following table on the basis of daily inspection. Prepare '*p*-chart' and state whether the production process is in control.**

Day	1	2	3	4	5	6	7	8	9	10
No. of needles inspected	90	60	70	100	120	50	100	110	100	100
No. of defective needles	5	12	7	3	6	5	10	6	8	25

Solution. Computation of Control limits for *p*-chart

Day	No. of needles Inspected	No. of defectives	Percentage defective needles
1	90	5	5·56
2	60	12	20·00
3	70	7	10·00
4	100	3	3·00
5	120	6	5·00

6	50	5	10·00
7	100	10	10·00
8	110	6	5·45
9	100	8	8·00
10	100	25	25·00
$k=10$	900	87	

$$\bar{p} = \frac{\text{Total No. of Defectiver Needles}}{\text{Total no. of Items Inspected}} = \frac{87}{900} = 0\cdot0967$$

$\bar{p}$ is the percentage form = 9·67

The value of $\bar{p}$ represents the central line

Control limit for *p*-chart

$$\bar{p} \pm 3\sqrt{\frac{\bar{p}\cdot\bar{q}}{n}}$$

where, $\bar{p} = 0\cdot0967$, $\bar{q} = 1 - 0\cdot0967 = 0\cdot9033$

$$n = \frac{\text{Total No. of Items Inspected}}{k} = \frac{900}{10} = 90$$

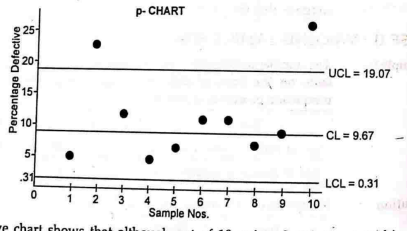
Substituting the values we get

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}\cdot\bar{q}}{n}} \qquad \text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}\cdot\bar{q}}{n}}$$

$$= 0\cdot0967 + 3\times\sqrt{\frac{0.0967\times0.9033}{90}} \qquad = 0\cdot0967 - 3\sqrt{\frac{0\cdot0967\times0\cdot9033}{90}}$$

$$= 0\cdot0967 + 3\times0\cdot0312 \qquad = 0\cdot0967 - 3\times0\cdot0312$$

$$= 0\cdot1903 \text{ or } 19\cdot03\% \qquad = 0\cdot0031 \text{ or } 0\cdot31\%$$



The above chart shows that although out of 10 points 8 points are within the control limits but the points of sample number 2 and 10 are outside the upper limit. This suggests that the process is not in control.

Example 9. **Construct a control chart for the proportion of defectives obtained in repeated samples of size 100 from a process which is considered to be under control when the average proportion of defective p is equal to 0·20. Draw the central line and the upper and lower control limits on graph paper.**

Solution. We are given :

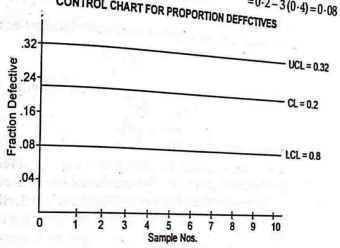
$\bar{p}$ = Average fraction defective = 0·2, n = 100, $\bar{q} = 1 - \bar{p} = 1 - 0\cdot2 = 0\cdot8$

Central Line = $\bar{p} = 0\cdot2$

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}\cdot\bar{q}}{n}} \qquad \text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}\cdot\bar{q}}{n}}$$

$$= 0\cdot2 + 3\sqrt{\frac{0\cdot20\times0\cdot80}{100}} \qquad = 0\cdot2 - 3\sqrt{\frac{0\cdot20\times0\cdot80}{100}}$$

$$= 0\cdot2 + 3(0\cdot04) = 0\cdot32 \qquad = 0\cdot2 - 3(0\cdot4) = 0\cdot08$$



Example 10. **A daily sample of 30 items was taken over a period of 14 days in order to establish control limits. If 21 defectives were found, what should be the upper and lower control limits for the proportion of defectives ?**

Solution. No. of samples (k) = 14

Size of the sample (n) = 30

Σd *i.e.*, number of defectives = 21

$$\bar{p} = \text{Average fraction defectives} = \frac{21}{14\times30} = 0\cdot05$$

$$\bar{q} = 1 - \bar{p} = 1 - 0\cdot5 = 0\cdot95$$

Control limits for *p*-chart :

Central Line = $\bar{p} = 0\cdot05$

$$\text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}\cdot\bar{q}}{n}} = 0\cdot05 + 3\sqrt{\frac{(\cdot05)(0\cdot95)}{30}} = 0\cdot17$$

$$\text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}\cdot\bar{q}}{n}} = 0\cdot05 - 3\sqrt{\frac{(\cdot05)(0\cdot95)}{30}} = -0.069.$$

The negative value of LCL is taken as zero.

(2) *np*-Chart (Number of Defective Chart)

This chart is constructed for controlling the quality standard of attributes in a process where the sample size is equal and it is required to plot the number of defectives (np) in samples instead of fraction defectives (p).

Procedure : The construction of np-chart involves the following steps :

(i) Find the average number of defectives ($n\bar{p}$)

$$n\bar{p} = \frac{\text{Total no. of Defectives}}{\text{Total no. of Samples}} = \frac{\Sigma d}{k}$$

The value of $n\bar{p}$ represents the central line

(ii) Find the value of $\bar{p}$ by using the formula :

$$\bar{p} = \frac{n\bar{p}}{n} \quad \Rightarrow \quad \bar{q} = 1 - \bar{p}$$

Aliter : The values of $\bar{p}$ can also be calculated by using the formula :

$$\bar{p} = \frac{\Sigma d}{n \times k}$$

(iii) Determine the control limits by using the formula :

$$\therefore \quad \text{UCL} = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}}$$
$$\text{L.C.L.} = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}}$$

The value of LCL cannot be negative and in such a case it would be reduced to zero.

(iv) Construct np-chart by plotting the sample number on x-axis and sample number of defectives, UCL, LCL and control line (CL) on the y-axis.

(v) Interpret np-chart. If all the sample number of defectives fall within the control limits, the process is in a state of cotnrol otherwise it is beyond control.

Note : The construction and interpretation of the number of defective chart *i.e.*, np chart is similar to that of p-chart. In np-chart, the central line is drawn at np instead of p and the actual number of defectives (np) in samples of fixed size n is plotted instead of fraction defectives.

Example 11. **An inspection of 10 samples of size 400 each from 10 lots reveal the following number of defectives :**

17 15 14 26 9 4 19 12 9 15

Calculate the control limits for the number of defective units. Plot on the graph and state whether the process is under control or not.

Solution. We are given, $n = 400$, $k =$ (No. of samples) $= 10$,

$$\bar{p} = \text{Average fraction defectives} = \frac{140}{10 \times 400} = 0{\cdot}035, \quad \bar{q} = 1 - 0{\cdot}35 = 0{\cdot}965$$

Also, $n = 400$

$$\therefore \quad n\bar{p} = 400 \times {\cdot}035 = 14$$

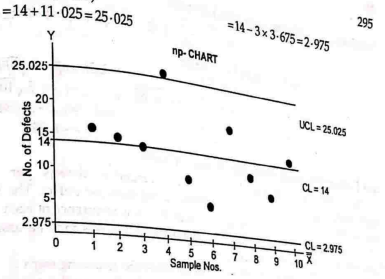
The value of $n\bar{p}$ represents the central line

Control Limits for np-chart

$$\text{UCL} = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}} = 14 + 3\sqrt{400 \times 0{\cdot}035 \times 0{\cdot}965}$$

$$\text{LCL} = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}} = 14 - 3\sqrt{400 \times 0{\cdot}035 \times 0{\cdot}965}$$

$$= 14 + 3(3{\cdot}675)$$
$$= 14 + 11{\cdot}025 = 25{\cdot}025$$
$$= 14 - 3 \times 3{\cdot}675 = 2{\cdot}975$$



The above chart shows that although out of 10 points 9 points are within the control limits but the point for sample 4 is outisde the UCL. This suggest that the process is not in control.

Example 12. **In a certain sampling inspection, the number of defectives found in 10 samples of 100 each are given below :**

16, 18, 11, 18, 21, 10, 20, 18, 17, 21

Do these indicate that the quality charactersitics inspected is under statistical control.

Solution. Here, we use $n\bar{p}$-chart to find whether quality characteristics under inspection is in a state of control or not.

We are given : $n = 100$ $k = 10$, $\Sigma d =$ Total no. of Defectives $= 170$

$$\bar{p} = \frac{170}{100 \times 10} = 0{\cdot}17, \; \bar{q} = 1 - 0{\cdot}17 = {\cdot}83$$

Also, $n = 100$,

Now, $n\bar{p} = 100 \times {\cdot}17 = 17$

The value of $n\bar{p}$ represents the central line

Control Limits for $n\bar{p}$ chart

$$\text{UCL} = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}} = 17 + 3\sqrt{100 \times 0{\cdot}17 \times 0{\cdot}83} = 17 + 11{\cdot}268 = 28{\cdot}268$$

$$\text{LCL} = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}} = 17 - 3\sqrt{100 \times 0{\cdot}17 \times 0{\cdot}83} = 17 - 11{\cdot}268 = 5{\cdot}732$$

Since, none of the points is lying outside the lower and upper control limits, the process is in a state of statistical control.

Example 13. **It was found that the production process is termed "Controlled" in a sample size of 10 units each when average number of defective is 1·2. What control limits you will establish for a control chart of a sample size of 10 units each ?**

Solution. We are given : $n\bar{p} = 1{\cdot}2$, $n = 10$,

$$\bar{p} = \frac{n\bar{p}}{n} = \frac{1.2}{10} = 0{\cdot}12, \qquad \bar{q} = 1 - \bar{p} = 1 - 0{\cdot}12 = 0{\cdot}88$$

Control Limits for $n\bar{p}$ chart

$$UCL = n\bar{p} + 3\sqrt{n\bar{p}\bar{q}} = 1{\cdot}2 + 3\sqrt{10 \times 0{\cdot}12 \times 0{\cdot}88} = 1{\cdot}2 + 3(1{\cdot}027) = 4{\cdot}281$$

$$LCL = n\bar{p} - 3\sqrt{n\bar{p}\bar{q}} = 1{\cdot}2 - 3{\cdot}081 = -1{\cdot}881$$

(iii) C-Chart (Number of Defects Per unit Chart)

This chart is used for the control of number of defects per unit say a piece of cloth/glass/paper/bottle which may contain more than one defect. The inspection unit in this chart will be a single unit of product. The probability of occurrence of each defect tends to remain very small. Hence, the distribution of the number of defects may be assumed to be a Poisson Distribution with Mean = Variance.

Procedure : The construction of *C*-chart involves the following steps :

(i) Determine the number of defects per unit (C) in the samples of equal size.

(ii) Find the mean of the number of defects counted in several units by using the formula :

$$\bar{C} = \frac{\Sigma C}{K}$$

where, K = Total No. of Units Inspected.

The value of $\bar{C}$ represents the central line of the *C*-chart.

(iii) Determine the control limits by using the formula :

Control Limits $= \bar{C} \pm 3\sqrt{\bar{C}}$.

$$UCL = \bar{C} + 3\sqrt{\bar{C}}$$

$$LCL = \bar{C} - 3\sqrt{\bar{C}}$$

The value of LCL cannot be negative and in such case it would be reduced to zero.

(iv) Construct *C*-chart by plotting the sample numbers on the *x*-axis and number of defects observed per unit, LCL, UCL and CL on the *y*-axis.

(v) Interpret *C*-Chart. If the observed values of the number of defects per unit fall within the control limits, the process is a state of control otherwise it is beyond the control.

USES OF C-CHART

Althoguh the application of *C*-chart is somewhat limited compared with $\bar{X}$ and *R* charts, yet a number of practical situation exist in many industry where *C*-chart is used. The following are the fields of applications of *C*-chart.

(1) Number of defects of all kinds of aircraft final assembly.

(2) Number of defects counted in a roll of coated paper, sheet of photographic film, bale (or pieces) of cloth, etc.

Example 14. **Ten pieces of cloth out of differnt rolls of equal length contained the follwoing number of defects :**

1, 3, 5, 0, 6, 0, 9, 4, 4, 3

Draw a control chart for the number of defects and state whether the process is in a state of statistical control.

Solution. We have $N = 10$, and C = No. of defects = 35

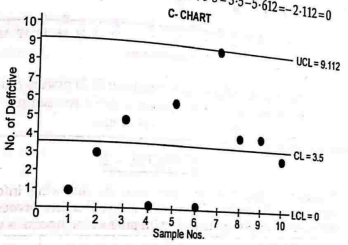
$$\bar{C} = \frac{\Sigma C}{N} = \frac{35}{10} = 3{\cdot}5$$

The value of $\bar{C}$ represents the central line.

Control Limits for C-Chart

$$UCL = \bar{C} + 3\sqrt{\bar{C}} = 3{\cdot}5 + 3\sqrt{3{\cdot}5} = 3{\cdot}5 + 5{\cdot}612 = 9{\cdot}112$$

$$LCL = \bar{C} - 3\sqrt{\bar{C}} = 3{\cdot}5 - 3\sqrt{3{\cdot}5} = 3{\cdot}5 - 5{\cdot}612 = -2{\cdot}112 = 0$$



The above chart shows that all the plotted points are within the two control limits. This suggests that the process is in control.

Example 15. **The number of defects of 20 items are gvien below :**

Item No.	1	2	3	4	5	6	7	8	9	10
No. of defects	2	0	4	1	0	8	0	1	2	0
Item No.	11	12	13	14	15	16	17	18	19	20
No. of defects	6	0	2	1	0	3	2	1	0	2

Devise a suitable control chart and draw your conclusion.

Solution. As the number of defects per unit is given, the suitable control chart is C-chart

We have $N = 20$, and C = No. of defects = 35

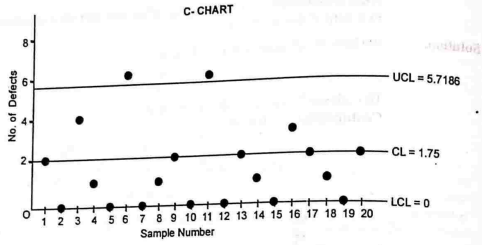
$$\bar{C} = \frac{\Sigma C}{N} = \frac{35}{20} = 1{\cdot}75$$

The value of $\bar{C}$ represents the central line.

Control Limits for *C*-chart

$$UCL = \bar{C} + 3\sqrt{\bar{C}} = 1{\cdot}75 + 3\sqrt{1{\cdot}75} = 5{\cdot}7186$$

$$LCL = \bar{C} - 3\sqrt{\bar{C}} = 1{\cdot}75 - 3\sqrt{1{\cdot}75} = -2{\cdot}218 = 0$$



The abvoe chart shows that although out of 20 plotted points 18 points are within the control limits but the points for sample 6 and sample 11 are outside the UCL. This suggests thet the process is not control.

EXERCISE – 2

1. Calculate the control lines for *p*-chart from the following information derived from inspection of 10 samples selected from the production process of an electric bulbs manufacturing industry. State whether the production process is under control ?

Sample No.	No. of Units Inspected	No. of Defective Units
1	10	2
2	40	4
3	100	8
4	50	5
5	60	12
6	100	10
7	100	3
8	30	3
9	80	4
10	60	3

[**Ans.** $\bar{p} = 0{\cdot}086$ or 8·6%, UCL = 19·4%, LCL = 0, The process is not within control]

2. Each of 20 lots of rubber belts contained 2000 rubber belts. Number of defective rubber belts in those lots are 410, 420, 324, 332, 292, 310, 282, 300, 320, 296, 392, 432, 294, 324, 220, 400, 258, 226, 460, 280.
Calculate control limits for fraction defective chart and give your conclusions.
[**Ans.** UCL = 0·1891; LCL = 0·1395]

3. If the average fraction defective of a large sample of products is 0·1537, caluclate the control limits for fraction defectives (Given that the size of each sample is 200)
[**Ans.** UCL = 0·17738, LCL = 0·12952 or 0]

4. An inspection of 9 samples of size 100 each from 9 lots reveal the following number of defective units :

Sample No. : (each of 100 items)	1	2	3	4	5	6	7	8	9
No. of defectives :	12	7	9	8	10	6	7	11	8

Construct a suitable cotnrol chart and give your conclusion.
[**Hint :** Use *np*-chart]
[**Ans.** CL = 8·66, UCL = 9·38, LCL = 7·94]

5. In a manufacturing concern of radio production lot of 250 items are inspected at a time. 20 samples taken are different in trades and defectives noted are given below. Draw a suitable control chart.

Lot No. (each of 250 items) :	1	2	3	4	5	6	7	8	9	10
No. of defectives :	25	47	23	36	24	34	39	32	35	22
Lot No. :	11	12	13	14	15	16	17	18	19	20
No. of defectives : (each of 250 Items)	45	40	32	35	21	40	15	28	23	42

[**Ans.** CL = = 31·9 UCL = 32·84, LCL = 16·16]

6. An inspection of 10 samples of size 100 each revealed the following number of defective units :

2, 1, 1, 3, 2, 3, 4, 2, 2, 0

Calculate control limits for the number of defective units. Plot the control limits and the observations and state whether the process is under control or not.
[**Ans.** *np*-chart, CL = 2, UCL = 6·2, LCL = 2·2]

7. In a certain sampling inspection, the number of defectives found in 21 samples of 100 each are given below :

5, 7, 9, 7, 8, 13, 8, 4, 8, 4, 3, 7, 7, 12, 15, 5, 13, 4, 3, 10, 8.

Does these indicate that the quality characteristics under inspection is under statistical control ? [**Ans.** UCL = 16·48, LCL = −0·48 = 0]

8. During an examination of equal length of cloth, the following number of defects are observed :
2, 3, 4, 0, 5, 6, 7, 4, 3, 2. Draw a control chart for the number of defects and comment whether the process is under control or not. [**Ans.** CL = $\bar{C}$ = 3·6, UCL = 9·292, LCL = 0]

9. A plant produces rolls of paper. The number of defects disclosed by the Inspection of 20 rolls are as follows :
12, 6, 18, 4, 5, 2, 4, 7, 12, 14, 8, 11, 14, 21, 21, 10, 12, 9, 13, 10. Comment on the state of control using C-Chart. [**Ans.** $\bar{C}$ = 10·65, UCL = 20·44, LCL = 0·86]

10. A manufacturer of transistors found the following number of defectives in 20 subgroups of 50 transistors :

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

Construct a control chart for the number of defectives.
[**Ans.** CL = 5, UCL = 11·708, LCL = 0]

11. The average number of defects per automobile battery at final inspection of twenty batteries is six per factory. Find the control limits for the *C*-chart.
[**Ans.** *C*-chart, UCL = 13·35, LCL = 0]

ACCEPTANCE SAMPLING

Another major area of statistical quality control is product control or acceptance sampling. Product control is concerned with the inspection of manufactured products. The items are inspected to know whether to accept a lot of items conforming to standards of quality or to reject a lot as non-conforming. Here the decision is arrived through sampling. That is why product control is called acceptance sampling. According to **Simpson and Kafka "Acceptance sampling is concerned with the decision to accept a mass of non-conforming to quality. The decision is reached through sampling."**

RISKS IN ACCEPTANCE SAMPLING OR PRODUCT CONTROL

There are following two types of risks in acceptance sampling or product quality control :

(i) Producer's Risk

(ii) Consumer's Risk

Let us discuss them briefly :

(i) Producer's Risk : Sometimes it happens that inspite of good quality, the sample taken may show defective units as such the lot will be rejected. Inspite of good quality the lot is rejected, such a type of risk of rejection is known as producer's risk. In other words, the probability of rejecting a lot which has actually been found satisfactory by the producer according to accptable quality level is known as producer's risk. Thus, the risk of rejecting a lot of good items is known as producer's risk.

