UNIT 3

Statistical Techniques

3.1 MOMENTS

Moments are statistical tools, used in statistical investigations. The moments of a distribution are the arithmetic means of the various powers of the deviations of items from some given number.

3.2 MOMENTS ABOUT MEAN (Central Moments)

3.2.1. For an Individual Series

If $x_1, x_2, ..., x_n$ are the values of the variable under consideration, the r^{th} moment μ_r about mean $\bar{x}$ is defined as

$$\mu_r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^r}{n} \; ; r = 0, 1, 2, ...$$

3.2.2. For a Frequency Distribution

If $x_1, x_2, ..., x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, ..., f_n$ respectively then r^{th} moment μ_r about the mean $\bar{x}$ is defined as

$$\mu_r = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^r}{N} \; ; r = 0, 1, 2, ... \quad \text{where} \quad N = \sum_{i=1}^{n} f_i$$

In particular, $$\mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x})^0 = \frac{1}{N}\sum_{i=1}^{n} f_i = \frac{N}{N} = 1$$

$\therefore$ For any distribution, $\boxed{\mu_0 = 1}$

For $r = 1$,

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i (x_i - \bar{x}) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - \bar{x}\left(\frac{1}{N}\sum_{i=1}^{n} f_i\right) = \bar{x} - \bar{x} = 0$$

$\therefore$ For any distribution, $\boxed{\mu_1 = 0}$

For $r = 2$,

$$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2 = (\text{S.D.})^2 = \text{Variance}$$

$\therefore$ For any distribution, μ_2 coincides with the variance of the distribution.

Similarly, $$\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^3 \,,\ \mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^4$$

and so on.

Note. In case of a frequency distribution with class intervals, the values of x are the mid-points of the intervals.

EXAMPLES

Example 1. *Find the first four moments for the following individual series:*

x	*3*	*6*	*8*	*10*	*18*

Sol. **Calculation of Moments**

S. No.	x	$x - \bar{x}$	$(x - \bar{x})^2$	$(x - \bar{x})^3$	$(x - \bar{x})^4$
1	3	– 6	36	– 216	1296
2	6	– 3	9	– 27	81
3	8	– 1	1	– 1	1
4	10	1	1	1	1
5	18	9	81	729	6561
$n = 5$	$\Sigma x = 45$	$\Sigma(x - \bar{x}) = 0$	$\Sigma(x - \bar{x})^2 = 128$	$\Sigma(x - \bar{x})^3 = 486$	$\Sigma(x - \bar{x})^4 = 7940$

Now, $$\bar{x} = \frac{\Sigma x}{n} = \frac{45}{5} = 9$$

$\therefore$ $$\mu_1 = \frac{\Sigma(x - \bar{x})}{n} = \frac{0}{5} = 0, \qquad \mu_2 = \frac{\Sigma(x - \bar{x})^2}{n} = \frac{128}{5} = 25.6$$

$$\mu_3 = \frac{\Sigma(x - \bar{x})^3}{n} = \frac{486}{5} = 97.2, \qquad \mu_4 = \frac{\Sigma(x - \bar{x})^4}{n} = \frac{7940}{5} = 1588.$$

Example 2. *Calculate μ_1, μ_2, μ_3, μ_4 for the following frequency distribution:*

Marks	*5–15*	*15–25*	*25–35*	*35–45*	*45–55*	*55–65*
No. of students	*10*	*20*	*25*	*20*	*15*	*10*

Sol. **Calculation of Moments**

Marks	*No. of students (f)*	*Mid-point (x)*	*fx*	$x - \bar{x}$	$f(x - \bar{x})$	$f(x - \bar{x})^2$	$f(x - \bar{x})^3$	$f(x - \bar{x})^4$
5–15	10	10	100	– 24	– 240	5760	– 138240	3317760
15–25	20	20	400	– 14	– 280	3920	– 54880	768320
25–35	25	30	750	– 4	– 100	400	– 1600	6400
35–45	20	40	800	6	120	720	4320	25920
45–55	15	50	750	16	240	3840	61440	983040
55–65	10	60	600	26	260	6760	175760	4569760
	N = 100		Σfx = 3400		$\Sigma f(x - \bar{x})$ = 0	$\Sigma f(x - \bar{x})^2$ = 21400	$\Sigma f(x - \bar{x})^3$ = 46800	$\Sigma f(x - \bar{x})^4$ = 9671200

Now, $\bar{x} = \dfrac{\Sigma fx}{N} = \dfrac{3400}{100} = 34$

$\therefore$ $\mu_1 = \dfrac{\Sigma f(x - \bar{x})}{N} = \dfrac{0}{100} = 0,$ $\mu_2 = \dfrac{\Sigma f(x - \bar{x})^2}{N} = \dfrac{21400}{100} = 214$

$\mu_3 = \dfrac{\Sigma f(x - \bar{x})^3}{N} = \dfrac{46800}{100} = 468,$ $\mu_4 = \dfrac{\Sigma f(x - \bar{x})^4}{N} = \dfrac{9671200}{100} = 96712.$

3.3 SHEPPARD'S CORRECTIONS FOR MOMENTS

While computing moments for frequency distribution with class intervals, we take variable x as the mid-point of class-intervals which means that we have assumed the frequencies concentrated at the mid-points of class-intervals.

The above assumption is true when the distribution is symmetrical and the no. of class-intervals is not greater than $\dfrac{1}{20}^{\text{th}}$ of the range, otherwise the computation of moments will have certain error called **grouping error.**

This error is corrected by the following formulae given by **W.F. Sheppard.**

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

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{2}h^2\mu_2 + \frac{7}{240}h^4$$

where h is the width of the class-interval while μ_1 and μ_3 require no correction.