(ii) Consumer's Risk : Sometimes it may happen that the quality of the lot is not good but the sample results show good quality units as such the consumer has to accept a defective lot. Such a risk is known as **consumer's risk.** In other words, the probability of accepting a lot which has actually been satisfactory by the consumer according to a pre-determined standard is known as consumer's risk. Thus, the risk of accepting a lot of bad items is known as consumer's risk.

The consumer and producer both decide the acceptance standard of the lot. This is as known of Acceptable Quality Level (AQL) or Lot Tolerance Percentage Defective (LTPD).

How to Conduct Acceptance Sampling ?

OR

Types of Sampling Inspection Plans

Acceptance sampling is based on sampling. After the inspection of samples, the decision is made about the acceptance or rejection of a lot. In acceptance sampling, the number of samples and their order plays a significant role. To frame the rules for acceptance or rejection of a lot acceptance sampling plan is prepared.

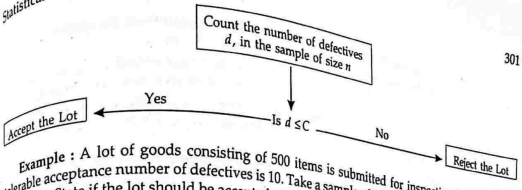
The following three types of sampling plan are frequently used in acceptance sampling :

(1) Single Sampling Plan

(2) Double Sampling Plan and

(3) Multiple or Sequential Sampling Plan

(1) Single Sampling Plan : Under single sampling plan, a sample of n items is first chosen at random from a lot of N items. If the sample contains , say, c or few defectives, the lot is accepted, while if it contains more than c defectives, the lot is rejected (c is known as 'acceptance number'). The single sampling plan is shown in the following chart :



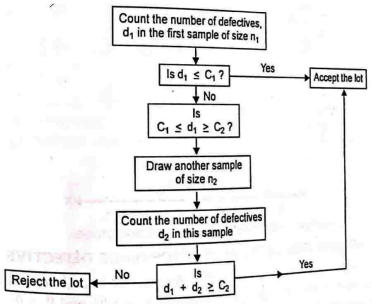
Example : A lot of goods consisting of 500 items is submitted for inspection for which the tolerable acceptance number of defectives is 10. Take a sample of 30 items you find that there are 8 defectives. State if the lot should be accepted or rejected for marketing.

Solution : We have c *i.e.*, acceptance number = 10 and d *i.e.*, the number of defective observed in the sample = 8.

Thus, $d < c$ (*i.e.*, $8 < 10$)

Since, $d < c$, the lot under consideration should be accepted.

(2) Double Sampling Plan : Under this sampling plan, a sample of n_1 items is first chosen at random from the lot of size N. If the sample contains, say, c_1 or few defectives, the lot is accepted; if it contains more than c_2 defectives, the lot is rejected. If however, the number of defectives in the sample exceeds c_1, but is not more than c_2, a second sample of n_2 items is taken from the same lot. If now, the total number of defectives in the two samples together does not exceed c_2, the lot is accepted; otherwise, it is rejected. (c_1 is known as acceptance number for the first sample and c_2, the acceptance number for both the samples taken together). The Double sampling plan is shown in the following chart :



Example : **A double sampling plan has the following facts as specified :**

$$N = 5{,}000,\ n_1 = 50,\ c_1 = 4,\ n_2 = 100 \text{ and } c_2 = 6$$ **Execute the plan.**

Solution. Execution of the double sampling plan involves the following steps :

(i) Inspect all the 50 items of the first sample after taking the same at random from the lot of 5,000 items.

(ii) Accept the lot the number of defectives observed from the sample (d_1) is less than or equal to 4 (*i.e.* c_1). if $d_1 > 6$ (*i.e.* c_2) reject the lot.

(iii) If the number of defectives in the sample thus observed *i.e.* d_1 is more than 4 (c_1) but not more than 6 (c_2), inspect all the 100 items of the second sample.

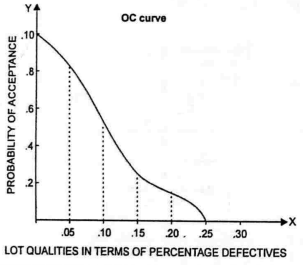
(iv) If now the total number of defectives ($d_1 + d_2$) observed in the combined sample of 150 items ($n_1 + n_2$) is less than 6 (*i.e.*, c_2), accept the lot. If it exceeds 6, than reject the lot.

(3) Multiple or Sequential Sampling Plan : Under this sampling plan, a decision to accept or reject a lot is taken after inspecting more than two samples of small size each. In this plan, units are examined one at a time and after examining each unit decision is taken. However, such plan are very complicated and hence rarely used in practice.

OPERATING CHARACTERISTIC CURVE OF AN ACCEPTANCE SAMPLING PLAN

This is a graphic measure of assessing the ability of a sampling plan in distinguishing between good and bad items. It depicts the relationship between the probability of acceptance of a lot Pα(p) for different lot quality expressed in terms of percentage defectives. In the construction of OC curve we take p *i.e.* lot qualities in terms of percentage defectives along the x-axis and p_2 (p) *i.e.*, probability of acceptance of a lot along the y-axis.

There is always an operating characteristic above (OC Curve) corresponding to any given sampling plan. A typical OC Curve has the following shape :



LOT QUALITIES IN TERMS OF PERCENTAGE DEFECTIVE

In the above figure, it has been assumed that the acceptable and rejectable qualities are measured as proportion of items that are defective and are $p_\alpha = 0{\cdot}05$; and $P_r = 0{\cdot}15$. From the OC curve, it must be seen that the probability of acceptance of a lot of the quality 0·05 is little less than 0·9, and the probability of rejection of a lot of the quality 0·15 is little more than 0·1. This shows that the chance of rejection of good products, which is producer's risk is little more than 0·1 and the chance of acceptance of bad products, which is the consumer's risk is little more than 0·1. Thus, the risks of both producer and consumer are more or less same.

The steepness of the curve depends upon the sample size. The larger the sample, the steeper the OC curve. The position of OC curve is determined by the maximum number of defective items allowable for acceptance, called the acceptance number. If the curve is shifted to left or right according as the acceptance number is made smaller or larger.

The following example illustrate the procedure of constructing an OC curve :

Example 16. **From the following data relating to a single sampling plan, determine the probability of acceptance at 0·5%, 7·5%, 1%, 2%, 5%, 10% and 15% defectives in the lot quality and fit a OC Curve to represent the data:**

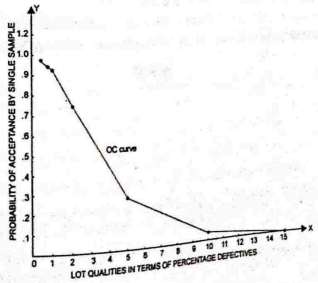
$N = 1,000$, $n = 50$, $C = 1$

Solution. It is case of single sampling plan. Since the number of tolerable defective or $c = 1$, the lot will be accepted if the sample gives 0 or 1 defective.

Computation of Cumulative Probabilities using Poisson Distribution

% defective in lot	Mean defective (m)	$P(0) = e^{-m}$	$P(1) = m.e^{-m}$	$P\alpha(p) = P(0) + P(1)$
0·50	$\frac{50}{100} \times 0{\cdot}5 = 0{\cdot}25$	0·7788	0·1947	0·9735
0·75	$\frac{50}{100} \times 0{\cdot}75 = 0{\cdot}38$	0·6839	0·2599	0·9438
1·00	$\frac{50}{100} \times 1{\cdot} = 0{\cdot}50$	0·6065	0·3033	0·9098
2·00	$\frac{50}{100} \times 2{\cdot}0 = 1{\cdot}00$	0·3678	0·3678	0·7358
5·00	$\frac{50}{100} \times 5 = 2{\cdot}50$	0·0821	0·2053	0·2874
10·00	$\frac{50}{100} \times 10 = 5{\cdot}00$	0·0070	0·0350	0·0420
15·00	$\frac{50}{100} \times 15 = 7{\cdot}50$	0·0006	0·0045	0·0051

Now we represent the probabilities of the acceptance of the lot with the given percentage defectives by a single sampling plan. This is drawn below :



The above OC curve indicates that out of the 1,000 items (Since $N = 1000$) inspected 974 (·9735 × 1000) items will be accepted and 26 items rejected with 0·5% defectives.

QUESTIONS

1. What is statistical quality control? Explain its merits and limitations.

   OR

   What is statistical quality control ? Explain it utility in industry.
2. Explain the followings:
   (a) Causes of variations in quality characteristics.
   (b) Purpose and logic of control charts.
3. What are control charts? Explain the purpose and logic of control charts.
4. Explain the basic concept and logic of construction of control charts.
5. Describe different types of control charts.
6. Discuss the basic principles uderlying control charts. Explain in brief the construction and uses of P-chart and c -chart.
7. Distinguish between process control and product control. How are control charts used in process control ?
8. Explain construction and uses of $\bar{X}$ - chart and R- chart.
9. Explain the meaning, purpose and types of various accepting sampling plans.
10. What is acceptance sampling plan ? Outline the procedure of single and double sampling plans ?
11. Explain the meaning and utility of OC curve.
12. State and discuss the significance of consumer's risk and producer's risk in statistical quality control.
13. Explain :
    (i) Purpose and logic of control charts.
    (ii) Causes of variations in quality characteristics.
    (iii) Consumer's risk and Producer's risk in acceptance sampling.
    (iv) Operating characteristic curve of an acceptance sampling plan.

❀❀❀

ASSOCIATION OF ATTRIBUTES

11

1. INTRODUCTION

In the chapter on correlation, we have studied the relationship between two such phenomena which are capable of direct quantitative measurement like weight, height, income etc. These are called statistics of variables. But all phenomena cannot be measured directly. There are certain phenomena like beauty, deafness, blindness, colour etc. which are qualitative or descriptive in nature and cannot be quantitative measured. These are called statistics of attributes.. Literally, an attribute means a quality or characteristics of the object under study. Theory of attributes deals with statistics of attributes. An attribute is observed by the persence or absence of some qualitative characteristic.

2. MEANING AND DEFINITION

Statistics of attributes are classified on the basis of attributes such as distribution of population into males and females, married and unmarried, educated and uneducated etc. The method of association employed to know the relationship between the two attributes. For example, if we want to investigate whether there is any association between the eyes colour of fathers and sons, literacy and criminality, we use the technique of association. Thus, association refers to a technique by which we can measure the relationship between the two attributes. Like correlation, association is a measure which deals with attributes rather than the variables.

DEFINITION OF ASSOCIATION

1. "Association measures the relationship between two such phenomena whose size cannot be measured". - **Walls and Roberts.**
2. "Association studies the nature of relationship between statics fo attributes". - **J.F. Kenny and E.S. Keeping**

Thus, association measures the relationship between two attributes.

3. DIFFERENCE BETWEEN CORRELATION AND ASSOCIATION

The main difference between correlation and association is as follows:

(1) Correlation is used to measure the relation between two variables whereas association is used to measure the relation between two attributes.

(2) In correlation, universe is classified in quantitative terms whereas in association of attributes, universe is classified on the basis of presence or absence of an attribute.

(3) Correlation is the analysis of the covariation between the two variables. For the association of attributes, the presence of the two attributes together is not only sufficient but they must appear together in a greater number of cases that is to be expected.

(4) The methods of finding correlation and association are different. It is relatively easy to determine association than correlation.

4. USE OF TERMS AND NOTATIONS

The following terms and notations are used in the study of association between the attributes:

(1) Positive and Negative Attributes: Presence of an attribute is known as positive attribute and absence of an attribute is known as negative attribute. Generally capital letters A,B,C etc. are used to denote the presence of an attribute and small letters a,b,c or Greek letters α (Alpha), β(Beta) etc. are used to denote the absence of an attribute. For example, if 'A' denote literate, then 'α' would denote illiterate. Similarly, if 'B' denote males, then 'β' would denote females.

(2) Combination of Attributes: When two attributes are studied together, we obtain their different group, which are called combination of attributes. In order to denote the combinations of different attributes the symbols related to these are written together. Thus, if A stands for males and B for graduates, then the combinations formed by them will be as under:

AB = Graduate males .

Aβ = Non-graduate males

αB = Graduate males

αβ = Non-graduate females.

(3) Class frequency: The number of observations falling in each class is called its class frequency and is denoted by enclosing the corresponding class symbol in brackets like (A), (α), (B), (Aβ) etc. Here (AB) represents the number of observations possessing both the attributes A and B simultaneously.

(4) Classes: Classes are of three types — (1) Positive classes, (2) Negative classes, and (3) Contrary classes. Classes expressed in capital letters represent presence of attribute and they are known as positive classes e.g. A, B, AB etc. Classes expressed in small letters or Greek letters represent absence of attribute e.g., α ,β,αβ etc. and they are known as negative classes. Classes formed by the combination of capital letters and small letters represent presence of one attribute and absence of another attribute, they are known as contrary classes e.g. $A\beta, \alpha B$ etc.

(5) Number of Classes: The total number of classes can be obtained by using the following formula:

No. of classes $= 3^n$ (where n stands for number of attributes)

If there is only one attributes under study, the total number of classes would be $3^1 = 3$ (N, A and α) and if there are two attributes the total number of classes would be $3^2 = 9$. These are: N, A, α, $B, \beta, AB, A\beta, \alpha B, \alpha\beta$.

(6) Order of Classes and Class Frequencies: On the basis of number of attributes, classes and class frequencies can be written according to different orders, such as 1st order, 2nd order, 3rd order etc. A class having only one attribute such as A, α, B, β are known as class of first order and classes having two attributes are known as classes of second order such as $AB, A\beta, \alpha B$ and $\alpha\beta$. The frequencies of the first order, second order and third order classes are called class frequencies of the 1st, 2nd and 3rd order. Class frequencies of the various order in case of two attributes are as follows:

N	Frequency of Zero Order
(A) (B)	Frequencies of First Order
(α) (β)	
(AB) (αB)	Frequencies of Second Order
$(A\beta)$ $(\alpha\beta)$	

(7) Ultimate Class Frequencies: If there are only two attributes, then the classes of the second order and is case of three attributes, classes of the third order are known as ultimate classes and the corresponding frequencies of such ultimate classes are called ultimate class frequencies.

To total number of ultimate classes can be obtained by using the formula:

No. of Ultimate Classes $= 2^n$ (where n stands for the number of attributes)

If the number of attributes are 2, then the number of ultimate classes would be $2^2 = 4$. The ultimate class frequency would be - $(AB), (A\beta), (\alpha B)$ *and* $(\alpha\beta)$.

5. DETERMINATION OF UNKNOWN CLASS FREQUENCIES

If we known some of the class frequencies, we can easily found out the frequencies of the remaining classes. In case of two attributes, the unknown class frequencies can be known from the following table which is known as Nine-square table (since nine squares are formed) or 2 × 2 association table.

	A	α	Total
B	(AB)	(αB)	(B)
β	(Aβ)	(αβ)	(β)
Total	(A)	(α)	N

From this table, certain relationships can be described.

Columnwise (Vertical):

$(AB) + (A\beta) = (A)$

$(\alpha B) + (\alpha\beta) = (\alpha)$

$B = (B) + (\beta) = N$

Row-wise (Horizontally) as well:

$(AB)+(\alpha B)=(B)$

$(A\beta)+(\alpha\beta)=(\alpha)$

$(A)+(\alpha)=N$

$N=(AB)+(A\beta)+(\alpha B)+(\alpha\beta)$

From these relationship, if we know any of the ultimate class frequencies and any other three values, we can find the frequencies of the remaining classes.

The following points are to be kept mind while determining the unknown class frequencies.

(i) Class frequencies of the first order are obtained by adding the class frequencies of the second order:

$(A) = (AB)+(A\beta) \Rightarrow (AB)=(A)-(A\beta)$

$(A\beta) = (A)-(AB)$

$(B) = (AB)+(\alpha B) \Rightarrow (AB)=(B)-(\alpha B)$

$(\alpha B) = (\beta)-(AB)$

$(\alpha) = (\alpha B)+(\alpha\beta) \Rightarrow (\alpha B)=(\alpha)-(\alpha\beta)$

$(\alpha\beta) = (\alpha)-(\alpha B)$

$(\beta) = (A\beta)+(\alpha\beta) \Rightarrow (A\beta)-(\alpha\beta)$

$(\alpha\beta) = (A\beta)-(\beta)$

(ii) If the ultimate class frequencies are known, the frequencies of the positive classes and negative classes can be obtained as follows:

$(A)=(AB)+(A\beta)$; $(\alpha)=(\alpha B)+(\alpha\beta)$

$(B)=(AB)+(\alpha B)$; $(\beta)=(A\beta)+(\alpha\beta)$

(iii) The sum of all ultimate class frequencies is always equal to the total number of observations. This,

$N=(AB)+(\alpha B)+(A\beta)+(\alpha\beta)$

The following examples illustrate the determination of unknown class frequencies.

Example 1. Find the missing frequencies from the following data:

(AB) = 100, (A) = 300, (B) = 600, N = 1000

Solution: Putting the given values in the nine-square table as follows:

	A	α	Total
B	(AB) 100	(αB) ?	(B) 600
β	$(A\beta)$?	$(\alpha\beta)$?	(β) ?
Total	(A) 300	(α) ?	N 1000

The missing frequencies $(A\beta)$, $(\alpha\beta)$ and (β) are calculated as follows:

$(A\beta)=(A)-(AB)=300-100=200$

$(\alpha)=N-(A)=1000-300=700$

$(\beta)=N-(B)=1000-600=400$

$(\alpha B)=(B)-(AB)=600-100=500$

$(\alpha\beta)=(\beta)-(A\beta)=400-200=200$

Thus, the missing frequencies are:

$(A\beta)=200, (\alpha B)=500, (\alpha\beta)=400, (\alpha)=700, (\beta)=400$

Example 2. Finding the missing frequencies from the following data:

$(\alpha\beta)=500, (B)=600, (\alpha)=500, (\beta)=1{,}000$

Solution: Putting the given data in the nine square table as follows:

	A	α	Total
B	(AB) ?	(α) ?	(B) 600
β	$(A\beta)$?	$(\alpha\beta)$ 500	(β) 1,000
Total	(A) ?	(α) 500	N ?

The missing frequencies are calculated as:

$N=(B)+(\beta)=600+1000=1600$

$(A) = N-(\alpha)=1600-500=1100$

$(A\beta) = (\beta)-(\alpha\beta)=1000-500=500$

$(AB) = (A)-(A\beta)=1100-500=600$

$(\alpha B) = (B)-(AB)=600-600=0$

Example 3. From the following ultimate class frequencies, find the frequencies of the positive and negative classes and the 'N':

$(AB)=20, (\alpha B)=80, (A\beta)=140, (\alpha\beta)=160$

Solution: Putting the given data in the nine square table as follows:

	A	α	Total
B	(AB) 20	(αB) 80	(B) ?
β	$(A\beta)$ 140	$(\alpha\beta)$ 160	(β) ?
Total	(A) ?	(α) ?	N ?

Missing frequencies of positive classes {(A), (B)} and negative classes {(α), (β)} and N are computed as follows:

Positive Classes: $(A) = (AB)+(A\beta)=20+140=160$

$(B) = (AB)+(\alpha B)=20+80=100$

Negative Classes: $(\alpha) = (\alpha B)+(\alpha\beta)=80+160=240$

$(\beta) = (A\beta)+(\alpha\beta)=140+160=300$

$N = (A)+(\alpha)=160+240=400$

or $N = (B)+(\beta)=100+300=400$

or $N = (AB)+(\alpha B)+(A\beta)+(\alpha\beta)$

$= 20+80+140+160=400$

Example 4. Find the frequencies of the ultimate classes if N = 1000, (A) = 300, (B) = 600 and (AB) = 100

Solution: No. of Ultimate Classes $= 2^n = 2^2 = 4$

These ultimate classes are: AB, $A\beta$, α B, $\alpha\beta$. The frequency of the ultimate class AB is known, the other frequencies are calculated as follows:

	A	α	Total
B	(AB) 100	(αB) ?	(B) 600
β	$(A\beta)$?	$(\alpha\beta)$?	(β) 400
Total	(A) 300	(α) 700	N 1000

$(\alpha B) = (B)-(AB)=600-100=500$

$(A\beta) = (A)-(AB)=300-100=200$

$(\alpha\beta) = (\alpha)-(\alpha\beta)=700-500=200$

Exercise 1

1. Find the missing frequencies from the following data:
   $(\alpha B) = 300, (B) = 600, (\alpha) = 800, (\beta) = 1000$
   [Ans. $(AB) = 100, (\alpha B) = 500, (A\beta) = 700, (A) = 800, N = 1600$]

2. Given the following frequencies of the positive order, find out the frequencies of the ultimate classes
   $N = 100$, $(A) = 70$, (B) = 40 and (AB) = 30
   [Ans. $(\alpha B) = 10, (A\beta) = 40, (\alpha\beta) = 20, (\beta) = 60, (\alpha) = 30$]

3. From the following ultimate class frequencies, find the frequencies of positive and negative classes and the 'N'.
   $(AB) = 160$, $(A\beta) = 80$ $(\alpha B) = 120, (\alpha\beta) = 40$
   [Ans. $(A) = 240, (\alpha) = 160, (B) = 280, (\beta) = 120, N = 400$]

4. From the following data, find out the missing frequencies:
   $(AB) = 35$, $(\alpha B) = 325$, $(A) = 383$, $N = 1500$
   [Ans. $(B) = 360, (A\beta) = 348, (\alpha\beta) = 729, (\beta) = 1140, (\alpha) = 1117$]

5. Find the frequencies of the ultimate classes if
   $N = 300$, $(A) = 100$, $(B) = 120$ and $(AB) = 40$
   [Ans. $(A\beta) = 60, (\alpha B) = 80, (\alpha\beta) = 120, (\alpha) = 200$]

6. In an examination at which 600 students appeared, boys outnumbered girls by 16% of all the candidates. The number of passed candidates exceeded those of the failed candidates by 310. Boys failing in the examination numbered 88. Construct a nine square table and find the unknown class frequencies.
   [Ans. No. of boys passed = 260; Girls passed = 195; Girls failed = 57]

6. CONSISTENCY OF DATA

In order to find out whether the given data is consistent or not, we apply a very simple test. The test is to find out whether any one or more of the ultimate class frequencies is negative or not. If none of the ultimate class frequencies is negative, we can conclude that the given data are consistent. On the other hand, if any of the ultimate class frequencies comes out to be negative, the given data are inconsistent. The necessary and sufficient condition for the consistency of the data is that no ultimate class frequencies is negative.

Procedure for Testing the Consistency of Data: In order to test the consistency of the data, the following procedure is adopted:

(i) Put the given data in the form of a nine square table and obtain the unknown class frequencies. It should be remembered that there are four ultimate class frequencies in case of two attributes such as (AB), (A β), (A β), (α β).

(ii) If any of the ultimate class frequencies such as (A B), (A β), (α B) and (α β) is negativ, then the given data are called inconsistent otherwise consistent.

The following examples clarify the consistency of data.

Example 5. From the following data find out whether the data are consistent or not:

$(A)=200,(B)=300,(AB)=285,N=1000$

Solution: Put the given values in the nine-square table, and then we can find the missing ultimate class frequencies such as $(A\beta),(\alpha B),(\alpha\beta)$

	A	α	Total
B	(AB) 285	(αB) 300 – 285 = 15	(B) 300
β	$(A\beta)$ 200 – 285 = –85	$(\alpha\beta)$ 700 –(–85) = 785	(β) 1000–300 = 700
Total	(A) 200	(α) 800	N 1000

From the table, the ultimate class frequencies are:

$(AB)=285,(A\beta)=-85,(\alpha B)=15,(\alpha\beta)=785$

Since one of the ultimate class frequencies is negative i.e. $(A\beta)=-85$, the given data are inconsistent.

Example 6. Examine the consistency of the following data:

$N=500,(\alpha B)=90,(A\beta)=40,(\alpha\beta)=310$

Solution: Put the values in the nine-square table, and then we can find the missing frequencies.

	B	β	Total
A	(AB) 100–40 = 60	$(A\beta)$ = 40	(A) 500–400 = 100
α	(αB) = 90	$(\alpha\beta)$ = 310	(α) = 400
Total	(β) = 150	(β) = 350	N = 500

Since all the ultimate class frequencies i.e. $(AB),(A\beta),(\alpha B),(\alpha\beta)$ are positive, the given data are consistent.

Example 7. Find ultimate class frequencies and test for consistency from the following data:

$N=100,(A)=76,(B)=60,(AB)=15$

Solution: Putting the given values in the nine-square table:

	A	α	Total
B	(AB) = 15	(αB) = 45	(B) 60
β	$(A\beta)$ = 61	$(\alpha\beta)$ = –21	(β) 40
Total	(A) = 76	(α) = 24	N = 100

The ultimate class frequencies are: (AB) = (given)

$(\alpha B)=(B)-(AB)=60-15=45$

$(A\beta)=(A)-(AB)=76-15=61$

$(\alpha\beta)=(\alpha)-(\alpha B)=24-45=-21$

Since one of the ultimate class frequencies is negative i.e. $(\alpha\beta)=-21$, the given data are inconsistent.

Example 8. In a report on consumer's preference, it was revealed that out of 500 persons surveyed, 410 preferred tea, 380 preferred coffee and 270 persons liked both. Are the data consistent?

Solution: Let A denote preference for Tea

B denote preference for coffee

Thus, the given data is:

$N=500,(A)=410,(B)=380,(AB)=270$

We can find the missing ultimate class frequencies by putting the values in the nine square table as follows.

	A	α	Total
B	(AB) 270	(αB) 110	(B) 380
β	$(A\beta)$ 140	$(\alpha\beta)$ –20	(β) 120
Total	(A) 410	(α) 90	N 500

Since one of the ultimate class frequency i.e. $(\alpha\beta)=-20$, so the data are inconsistent.

Exercise 2

1. Test the consistency of the following data:
$N=1000,(A)=600,(B)=500,(AB)=50$

[Ans. $(\alpha\beta)=-50$, Inconsistent]

2. Test the consistency of the following data:
$(\alpha B)=100,(A\beta)=125,(\alpha\beta)=80,N=300$

[Ans. (AB) = −5, Inconsistent]

3. 1,000 persons are living in a locality. The number of educated persons is 750 and that of unemployed persons is 400. Of the unemployed persons, 410 are educated. Is there any inconsistency in the information ?

[Ans. $(\alpha\beta)=-10$, Inconsistent]

4. An enquiry of 30 persons was conducted regrading their food habits. It was found that 25 of them were vegetarians and 20 of them liked boiled vegetables. Another 10 were vegetarians and liking boiled vegetables. Show that the data are inconsistent.

[Ans. $(\alpha\beta)=-5$, Inconsistent]

5. In a report on consumer's preference, it was given that out of 500 persons surveyed, 410 preferred Coca Cola, 380 preferred Pepsi Cola and 270 persons liked both. Are the data consistent?

[Ans. $(\alpha\beta)=-20$, Inconsistent]

6. Find ultimate class frequencies and test for consistency of the data:
$(A)=40,(B)=60,(AB)=30,N=130$

[Ans. $(\alpha B)=30,(A\beta)=10,(\alpha\beta)=60$, consistent]

7. ASSOCIATION OF ATTRIBUTES

Generally, when one attribute appear in a number of cases along with the other attribute, then we find mutual association between them. But in statistics, it has a special meaning. In statistics, two attributes are said to be associated when both the attribute are more commonly found together than is ordinarily expected. In the words of *Yule and Kendel, "In statistics A and B are associated only if they together in a greater number of cases then is to be expected if they are independent"*. Evidently A and B are disassociated if this number is less than expected for independence.

7.1 KINDS OF ASSOCIATION

Association of attributes can be positive, negative or independent. It may be of the following three forms:

(1) Positive Association: When two attributes are found to be present or absent together, they are said to be **positively Associated or merely Associated.** Such association is found between smoking and cancer, illiteracy and criminality, education and unemployment etc. In such a situation the observed frequency is found to be more than the expected frequency. Symbolically:

$$(AB) > \frac{(A)(B)}{N}$$

(2) Disassociation or Negative Association: When presence of one attribute is associated with the absence of other attribute, they are said to negative associated. Negative association may also be termed as disassociation. Negative Association is found between literacy and criminality, vaccination and attack of small pox, education and dishonesty etc. In such a situation, the observed frequency is found to be less then the expected frequency. Symbolically:

$$(AB) < \frac{(A)(B)}{N}$$

(3) Independence: When the two attributes have not a tendency of being present together or not that a tendency of one attribute being absent when another is present, then they are said to be independent of each other. In such a situation, the observed frequency is equal to be expected frequency. Symbolically:

$$(AB) = \frac{(A)(B)}{N}$$

7.2 METHODS OF DETERMINING ASSOCIATION

Following are the main methods for determining the association between the two attributes:

(1) **Frequency Method**
(2) **Proportion Method.**
(3) **Yule's Coefficient of Association.**
(4) **Coefficient of Colligating.**
(5) **Coefficient of Contingency.**

Let us discuss them in detail

(1) Frequency Method

The method is also called **comparison of observed and expected frequency method.** Under this method, we determine the nature of association between the two attributes by comparing the observed frequency with expected frequency. If A and B are two attributes, we compare the observed frequency of AB with the expected frequency of AB. Expected frequency of AB is calculated by using the following formula:

$$E(AB) = \frac{(A).(B)}{N}$$

When N = Total number of observations

(A) = Frequency of attribute A

(B) = Frequency of attribute B

Symbolically, two attributes A and B are:

(i) **Independent** if O (AB) = i.e. (AB) $= \frac{(A).(B)}{N}$

(ii) **Positively associated** if O(AB) > E(AB) i.e. (AB) > $\frac{(A).(B)}{N}$

(iii) **Negatively associated** if O (AB) < E(AB) i.e. (AB) < $\frac{(A).(B)}{N}$

Similarly, we can determine the nature of association for A and β, α and B and α and β by comparing the observed frequency with the expected frequency.

The frequency method can be summarised in the table below:

Attributes	Independent	Positive Association	Negative Association
A and B	$(AB)=\frac{(A)\times(B)}{N}$	$(AB)>\frac{(A)\times(B)}{N}$	$(AB)<\frac{(A)\times(B)}{N}$
A and β	$(A\beta)=\frac{(A)\times(\beta)}{N}$	$(A\beta)>\frac{(A)\times(\beta)}{N}$	$(A\beta)<\frac{(A)\times(\beta)}{N}$
α and β	$(\alpha B)=\frac{(\alpha)\times(B)}{N}$	$(\alpha B)>\frac{(\alpha)\times(B)}{N}$	$(\alpha B)<\frac{(\alpha)\times(B)}{N}$
α and β	$(\alpha\beta)=\frac{(\alpha)\times(\beta)}{N}$	$(\alpha\beta)>\frac{(\alpha)\times(\beta)}{N}$	$(\alpha\beta)<\frac{(\alpha)\times(\beta)}{N}$

Example 9. Find if A and B are independent, positively associated or negatively associated from the data given below:

$$(A)=470,(B)=620,(AB)=320,N=1000$$

Solution: O (AB) = 320 (Given)

$$E(AB)=\frac{(A).(B)}{N}$$

$$=\frac{470\times620}{1000}=291.4$$

Since the observed frequency of AB (320) is more than the expected frequency of AB (291.4), attribute A and B are positively associated.

Example 10. Given (A) = 80, (B) = 60, (AB) = 40, N = 200.

Study the association between A and B; α and β; A and β and α and B.

Solution: Putting the given values in a nine square table:

	A	α	Total
B	(AB) = 40	(αB) = 20	(B) = 60
β	(Aβ) = 40	(αβ) = 100	(β) = 140
Total	(A) = 80	(α) = 120	N = 200

We can calculate the expected frequencies as follows.

(i) $E(AB) = \frac{(A).(B)}{N}=\frac{80\times60}{200}=24$

$O(AB)$ = 40 (Given)

$\because$ $O(AB) > E(AB)$ i,e., 40 > 24

$\therefore$ A and B are positively associated.

(ii) $E(\alpha\beta) = \frac{(\alpha).(\beta)}{N}=\frac{120\times80}{200}=84$

O(αβ) = 100(Given)

$\because$ O(αβ) > $E(\alpha\beta)$

$\therefore$ 100 > 84

$\therefore$ α and β are positively associated.

(iii) $E(A\beta) = \frac{(A).(B)}{N}=\frac{80\times140}{200}=56$

O (A β) = 40 (Given)

$\because$ $O(A\beta) < E(A\beta)$ i.e., 40 < 56

$\therefore$ A and β are negatively associated.

(iv) $E(\alpha\beta) = \frac{(\alpha).(B)}{N}=\frac{120\times60}{200}=36$

$\because$ O(α B) = 20(*Given*)

$O(\alpha B) < E(\alpha B)$ i.e., 20 < 36

$\therefore$ α and B are negatively associated.

Example 11. Out of total number of 900 members, 300 have A and 280 have B attributes, 180 persons have both A and B attributes. Are A and B independent?

Solution: Using frequency method, two attributes A and B are independent if

$$(AB) = \frac{(A)\times(B)}{N}$$

According to given data, (AB) = 180, (A) = 300, (B) = 280

and $\frac{(A)\times(B)}{N} = \frac{300\times 280}{900} = 93$ approx.

Thus $(AB) > \frac{(A).(B)}{N}$

Hence A and B are not independent but positively associated.

Limitation of Frequency Method

This method determines only the nature of association (i.e. whether there is positive or negative association or no association) and does not tell us anything about the degree of association (i.e.whether the association is high or low) between the attributes.

Exercise 3

1. Show that whether A and B are independent, positively associated or negatively associated in each of the following cases:
   (i) N = 300, (A) = 48, (B) = 100 and (AB) = 16
   (ii) N = 120, (A) = 24, (B) = 100 and (AB) = 52

   [Ans. (i) Independent (ii) Positively associated]

2. Find if A and B are independent, positively associated or negatively associated from the data given below:
   $(AB) = 256, (\alpha B) = 768, (A\beta) = 48, (\alpha\beta) = 144$

   [Ans. Independent]

3. From the following data, find out whether the attributes (i) A and B (ii) A and β (iii) α and B and (iv) α and β are independent, positively associated or negatively associated.
   N = 100, (A) = 40, (B) = 80, (AB) = 30

   [Ans. (i) –vely associated (ii) + vely associated (ii) + vely associated and (iv) –vely associated]

4. Given (A) = 12, (B) = 25, (AB) = 4, N = 75,
   Are A and B independent?

   [Ans. Independent]

(2) Proportion Method

Under this method, if A and B are two attributes, two proportions are calculated:

(i) Proportion of B's in A's $= \left[\frac{(AB)}{(A)}\right]$

(ii) Proportion of B's in α's $= \left[\frac{(\alpha B)}{(\alpha)}\right]$

These two proportions are compared to find out the association between A and B. Two attributes A and B are:

(i) **Independent** if $\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$

(ii) **Positively associated** if $\frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}$

(iii) **Negatively associated** if $\frac{(AB)}{(A)} < \frac{(\alpha B)}{(\alpha)}$

The association between A and B can also be found by comparing the proportion of A in B and β. The two proportions are calculated as:

Proportion of A's in $B = \left[\frac{(AB)}{(B)}\right]$

Proportion of A's in βs $= \left[\frac{(A\beta)}{(\beta)}\right]$

The same results hold good.

Similarly we can determine the nature of association for α and B by comparing the proportion of α in B and β.

The proportion method can be summarized in the table below:

Attributes	Independence	Positive Association	Negative Association
A in B and β	$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)}$	$\frac{(AB)}{(B)} > \frac{(A\beta)}{(\beta)}$	$\frac{(AB)}{(B)} < \frac{(A\beta)}{(\beta)}$
B in A and α	$\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$	$\frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}$	$\frac{(AB)}{(A)} < \frac{(\alpha B)}{(\alpha)}$
α in B and β	$\frac{(\alpha B)}{(B)} = \frac{(\alpha\beta)}{(\beta)}$	$\frac{(\alpha B)}{(B)} > \frac{(\alpha\beta)}{(\beta)}$	$\frac{(\alpha B)}{(B)} < \frac{(\alpha\beta)}{(\beta)}$
β in A and α	$\frac{(A\beta)}{(A)} = \frac{(\alpha\beta)}{(\beta)}$	$\frac{(A\beta)}{(A)} > \frac{(\alpha\beta)}{(\alpha)}$	$\frac{(A\beta)}{(A)} < \frac{(\alpha\beta)}{(\alpha)}$

Example 12. Show whether A and B are independent, positively associated or negatively associated by the method of proportions from the following data:

$(A) = 430, (\alpha) = 570$

$(AB) = 294, (\alpha B) = 380$

Solution: Given $(A) = 430, (\alpha) = 570$

$(AB) = 294, (\alpha B) = 380$

Using method of proportions,

Proportion of B in $A = \frac{(AB)}{(A)} = \frac{294}{430} = 0.68$ or 68%

Proportion of B in $\alpha = \frac{(\alpha B)}{(\alpha)} = \frac{380}{570} = 0.66$ or 66%

As $\frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}$

i.e. 68% > 66%

$\therefore$ A and B are positively associated.