These formulae are known as **Sheppard's corrections.**

Example 3. *Find the corrected values of the following moments using Sheppard's correction. The width of classes in the distribution is 10:*

$\mu_2 = 214,$ $\mu_3 = 468,$ $\mu_4 = 96712.$

Sol. We have $\mu_2 = 214,$ $\mu_3 = 468,$ $\mu_4 = 96712,$ $h = 10.$

Now, $\mu_2 \text{ (corrected)} = \mu_2 - \dfrac{h^2}{12} = 214 - \dfrac{(10)^2}{12} = 214 - 8.333 = 205.667.$

$\mu_3 \text{ (corrected)} = \mu_3 = 468$

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{2}h^2\mu_2 + \frac{7}{240}h^4 = 96712 - \frac{(10)^2}{2}(214) + \frac{7}{240}(10)^4$$
$$= 96712 - 10700 - 291.667 = 86303.667.$$

3.4 MOMENTS ABOUT AN ARBITRARY NUMBER (Raw Moments)

If $x_1, x_2, x_3, ..., x_n$ are the values of a variable x with the corresponding frequencies $f_1, f_2, f_3, ..., f_n$ respectively then r^{th} moment μ_r' about the number $x = A$ is defined as

$$\mu_r' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^r \; ; r = 0, 1, 2,... \quad \text{where,} \quad N = \sum_{i=1}^{n} f_i$$

For $r = 0$, $\mu_0' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^0 = 1$

For $r = 1$, $\mu_1' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A) = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - \frac{A}{N}\sum_{i=1}^{n} f_i = \bar{x} - A$

For $r = 2$, $\mu_2' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^2$

For $r = 3$, $\mu_3' = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^3$ and so on.

In calculation work, if we find that there is some common factor h (> 1) in values of $x - A$, we can ease our calculation work by defining $u = \frac{x - A}{h}$. In that case, we have

$$\mu_r' = \frac{1}{N}\left(\sum_{i=1}^{n} f_i {u_i}^r\right) h^r \; ; r = 0, 1, 2, ...$$

Note. For an individual series,

1. $\mu_r' = \frac{1}{n}\sum_{i=1}^{n} (x_i - A)^r \; ; r = 0, 1, 2, ...$

2. $\mu_r' = \frac{1}{N}\left(\sum_{i=1}^{n} {u_i}^r\right) h^r \; ; r = 0, 1, 2, ...$ for $u = \frac{x - A}{h}$

3.5 MOMENTS ABOUT THE ORIGIN

If $x_1, x_2, ..., x_n$ be the values of a variable x with corresponding frequencies $f_1, f_2, ..., f_n$ respectively then r^{th} moment about the origin ν_r is defined as

$$\nu_r = \frac{1}{N}\sum_{i=1}^{n} f_i {x_i}^r \; ; r = 0, 1, 2,... \quad \text{where,} \quad N = \sum_{i=1}^{n} f_i$$

For $r = 0$, $\nu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^{\,0} = \frac{N}{N} = 1$

For $r = 1$, $\nu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i = \bar{x}$

For $r = 2$, $\nu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^{\,2}$ and so on.

3.6 RELATION BETWEEN μ_r AND μ_r'

We know that,

$$\mu_r = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^r}{N} = \frac{1}{N}\sum_{i=1}^{n} f_i \,[(x_i - A) - (\bar{x} - A)]^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \,[(x_i - A) - \mu'_1\,]^r \qquad | \because \quad \mu_1' = \bar{x} - A$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i \,[(x_i - A)^r - {}^rc_1 (x_i - A)^{r-1}\,\mu_1' + {}^rc_2 (x_i - A)^{r-2}\,\mu_1'^2 - \ldots\ldots + (-1)^r\,\mu_1'^r]$$

| Using binomial theorem

$$\Rightarrow \qquad \mu_r = \mu_r' - {}^rc_1\,\mu'_{r-1}\mu_1' + {}^rc_2\,\mu'_{r-2}\mu_1'^2 - \ldots\ldots + (-1)^r\,\mu_1'^r$$

Putting $r = 2, 3, 4$, we get

$$\mu_2 = \mu_2' - 2\mu_1'^2 + \mu_1'^2 = \mu_2' - \mu_1'^2 \qquad | \because \quad \mu_0' = 1$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 3\mu_1'^3 - \mu_1'^3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$

Hence, we have the following relations:

$$\boxed{\mu_1 = 0}$$

$$\boxed{\mu_2 = \mu_2' - \mu_1'^2}$$

$$\boxed{\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3}$$

and

$$\boxed{\mu_4 = \mu_4' - 4\mu'_3\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4}$$

3.7 RELATION BETWEEN ν_r AND μ_r

We know that,

$$\nu_r = \frac{1}{N}\sum_{i=1}^{n} f_i x_i^{\,r}\,;\ r = 0, 1, 2, \ldots$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A + A)^r$$

$$= \frac{1}{N}\sum_{i=1}^{n} f_i\,[(x_i - A)^r + {}^rc_1(x_i - A)^{r-1}\,.\,A + \ldots + A^r]$$

$$= \mu_r' + {}^rc_1\mu_{r-1}\,A + \ldots + A^r$$

If we take, $A = \bar{x}$ (for μ_r) then

$$\nu_r = \mu_r + {}^rc_1\mu_{r-1}\,\bar{x} + {}^rc_2\mu_{r-2}\,\bar{x}^2 + \ldots + \bar{x}^r \qquad \ldots(1)$$

Putting, $r = 1, 2, 3, 4$ in (1), we get

$$\nu_1 = \mu_1 + \mu_0\bar{x} = \bar{x} \qquad | \because \mu_1 = 0, \mu_0 = 1$$

$$\nu_2 = \mu_2 + {}^2c_1\mu_1\bar{x} + {}^2c_2\,\mu_0\bar{x}^2 = \mu_2 + \bar{x}^2$$

$$\nu_3 = \mu_3 + {}^3c_1\mu_2\bar{x} + {}^3c_2\mu_1\bar{x}^2 + {}^3c_3\mu_0\bar{x}^3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3$$

$$\nu_4 = \mu_4 + {}^4c_1\mu_3\bar{x} + {}^4c_2\mu_2\bar{x}^2 + {}^4c_3\mu_1\bar{x}^3 + {}^4c_4\mu_0\bar{x}^4$$

$$= \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4$$

Hence we have the following relations:

$$\boxed{\nu_1 = \bar{x}} \qquad \boxed{\nu_2 = \mu_2 + \bar{x}^2}$$

$$\boxed{\nu_3 = \mu_3 + 3\mu_2\,\bar{x} + \bar{x}^3} \quad \text{and} \quad \boxed{\nu_4 = \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4.}$$