Example 13. Out of 70,000 literates in a district, number of criminals were found to be 500. Out of 9,30,000 illiterates in the same district, the number of criminals was 15,000. On the basis of data do you find any association between illiteracy and criminality.

Solution: Let A denote illiteracy and as such α represents literacy. Let B denote criminality and as such β would denote non-criminality.

We are given

$(\alpha) = 70,000, (\alpha B) = 500, (A) = 9,30,000, (AB) = 15,000$

Using the proportion method,

Proportion of criminals (B) among illiterate (A) –

$\frac{(AB)}{(A)} = \frac{15,000}{9,30,000} = .016$ or 1.6%

Proportion of criminals (B) among literate (α) –

$\frac{(\alpha B)}{(\alpha)} = \frac{500}{70,000} = 0.0071$ or .71%

As $\frac{(AB)}{(A)} > \frac{(\alpha B)}{(\alpha)}$

i.e. 1.6% > .71%

Thus, there is a positive association between criminality and illiteracy.

Example 14. In a population of 500 students the number of married is 200. Out of 150 students who failed 60 belonged to the married group. It is required to find out whether the attributes marriage and failure are independent, positively associated or negatively associated.

Solution: Let A denote married students and as such α would denote unmarried Let B denote number of failures and as such β would denote non-failures. Putting the information in a nine-square table:

	A (Married)	α (Unmarried)	Total
B (Failure)	$(AB) = 60$	$(\alpha B) = 90$	$(B) = 150$
β (Non-Failure)	$(A\beta) = 140$	$(\alpha\beta) = 210$	$(\beta) = 350$
Total	$(A) = 200$	$(\alpha) = 300$	$N = 500$

Using the proportion method,

Percentage of failed students (B) among married (A)

$\frac{(AB)}{(A)} \times 100 = \frac{60}{200} \times 100 = 30\%$

Percentage of failed students (B) among unmarried (α)

$\frac{(\alpha\beta)}{(\alpha)} \times 100 = \frac{90}{300} \times 100 = 30\%$

Since the two percentages (or proportions) are same, we conclude that the attributes marriage and failure are independent.

Limitation of Proportion Method

Like frequency method, this method determines only the nature of association and does not tell us anything about the degree of association.

Exercise 4

1. In a certain study the following data were reported:

$(AB) = 216, (A\beta) = 25, N = 400, (B) = 300$

Determine the association between A and B by the method of proportions.

[Ans.: Positively associated]

2. Out of 3,000 unskilled workers of a factory, 2,000 come from rural areas and out of 1,200 skilled workers, 300 come from rural areas. Determine the association between skill and residence in rural areas by the method of proportions.

[Ans.: Negatively associated]

3. Out of 900 persons, 300 were literate and 400 had travelled beyond the limits of their district. Of the literate persons 200 were among those who travelled. Is there any relation between literacy and travelling?

[Ans.: Positively associated]

4. 200 candidates appeared for a competitive examination and 60 of them succeeded, 35 received special coaching and out of them 20 candidates succeeded. By using proportion method, discuss whether coaching is effective or not.

[Ans.: Special coaching is effective]

5. Out of 5 lakh literates in a particular district of India, no. of criminals was 2,000. Out of 50 lakh illiterates in the same district, no. of criminals was 80,000. On the basis of these figures, do you find any association between illiteracy and criminality?

[Ans.: Positive association]

(3) Yule's Coefficient of Association

Prof. Yule has propounded a coefficient of association to find out degree of association between two attributes. The coefficient of association gives us the direction and the degree of association between the two attributes. The coefficient of association is denoted by Q and is calculated by applying the following formula:

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

Where Q = Yule's coefficient of Association.

(AB), $(\alpha\beta)$, $(A\beta)$ and $(\alpha\beta)$ = Ultimate class frequencies.

Interpretation of Coefficient of Association

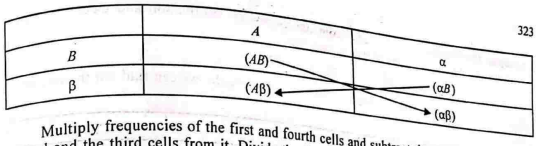
Like coefficient of correlation, the value of Q lies between -1 and $+1$. Its interpretation is as follows:

(i) If $Q = 0$, there is no association between the attributes i.e. they are independent.

(ii) If $Q = +1$, there is a perfect positive association

(iii) If $Q = -1$, there is a perfect negative association.

Remark: Yule's coefficient of association can be remembered with the help of the following table:



Multiply frequencies of the first and fourth cells and subtract the multiplication of the second and the third cells from it. Divide the figure so obtained by the addition of the two multiplications.

Example 15. Eighty-eight students residents of an Indian city, who were interviewed during a sample survey are classified below according to their smoking and tea drinking habits. Calculate Yule's coefficient of Association and comment on its value:

	Smokers	Non-smokers
Tea Drinkers	40	33
Non-Tea-Drinkers	3	12

Solution: Let A = Tea drinkers α = Non-tea drinkers

B = Smokers β = Non-smokers

Tabulating the given data:

	B	β
A	$(AB) = 40$	$(A\beta) = 33$
α	$(\alpha B) = 3$	$(\alpha\beta) = 12$

Applying Yule's Coefficient of Association:

$$Q = \frac{(AB)(\alpha\beta) - (\alpha B)(A\beta)}{(AB)(\alpha\beta) + (\alpha B)(A\beta)}$$

$$= \frac{40 \times 12 - 3 \times 33}{40 \times 12 + 3 \times 33}$$

$$= \frac{381}{579} = +0.65$$

Thus, there is positive association between tea drinkers and smokers.

Example 16. Prepare a 2 × 2 table from the following information and calculate Yule's coefficient of Association and interpret the result:

$N = 1500, (A) = 383, (B) = 360, (AB) = 35$

Solution: By putting the know values in the nine square table, we can find out the unknown values.

	A	α	Total
B	(AB) = 35	(αB) = 325	(B) = 360
β	$(A\beta)$ = 348	$(\alpha\beta)$ = 792	(β) = 1140
Total	(A) = 383	(α) = 1117	N = 1500

Applying Yule's Coefficient of Association:

$$Q = \frac{(35)(792)-(348)(325)}{(35)(792)+(348)(325)}$$

$$= \frac{27720-113100}{27720+113100} = \frac{-85380}{140820} = -0.606$$

Thus, there is negative association between A and B.

Example 17. Investigate the association between the temperament of brothers and sisters from the following data:

Good natured brothers and good natured sisters: 1040

Good natured brothers and sullen sisters: 160

Sullen brothers and good natured sisters: 180

Sullen brothers and sullen sisters: 120

Solution: Let A = good natured brothers; α = sullen brothers

B = good natured sisters; β = sullen sisters

Tabulating the given data in a nine square table:

	B	β
A	(AB) = 1040	$(A\beta)$ = 160
α	(αB) = 180	$(\alpha\beta)$ = 120

Applying Yule's method: $Q = \frac{(AB)(\alpha\beta)-(\alpha B)(A\beta)}{(AB)(\alpha\beta)+(\alpha B)(A\beta)}$

$$Q = \frac{1040\times120-160\times180}{1040\times120+160\times180}$$

$$= \frac{1,24,800-28,800}{1,24,800+28,800} = \frac{96000}{153600} = +0.625$$

Thus, there is positive association between temperament of brothers and sisters.

Example 18. Calculate Yule's coefficient of association between marriage and failure of students from the following data pertaining to 525 students:

	Passed	Failed	Total
Married	90	65	155
Unmarried	260	110	370

Solution: Let A denote married persons and as such α will denote unmarried persons.

Let B denote those who failed and as such β will denote those passed. Thus,

$(A\beta) = 90, (\alpha\beta) = 260, (AB) = 65\ (\alpha B) = 110$

These figures can be tabulated as:

	B	β
A	(AB) = 65	$(A\beta)$ = 90
α	(αB) = 110	$(\alpha\beta)$ = 260

Applying Yule's coefficient of association:

$$Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha\beta)}{(AB)(\alpha\beta)+(A\beta)(\alpha\beta)} = \frac{(65)(260)-(110)(90)}{(65)(260)+(110)(90)}$$

$$= \frac{16900-9900}{16900+9900} = \frac{7000}{26800} = +0.261$$

Example 19. Calculate the Coefficient of Association between extravagance in fathers and sons from the following data:

Extravagant fathers with extravagant sons	:	327
Extravagant fathers with miserly sons	:	545
Miserly fathers with extravagant sons	:	741
Miserly fathers with miserly sons	:	235

Solution: Let A = extravagant fathers; α = miserly fathers
B = extravagant sons; β = miserly sons

Tabulating the given data in the nine-square table:

	B	β
A	(AB) = 327	$(A\beta)$ = 545
α	(αB) = 741	$(\alpha\beta)$ = 235

Applying Yule's coefficient of Association:

$$Q = \frac{(327)\times(235)-(741)\times(545)}{(327)\times(235)+(741)\times(545)}$$

There is, thus negative association between extravagant in fathers and sons.

Example 20. From the data given in the following table, compare the association between literacy and unemployment in the urban and rural areas;

	Urban	Rural
Total Adult Males	25 lakh	200 lakh
Literate Male	10 lakh	40 lakh
Unemployed Male	5 lakh	12 lakh
Literate and Unemployed males	3 lakh	4 lakh

Solution: Let A = Literate males; α = illiterate males
B = Unemployed males; β = Employed males

Urban Area:

Tabulating the given data in a nine square table:

	B	β	
A	(AB) = 3	$(A\beta)$ = 7	(A) = 10
α	(αB) = 2	$(\alpha\beta)$ = 13	(α) = 15
	(B) = 5	(β) = 20	N = 25

$$Q = \frac{(3\times13)-(2\times7)}{(3\times13)+(2\times7)} = \frac{39-14}{39+14} = \frac{25}{53} = +0.47$$

Rural Area:

Tabulating the given data in a nine square table:

	B	β	Total
A	(AB) = 4	$(A\beta)$ = 36	(A) = 40
α	(αB) = 8	$(\alpha\beta)$ = 152	(α) = 160
Total	(B) = 12	(β) = 188	N = 200

$$Q = \frac{(4\times152)-(8\times36)}{(4\times152)+(8\times36)}$$

$$= \frac{608-288}{608+288} = \frac{320}{896}$$

$$= +0.35714$$

Thus, the coefficient of association between literacy and unemployment in the urban area is + .47 and in rural area + .35

Example 21. 200 candidates appeared for a competitive examination and 60 of them succeeded 35 received special coaching and out of them 20 candidates succeeded. Using Yule's coefficient of association discuss whether special coaching is effective or not.

Solution: Let A = Successful candidates;
α = Unsuccessful candidates;
B = who received special coaching;
β = those who did not receive special coaching

The data given are:

	A	α	Total
B	(AB) = 20	–	(B) = 35
β	–	–	–
Total	(A) = 60	–	N = 200

The complete data will be as follows:

2 × 2 Association Table

	A	α	Total
B	(AB) = 20	(αB) = 15	(B) = 35
β	$(A\beta)$ = 40	$(\alpha\beta)$ = 125	(β) = 165
Total	(A) = 60	(α) = 140	N + 200

Yule's coefficient of Association is given by -

$$Q = \frac{(20\times125)-(15)(60)}{(20\times125)-(15)(60)}$$

$$= \frac{1900}{3100} = \frac{19}{31} = 0.65$$

There is thus a moderate degree of positive association between A and B. Hence the special coaching is effective.

IMPORTANT TYPICAL EXAMPLES

Example 22. In an examination at which 600 candidates appeared, boys out numbered girls by 16 per cent of all candidates. Number of passed candidates exceeded the number of failed candidates by 310. Boys failing in the examination numbered 88. Calculate coefficient of association between male sex and success in the examination.

Solution: Denoting boys by A and girls by α; Success by B and failure by β, the given values will by presented like this:

$(A) + (\alpha) = 600$ ...(i)

$(A) - (\alpha) = 96$ ...(ii)

By adding (i) & (ii), we ge

$2(A) = 696$

$\therefore (A) = 348$

$(B) + (\beta) = 600$ ...(iii)

$(B) - (\beta) = 310$ ... (iv)

By adding (iii) & (iv), we get

$(2B) = 910$

$\therefore (B) = 455$

Now we have:

$(\alpha) = 600 - 348 = 252, (\beta) = 600 - 455 = 145$

We are also given $(A\beta) = 88$

Other values can be obtained from the nine square table as follows:

	A	α	Total
B	(AB) = 260	(αB) = 195	(B) = 455
β	$(A\beta)$ = 88	$(\alpha\beta)$ = 57	(β) = 145
Total	(A) = 348	(α) = 252	N = 600

Applying Yule's Coefficient of Association:

$$Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)-(A\beta)(\alpha B)}$$

$$= \frac{260\times57-88\times195}{260\times57+88\times195} = \frac{14820-17160}{14820+17160} = \frac{-2340}{31980} = -0.073$$

There is, thus, negative association between the two attributes i.e. male examines and success in the examination.

Example 23. The following table gives the distribution of students according to age in completed years and regular player among them:

Age in years:	15	16	17	18	19	20
No. of students:	250	200	150	120	100	80
Regular players:	200	150	90	48	30	12

Calculate coefficient of association between maturity and playing habits on the assumption that maturity is attained in the 18th year of age.

Solution: Let 'A' denote maturity (18 to 20 years) and 'α' minority (15 to 17 years)

'B' denote regular players and 'β not regular player

No. of minor students $(\alpha) = 250 + 200 + 150 = 600$

No. of major students $(A) = 120 + 100 + 80 = 300$

No. of minor regular player $(\alpha B) = 200 + 150 + 90 = 440$

No. of major regular players $(AB) = 48 + 30 + 12 = 90$

Putting the values in a 2 × 2 association table as follows:

	A	α	Total
B	(AB) = 90	(αB) = 440	(B) = 530
β	$(A\beta)$ = 210	$(\alpha\beta)$ = 160	(β) = 370
Total	(A) = 300	(α) = 600	N = 900

Applying Yule's coefficient of association:

$$Q = \frac{(AB)(\alpha\beta)-(\alpha B)(A\beta)}{(AB)(\alpha\beta)+(\alpha B)(A\beta)}$$

$$= \frac{90\times160-440\times210}{90\times160+440\times210}$$

$$= \frac{14400-92400}{14400+92400} = \frac{-78000}{106800} = -0.73$$

Thus, there is negative association between maturity and playing habits among the students.

Example 24. Using Yule' coefficient of association, investigate the association between eye colour of husbands and eye colour of wives from the data given below:

Husbands with light eyes and wives With light eyes = 309

Husbands with light eyes and wives with not light eyes = 214

Husbands with not light eyes and wives with light eyes = 132

Husbands with not light eyes and wives with not light eyes = 119

Solution: Denoting husbands with light eyes by A and husbands with not light eyes byα; wives with light eyes by B and wives with not light eyes by β.

Putting the values in a 2 × 2 association table as follows.

	B	β
A	(AB) = 309	$(A\beta)$ = 214
α	(αB) = 132	$(\alpha\beta)$ = 119

Applying Yule's Coefficient of association:

$$Q = \frac{(AB)(\alpha\beta)-(\alpha B)(A\beta)}{(AB)(\alpha\beta)+(\alpha B)(A\beta)}$$

$$= \frac{309\times119-132\times214}{309\times119+132\times214} = \frac{8523}{6519} = 0.13$$

There is, thus, little association between the eye colour of husbands and that of wives.

Exercise 5

1. Find out the coefficient of association between the type of teaching and success in teaching from the following data:

	Successful	Unsuccessful
College	58	42
University	49	51

[Ans.: Q=+.179]

2. Investigate the association between eye colour of fathers and eye colour of sons from the data given below:

Fathers with black eyes and sons with black eyes : 200

Fathers with black eyes and sons with not black eyes : 100

Fathers with not black eyes and sons with black eyes : 200

Fathers with not black eyes and sons with not black eyes : 400

[Ans.: Q=.6, Positive Association]

3. In a study to find whether tall husbands tend to marry tall wives, the following information about the wives of 250 tall and 250 short statu red husbands were published. Find the coefficient of association between the stature of wives and husbands:

	Tall husbands	Short husbands
Tall wives	112	26
Short wives	22	90

[Ans.Q=+0.89]

4. Find the association between literacy and unemployment from the following figures:

Total Adults : 10,000
Literate : 1,290
Unemployed : 1,390
Literate unemployed : 820

Comment on the result.

[Ans. Q=0.923]

5. Prepare a 2 × 2 table from the following information, calculate Yule's coefficient association and interpret the result:

$N = 500, (\alpha) = 300, (B) = 125, (AB) = 25$

[Ans. Q=–.55, Negative association]

6. A teacher examined 280 students in Economics and Auditing and found that 160 failed in Economics, 140 failed in Auditing and 80 failed in both the subjects. Is there any association between failure in Economics and Auditing?

[Ans. $Q = 0$, No Association]

7. In an experiment on immunization of cattle from tuberculosis, the following results were obtained:

	Died or affected	Unaffected
Inoculated	12	26
Not Inoculated	16	6

Examine the effect of vaccine in controlling susceptibility to tuberculosis.

[Ans. $Q = -.70$, effective]

8. 1660 candidates appeared for a competitive examination and of these 422 were successful, 256 had attended a coaching class and of these 150 came out successful. Examine the utility of the coaching class.

[Ans. $Q = +.7096$ Coaching is effective]

(4) Coefficient of Colligation

Prof. Yule has given another coefficient known as Coefficient of Colligation to find the degree of association between the two attributes. It is denoted by symbol γ (Gamma) and is calculated by applying the following formula:

$$\gamma = \frac{\sqrt{(AB)(\alpha\beta)} - \sqrt{(A\beta)(\alpha B)}}{\sqrt{(AB)(\alpha\beta)} + \sqrt{(A\beta)(\alpha B)}} = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

where γ = Coefficient of Collignation.

Relationship between Coefficient of Colligation and Coefficient of Association: Yule coefficient of association can also be obtained from the coefficient of Colligation by using the formula:

$$Q = \frac{2\gamma}{1+\gamma^2}$$

Coefficient of Colligation is not widely used in practice.

Example 25. Form the data given below, find the coefficient of Colligation

$(AB) = 80, (A\beta) = 20, (\alpha B) = 220, (\alpha\beta) = 180$

Hence or otherwise find Yule's coefficient of association.

Solution: Tabulating the given data in a nine square table:

	A	
B	$(AB) = 80$	α
β	$(A\beta) = 20$	$(\alpha B) = 220$
		$(\alpha\beta) = 180$

$$\text{Coefficient of Colligation } (\gamma) = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

$$= \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}} = \frac{1 - \sqrt{\frac{20\times220}{80\times180}}}{1 + \sqrt{\frac{20\times220}{80\times180}}}$$

$$= \frac{1 - \sqrt{\frac{4400}{14400}}}{1 + \sqrt{\frac{4400}{14400}}} = \frac{1 - \sqrt{3055}}{1 + \sqrt{3055}} = \frac{1 - .5527}{1 + .5527} = \frac{.4473}{1.5527} = +0.288$$

Calculation of Yule's coefficient of Association

Yule's coefficient of Association can be obtained from coefficient of Colligation using the formula:

$$Q = \frac{2\gamma}{1+\gamma^2} = \frac{2\times.288}{1+(.288)^2} = \frac{.576}{1+.083}$$

$$= \frac{.576}{1.083} = +.532$$

Example 26. From the following data relating to sanity and deafness of 135 persons, find out the coefficient of association between sanity and deafness using coefficient of Colligation:]

	Sane	Insane	Total
Deaf	20	40	60
Not Deaf	50	25	75
Total	70	65	135

Solution: Let A = Sane; α = Insane
B = Deaf; β = Not Deaf

Tabulating the given data in a nine square table:

	A	α
B	(AB) = 20	(αB) = 40
β	$(A\beta)$ = 50	$(\alpha\beta)$ = 25

$$\text{Coefficient of Colligation } (\gamma) = \frac{1-\sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1+\sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

$$= \frac{1-\sqrt{\frac{50\times 40}{20\times 25}}}{1+\sqrt{\frac{50\times 40}{20\times 25}}} = \frac{1-\sqrt{\frac{2000}{500}}}{1+\sqrt{\frac{2000}{500}}}$$

$$= \frac{1-\sqrt{4}}{1+\sqrt{4}} = \frac{1-2}{1+2} = \frac{-1}{3} = -0.333$$

$$\text{Coefficient of Association } (Q) = \frac{2\gamma}{1+\gamma^2} = \frac{2(-333)}{1+(-333)^2} = \frac{-.666}{1.110} = -0.6$$

There is negative association between sanity and deafness.

Exercise 6

1. Find the coefficient of colligation and Yule's coefficient of association from the following data:

   N = 1,000, (A) = 380, (B) = 380, (AB) = 230

   [Ans. γ = .127, Q = .25]

2. Determine the coefficient of colligation and coefficient of association.

   (AB) = 100, (αB) = 30, $(A\beta)$ = 20, $(\alpha\beta)$ = 10

   [Ans. γ = .127, Q = .25]

3. In an investigation whether tall husbands tend to marry tall wives, the following results were obtained:

	Tall Husbands	Short Husbands
Tall wives	56	13
Short wives	11	45

Calculate the coefficient of colligation and also show its relationship with Yule's coefficient of Association. Verify your answer with direct calculation.

[Ans. γ = .6155, Q = + .893, Verified]

(5) Coefficient of Contingency

Prof. Karl Person has propounded coefficient of contingency to find the degree of association in a 2 × 2 or 3 × 3 etc. association table. It is denoted by 'C' and is calculated by **applying the following formula:**

$$\text{Coefficient of contingency } (C) = \sqrt{\frac{\chi^2}{N+\chi^2}}$$

Where χ^2 = Chi-square quantity (pronounced as Ki square)

N = Number of observations.

Calculation of χ^2: The calculation of χ^2 involves the following steps:

(i) Obtain the expected frequency for each cell. For example, the expected value of the cell $E(AB) = \frac{(A)(B)}{N}$

(ii) Calculate the difference between the observed frequencies and expected frequencies i.e., calculate $(O-E)$.

(iii) Square the difference between the observed and expected frequencies in each cell i.e. calculate $(O-E)^2$

(iv) Divide the squared differences by corresponding expected frequencies i.e. calculate $(O-E)^2/E$.

(v) Obtain $\sum\left(\frac{(O-E)^2}{E}\right)$ to get the value of χ^2

$$\text{Thus, } \chi^2 = \sum\left(\frac{(O-E)^2}{E}\right)$$

The following examples illustrate the calculation of coefficient of contingency.

Example 27. Find out the **coefficient of contingency** between the type of institution and the success in teaching from the following table:

Institution	Successful	Unsuccessful	Total
College	58	42	100
University	49	51	100
Total	107	93	200

Solution: Let A = College Education;
α = University Education
B = Successful;
β = Unsuccessful

Putting the given values in a 2 × 2 association table (or contingency table) as follows:

	B	β	Total
A	(AB) = 58	$(A\beta)$ = 42	(A) = 100
α	$((\alpha B)$ = 49	$(\alpha\beta)$ = 51	(α) = 100
Total	(B) = 107	(β) 93	N = 200

From this we calculate the expected frequency for each cell as follows:

$$E(AB) = \frac{(A).(B)}{N} = \frac{100 \times 107}{200} = 53.5$$

$$E(\alpha B) = \frac{(\alpha).(B)}{N} = \frac{100 \times 107}{200} = 53.5$$

or $E(\alpha B) = 107 - 53.5 = 53.5$

$$E(A\beta) = \frac{(A).(\beta)}{N} = \frac{100 \times 93}{200} = 46.5$$

or $E(A\beta) = 100 - 53.5 = 46.5$

$$E(\alpha\beta) = \frac{(\alpha).(\beta)}{N} = \frac{100 \times 93}{200} = 46.5$$

or $E(\alpha\beta) = 100 - 53.5 = 46.5$

Calculation of χ^2

	O	E	$(O-E)$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
(AB)	58	53.5	+4.5	20.25	0.378
(αB)	49	53.5	−4.5	20.25	0.378
$(A\beta)$	42	46.5	−4.5	20.25	0.435
$(\alpha\beta)$	51	46.5	+4.5	20.25	0.35
					$\chi^2 = \sum\frac{(O-E)^2}{E} = 1.626$

Coefficient of Contingency $(C) = \sqrt{\frac{\chi^2}{N+\chi^2}} = \sqrt{\frac{1.626}{200+(1.626)}} = \sqrt{\frac{1.626}{201.626}} = 0.087$

Since the coefficient of contingency is 0.087, there is, thus, poor association between the two attributes.

Example 28. Calculate the coefficient of contingency from the following data:

Intelligence → / Social Status ↓	Dull	Intelligent	Brilliant	Total
Lower	22	35	23	80
Middle	38	70	32	140
Upper Middle	60	20	20	100
Total	120	125	75	320

Solution: The contingency table will be as follows:

Intelligence → / Social Status ↓	Dull A_1	Intelligent A_2	Brilliant A_3	Total
Lower Middle B_1	(A_1B_1) 22	(A_2B_1) 35	(A_3B_1) 23	(B_1) 80
Middle B_2	(A_1B_2) 38	(A_2B_2) 70	(A_3B_2) 32	(B_2) 140
Upper Middle B_3	(A_1B_3) 60	(A_2B_3) 20	(A_3B_3) 20	(B_3) 100
Total	(A_1) 120	(A_2) 125	(A_3) 75	N 320

The expected frequency of each cell is calculated as follows:

$$E(A_1B_1)=\frac{(A_1).(B_1)}{N}=\frac{120\times80}{320}=30$$

$$E(A_1B_2)=\frac{(A_1).(B_2)}{N}=\frac{120\times140}{320}=52.5$$

$$E(A_1B_3)=\frac{(A_1)(B_3)}{N}=\frac{120\times100}{320}=37.5$$

$$E(A_2B_1)=\frac{(A_1)(B_1)}{N}=\frac{125\times80}{320}=31.25$$

$$E(A_2B_2)=\frac{(A_2).(B_3)}{N}=\frac{125\times140}{320}=54.7$$

$$E(A_2B_3)=\frac{(A_2).(B_3)}{N}=\frac{125\times100}{320}=39.1$$

$$E(A_3B_2)=\frac{(A_3).(B_3)}{N}=\frac{75\times80}{320}=18.75$$

$$E(A_3B_2)=\frac{(A_3).(B_3)}{N}=\frac{75\times140}{320}=32.8$$

$$E(A_3B_3)=\frac{(A_3).(B_3)}{N}=\frac{75\times100}{320}=23.4$$

Calculation of γ^2

	O	E	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
(A_1B_1)	22	30	−8	64	2.133
(A_1B_2)	38	52.5	−14.5	210.25	4.004
(A_1B_3)	60	37.5	22.5	506.25	13.50
(A_2B_1)	35	31.25	3.75	14.0625	.45
(A_2B_2)	70	54.7	15.3	234.09	4.279
(A_2B_3)	20	39.1	−19.1	364.81	9.330
(A_3B_1)	23	18.75	+4.25	18.0625	.963
(A_3B_2)	32	32.8	−0.8	0.64	.019
(A_3B_3)	20	23.4	−3.4	11.56	.494
					$\chi^2=\sum\frac{(O-E)^2}{E}$ = 35.172

Coefficient of contingency $(C)=\sqrt{\frac{\chi^2}{N+\chi^2}}$

$=\sqrt{\frac{35.172}{320+35.172}}=\sqrt{0.991}=.315$ approx.

Exercise 7

1. Calculate the coefficient of contingency from the following data;

	Employed	Unemployed
illiterate	31	469
Literate	185	1315

[Ans. $C=.085$]

2. Calculate the coefficient of contingency from the following data:

Intelligence / Height	Good	Bad	Indifferent
Tall	10	5	5
Middle	20	5	5
Short	30	10	10

[Ans. $C=0.1$]

MISCELLANEOUS SOLVED EXAMPLES

Example 29. Given the following data, calculate Yule's coefficient of association.

	A	α
B	75	23
β	5	42

Solution: Given 2 × 2 Association Table

	A	α
B	75	23
β	5	42

Yule's Coefficient of Association is given by:

$$Q = \frac{75\times 42 - 5\times 23}{75\times 42 + 5\times 23}$$

$$= \frac{3150-115}{3150+115} = \frac{3035}{3265} = +0.929$$

Example 30. Out of 2000 people exposed to a small-pox in a village, 450 were attacked. Among the people 365 were vaccinated and out of them 50 were attacked. Form a nine square table and conclude there from whether vaccination can be regarded as a good preventive or not.

Solution: Let vaccination be denoted by B and not attacked by A. Putting the given information in a nine square table:

Attributes	Not-Attacked (A)	Attacked (α)	Total
Vaccinated (B)	(AB) 315	(αB) 50	(B) 365
Not Vaccinated (β)	$(A\beta)$ 1235	$(\alpha\beta)$ 50	(β) 1635
	(A) 1550	(α) 450	$N=$ 2000

Using Yule's coefficient of association:

$$Q = \frac{(AB)(\alpha\beta)-(\alpha B)(A\beta)}{(AB)(\alpha\beta)+(\alpha B)(A\beta)}$$

$$= \frac{315\times 400 - 1235\times 50}{315\times 400 + 1235\times 50} = \frac{64{,}250}{1{,}87{,}750} = +0.342$$

Thus, vaccination can be regarded as a satisfactory preventive measure, though not very good.

Example 31. 800 candidates of both sexes appeared at an examination. The boys outnumbered the girls by 15% of the total. The number of candidates who passed exceeded the number failed by 480. Equal number of boys and girls failed in the examination. Prepare a 2 × 2 table and find the coefficient of association.

Solution: Denoting boys by A and girls by α. Success by B and failure by β, the given values will be calculated like this:

$$(A)+(\alpha)=800 \quad \ldots (i)$$

$$(A)-(\alpha)=120 \quad \ldots (ii)$$

By adding $2(A) = 920$

$\therefore$ $(A) = 460$

Putting the value of (A) in (i), we get

$460 + (\alpha) = 800$

$\therefore$ $(\alpha) = 340$

$\therefore$ $(A) = 460, (\alpha) = 340$

$(B)+(\beta)=800$ ... (iii)

$(B)-(\beta)=480$ ... (iv)

By adding $(2B) = 1280$

$\therefore$ $(B) = 640$

Putting the value of (B) in (iii), we get

$640 + (\beta) = 800$

$\therefore$ $(\beta) = 160$

$\therefore$ $(B) = 640, (\beta) = 160$

Also we are given $(A\beta) = (\alpha\beta) = \frac{160}{2} = 80$

We can now present the above information in the form of a 2 × 2 table as follows:

	A	α	Total
B	(AB) 380	(αB) 260	(B) 640
β	$(A\beta)$ 80	$(\alpha\beta)$ 80	(β) 160
Total	(A) 460	(α) 340	$N=$ 800

Yule's coefficient of association is given by:

$$Q = \frac{(AB)(\alpha\beta)-(\alpha B)(A\beta)}{(AB)(\alpha\beta)+(\alpha B)(A\beta)}$$

$$= \frac{380\times 80 - 80\times 260}{380\times 80 + 80\times 260} = \frac{30400-20800}{30400-20800} = \frac{9600}{51200} = +.1875$$

Thus, the coefficient of association shows positive association of a low degree between sex and success in examination.

Example 32. 1,000 candidates appeared in a certain examination. Boys outnumbered girls by 20% of all candidates who appeared in the examination. Number of passed candidates exceeded the number of failed candidates by 166. Girls failing in the examination numbered 58. Construct 2 × 2 table and then find the coefficient of association. Also interpret the coefficient.

Solution: Denoting boys by A and girls by α ; Success by B and failure by β. We are given;

$(A)+(\alpha)=1000$ .... (i)

$(A)-(\alpha)=200$ ..... (ii)

Adding (i) and (ii), we get $(A)=600$ and $(\alpha)=400$

$(B)+(\beta)=1000$ ..... (iii)

$(B)-(\beta)=166$ .... (iv)

Adding (iii) and (iv), we get $(B)=583$ and $(\beta)=417$

Also we are given $(\alpha\beta)=58$

We can now present the above information in the form of a 2 × 2 table as follows:

	A	α	
B	(AB) 241	$(\alpha\beta)$ 342	(B) 583
β	$(A\beta)$ 241	$(\alpha\beta)$ 58	(β) 417
	(A) 600	(α) 400	$N=$ 1000

Yule's coefficient of association is given by:

$$Q=\frac{(AB)(\alpha\beta)-(\alpha B)(A\beta)}{(AB)(\alpha\beta)+(\alpha B)(A\beta)}$$

$$=\frac{241\times58-359\times342}{241\times58+359\times342}=\frac{-108800}{136756}=-0.796$$

This shows that there is a negative association between male sex and success in the examination.

Example 33. In a class test in which 135 candidates were examined for proficiency in English and Economics. It was discovered that 75 students failed in English, 90 failed in Economics and 50 failed in both. Find if there is any association between failure in English and Economics.

Solution: Denoting those whose failed in English by A and passed and passed in English by α. Failed in Economics by B and passed in Economics by β.

Putting the given information in a nine-square table, we have

	A	α	
B	(AB) 50	(αB) 40	(B) 90
β	$(A\beta)$ 25	$(\alpha\beta)$ 20	(β) 45
	(A) 75	(α) 60	$N=$ 135

Yule's coefficient of association is given by:

$$Q=\frac{(AB)(\alpha\beta)-(\alpha B)(A\beta)}{(AB)(\alpha\beta)+(\alpha B)(A\beta)}=\frac{(50)(20)-(25)(40)}{(50)(20)+(25)(40)}=\frac{1000-1000}{1000+1000}=0$$

The coefficient allows that the attributes are independent.

Example 34 . In a population of 1000 students, the number of married is 400. Out of the 300 students who failed, 120 belonged to married group. Using Yule's coefficient of association, find out extent of association of the attributes, marriage and failure.

Solution: Denoting married students by A and unmarried by α; Failed by B and passed by β.

Putting the given information in the form of nine square table:

	(A)	(α)	
(B)	(AB) 120	(αB) 280	(B) 300
(β)	$(A\beta)$ 280	$(\alpha\beta)$ 320	(β) 700
	(A) 400	(α) 600	$N=$ 1000

Yule's coefficient of association is given by:

$$Q=\frac{(AB)(\alpha\beta)-(\alpha B)(A\beta)}{(AB)(\alpha\beta)+(\alpha B)(A\beta)}$$

$$=\frac{(120)(320)-(280)(280)}{(120)(320)+(280)(280)}=\frac{38400-78400}{38400+78400}=\frac{-40000}{1168300}=-0.3424$$

The coefficient shows that there exists low degree of negative association between marriage and failure.

Example 35. In a population of 1000 students, 40 percent students are married. Out of 40 percent students who failed, 300 belonged to the married group. Prepare a 2 × 2 table and using Yule's coefficient of association, whether there is any association between the two attributes marriage and failure.

Solution: Denoting married student by A, and unmarried by α; students failed by B and student passed by β

	Married (A)	Unmarried (α)	
Failed (B)	(AB) 300	(αB) 100	(B) 400
Passed (β)	($A\beta$) 100	($\alpha\beta$) 500	(β) 600
	(A) 400	(α) 600	N= 1000

Yule's coefficient of association is given by:

$$Q = \frac{300\times500-100\times100}{300\times500+100\times100} = \frac{140000}{16000} = +.875$$

The coefficient shows that there exists high degree of positive association between marriage and failure.

Example 36. A distribution according to age group and martial status of girls studying in a particular collage is given below:

Age:	15	16	17	18	19	20	21	22
No. of girls:	15	18	22	25	20	23	27	30
No. of married girls:	1	2	3	4	5	7	8	10

Obtain the value of coefficient of association between the adult girls and married girls if it is assumed that adulthood is attained after 18 years of age.

Solution: As the adulthood is attained after 18 years of age, so

20 + 23 27 30 = 100 adult girls,

15 + 18 + 22 + 25 = 80 minor girls,

5 + 7 + 8 + 10 = 30 adult married girls, and

1 + 2 + 3 + 4 = 10 minor married girls.

Let A denote adult girls

$\therefore$ α will denote minor girls

Let B denote those who are married

$\therefore$ β will denote those who are unmarried.

Putting the given information in a nine square table:

	Married (B)	Unmarried (β)	
Adult Girls (A)	(AB) 30	($A\beta$) 70	(A) 100
Minor Girls (α)	(αB) 10	($\alpha\beta$) 70	(α) 80
	(B) 40	(β) 140	N= 180

Yule's coefficient of association is given by:

$$Q = \frac{30\times70-70\times10}{30\times70+70\times10} = \frac{2100-700}{2100+700} = \frac{1400}{2800} = +.50$$

Thus, we find a positive association between maturity and marriage.

Example 37. The male population of U.P. is 250 lakhs. The number of literate males is 20 lakhs and total number of criminals is 26 thousands. The number of literate criminals is 2 thousand. Do you find any association between literacy and criminality?

Solution: Let A denote literate males

$\therefore$ α will denote literate males

Let B denote male criminals

$\therefore$ β will denote male non-criminals.

The given information can be put in a nine square table:

	Criminality (B)	Non-criminality (β)	Total
Literate (A)	(AB) 2	($A\beta$) 18	(A) 20
Illiterate (α)	(αB) 24	($\alpha\beta$) 206	(α) 230
Total	(B) 26	(β) 224	N= 250

Yule's coefficient of association is given by:

$$Q = \frac{2\times206-18\times24}{2\times206+18\times24} = \frac{412-432}{412+432} = \frac{-20}{844} = -0.023$$

The coefficient shows that the attributes literacy and criminality are negatively association i.e. literacy checks criminality.