3.8 KARL PEARSON'S β AND γ COEFFICIENTS

Karl Pearson defined the following four coefficients based upon the first four moments of a frequency distribution about its mean:

$$\left.\begin{aligned}\beta_1 &= \frac{\mu_3^2}{\mu_2^3}\\ \beta_2 &= \frac{\mu_4}{\mu_2^2}\end{aligned}\right\} \quad (\beta\text{-coefficients})$$

$$\left.\begin{aligned}\gamma_1 &= +\sqrt{\beta_1}\\ \gamma_2 &= \beta_2 - 3\end{aligned}\right\} \quad (\gamma\text{-coefficients})$$

The practical use of these coefficients is to measure the skewness and kurtosis of a frequency distribution. These coefficients are pure numbers independent of units of measurement.

EXAMPLES

Example 1. *The first three moments of a distribution, about the value '2' of the variable are 1, 16 and – 40. Show that the mean is 3, variance is 15 and $\mu_3 = -86$.*

Sol. We have $A = 2$, $\mu_1' = 1$, $\mu_2' = 16$, and $\mu_3' = -40$

We know that $\mu_1' = \bar{x} - A \Rightarrow \bar{x} = \mu_1' + A = 1 + 2 = 3$

$$\text{Variance} = \mu_2 = \mu_2' - \mu_1'^2 = 16 - (1)^2 = 15$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -40 - 3(16)(1) + 2(1)^3 = -40 - 48 + 2 = -86.$$

Example 2. *The first four moments of a distribution, about the value '35' are – 1.8, 240, – 1020 and 144000. Find the values of* $\mu_1, \mu_2, \mu_3, \mu_4$.

Sol. $\mu_1 = 0.$

$\mu_2 = \mu_2' - \mu_1'^2 = 240 - (-1.8)^2 = 236.76$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -1020 - 3(240)(-1.8) + 2(-1.8)^3 = 264.36$

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

$= 144000 - 4(-1020)(-1.8) + 6(240)(-1.8)^2 - 3(-1.8)^4 = 141290.11.$

Example 3. *Calculate the variance and third central moment from the following data:*

x_i	0	1	2	3	4	5	6	7	8
f_i	1	9	26	59	72	52	29	7	1

(U.P.T.U. 2006)

Sol. **Calculation of Moments**

x	f	$u = \frac{x - A}{h}$ $A = 4, h = 1$	fu	fu^2	fu^3
0	1	– 4	– 4	16	– 64
1	9	– 3	– 27	81	– 243
2	26	– 2	– 52	104	– 208
3	59	– 1	– 59	59	– 59
4	72	0	0	0	0
5	52	1	52	52	52
6	29	2	58	116	232
7	7	3	21	63	189
8	1	4	4	16	64
	$N = \Sigma f = 256$		$\Sigma fu = -7$	$\Sigma fu^2 = 507$	$\Sigma fu^3 = -37$

Now, moments about the point $x = A = 4$ are

$$\mu_1' = \left(\frac{\Sigma fu}{N}\right) h = \frac{-7}{256} = -0.02734$$

$$\mu_2' = \left(\frac{\Sigma fu^2}{N}\right) h^2 = \frac{507}{256} = 1.9805$$

$$\mu_3' = \left(\frac{\Sigma fu^3}{N}\right) h^3 = \frac{-37}{256} = -0.1445$$

Moments about mean

$\mu_1 = 0$

$\mu_2 = \mu_2' - \mu_1'^2 = 1.9805 - (-.02734)^2 = 1.97975$

$\therefore$ Variance = 1.97975

Also, $\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$

$= (-0.1445) - 3(1.9805)(-.02734) + 2(-.02734)^3$

$= 0.0178997$

$\therefore$ Third central moment = 0.0178997.

Example 4. *The first three moments of a distribution about the value 2 of the variable are 1, 16 and – 40 respectively. Find the values of the first three moments about the origin.*

Sol. We have $A = 2, \quad \mu_1' = 1, \quad \mu_2' = 16, \ \mu_3' = -40$

$\therefore \quad \nu_1 = \bar{x} = A + \mu_1' = 2 + 1 = 3$

$\nu_2 = \mu_2 + \bar{x}^2 = 15 + (3)^2 = 24$

$\nu_3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3 = -86 + 3(15)(3) + (3)^3 = 76.$

Example 5. *The first four moments of a distribution about x = 2 are 1, 2.5, 5.5 and 16. Calculate the first four moments about the mean and about origin.*

Sol. We have $A = 2, \quad \mu_1' = 1, \quad \mu_2' = 2.5, \quad \mu_3' = 5.5, \quad \mu_4' = 16.$

Moments about mean

$\mu_1 = 0$

$\mu_2 = \mu_2' - (\mu_1')^2 = 2.5 - (1)^2 = 1.5$

$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 5.5 - 3(2.5)(1) + 2(1)^3 = 0$

$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4 = 16 - 4(5.5)(1) + 6(2.5)(1)^2 - 3(1)^4 = 6.$

Moments about origin

$\nu_1 = \bar{x} = A + \mu_1', \qquad \nu_2 = \mu_2 + \bar{x}^2$

$\nu_3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3, \qquad \nu_4 = \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4$

$\therefore \quad \nu_1 = \bar{x} = 2 + 1 = 3, \qquad \nu_2 = 1.5 + (3)^2 = 10.5$

$\nu_3 = 0 + 3(1.5)(3) + (3)^3 = 40.5, \qquad \nu_4 = 6 + 4(0)(3) + 6(1.5)(3)^2 + (3)^4 = 168.$