IMPORTANT FORMULAE

1. **Consistency of Data**

The criterion of consistency of data is that no ultimate class frequency should be negative.

2. **Frequency Method**

Attributes A and B are said to be:

(i) Independent if $AB = \frac{(A)\times(B)}{N}$

(ii) Positively associated if $(AB) > \frac{(A)\times(B)}{N}$.

(iii) Negatively associated if $(AB) < \frac{(A)\times(B)}{N}$

3. **Proportion Method**

Attributes A and B are said to be:

(i) Independent if $\frac{(AB)}{(A)} = \frac{(\alpha B)}{(\alpha)}$

(ii) Positively associated if $\frac{(AB)}{(A)} < \frac{(\alpha B)}{(\alpha)}$

4. **Yule's Coefficient of Association**

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

5. **Coefficient of Colligation**

$$\gamma = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}; \qquad \text{Also } Q = \frac{2\gamma}{1+\gamma^2}$$

6. **Coefficient of Contingency**

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

QUESTIONS

1. Define Association of Attributes. Discuss the various types of association.
2. Explain the difference between association and correlation.
3. (a) Explain the terms 'Association' and Dis-association between two attributes with examples.

   (b) State Yule's coefficient of association and its range. State its limitations.
4. What is meant by association of attributes? Explain briefly the various methods of measuring association between two attributes.
5. Write Short notes on: (i) Consistency of Data; (ii) Association of Attributes.

□□□

Advanced Statistical Tables

LOGARITHMS

n	1	2	3	4	5	6	7	8	9	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	30	34	38
11	0411	0453	0492	0531	0569	0607	0645	0682	0719	0755	4	8	12	16	20	24	28	32	36
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106	4	7	11	15	18	22	26	29	33
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430	3	7	10	14	17	20	24	27	31
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732	3	6	10	13	16	19	22	25	29
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014	3	6	9	12	15	19	22	25	28
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279	3	6	9	11	14	17	20	23	26
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529	3	6	8	11	14	17	19	22	25
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765	3	5	8	11	14	16	19	22	24
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989	3	5	8	10	13	16	18	21	24
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201	3	5	7	10	13	15	18	20	23
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404	2	4	7	9	12	14	17	19	22
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598	2	4	7	9	11	14	16	18	21
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784	2	4	6	8	11	13	15	17	20
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962	2	4	6	8	10	12	14	16	18
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133	2	3	5	7	9	10	12	14	15
26	4150	4165	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	5599	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	6444	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	7	8
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712	1	2	3	4	5	6	7	8	8
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803	1	2	3	4	4	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						

	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067	1	2	3	4	5	6	7	8	9
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152	1	2	3	4	5	6	7	8	9
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235	1	2	3	4	5	6	7	8	9
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316	1	2	3	4	5	6	7	8	9
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474	1	2	3	4	5	6	7	8	9
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551	1	2	3	4	5	6	7	8	9
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627	1	2	3	4	5	6	7	8	9
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701	1	2	3	4	5	6	7	8	9
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774	1	2	3	4	5	6	7	8	9
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846	1	1	2	3	4	4	5	6	6
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917	1	1	2	3	4	4	5	6	6
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987	1	1	2	3	3	4	5	6	6
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055	1	1	2	3	3	4	5	5	6
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122	1	1	2	3	3	4	5	5	6
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189	1	1	2	3	3	4	5	5	6
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254	1	1	2	3	3	4	5	5	6
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319	1	1	2	3	3	4	4	5	6
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382	1	1	2	2	3	4	4	5	6
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445	1	1	2	2	3	4	4	5	6
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506	1	1	2	2	3	4	4	5	5
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567	1	1	2	2	3	4	4	5	5
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627	1	1	2	2	3	4	4	5	5
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686	1	1	2	2	3	4	4	5	5
74	8692	8698	8704	8710	8716	8722	8727	8733	8739	8745	1	1	2	2	3	4	4	5	5
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802	1	1	2	2	3	3	4	5	5
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859	1	1	2	2	3	3	4	5	5
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915	1	1	2	2	3	3	4	4	5
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971	1	1	2	2	3	3	4	4	5
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025	1	1	2	2	3	3	4	4	5
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079	1	1	2	2	3	3	4	4	5
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133	1	1	2	2	3	3	4	4	5
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186	1	1	2	2	3	3	4	4	5
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238	1	1	2	2	3	3	4	4	5
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289	1	1	2	2	3	3	4	4	5
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340	1	1	2	2	3	3	4	4	5
86	9345	9350	9355	9360	9365	9170	9375	9380	9385	9390	1	1	2	2	3	3	4	4	5
87	9395	9400	9405	9410	9415	91201	9425	9430	9435	9440	0	1	1	2	2	3	3	4	4
88	9445	9450	9455	9460	9265	9469	9474	9479	9484	9489	0	1	1	2	2	3	3	4	4
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538	0	1	1	2	2	3	3	4	4
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586	0	1	1	2	2	3	3	4	4
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633	0	1	1	2	2	3	3	4	4
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680	0	1	1	2	2	3	3	4	4
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727	0	1	1	2	2	3	3	4	4
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773	0	1	1	2	2	3	3	4	4
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818	0	1	1	2	2	3	3	4	4
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863	0	1	1	2	2	3	3	4	4
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908	0	1	1	2	2	3	3	4	4
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952	0	1	1	2	2	3	3	4	4
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996	0	1	1	2	2	3	3	4	4

ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	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	0	1	1	1	1	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	2
.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	1	2	2	2	3
.12	1318	1321	1324	1327	1330	1334	1337	1340	1343	1346	0	1	1	1	1	2	2	2	3
.13	1349	1352	1355	1358	1361	1365	1368	1371	1374	1377	0	1	1	1	1	2	2	2	3
.14	1380	1384	1387	1390	1393	1396	1400	1403	1406	1409	0	1	1	1	1	2	2	2	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	2	3	3
.20	1585	1589	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	3	3	3	4
.25	1778	1782	1786	1791	1795	1799	1803	1807	1811	1816	0	1	1	2	2	3	3	4	4
.26	1820	1824	1828	1832	1837	1841	1845	1849	1854	1858	0	1	1	2	2	3	3	4	4
.27	1862	1866	1871	1875	1879	1884	1888	1892	1897	1901	0	1	1	2	2	3	3	4	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	6	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	7

ANTILOGARITHMS

	0	1	2	3	4	5	6	7	8	9	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	7	7
.55	3548	3556	3565	3573	3581	3589	3597	3606	3614	3622	1	2	2	3	4	5	6	7	8
.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	8	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	11
.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	12
.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	6714	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	13	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	9595	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

BINOMIAL COEFFICIENTS

n	$\binom{n}{0}$	$\binom{n}{1}$	$\binom{n}{2}$	$\binom{n}{3}$	$\binom{n}{4}$	$\binom{n}{5}$	$\binom{n}{6}$	$\binom{n}{7}$	$\binom{n}{8}$	$\binom{n}{9}$	$\binom{n}{10}$
0	1										
1	1	1									
2	1	2									
3	1	3	3	1							
4	1	4	6	4	1						
5	1	5	10	10	5	1					
6	1	6	15	20	15	6	1				
7	1	7	21	35	35	21	7	1			
8	1	8	28	56	70	56	28	8	1		
9	1	9	36	84	126	126	84	36	9	1	
10	1	10	45	120	210	252	210	120	45	10	1
11	1	11	55	165	330	462	462	330	165	55	11
12	1	12	66	220	495	792	924	792	495	220	66
13	1	13	78	286	715	1287	1716	1716	1287	715	286
14	1	14	91	364	1001	2002	3003	3432	3003	2002	1001
15	1	15	105	455	1365	3003	5005	6435	6435	5005	3003
16	1	16	120	560	1820	4368	8008	11440	12870	11440	8008
17	1	17	136	680	2380	6188	12376	19448	24310	24310	19448
18	1	18	153	816	3060	8568	18564	31824	43758	48620	43758
19	1	19	171	969	3876	11628	27132	50388	75582	92378	92378
20	1	20	190	1140	4845	15504	38760	77520	125970	167960	184750

(VOLUE OF e^{-m} (For Computing Possion Probabilities) (0<m<1)

m	0	1	2	3	4	5	6	7	8	9
0·0	1·0000	·9900	·9802	·9704	·9608	·9512	·9418	·9324	·9231	·9139
0·1	0·9048	·8958	·8860	·8781	·8694	·8607	·8521	·8437	·8353	·8270
0·2	0·8187	·8106	·8025	·7945	·7866	·7788	·7711	·7634	·7558	·7483
0·3	0·7408	·7334	·7261	·7189	·7118	·7047	·6970	·6907	·6839	·6771
0·4	0·6703	·6636	·6570	·6505	·6440	·6376	·6313	·6250	·6188	·6125
0·5	0·6065	·6005	·5945	·5886	·5827	·5770	·5712	·5655	·5599	·5543
0·6	0·5448	·5434	·5379	·5326	·5278	·5220	·5160	·51127	·5066	·5016
0·7	0·4966	·4916	·4868	·4810	·4771	·4724	·4670	·4630	·4584	·4538
0·8	0·4493	·4449	·4404	·4360	·4317	·4274	·4232	·4190	·4148	·4107
0·9	0·4066	·4025	·3985	·3946	·3906	·3867	·3829	·3791	·3753	·3716

(m=1, 2, 3.....10)

m	1	2	3	4	5	6	7	8	9	10
e^{-m}	·36788	·13534	·04979	·01832	·00698	·00279	·00092	·000395	·000121	000045

Note : To obtain values of e^{-m} for other values of m, use the laws of exponents.

Example. $e^{-2.35} = \left(e^{-2.00}\right)\left(e^{-.35}\right) = (.13534)(.7047) = .095374$

AREA UNDER THE NORMAL CURVE

The entries in the table are the probabilities that a random variable having the standard normal distribution assumes a value between 0 and z; they are given by the area under the curve shaded in the figure shown on the right hand side.

The Standard Normal Distribution.

Area under Normal Curve

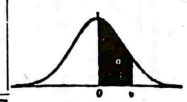
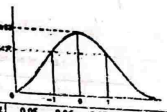


Table of Area

z	0	1	2	3	4	5	6	7	8	9
0	·0000	·0040	·0080	·0120	·0160	·0199	·0239	·0279	·0319	·0359
1	·0398	·0438	·0478	·0517	·0557	·0596	·0636	·0675	·0714	·0753
2	·0793	·0832	·0871	·0910	·0948	·0987	·1026	·1064	·1103	·1141
3	·1179	·1217	·1255	·1293	·1331	·1368	·1406	·1443	·1480	·1517
4	·1554	·1591	·1628	·1664	·1700	·1736	·1772	·1808	·1844	·1879
5	·1915	·1950	·1985	·2019	·2054	·2083	·2123	·2157	·2190	·2224
6	·2257	·2291	·2324	·2357	·2389	·2422	·2454	·2486	·2517	·2549
7	·2580	·2611	·2642	·2673	·2703	·2734	·2764	·2794	·2823	·2852
8	·2881	·2910	·2939	·2967	·2995	·3023	·3051	·3078	·3106	·3133
9	·3159	·3186	·3212	·3238	·3264	·3289	·3315	·3340	·3365	·3389
10	·3413	·3438	·3461	·3485	·3508	·3531	·3554	·3577	·3599	·3621
11	·3643	·3665	·3686	·3708	·3729	·3749	·3770	·3790	·3810	·3830
12	·3849	·3869	·3888	·3907	·3925	·3944	·3962	·3980	·3997	·4015
13	·4032	·4049	·4066	·4082	·4099	·4115	·4131	·4147	·4162	·4177
14	·4192	·4207	·4222	·4236	·4251	·4265	·4279	·4292	·4306	·4319
15	·4332	·4345	·4357	·4370	·4382	·4394	·4406	·4418	·4429	·4441
16	·4452	·4463	·4474	·4484	·4495	·4505	·4515	·4525	·4535	·4545
17	·4554	·4564	·4573	·4582	·4591	·4599	·4608	·4616	·4625	·4633
18	·4641	·4649	·4656	·4664	·4671	·4678	·4686	·4693	·4699	·4706
19	·4713	·4719	·4726	·4732	·4738	·4744	·4750	·4756	·4761	·4767
20	·4772	·4778	·4783	·4788	·4793	·4798	·4803	·4808	·4812	·4817
21	·4821	·4826	·4830	·4834	·4838	·4842	·4846	·4850	·4854	·4857
22	·4861	·4864	·4868	·4871	·4875	·4878	·4881	·4884	·4887	·4890
23	·4893	·4896	·4898	·4901	·4904	·4906	·4909	·4911	·4913	·4916
24	·4918	·4920	·4922	·4925	·4927	·4929	·4931	·4932	·4934	·4936
25	·4938	·4940	·4941	·4943	·4945	·4946	·4948	·4959	·4951	·4952
26	·4953	·4955	·4956	·4957	·4959	·4960	·4961	·4962	·4963	·4964
27	·4965	·4966	·4967	·4968	·4969	·4970	·4071	·4972	·4973	·4974
28	·4974	·4975	·4976	·4977	·4977	·4978	·4979	·4979	·4980	·4981
29	·4981	·4982	·4982	·4983	·4984	4984	·4985	·4985	·4986	·4986
30	·4987	·4987	·4987	·4988	·4988	·4989	·4989	·4989	·4990	·4990
31	·4990	·4991	·4991	·4991	·4992	·4992	·4992	·4992	·4990	·4993
32	·4993	·4993	·4994	·4994	·4994	·4994	·4994	·4995	·4995	·4995
33	·4995	·4995	·4995	·4996	·4996	·4996	·4996	·4996	·4967	·4967
34	·4997	·4997	·4997	·4997	·4997	·4997	·4997	·4997	·4997	·4978
35	·4998	·4998	·4998	·4998	·4999	·4998	·4998	·4998	·4998	·4998
36	·4998	·4998	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999
37	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999
38	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999	·4999
39	·5000	·5000	·5000	·5000	·5000	·5000	·5000	·5000	·5000	·5000

ORDINATES (Y) of the STANDARD NORMAL CURVE



0	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0·0	·3989	·3989	·3989	·3988	·3986	·3984	·3982	·3980	·3977	·3973
0·1	·3970	·3965	·3961	·3956	·3951	·3945	·3939	·3932	·3925	3918
0·2	·3910	·3902	·3894	·3885	·3876	·3867	·3857	·3847	·3836	·3825
0·3	·3814	·3802	·3790	·3778	·3765	·3752	·3739	·3725	·3712	·3697
0·4	·3683	·3668	·3653	·3637	·3621	·3605	·3589	·3572	·3555	·3538
0·5	·3521	·3503	·3485	·3467	·3448	·3429	·3410	·3391	·3372	·3352
0·6	·3332	·3312	·3292	·3271	·3251	·3230	·3209	·3187	·3166	·3144
0·7	·3123	·3101	·3079	·3056	·3034	·3011	·2989	·2966	·2943	·2920
0·8	·2897	·2874	·2850	·2827	·2803	·2780	·2756	·2732	·2709	·2685
0·9	·2661	·2637	·2613	·2589	·2565	·2541	·2516	·2492	·2468	·2444
1·0	·2420	·2396	·2371	·2347	·2323	·2299	·2275	·2251	·2227	·2203
1·1	·2179	·2155	·2131	·2107	·2083	·2059	·2036	·2012	·1989	·1965
1·2	·1942	·1919	·1895	·1872	·1849	·1826	·1804	·1781	·1758	·1736
1·3	·1714	·1691	·1669	·1647	·1626	·1604	·1582	·1561	·1539	·1518
1·4	·1497	·1476	·1456	·1435	·1415	·1394	·1374	·1354	·1334	·1315
1·5	·1295	·1276	·1257	·1238	·1219	·1200	·1182	·1163	·1145	·1127
1·6	·1109	·1092	·1074	·1057	·1040	·1025	·1006	0989	·0973	0957
1·7	·0940	·0925	·0909	·0893	·0878	0863	·0848	0833	·0818	0804
1·8	0790	·0775	·0761	·0748	·0734	·0721	·0707	·0694	·0681	·0669
1·9	·0656	·0644	·0632	·0620	0608	0596	0584	·0573	·0562	·0551
2·0	·0540	·0529	·0519	·0508	·0498	·0488	·0478	·0468	·0459	·0449
2·1	·0440	·0431	·0422	·0413	·0404	·0396	·0387	·0379	·0371	·0363
2·2	·0355	·0347	·0339	·0332	·0325	·0317	·0310	·0303	·0297	·0290
2·3	·0283	·0277	·0270	·0264	·0258	·0252	·0246	·0241	·0235	·0229
2·4	·0224	·0219	·0213	·0208	·0203	·0108	·0194	·0189	·0184	0180
2·5	·0175	·0171	·0167	·0163	·0158	·0154	·0151	·0147	·0143	·0139
2·6	·0136	·0132	·0129	·0126	·0122	·0119	·0116	·0113	0110	·0107
2·7	0104	·0101	·0099	·0096	·0093	·0091	·0088	·0086	·0084	·0081
2·8	·0079	·0077	·0075	·0073	·0071	0099	0067	0065	0063	0061
2·9	·0060	·0058	·0056	·0055	·0053	·0051	·0050	·0048	·0047	0046
3·0	·0044	·0043	·0042	·0040	·0039	·0038	·0037	·0036	·0035	0034
3·1	0033	·0032	0031	·0030	0029	0028	·0027	·0026	·0025	·0250
3·2	·0024	·0023	·0022	·0022	·0021	·0020	·0020	0019	·0018	·0018
3·3	·0017	·0017	·0016	·0016	·0015	·0015	·0014	·0014	·0013	·0013
3·4	·0012	·0012	·0012	·0011	·0011	·0010	·0010	·0010	·0009	·0009
3·5	0009	·0008	0008	·0008	·0008	·0007	0007	·0007	·0007	·0006
3·6	·0006	·0006	0006	·0005	0007	·0005	0005	·0005	0005	·0004
3·7	0004	·0004	·0004	·0004	0004	0004	0003	0003	0003	·0003
3·8	0003	·0003	0003	0003	0003	0002	0002	0002	0002	0002
3·9	0002	·0002	0002	0002	0002	0002	0002	0002	0001	0001

CRITICAL VALUES OF STUDENT'S *t*-DISTRIBUTION

d.f.	Level of significance for two-tailed test					d.f.
	0.20	0.10	0.05	0.02	0.01	
	Level of significance for one-tailed test					
	0.10	0.05	0.025	0.01	0.005	
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.731	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.703	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
Infinity	1.282	1.645	1.960	2.326	2.576	Infinity

Values of *F* for *F* Distribution at 5% Points

Degrees of freedom for numerator

	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Degrees of freedom for denominator