Example 6. *For a distribution, the mean is 10, variance is 16, γ_1 is 1, and β_2 is 4. Find the first four moments about the origin.*

Sol. $\bar{x} = 10, \quad \mu_2 = 16, \quad \gamma_1 = 1, \quad \beta_2 = 4$ | given

Now, $\gamma_1 = 1$

$\Rightarrow \quad \beta_1 = 1$ | $\because \ \gamma_1 = \sqrt{\beta_1}$

$\Rightarrow \quad \dfrac{\mu_3^2}{\mu_2^3} = 1 \quad \Rightarrow \quad \mu_3^2 = \mu_2^3 = (16)^3 = (64)^2$

$\Rightarrow \quad \mu_3 = 64$

and $\beta_2 = 4$

$\Rightarrow \quad \dfrac{\mu_4}{\mu_2^2} = 4 \qquad \Rightarrow \quad \mu_4 = 4(16)^2 = 1024$ | $\because \quad \mu_2 = 16$

Moments about the origin

$\nu_1 = \bar{x} = 10$

$\nu_2 = \mu_2 + \bar{x}^2 = 16 + 100 = 116$

$\nu_3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3 = 64 + 480 + 1000 = 1544$

$\nu_4 = \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4 = 1024 + 4(64)(10) + 6(16)(100) + (10)^4$

$= 22184$

Example 7. *In a certain distribution, the first four moments about the point x = 4 are – 1.5, 17, – 30 and 308. Find the moments about mean and about origin. Also, Calculate β_1 and β_2.* (U.P.T.U. 2014)

Sol. We have, $A = 4, \mu'_1 = -1.5, \mu'_2 = 17, \mu'_3 = -30, \mu'_4 = 308$

Moments about mean

$$\mu_1 = 0$$
$$\mu_2 = \mu'_2 - \mu_1'^2 = 17 - (-1.5)^2 = 14.75$$
$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu_1'^3 = -30 - 3(17)(-1.5) + 2(-1.5)^3 = 39.75$$
$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu_1'^2 - 3\mu_1'^4$$
$$= 308 - 4(-30)(-1.5) + 6(17)(-1.5)^2 - 3(-1.5)^4 = 342.3125$$

Moments about origin

$$\nu_1 = \bar{x} = \mu'_1 + A = -1.5 + 4 = 2.5$$
$$\nu_2 = \mu_2 + \bar{x}^2 = 14.75 + (2.5)^2 = 21$$
$$\nu_3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3 = 166$$
$$\nu_4 = \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4 = 1332$$

Calculation of β_1 and β_2

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0.492377 \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} = 1.573398$$

Example 8. *The first four moments of a distribution about the value '4' of the variable are – 1.5, 17, – 30 and 108. Find the moments about mean, about origin ; β_1 and β_2. Also find the moments about the point x = 2.* (U.P.T.U. 2007)

Sol. We have $A = 4, \mu'_1 = -1.5, \mu'_2 = 17, \mu'_3 = -30, \mu'_4 = 108$

Moments about mean

$$\mu_1 = 0$$
$$\mu_2 = \mu'_2 - \mu_1'^2 = 14.75$$
$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu_1'^3 = 39.75$$
$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu_1'^2 - 3\mu_1'^4 = 142.3125$$

Also, $\bar{x} = \mu'_1 + A = -1.5 + 4 = 2.5$

Moments about origin

$$\nu_1 = \bar{x} = 2.5$$
$$\nu_2 = \mu_2 + \bar{x}^2 = 14.75 + (2.5)^2 = 21$$
$$\nu_3 = \mu_3 + 3\mu_2\bar{x} + \bar{x}^3 = 166$$
$$\nu_4 = \mu_4 + 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 + \bar{x}^4 = 1132$$

Calculation of β_1 and β_2

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0.492377 \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} = 0.654122$$

Moments about the point x = 2

$$\mu'_1 = \bar{x} - A = 2.5 - 2 = 0.5$$
$$\mu'_2 = \mu_2 + \mu_1'^2 = 14.75 + (.5)^2 = 15$$
$$\mu'_3 = \mu_3 + 3\mu'_2\mu'_1 - 2\mu_1'^3 = 39.75 + 3(15)(.5) - 2(.5)^3 = 62$$
$$\mu'_4 = \mu_4 + 4\mu'_3\mu'_1 - 6\mu'_2\mu_1'^2 + 3\mu_1'^4 = 244$$

ASSIGNMENT

1. (*i*) Calculate first four moments about mean, for the following individual series:

5, 5, 5, 5, 5, 5.

(*ii*) Find the first four moments about the mean of the following series:

1, 3, 7, 9, 10.

(*iii*) Calculate μ_1, μ_2, μ_3, μ_4 for the series : 4, 7, 10, 13, 16, 19, 22.

2. (*i*) Find the first four moments for the following frequency distribution:

x	1	2	3	4	5	6	7	8	9
f	1	2	3	4	5	4	3	2	1

(*ii*) Calculate the first four moments of the following distribution about the mean and hence find β_1 and β_2.

x	0	1	2	3	4	5	6	7
f	1	8	28	56	70	56	28	8

(*iii*) The number of flowers on Sunflower plants are given below:

No. of flowers	3	6	12	16	25
No. of plants	1	2	3	4	5

Calculate the first four moments about mean. [*M.T.U. (B. Pharma) 2011*]

3. (*i*) Find the first four moments about mean for the following frequency distribution :

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

(*ii*) Calculate the first four moments about the mean for the following:

Classes	5–15	15–25	25–35	35–45	45–55
f	14	22	36	18	10

(*iii*) Calculate the first four moments about the mean for the following data:

Class-interval	0–10	10–20	20–30	30–40	40–50
f	10	20	40	20	10

(*M.T.U. 2014*)

4. Calculate the first four moments about $x = 15$ and hence find the moments about the mean of the following distribution :

x	10	11	12	13	14	15	16	17	18	19	20	21
f	9	36	75	105	116	107	88	66	45	30	18	5

5. (*i*) The first three moments of a distribution about the value 4 of the variable are 1.5, 17 and – 30. Find the moments about mean.

(*ii*) The first four moments of a distribution about $x = 4$ are 1, 4, 10 and 45. Show that the mean is 5, the variance is 3, μ_3 is 0 and μ_4 is 26.

(*iii*) The first four moments of a distribution about the value 5 of the variable are 2, 20, 40 and 50. Calculate mean, variance, μ_3 and μ_4. (*G.B.T.U. 2011*)

6. If the first four moments of a distribution about the value 5 are equal to – 4, 22, – 117 and 560. Determine the corresponding moments :
(*i*) about the mean, and (*ii*) about zero.

7. Compute first four moments of the data 3, 5, 7, 9 about the mean. Also, compute the first four moments about the point 4.

8. In a frequency distribution, the mean is 1.5, variance is 0.64, β_2 is 2.5 and γ_1 is 0.3. Find μ_3 and μ_4 and also the first four moments about the origin.

9. The first four moments of a distribution about the value '0' are – 0.20, 1.76, – 2.36 and 10.88. Find the moments about the mean. (*U.P.T.U. 2009*)

Answers

1. (*i*) $\mu_1 = 0, \mu_2 = 0, \mu_3 = 0, \mu_4 = 0$ (*ii*) $\mu_1 = 0, \mu_2 = 12, \mu_3 = -12, \mu_4 = 208.8$
(*iii*) $\mu_1 = 0, \mu_2 = 36, \mu_3 = 0, \mu_4 = 2268$

2. (*i*) $\mu_1 = 0, \mu_2 = 4, \mu_3 = 0, \mu_4 = 37.6$ (*ii*) $\mu_1 = 0, \mu_2 = 2, \mu_3 = 0, \mu_4 = 11, \beta_1 = 0, \beta_2 = 2.75$
(*iii*) $\mu_1 = 0, \mu_2 = 54.8, \mu_3 = -49.6, \mu_4 = 5475.6$

3. (*i*) $\mu_1 = 0, \mu_2 = 125, \mu_3 = -300, \mu_4 = 37625$ (*ii*) $\mu_1 = 0, \mu_2 = 134.56, \mu_3 = 126.14, \mu_4 = 41840.82$
(*iii*) $\mu_1 = 0, \mu_2 = 120, \mu_3 = 0, \mu_4 = 36000$

4. $\mu_1 = 0, \mu_2 = 5.5, \mu_3 = 4.4, \mu_4 = 77.8$

5. (*i*) $\mu_1 = 0, \mu_2 = 14.75, \mu_3 = -99.75$ (*iii*) 7, 16, – 64, 162

6. (*i*) 0, 6, 19, 32 (*ii*) 1, 7, 38, 145

7. $\mu_1 = 0, \mu_2 = 5, \mu_3 = 0, \mu_4 = 41$; $\mu_1' = 2, \mu_2' = 9, \mu_3' = 38.25, \mu_4' = 177$

8. $\mu_3 = 0.1536, \mu_4 = 1.024, \nu_1 = 1.5, \nu_2 = 2.89, \nu_3 = 6.4086, \nu_4 = 15.6481$

9. $\mu_1 = 0, \mu_2 = 1.72, \mu_3 = -1.32, \mu_4 = 9.4096$.

3.9 MOMENT GENERATING FUNCTION

[*U.P.T.U. 2014 ; G.B.T.U. 2012, M.T.U. 2013*]

For certain theoretical developments, an indirect method for computing moments is used. The method depends on the finding of the moment generating function.

3.9.1. In Case of a Continuous Variable *x*, it is defined as

$$M(t) = \int_a^b e^{tx} f(x)\, dx \qquad ...(1)$$

where integral is a function of parameter t only. The limits a, b can be $-\infty$ and ∞ respectively. It is possible to associate a moment generating function with the distribution only when all the moments of the distribution are finite.

Let us see how $M(t)$ generates moments. For this, let us assume that $f(x)$ is a distribution function for which the integral given by (1) exists.

Then e^{tx} may be expanded in a power series and the integration may be performed term by term. It follows that

$$M(t) = \int_a^b \left(1 + tx + \frac{t^2}{2!}x^2 + \ldots\right) f(x)\, dx$$

$$= \int_a^b f(x)\, dx + t\int_a^b xf(x)\, dx + \ldots$$

$$= \nu_0 + \nu_1 t + \nu_2 \cdot \frac{t^2}{2!} + \ldots \qquad \ldots(2)$$

Obviously, the coefficient of $\frac{t^r}{r!}$ in (2) is the r^{th} moment about the origin.

Also, $$\left| \frac{d^r}{dt^r} \mathrm{M}(t) \right|_{t=0} = \left| \frac{\nu_r}{r!} r! + \nu_{r+1} t + \ldots \right|_{t=0} = \nu_r \qquad \ldots(3)$$

Thus, ν_r about origin = r^{th} derivative of $\mathrm{M}(t)$ with $t = 0$.

Although the moment generating function (m.g.f.) has been defined for the variable x only, the definition can be generalized so that it holds for a variable z where z is a function of x. *e.g.,* if $z = x - m$ (m is mean), the r^{th} moment about z will give r^{th} moment of x about the mean m.

Moment generating function for z will clearly be given as

$$\mathrm{M}_z(t) = \int_a^b e^{tz} f(x)\, dx$$

$$\mathrm{M}_{x-m}(t) = \int_a^b e^{t(x-m)} f(x)\, dx = e^{-mt} \int_a^b e^{tx} f(x) dx = e^{-mt}\, \mathrm{M}_x(t).$$

3.9.2. In Case of Discrete Distribution of the Variable *x*

We know that, for a variable x,

$$\nu_r = \Sigma\, x^r \,.\, \mathrm{P}$$

where P is the probability that the variable takes on the value x.