Values of F for F Distribution at 1% Points

Degrees of freedom for numerator

d.f.	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.68	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.30
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.04	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

PERCENTAGE POINTS OF X^2 DISTRIBUTION

d.f. \ α	.995	.990	.975	.950	.050	.025	.010	.005
1	392704×10^{-10}	157088×10^{-9}	982069×10^{-9}	393214×10^{-8}	3.84146	5.02389	6.63490	7.87944
2	.0100251	.0201007	.0506356	102587	5.99147	7.37776	9.21034	10.5966
3	.0717212	.114832	.215795	.351846	7.81473	9.34840	11.3449	12.8381
4	.206990	.297110	.484419	.710721	9.48773	11.1433	13.2767	14.8602
5	.411740	.554300	.831211	1.145476	11.0705	12.8325	15.0863	16.7496
6	.675727	.872085	1.237347	1.63539	12.5916	14.4494	16.8119	18.5476
7	.989265	1.239043	1.68987	2.16745	14.0671	16.0128	18.4753	20.2777
8	1.344419	1.646482	2.17973	2.73264	15.5073	17.5346	20.0902	21.9550
9	1.734926	2.087912	2.70039	3.32511	16.9190	19.0228	21.6660	23.5893
10	2.15585	2.55821	3.24697	3.94030	18.3070	20.4831	23.2093	25.1882
11	2.60321	3.05347	3.81575	4.57481	19.6751	21.9200	24.7250	26.7569
12	3.07382	3.57056	4.40379	5.22603	21.0261	23.3367	26.2170	28.2995
13	3.56503	4.10691	5.00874	5.89186	22.3621	24.7356	27.6883	29.8194
14	4.07468	4.66043	5.62872	6.57063	23.6848	26.1190	29.1413	31.3193
15	4.60094	5.22935	6.26214	7.26094	24.9958	27.4884	30.5779	32.8013
16	5.14224	5.81221	6.90766	7.96164	26.2962	28.8454	31.9999	34.2672
17	5.69724	6.40776	7.56418	8.67176	27.5871	30.1910	33.4087	35.7185
18	6.26481	7.01491	8.23075	9.39046	28.2693	31.5264	34.8053	37.1564
19	6.84398	7.63273	8.90655	10.1170	30.1435	32.8523	36.1908	38.5822

(*Continued*)

20	7.43386	8.26040	9.59083	10.8508	31.4104	34.1696	37.5662	39.9968
21	8.03366	8.89720	10.28293	11.5913	32.6705	35.4789	38.9321	41.4010
22	8.64272	9.54249	10.9823	12.3380	33.9244	36.7807	40.2894	42.7956
23	9.26042	10.19557	11.6885	13.0905	35.1725	38.0757	41.6384	44.1813
24	9.88623	10.8564	12.4011	13.8484	36.4151	39.3641	42.9798	45.5585
25	10.5197	11.5240	13.1197	14.6114	37.6525	40.6465	44.3141	46.9278
26	11.1603	12.1981	13.8439	15.3791	38.8852	41.9232	45.6417	48.2899
27	11.8076	12.8786	14.5733	16.1513	40.1133	43.1944	46.9630	49.6449
28	12.4613	13.5648	15.3079	16.9279	41.3372	44.4607	48.2782	50.9933
29	13.1211	14.2565	16.0471	17.7083	42.5569	45.7222	49.5879	52.3356
30	13.7867	14.9535	16.7908	18.4926	43.7729	46.9792	50.8922	53.6720
40	20.7065	22.1643	24.4331	26.5093	55.7585	59.3417	63.6907	66.7659
50	27.9907	29.7067	32.3574	34.7642	67.5048	71.4202	76.1539	79.4900
60	35.5346	37.4848	40.4817	43.1879	79.0819	83.2976	88.3794	91.9517
70	43.2752	45.4418	48.7576	51.7393	90.5312	95.0231	100.425	104.215
80	51.1720	53.5400	57.1532	60.3915	101.879	106.629	112.329	116.321
90	59.1963	61.7541	65.6466	69.1260	113.145	118.136	124.116	128.299
100	67.3276	70.0648	74.2219	77.9295	124.342	129.561	135.807	140.169

Critical Values of *T* In The Wilcoxon Matched Paired Test

n	Level of significance for one-tailed test .05	Level of significance for two-tailed test .10	Level of significance for one-tailed test .025	Level of significance for two-tailed test .05	Level of significance for one-tailed test .01	Level of significance for two-tailed test .02
5	1	1	-	-	-	-
6	2	2	1	1	-	-
7	4	4	2	2	0	0
8	6	6	4	4	2	2
9	8	8	6	6	3	3
10	11	11	8	8	5	5
11	14	14	11	11	7	7
12	17	17	14	14	10	10
13	21	21	17	17	13	13
14	25	25	21	21	16	16
15	30	30	25	25	20	20
16	35	35	30	30	24	24
17	40	40	35	35	28	28
18	46	46	40	40	33	33
19	52	52	46	46	38	38
20	59	59	52	52	43	43
21	66	66	59	59	49	49
22	73	73	66	66	56	56
23	81	81	73	73	62	62
24	89	89	81	81	69	69
25			89	89	77	77

					1	2
					2	4
					4	6
					6	8
					8	11
					11	14
					14	17
					17	21
					21	24
					24	28
					28	32
					32	36
					36	40
					40	44
					44	49
					49	54
					54	59
					59	64
					64	70
					70	75
					75	81
					81	87
					87	93
					93	99
					99	106
					106	112
					112	119
					119	126
					126	132
					132	139
					139	146
					146	153
					153	160
					160	167
					167	174
					174	182
					182	189
					189	197
					197	204
					204	212
					212	219
					219	227
					227	235
					235	242
					242	250

FACTORS USEFUL IN THE CONSTRUCTION OF CONTROL CHARTS

Sample size	Mean-chart		Standard deviation chart				Range chart			
	Factors for control limit		Factors for central line	Factors for control limits			Factors for central line	Factors for control limits		
n	A	A_1	A_2	C_2	B_1	B_2	B_4	d_2	D_1	D_2	D_3	D_4
2	2.121	3.760	1.880	0.5642	0	1.843	3.267	1.128	0	3.686	0	3.267
3	1.732	2.394	1.023	0.7236	0	1.858	2.568	1.693	0	4.358	0	2.575
4	1.500	1.880	0.729	0.7979	0	1.808	2.266	2.059	0	4.698	0	2.282
5	1.342	1.596	0.577	0.8407	0	1.756	2.089	2.326	0	4.918	0	2.115
6	1.225	1.410	0.483	0.8686	0.026	1.711	1.970	2.534	0	5.078	0	2.004
7	1.134	1.277	0.419	0.8882	0.105	1.672	1.882	2.704	0.205	5.203	0.076	1.924
8	1.061	1.175	0.373	0.9027	0.167	1.638	1.815	2.847	0.387	5.307	0.136	1.864
9	1.000	1.094	0.337	0.9139	0.219	1.609	1.761	2.970	0.546	5.394	0.184	1.816
10	0.949	1.028	0.308	0.9227	0.262	1.584	1.716	3.078	0.687	5.469	0.223	1.777
11		0.973	0.285	0.9300	0.299	1.561	1.679	3.173	0.812	5.534	0.256	1.744
12		0.925	0.266	0.9359	0.331	1.541	1.646	3.258	0.924	5.592	0.284	1.716
13		0.884	0.249	0.9410	0.359	1.523	1.618	3.336	1.026	5.646	0.308	1.692
14		0.848	0.235	0.9453	0.384	1.507	1.594	3.407	1.121	5.693	0.329	1.671
15		0.816	0.223	0.9490	0.406	1.492	1.572	3.472	1.207	5.737	0.348	1.652
16		0.788	0.212	0.9523	0.427	1.478	1.552	3.532	1.285	5.779	0.365	1.636
17		0.762	0.203	0.9551	0.445	1.465	1.534	3.588	1.359	5.817	0.379	1.621
18		0.738	0.194	0.9576	0.461	1.454	1.518	3.640	1.426	5.854	0.401	1.608
19		0.717	0.187	0.9599	0.477	1.443	1.503	3.689	1.490	5.888	0.404	1.596
20		0.697	0.180	0.9619	0.491	1.433	1.490	3.735	1.548	5.922	0.414	1.585
21		0.679	0.173	0.9638	0.504	1.424	1.477	3.778	1.606	5.950	0.425	1.575
22		0.662	0.167	0.9655	0.516	1.415	1.466	3.819	1.659	5.979	0.434	1.565
23		0.647	0.162	0.9670	0.527	1.407	1.455	3.858	1.710	6.006	0.443	1.557
24		0.633	0.157	0.9684	0.538	1.399	1.445	3.895	1.759	6.031	0.452	1.548
25		0.619	0.153	0.9696	0.548	1.392	1.435	3.931	1.804	6.058	0.459	1.541

STATISTICS FOR M.B.A. STUDENTS

APPENDIX

Descriptive Questions (Very Short Answer/Short Answer Questions)

DATA ANALYSIS

MEASURES OF CENTRAL TENDENCY

1. Define an average or a measure of central tendency.
2. What is the significance of studying average ?
3. List the characteristics (or properties) of a good average
4. What are the essentials of an ideal average ?
5. Give various measures of central tendency.
6. Write down two mathematical properties of arithmetic mean.
7. Define mean. Give its merits.
8. Explain Weighted Arithmetic Mean.
9. Explain : Median.
10. Explain : Mode.
11. Explain : Quartiles. and Percentile.
12. Explain Geometric Mean.
13. Find the geometric mean of 1, 4 and 9.
14. Explain Harmonic Mean.
15. Explain the relationship between meen, median and mode.
16. Find the missing figure : $M = Z + ?\ (\bar{X} - Z)$.
17. Show graphicaly the position of $\bar{X}$, M and Z in a positively and negatively skewed curves.
18. What is the relationship between $\bar{X}$, M and Z in a symertical distribution ?
19. Explain the relationship between $\bar{X}$, M and Z.

MEASURES OF DISPERSION

20. What is meant by dispersion ?
21. List the characteristics of a good measure of dispersion.
    or What are the essentials of a good measure of variation ?
22. Distinguish between central value and dispersion.
    or Distinguish between measures of central tendency and dispersion.
23. Distinguish between absolute and relative measure of dispersion.
24. State the various methods of measuring dispersion or state the various measures of dispersion known to you.
25. What are the merits of dispersion. *or* What are objects of measuring dispersion ?

26. What is standard deviation ?
27. Why standard deviation is considered a better method of variation as compared to mean deviation ?
28. Distinguish between mean deviation and standard deviation.
29. Define (i) Range and (ii) Quartile Deviation.
30. Define (i) Mean Deviation and (ii) Standard Deviation.
31. What is coefficient of variation ? What purpose does it serve ?
32. If a constant is subtracted from each score in a series, what will be its effect on mean and standard deviation ?

SKEWNESS

33. What is skewness ?
34. Distinguish between positive and negative skewness.
35. What is the significance of skewness ?
36. Distinguish between dispersion and skewness.
37. What are the various tests of skewness ?
38. State various measures of skewness.
39. What is Bowley measure of skewness ?
40. Mention the formulae of Karl Pearson and Bowley for the coefficient of skewness.

MOMENTS AND KURTOSIS

41. Define moments.
42. Distinguish between central moments and raw moments.
43. Define Kurtosis. Give its three types used diagrams.
44. Distinguish between skewness and kurtosis. or Define (i) Skewness and (ii) Kurtosis.
45. Give formula of measuring kurtosis.
46. Give the measures of skewness and kurtosis by using moments.

CORRELATION AND REGRESSION

CORRELATION

1. What is correlation ? What is its significance ?
2. Explain positive and negative correlation.
3. Does correlation always signify cause and effect relationship between the variables ?
4. Define covariance.
5. Define Karl Pearson's Coefficient of Correlation.
6. State the properties of Karl Pearson's Coefficient of Correlation. (or Mention any two properties of coefficient of correlation).
7. What is the maximum and minimum value of the coefficient of correlation ? or What are the limits of the coefficient of correlation ?
8. What is the nature of correlation when the value of r is +1 and –1.

or If r = + 1 and r = – 1, what kind of relationship exist between x and y ?

9. State the formula for calculating Karl Pearson's coefficient of correlation if the deviations are taken from (a) actual means and (b) assumed means.
10. Name two methods of studying correlation.
11. What does the value of r = 0 imply ?
12. Define rank correlation coefficient.

or

Give the formula to calculate rank correlation coefficient with (a) non-repeated ranks and (b) repeated ranks.

13. Define concurrent correlation. *or* Write a short note on scatter diagram.
14. Define (a) Probable Error, and (b) Standard Error.

REGRESSION

15. Distinguish between correlation and regression.
16. What is regression ? *or* Define regression and explain its significance.
17. Why are there two regression lines in general ?
18. Where do the two lines of regression of X on Y and Y on X cross each other ?
19. Under what conditions the two regression lines (a) concide and (b) intersect each other at 90° ?

or Under what condition there will be one regression line ?

20. What is the nature of regression lines when (i) r = +1 and (ii) r = 0.
21. What are regression coefficients ?
22. What is the significance of regression coefficient ?
23. State the important properties of regression coefficients.
24. Explain the relationship between correlation coefficient and regression coefficients ?

or Mathematicaly, prove that $r = \sqrt{byx \cdot bxy}$

25. If $b_{yx} = 1.5$ and $b_{xy} = 0.2$, find correlation coefficient.
26. If $b_{yx} = -3/2$ and $b_{xy} = -1/6$, what is the value of correlation coefficient ?
27. State the uses of regression or what is the significance of regression ?
28. Write a brief note on standard error of estimate.
29. If the two lines of regression are : $4x - 5y + 30 = 0$ and $20x - 9y - 107 = 0$, which of these is the line of regression of x on y ?
30. From the following regression equations : $20x - 9y = 107$, $4x - 8y = -33$, calculate $\bar{X}$ and $\bar{Y}$.
31. Given the regression equation of y on x and x on y as $y = x$ and $4x - y = 3$, find 'r'.
32. Given two lines of regression, explain how will you find the values of $\bar{X}$ and $\bar{Y}$.

INDEX NUMBERS & TIME SERIES

Index Numbers

1. What are Index numbers ?
2. What are the uses of Index numbers ?
3. Point any two limitations of Index numbers.
4. Distinguish between weighted and unweighted Index numbers.
5. Mention any two problems in the construction of an Index numbers.
6. What are the desirable properties of the base period ?

   or

   How will you choose a base year for constucting Index numbers ?
7. Distinguish between Laspreye's and Paasche's Index.

   or

   Define (i) Laspreye's Index and (ii) Paasche's Index.
8. What is consumer Price Index ? What is its significance ?
9. What is the difference between price Index and quantity Index ?
10. Give the formula of constructing weighted Index using (i) Fisher's method and (ii) Weighted average of relative method.
11. What is Fisher's ideal index ? Why is it called an ideal ?
12. Explain : Chain Base Index.

    or

    Explain how chain base index are constructed ?
13. Distinguish between CB I and FBI.
14. Write formulae to convert (i) CBI into FBI and (II) FBI into CBI.
15. Why are Index numbers called economic barometers ?
16. Explain the meaning of (i) Base shifting (ii) Splicing and (iii) Deflating
17. Explain the meaning of Splicing of Index Numbers.
18. What is an ideal index number ? What properties should it have ?
    or List the characteristics of an ideal Index number.
    or Explain : (i) Time Reversal Test, (ii) Factor Reversal Test and (iii) Circular Test.
19. Show how Fisher's formula of Index numbers satisfy TRT and FRT.
20. Define deflating of Index number.
21. Explain deflating. How real wages Index are computed ?
22. What are value Index numbers ?
23. How many types are the price Index numbers ?
24. Write Kelley's formula of construction of Index Number.

ANALYSIS OF TIME SERIES

25. What is time series ? Discuss its importance or utility.
26. What are components of time series ? Define any one of them.
27. What is a time series ? What are its main components ?
28. What is secular trend.

    or

    What is meant by trend in a time seris ?
29. Exlain Linear Trend.
30. What are seasonal variations ?
31. What are cyclical variations or cyclical fluctuations ?
32. What are irregular variations ?
33. State additive and multiplicative models of analysing time series.
34. What is moving average method ?
35. What is semi-averge method ?
36. Distinguish between secular trend and periodic variations.
37. Distinguish between seasonal variations and cyclical variations.
38. Explain cyclical and irregular variations.
39. How would you measure trend by the method of least squares ?
40. (a) Write the normal equations to determine the value of a and b in the trend equation $y = a + bx$, given the n observations.
    (b) You are given the following trend equation : $Y = 45 + 5X$ (origin = 1990, X unit = 1 year) Shift the origin to (i) 1988 & (ii) 1993.
41. With what characteristic component of a time series should each of the following be associated ?
    (i) A fire in a factory delaying production for three weeks.
    (ii) Arena of prosperity.
    (iii) Sales of a textile firm during Deepawali.
    (iv) A need for increased wheat production due to constant increase in population.

PROBABILITY & PROBABILITY DISTRIBUTIONS

PROBABILITY

1. Define Probability.
2. Give classical definition of probability.
3. Give statistical definition of probability.
4. Define (i) Mutually exclusive events (ii) Independent events (iii) Dependent Events and (iv) Equally likely events (v) Non-mutually exlusive events
5. Define Joint Probability.
6. State addition theorem of probability.
7. State multiplication theorem of probabilty.

8. Explain the concept of conditional probability.
9. Give the statement of Bayes' Theorem.
10. State the addition theorem of probability for two events which are (a) mutually exclusive and (b) non-mutually exclusive.
11. State the multiplication theorem of probabilty for two events which are (a) independent and (b) non-independent.
12. Write the formula for the calculation of probability at least one event in case of independent events.
13. State the axioms of probability.
14. What is mathematical expectation of a random variable ?
15. Define random variable and its expectation.

PROBABILITY DISTRIBUTIONS.

16. What is Binomial Distribution ?
17. Give properties of binomial distribution.
18. Is there any fallacy in the statement : The mean of Binomial Distrubution is 20 and its standard deviation is 7 ?
19. Discuss the conditions for the applications of Binomial Distribution.
20. A binomial distribution has n = 20 and p = 0.3. What are the mean and variance of the distribution ?
21. The mean of the Binomial Distrubution is 20 and standard deviation is 4. Calculate n, p and q.
22. What is Poisson Distribution ?
23. Give the properties *or* characteristics of Poisson Distribution.
24. State the conditions under which the Binomial Distribution tends to Poisson distribution
25. Give six examples where Poisson Distrubution can be applied.
26. Write the probability function of Binomial and Poisson Distributions.
27. To which probabilty distribution, mean and variance are equal ?
28. Comment on the following : For a Poisson distribution, Mean = 8, Variance = 7.
29. Define Poisson distribution and state the conditions under which this distribution is to be used.
30. What is normal distribution ?

or

Write the p.d.f. of General and Standard Normal Distribution.

31. Explain the main properties of normal curve/normal distribution.

or Give the chief characteristics of Normal Distribution.

32. Give the applications of Normal Distribution.
33. Give the area property of normal curve or normal distribution.
34. Indicate the area of normal distribution covered by : (i) $\bar{X} \pm 1\sigma$ (ii) $\bar{X} \pm 2\sigma$ (iii) $\bar{X} \pm 3\sigma$.
35. Under what conditions Poisson Distribution tends to normal distribution ?
36. Under what conditions Binomial Distribution will tend to normal distribution ?
37. How does normal distribution differ from binomial distribution ?

SAMPLING THEORY

SAMPLING AND SAMPLING METHODS

1. Explain the meaning of (i) Population (or universe) and (ii) Sample
2. Distinguish between Census and sample methods.
3. What is sampling ?
4. Point out two uses of sampling.
5. Name any four well known methods of sampling. Explain anyone of them.
6. Define simple random sampling ?
7. What is stratified random sampling ?
8. State the situation where stratified random sampling is preferred to simple random sampling.
9. What is systematic sampling ?
10. Distiniguish between simple random sample and stratified sampling.
11. Define random sampling and discuss various methods of selecting a random sampling.
12. Point out any two advantages of sampling.
13. Point out any two limitations of sampling.
14. Discuss the two advantages of sampling over census method.
15. Explain the following : (i) Random Sampling (ii) Cluster Sampling and (iii) Deliberate Sampling.
16. What is sampling error ?
17. Explain the meaning of sampling error and non-sampling error or Distinguish between sampling and non-sampling errors.

SAMPLING DISTRIBUTION AND STANDARD ERROR

18. Distiniguish between Statistics and Parameters.
19. Explain the concept of sampling distribution of a statistic.
20. Define standard error of a stastic.
21. Define standard error of mean.
22. State Central Limit Theorem.
23. State Law of Large Numbers.
24. Find the number of all possibile samples of size n = 4 from a population of size N=8 where (a) sampling is with replacement and (b) sampling is without replacement.

INFERENTIAL STATISTICS

THEORY OF ESTIMATION

1. What is inferential statistics ? *or* What is estimation
2. Distinguish between point estimation and interval estimation.
3. Explain the concept of interval estimation.
4. Explain the conept of point estimation.
5. Explain the important properties of a good estimator. or State the properties of a good estimator.

TESTS OF HYPOTHESIS.

6. Explain the procedure for testing a hypothesis.
7. Distinguish between large and small samples.
8. Distinguish between null hypothesis and alternative hypothesis.
9. Define Type I and Type II errors in testing of hypothesis.
10. Point out the difference between one tailed and two tailed tests.
11. Explain the term : Acceptance and Rejection (or critical) region.

or

Distinguish between critical region and acceptance region.

12. Explain the term level of significance as used in the tests of significance.
13. Explain the term 'degrees of freedom'.

PARAMETRIC TESTS

14. Define Fisher's Z-transformation.
15. Define t-test.
16. What is F-test ? *or* F-test of testing of hypothesis ?
17. Explain the procedure for testing of the hypothesis concerning the difference between two population proportions based on samples taken from each of two population.
18. What are the assumptions of ANOVA ? Explain briefly.
19. Distinguish between paired t-test and t-test for independent samples.

Or

Write a brief note on paired t-tests.

20. Describe the large sample testing procedure.
21. Explain the procedure in testing equality of two means through t-test.

Non-Parametric Tests

22. Define non-parametric tests or what are non-parametric tests ?
23. Name two non-parametric test.
24. Explain sign test or what is a sign test ?
25. Define Mann-Whitney U-test.
26. What is χ^2 test of independence of attributes.
27. Describe Yates' corrections of 2 x 2 contingency table.
28. What is χ^2-test of goodness of fit ?
29. What is Chi-square test. Explain the uses to which χ^2-test can be applied.
30. What are the conditions for the validity of Chisquare test ?
31. Distinguish between parametric and non-parametric tests.

STATISTICAL QUALITY CONTROL

1. What is statistical quality control ?
2. What is a control chart ?
3. What is an O.C. curve ?
4. How are control limits set up in C-Chart ?
5. What is acceptance sampling ?
6. What is the utility of statistical quality control ?

KURUKSHETRA UNIVERSITY, KURUKSHETRA

Business Statistics

Paper: CP-102

MBA

Ist Semester (Dec./Jan. 2009-10)

Time: Three Hours

Max. Marks: 70

Note: Attempt any five questions in all. Question No. 1 is compulsory. All questions carry equal marks.

1. Write brief explanation of the following: 2×7
   (i) Explain inferential statistics.
   (ii) Discuss the utility of diagrammatic presentation.
   (iii) Which is the best average for the manufacturer of garments?
   (iv) Distinguish between linear and curvilinear correlation.
   (v) Central Limit Theorem.
   (vi) Level of Significance.
   (vii) Objectives of measuring trend.
2. Suppose that samples of polythene bags from two manufactures, A and B are tested by a prospective buyer for bursting pressure with the following results:

Bursting Pressure (Lbs)		5.0-9.9	10.0-14.9	15.0-19.9	20.0-24.9	25.0-29.9	30.0-34.9	Total
Number of Bags	A	2	9	29	54	11	5	110
	B	9	11	18	32	27	13	110

Which set of bags has the highest average bursting pressure? Which has more uniform pressure? If prices are same, which manufacture's bags would be preferred by the buyer? Why? 14

3. An investment consultant predicts that the odds against the prices of a certain stock going up are 2:1 and the odds in favour of the prices remaining the same are 1:3. What is the probability that price of the stock will go down? 14
4. Define Poisson distribution and state the conditions under which this distribution is used for solving business problems. 14
5. (a) Discuss briefly the importance of estimation theory in decision making in the face of uncertainty.
   (b) Explain the regression coefficients. 7+7=14
6. XYZ physical fitness centre claims that completion of their weight loss programme will result in a weight loss. To test this claim, six persons were selected at random and they were put through the weight loss programme and their weights, before and after the programme, were recorded. Test the claim of the fitness centre at $\alpha = 0.05$. The weights in pounds of these six persons recorded before and after the programme are as follows:

Person	Weight (before) (in pounds)	Weight (after) (in pounds)
1	145	143
2	200	190
3	160	165
4	185	183
5	164	160
6	175	176

14

7. The following table gives the cost of living index numbers for different groups with their respective weights for the year, 1992. (base year: 1982)

Group	Cost of Living Index	Weight
Food	525	40
Clothing	325	16
Light and Fuel	240	15
Rent	180	20
Others	200	9

Calculate the overall cost of living Index Number. Mr. Bose got a salary of Rs. 550 in 1982. Determine how much he should have to receive as salary in 1992 to maintain his same standard of living as in 1982. 14

8. The following are the mean lengths and ranges of lengths of a finished product from 10 samples each of size 5. The specification limits for lengths are 200 ± 5 cm. Construct $\bar{X}$ and R charts and examine whether the process is under control and state your recommendations.

Sample No.	1	2	3	4	5	6	7	8	9	10
Mean $\bar{X}$	201	198	202	200	203	204	199	196	199	201
Range R	5	0	7	3	4	7	2	8	5	6

Assume for $n = 5, A_2 = 0.577, D_3 = 0$ and $D_4 = 2.115$. 14

MAHARISHI DAYANAND UNIVERSITY, ROHTAK

Quantitative Analysis

Paper: 2104

MBA

Ist Semester, Dec./Jan. 2009-10

Time: Three Hours

Max. Marks: 70

Note: Attempt any five questions, selecting at least one question from each unit. All questions carry equal marks.

Unit - I

1. Calculate mean, median and mode from the following data:

Marks	No. of Students	Marks	No. of Students
10-20	4	10-60	124
10-30	16	10-70	137
10-40	56	10-80	140
10-50	95	10-90	150

14

2. Define skewness. Explain briefly the different methods of measuring skewness. 14

Unit - II

3. From the following information, calculate the line of regression of Y on X:

	X	Y
Mean	40	60
Standard Deviation (SD)	10	15
Correlation Coefficient	(r) = 0.7	

14

4. Fit a straight line trend by method of least squares to the following data:

Years:	1990	1991	1992	1993	1994	1995	1996	1997
Sales:	38	40	65	72	69	67	95	104

14

PUNJAB TECHNICAL UNIVERSITY, JALANDHAR

Quantitative Techniques

MBA

Ist Semester, Dec. 2009

Time: Three Hours Max. Marks: 60

Instructions to candidates:

(i) Section-A is compulsory

(ii) Attempt any Four questions from Section-B

Section -A

1. (a) Prove that $\log 2 + 2\log 5 - \log 3 - 2\log 7 = \log \frac{50}{147}$

   (b) Find the sum of first 35 terms of an A.P. if $t_2 = 2$ and $t_7 = 22$.

   (c) Find Geometric mean of 2, 4, 6, 8, 10.

   (d) Find Mode of the data: 3, 6, 9, 12, 15, 18, 21, 12, 9, 15, 12, 6, 15.

   (e) Give formula of Rank's coefficient of correlation.

   (f) Show that the coefficient of correlation is G.M. of coefficient of regression.

   (g) Give different ways in which index numbers can be constructed.

   (h) What is the chance that a leap year will have 53 Mondays.

   (j) Give 5 properties of Normal distribution.

   (j) Briefly explain F-test.

2. (a) A machine depreciates in value in a year by 6% of its value at the beginning of the year. If value of new machine is Rs. 62,500, using logarithms, find its depreciated value after 7 years.

   (b) If α and β are the roots of $2x^2 - 3x - 6 = 0$, find the equation whose roots are $\alpha^2 + 2$ and $\beta^2 + 2$.

3. (a) Find the 7th term in the expansion of $\left(3X^2 - \frac{1}{X^3}\right)$

   (b) Find mean and standard deviation for the data:

Class:	0-7	7-14	14-21	21-28	28-35	35-42	42-49
Frequency:	19	25	36	72	51	43	28

4. (a) Calculate first four moments about mean of the distribution:

X:	2.0	2.5	3.0	3.5	4.0	4.5	5.0
f:	5	38	65	92	70	40	10

Also calculate β_1 and β_2.

(b) Calculate the coefficient of correlation between X and Y:

X:	1	3	5	7	8	10
Y:	8	12	15	17	18	20

5. (a) If θ is the acute angle between two regression lines in case of two variables x and y, show that $\tan\theta = \frac{1-r^2}{r} \times \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$, r, σ_x, σ_y have their usual meaning.

   (b) What are trend values? Fit a trend line by method of least squares to the following data and obtain trend values.

Year:	1941	1942	1943	1944	1945	1946	1947
Sales ('000 Rs.):	80	90	92	83	94	99	92

6. (a) Estimate changes in cost of living figures of 1992 as compared to 1991.

Expenses on	Food 35%	Rent 15%	Clothing 20 %	Fuel 10%	Miscellaneous 20%
Prices 1991:	1500	300	750	250	400
Prices 1992:	2000	300	650	230	450

   (b) A can hit a target 4 times in 5 shots, B 3 times in 4 shots and C 2 times in 3 shots. They fire valley. What is the probability that at least 2 shots hit the target?

7. (a) Fit Poisson's distribution and calculate theoretical frequencies:

Deaths:	0	1	2	3	4
Frequencies:	122	60	15	2	1

   (b) 5 dice were thrown 96 times and numbers 4, 5 or 6 were thrown as given below:

No. of dice throwing 4, 5, or 6:	5	4	3	2	1	0
f:	7	19	35	24	8	3

Calculate χ^2.

KURUKSHETRA UNIVERSITY, KURUKSHETRA

BUSINESS STATISTICS

MBA

Semester–I

2012

Time Allowed: 3 Hours] **[Maximum Marks: 70**

Note: Attempt ***eight*** questions from **Part-A** of 5 marks each and three questions of 10 marks each from **Part-B**.

PART–A

1. List out applications of probability in business decision making.
2. Explain with examples difference between classical approach and relative frequency approach of probability.
3. Explain with example multiplication probability model.
4. What is Central Limit Theorem?
5. Explain the terms 'Sampling distribution' and 'standard error' of a statistic.
6. Explain cluster sampling and simple random sampling techniques.
7. Describes the difference between Parametric Test and Non-Parametric Test. Also explain Wilcoxon signed Test.
8. Briefly explain statistical estimation.
9. Describes the use of Microsoft Excel in data analysis.
10. If two dice are thrown, what is the probability that the some of the numbers on the dice is (i) Greater than 8 and (ii) Neither 7 nor 11?

PART–B

11. Probability that a man will be alive 25 years hence is 0.3 and the probability that his wife will be alive 25 years hence is 0.4. Find the Probability that 25 years hence.

    (i) Both will be alive

    (ii) Only the man will be alive

    (iii) Only the woman will be alive

    (iv) none will be alive

    (v) At least one of them will be alive.

12. The average daily sales of 500 branch offices was ₹ 150 thousands and the standard deviation ₹ 15 thousand. Assuming the distribution to be normal, indicate how many branches have sales between.

    (i) ₹ 120 thousands and ₹ 145 thousand.

    (ii) ₹ 140 thousands and ₹ 165 thousand.

13. State briefly the reasons for the increasing popularity of sampling methods. Explain briefly and two methods of sampling which help us to obtain a representative sample.

14. A daily sample of 30 times was taken over a period of 14 days in order in establish attributes control limits. If 21 defective were found, what should be the upper and lower control limits of the proportion of defective ?

15. It is found that 35 of 250 housewives in Delhi, 22 of 220 housewives in Mumbai and 39 of 300 housewives in Chandigarh watch at least one talk show everyday. At the 0.05 level of significance, test that there is no difference between the true proportions of housewives who watch talk show in these cities.

KURUKSHETRA UNIVERSITY, KURUKSHETRA

BUSINESS STATISTICS

MBA

Semester–I

2013

Time Allowed: 3 Hours] **[Maximum Marks: 70**

Note: Attempt any ***eight*** questions from **Part-A** of 5 marks each and three questions of 10 marks each from **Part-B**.

PART–A

1. What is the importance of Probability in business decision making ?
2. State the Multiplicative theorem of Probability.
3. Explain Bayes's theorem.
4. What do you mean by Non-sampling error?
5. What is Central Limit Theorem ?
6. Define Statistical estimation.
7. In test of hypothesis, how p-value is interpreted ?
8. Explain the concept of Standard error.
9. What do you understand by Non-parametric methods ?
10. Explain how hypothesis testing is useful to decision-makers.

PART–B

11. How does a Normal distribution differ from Binomial distribution ? What are the important properties of normal distribution and how are they useful in business decision-making ?
12. How would you plan a survey to study the employment pattern of MBA students of your university? Draft a Questionnaire giving at least 10 questions.
13. A production supervisor is interested in knowing if number of breakdowns on four machines is independent of the shift using the machines. Test this hypothesis based on the following sample information:

Shift	Machine			
	A	B	C	D
Morning	15	10	18	12
Evening	12	8	15	10

14. A company manufactures tyres. A quality control engineer is responsible to ensure that the tyres turned out are fit for use up to 40,000 km. He monitors the life of the output from the production process. From each of the 10 batches of 900 tyres, he has tested 5 tyres and recorded the following data, with and measured in thousands of km.

Batch	1	2	3	4	5	6	7	8	9	10
$\bar{X}$	40.2	43.1	42.4	39.8	43.1	41.5	40.7	39.2	38.9	41.9
$\bar{R}$	1.3	1.5	1.8	0.6	2.1	1.4	1.6	1.1	1.3	1.5

Construct an chart using the above data. Do you think that the production process is in control? Explain.

15. Write a detailed note on SPSS for the purpose of descriptive analysis of the data.

KURUKSHETRA UNIVERSITY, KURUKSHETRA

BUSINESS STATISTICS

MBA

Semester–I

2014

Time Allowed: 3 Hours **Maximum Marks: 70**

Note: Attempt any **eight** questions from Part A of 5 marks each and **three** questions of 10 marks each from Part B.

PART–A

1. Explain the use of probability distributions in Business decision-making.
2. Describe Addition Probability theorem by giving example.
3. Explain Baye's theorem with example.
4. What are sampling errors and non-sampling errors?
5. Explain meaning and characteristics of sampling distribution of sample mean.
6. Show the difference between Point estimation and Interval estimation of Population mean.
7. Write a note on Kruskal-Wallis test.
8. Write down properties and applications of T-Test and F-Test.
9. What are the uses of SPSS software in Data analysis?
10. Explain the purpose and logic of constructing Quality Control Charts.

PART–B

11. Explain probability sampling methods and non-probability sampling methods.
12. Suppose the waist measurements W of 800 Girls are normally distributed with mean 66 cms and standard deviation 5 cms. Find the number N of Girls with waists:

    (a) Between 65 and 70 cms

    (b) Greater than or equal to 72 cms.

13. A problem in Statistics is given to two Students A and B. The odds in Favour of A solving the problem are 6 to 9 against B solving the problem are 12 to 10. If both A and B attempt, find the probability of the problem being solved.

14. Two Researchers adopted different sampling techniques while investigating the same group of Students to find the number of students falling in different intelligence levels. The results are as follows:

Researcher	No. of Students is each level				Total
	Below average	Average	Above average	Genuine	
X:	86	60	44	10	200
Y:	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two Researchers are significantly different? (Given 5% values of X^2 for 3 d.f. and 4 d.f. are 7.82 and 9.49 respectively).

15. An inspection of 10 samples of size 400 each from 10 lots revealed the following number of defective units:

17, 15, 14, 26, 9, 4, 19, 12, 9, 15

Calculate control limits for the number of defective units. Plot the control limits and the observations and state whether the process is under control or not.

KURUKSHETRA UNIVERSITY, KURUKSHETRA

BUSINESS STATISTICS

MBA

2015

Time: 3 Hours **Maximum Marks: 70**

Note: *Attempt any **five** questions in all. Question No. 1 is compulsory. All questions carry equal marks.*

Compulsory Questions

1. Explain briefly of the following:
   (a) Descriptive Statistics.
   (b) Mutually Exclusive events
   (c) Conditional Probability
   (d) Range
   (e) Random sampling
   (f) Type-II error
   (g) Concept of Splicing.
2. Explain the concept of Conditional Probability. Also give proof of Baye's theorem. 14
3. Explain various sampling methods used for Data collection. Also discuss sampling and non-sampling errors. 14
4. Write notes on the following:
   (a) Interval Estimation
   (b) Statistical Quality Control 14
5. Write notes on the following:
   (a) Differentiate Correlation and Regression. Also write the properties of Regression coefficients.
   (b) Write a note on Coefficient of Determination. 14
6. Samples of two different types of bubls were tested for length of life and the following data were obtained:

	Type I	Type II
Sample Size	8	7
Sample Mean	1234 hrs.	1136 hrs.
Sample S.D.	36 hrs.	40 hrs.

Is the difference in Mean life of two bubls significant? 14

7. Write notes on the following:

   (a) Index numbers, uses and problems of Index numbers?

   (b) Method of moving average for the determination of trend in a time series. 14

8. Comprehensively explain various types of Control Charts related to Variables and Attributes. 14