If z is any function of x, we get r^{th} moment for z by the relation

$$\nu_r = \Sigma z^r \mathrm{P}$$

and the moment generating function is given by

$$\mathrm{M}_z(t) = \Sigma e^{tz}\, \mathrm{P} \qquad \ldots(1)$$

To verify that this function generates moments, we will expand e^{tz} and then sum term by term,

$$\therefore \qquad \mathrm{M}_z(t) = \sum \left(1 + tz + \frac{t^2}{2!} z^2 + \ldots\right) \mathrm{P} = \Sigma \mathrm{P} + t \Sigma z \mathrm{P} + \frac{t^2}{2!} \Sigma z^2 \,.\, \mathrm{P} + \ldots$$

$$= \nu_0 + t\nu_1 + \frac{t^2}{2!} \nu_2 + \ldots$$

In this case, we can also show that $\nu_r = \left| \frac{d^r}{dt^r} \mathrm{M}_z(t) \right|_{t=0}$

$\mathrm{M}(t)$ is clearly the expected value of e^{tx} and hence can be written as $\mathrm{E}(e^{tx})$ which gives the moment generating function incase of discrete as well as continuous variables.

Expectation of any function $\phi(x)$ is given by

$$\mathrm{E}\{\phi(x)\} = \sum_i \phi(x_i) f(x_i) \qquad \text{| for discrete distribution}$$

or, $$\mathrm{E}\{\phi(x)\} = \int_{-\infty}^{\infty} \phi(x) f(x)\, dx \qquad \text{| for continuous distribution}$$

Eqn. (1) can also be rewritten as

$$M_{x-a}(t) = E\,[e^{t(x-a)}] = \sum_i e^{t(x_i - a)}\, P_i = e^{-at} \sum_i e^{tx_i}\, P_i = e^{-at}\, M_0(t)$$

Therefore the moment generating function about the point 'a' is equal to e^{-at} times the moment generating function about the origin.

Note. m.g.f. is not always defined since $E\{|\ e^{tx}\ |\}$ does not always exist for all values of t.

e.g., if $f(x) = \dfrac{6}{\pi^2 x^2}$, $x = 1, 2, 3, \ldots$ then m.g.f. does not exist.

m.g.f. always exists for $t = 0$ since $M_{x=0}(0) = 1$.

3.9.3. Properties of Moment Generating Function (*M.T.U. 2013*)

(1) *The moment generating function of the sum of two independent chance variables is the product of their respective moment generating functions.*

Symbolically, $M_{x+y}(t) = M_x(t) \times M_y(t)$ provided that x and y are independent random variables.

Proof. Let x and y be two independent random variables so that $x + y$ is also a random variable.

The m.g.f. of the sum $x + y$ w.r.t. origin is

$$M_{x+y}(t) = E\{e^{t(x+y)}\} = E(e^{tx} \cdot e^{ty}) = E(e^{tx}) \cdot E(e^{ty})$$

Since x and y are independent variables and so are e^{tx} and e^{ty}.

$$\therefore \qquad M_{x+y}(t) = M_x(t) \cdot M_y(t)$$

Hence the theorem.

(2) *Effect of change of origin and scale on m.g.f.*

$$\boxed{M_u(t) = e^{-at/h}\, M_x(t/h)} \qquad \text{where } u = \frac{x-a}{h}$$

Proof. Let u be a new random variable given by $u = \dfrac{x-a}{h}$ so that $x = a + hu$

then by definition, the effect of linear transformation on m.g.f. is governed by

$$M_x(t) = E(e^{tx}) = E[e^{t(a+hu)}] = E(e^{at} \cdot e^{thu})$$
$$= e^{at}\, E(e^{thu}) = e^{at}\, M_u(th)$$

Also, $$M_u(t) = E(e^{tu})$$

$$= E\left[e^{t\left(\frac{x-a}{h}\right)}\right] = e^{-\frac{at}{h}}\, M_x\left(\frac{t}{h}\right)$$

(3) $M_{cx}(t) = M_x(ct)$, *c being a constant.*

Proof. By definition,

$$\text{LHS} = M_{cx}(t) = E(e^{tcx}) = M_x(ct) = \text{RHS}$$

Hence the result.

EXAMPLES

Example 1. *Find the moment generating function of the exponential distribution*

$$f(x) = \frac{1}{c}\, e^{-x/c}\ ;\ 0 \le x \le \infty,\ c > 0$$ (M.T.U. 2014)

Hence find its mean and standard deviation.

Sol. Moment generating function about the origin is given by

$$M_x(t) = \int_0^\infty e^{tx} \cdot \frac{1}{c} e^{-x/c}\, dx$$

$$= \frac{1}{c}\int_0^\infty e^{\left(t-\frac{1}{c}\right)x}\, dx = \frac{1}{c}\left[\frac{e^{\left(t-\frac{1}{c}\right)x}}{\left(t-\frac{1}{c}\right)}\right]_0^\infty$$

$$= (1 - ct)^{-1} = 1 + ct + c^2t^2 + c^3t^3 + \ldots$$

$\therefore$ $$\nu_1 = \left[\frac{d}{dt} M_x(t)\right]_{t=0} = (c + 2c^2t + 3c^3t^2 + \ldots)_{t=0} = c$$

and $$\nu_2 = \left[\frac{d^2}{dt^2} M_x(t)\right]_{t=0} = 2c^2$$

Now, mean $\bar{x} = \nu_1 = c$

Variance $\mu_2 = \nu_2 - \bar{x}^2 = \nu_2 - \nu_1^2 = 2c^2 - c^2 = c^2$

$\therefore$ Standard deviation $= \sqrt{\mu_2} = c$.

Example 2. *Obtain the moment generating function of the random variable x having probability distribution*

$$f(x) = \begin{cases} x, & \text{for } 0 < x < 1 \\ 2 - x, & \text{for } 1 \le x < 2 \\ 0, & \text{elsewhere} \end{cases}$$

[M.T.U. 2012; G.B.T.U. (C.O.) 2011; G.B.T.U. 2013]

Also determine mean ν_1, ν_2 *and variance* μ_2.

Sol. $M_x(t) = E(e^{tx})$

$$= \int_0^1 x \cdot e^{tx}\, dx + \int_1^2 (2 - x)\, e^{tx}\, dx + \int_2^\infty 0 \cdot e^{tx}\, dx$$

$$= \left(\frac{xe^{tx}}{t} - \frac{e^{tx}}{t^2}\right)_0^1 + \left(\frac{2e^{tx}}{t} - \frac{xe^{tx}}{t} + \frac{e^{tx}}{t^2}\right)_1^2$$

$$= \frac{e^t}{t} - \frac{e^t}{t^2} + \frac{1}{t^2} + \left[\left(\frac{2e^{2t}}{t} - \frac{2e^{2t}}{t} + \frac{e^{2t}}{t^2}\right) - \left(\frac{2e^t}{t} - \frac{e^t}{t} + \frac{e^t}{t^2}\right)\right] = \frac{e^{2t} - 2e^t + 1}{t^2}$$

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

$$\text{Mean} = \nu_1 = \left[\frac{d}{dt} M_x(t)\right]_{t=0} = 1$$

Similarly, $\nu_2 = 2$, $\mu_2 = \nu_2 - \bar{x}^2 = \nu_2 - \nu_1^2 = 2 - (1)^2 = 1$ = Variance.

Example 3. *Find the moment generating function of the random variable whose moments are* $\nu_r = (r + 1)\,!\,2^r$.

Sol.
$$M_x(t) = E(e^{tx}) = \sum_{x=0}^{\infty} e^{tx}\, P(X = x)$$
$$= \sum_{r=0}^{\infty} \frac{t^r}{r\,!}\nu_r = \sum_{r=0}^{\infty} \frac{t^r}{r\,!}(r+1)\,!\,.\,2^r = \sum_{r=0}^{\infty} (r+1)\,(2t)^r$$
$$= 1 + 2.2t + 3\,.\,(2t)^2 + \ldots. = (1 - 2t)^{-2}.$$

Example 4. *Find the moment generating function of the probability distribution function* $f(z) = e^{-z}\,(1 + e^{-z})^{-2},\ -\infty < z < \infty$.

Sol.
$$M_z(t) = E(e^{tz})$$
$$= \int_{-\infty}^{\infty} e^{tz}\,.\,e^{-z}(1+e^{-z})^{-2}\,dz$$
$$= \int_1^{\infty} u^2\,(u-1)^{-t}\,du \quad \text{where } 1 + e^{-z} = u \ \Rightarrow\ -e^{-z}\,dz = du$$
$$= \int_0^1 v^{-t}\,(1-v)^t\,dv \quad \text{where } v = 1 - \frac{1}{u} \ \Rightarrow\ dv = \frac{1}{u^2}\,du$$
$$= \beta(1-t,\ 1+t)\ ;\ 1 - t > 0$$
$$= \pi t \operatorname{cosec} \pi t,\ t < 1.$$

Example 5. *Find the moment generating function of the negative exponential function* $f(x) = \lambda e^{-\lambda x}\ ;\ x,\ \lambda > 0$.

Sol.
$$M_x(t) = \lambda\int_0^{\infty} e^{tx}\,.e^{-\lambda x}\,dx = \lambda\int_0^{\infty} e^{(t-\lambda)x}\,dx = \lambda\int_0^{\infty} e^{-(\lambda - t)x}\,dx$$
$$= \frac{\lambda}{\lambda - t} = \left(1 - \frac{t}{\lambda}\right)^{-1} = \sum_{r=0}^{\infty}\left(\frac{t}{\lambda}\right)^r\ ;\ \lambda > t$$

Example 6. *Find the moment generating function of the discrete binomial distribution given by*

$$P(x) = {}^nc_x\, p^x\, q^{n-x} \quad (\text{where } q = 1 - p)$$

Also find the first and second moments about the mean. (U.P.T.U. 2008)

Sol. Moment generating function about the origin is given by

$$M_x(t) = \Sigma\, e^{tx}\,.\,{}^nc_x\,.\,p^x q^{n-x}$$
$$= \Sigma\, {}^nc_x\,(pe^t)^x\, q^{n-x} = (q + pe^t)^n$$
$$\nu_1 = \left[\frac{d}{dt} M_x(t)\right]_{t=0} = [n(q + pe^t)^{n-1}\,.\,pe^t\,]_{t=0} = np \qquad |\ \text{Since } q + p = 1$$
$$\nu_2 = \left[\frac{d^2}{dt^2} M_x(t)\right]_{t=0}$$
$$= [np\{e^t\,.\,(n-1)\,(q + pe^t)^{n-2}\,pe^t + (q + pe^t)^{n-1}\,.\,e^t\}]_{t=0}$$

$$= [np\,(q + pe^t)^{n-2}\,.\,e^t\,\{(n-1)pe^t + (q + pe^t)\}]_{t=0}$$
$$= [np(q + pe^t)^{n-2}\,.\,e^t(q + npe^t)]_{t=0}$$
$$= np(q + np) \qquad | \because \quad q + p = 1$$
$$= npq + n^2 p^2$$

Hence first and second moments about the mean are given by

$$\mu_1 = 0$$

Since $\bar{x} = \nu_1 = np$

$\therefore$ $\mu_2 = \nu_2 - \bar{x}^2 = \nu_2 - \nu_1^2 = npq + n^2p^2 - n^2p^2 = npq$

Hence, mean = np, S.D. = $\sqrt{\mu_2} = \sqrt{npq}$.

Example 7. *Find the moment generating function of the discrete Poisson distribution given by* $P(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$. *Also find the first and second moments about the mean.*

(M.T.U. 2013)

Sol. Moment generating function about the origin is given by

$$M_x(t) = \Sigma e^{tx}\,.\,e^{-\lambda}\,.\,\frac{\lambda^x}{x!} = e^{-\lambda}\,\Sigma\,\frac{(\lambda e^t)^x}{x!} = e^{-\lambda}\,.\,e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

$$\nu_1 = \left[\frac{d}{dt}\,M_x(t)\right]_{t=0} = [e^{\lambda(e^t-1)}\,\lambda e^t]_{t=0} = \lambda$$

$$\nu_2 = \left[\frac{d^2}{dt^2}\,M_x(t)\right]_{t=0} = [\lambda\,\{e^t\,.\,e^{\lambda(e^t-1)}\,.\,\lambda e^t + e^{\lambda(e^t-1)}e^t\}]_{t=0}$$

$$= [\lambda e^{\lambda(e^t-1)}\,e^t\,(\lambda\,e^t + 1)]_{t=0} = \lambda(\lambda + 1)$$

Hence first and second moments about the mean are given by

$$\mu_1 = 0$$

Since $\nu_1 = \bar{x} = \lambda$

$\therefore$ $\mu_2 = \nu_2 - \bar{x}^2 = \nu_2 - \nu_1{}^2 = \lambda\,(\lambda + 1) - \lambda^2 = \lambda$

Example 8. *Find the moment generating function of the continuous normal distribution given by* $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$; $-\infty < x < \infty$.

Sol. Moment generating function about the origin is defined as

$$M_x(t) = E\,(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} \cdot \frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\,dx$$

$$= \frac{e^{\mu t}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\,.\,e^{-t\sigma z}\,dz \qquad \text{where } z = \frac{x-\mu}{\sigma}$$

$$= \frac{1}{\sqrt{2\pi}}\,e^{\left(\mu t + \frac{1}{2}t^2\sigma^2\right)} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(z - t\sigma)^2}\,dz$$

$$= e^{\mu t + \frac{1}{2}t^2\sigma^2} \cdot 1 = e^{\mu t + \frac{1}{2}t^2\sigma^2}. \qquad \left| \because \int_0^\infty e^{-z^2}\, dz = \frac{\sqrt{\pi}}{2} \right.$$

Example 9. *The random variable X assuming only non-negative values has a Gamma probability distribution if its probability distribution is given by*

$$f(x) = \begin{cases} \dfrac{\alpha^\beta}{\Gamma\beta} x^{\beta-1} e^{-\alpha x} \; ; x > 0, \alpha > 0, \beta > 1 \\ 0, \qquad\qquad \text{elsewhere} \end{cases}$$

Find the moment generating function of Gamma probability distribution.

Sol. $M_x(t) = E(e^{tx})$

$$= \int_0^\infty e^{tx} \cdot \frac{\alpha^\beta}{\Gamma\beta} \cdot x^{\beta-1} e^{-\alpha x}\, dx = \frac{\alpha^\beta}{\Gamma\beta} \int_0^\infty x^{\beta-1} e^{-x(\alpha - t)}\, dx$$

$$= \frac{\alpha^\beta}{(\alpha - t)^\beta \Gamma\beta} \int_0^\infty y^{\beta-1} e^{-\beta}\, dy \quad | \text{ where } y = x(\alpha - t) \text{ so that } dy = (\alpha - t)\, dx$$

$$= \frac{1}{\left(1 - \frac{t}{\alpha}\right)^\beta} \cdot \frac{1}{\Gamma\beta} \Gamma\beta = \left(1 - \frac{t}{\alpha}\right)^{-\beta} ; \; | t | < \alpha.$$

Example 10. *Let the random variable X assume the value 'n' with the probability law $p(X = n) = pq^{n-1}$, n = 1, 2, 3, ... Find the moment generating function and hence mean and variance.* (G.B.T.U. 2010)

Sol. The given distribution is a discrete distribution.

$$M_n(t) = \Sigma\, e^{tn}\, pq^{n-1} = \frac{p}{q} \Sigma (e^t q)^n = \frac{p}{q} (1 - e^t q)^{-1} = \frac{p}{q(1 - qe^t)}$$

which is the moment generating function.

$$\nu_1 = \left[\frac{d}{dt} M_n(t)\right]_{t=0} = \frac{p}{q}\left[\frac{q e^t}{(1 - q e^t)^2}\right]_{t=0} = \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}$$

$$\nu_2 = \left[\frac{d^2}{dt^2} M_n(t)\right]_{t=0} = p\left[\frac{d}{dt}\left\{\frac{e^t}{(1 - qe^t)^2}\right\}\right]_{t=0}$$

$$= p\left[\frac{(1 - qe^t)^2 \cdot e^t - e^t \cdot 2(1 - qe^t)(-qe^t)}{(1 - qe^t)^4}\right]_{t=0}$$

$$= p\left[\frac{(1-q)^2 + 2q(1-q)}{(1-q)^4}\right] = \frac{1}{p} + \frac{2q}{p^2}$$

$$\text{Mean} = \bar{x} = \nu_1 = \frac{1}{p}$$

$$\text{Variance} = \mu_2 = \nu_2 - \bar{x}^2 = \frac{1}{p} + \frac{2q}{p^2} - \frac{1}{p^2} = \frac{q}{p^2}$$

ASSIGNMENT

1. Define moment generating function. Find the moment generating function of a random variable X whose probability function is given by:

$$P(X = x) = p(1-p)^x,\ x = 0, 1, 2, \ldots, \infty$$ (*G.B.T.U. 2012*)

2. Define moment generating function and two properties of moment generating function with proof. (*M.T.U. 2013*)

3. The probability density function of the random variable X is $f(x) = \frac{1}{2\theta} \exp\left(-\frac{|x-\theta|}{\theta}\right), -\infty < x < \infty$.

Find moment generating function of X. Hence find the mean E(X) and variance V(X). (*M.T.U. 2013*)

$$\left[\textbf{Hint: } M_x(t) = \frac{1}{2\theta}\int_{-\infty}^{\theta} \exp\left(-\frac{\theta - x}{\theta}\right) e^{tx}\,dx + \frac{1}{2\theta}\int_{\theta}^{\infty} \exp\left(-\frac{x-\theta}{\theta}\right) e^{tx}\,dx\right]$$

4. Show that the moment generating function of random variable X having the p.d.f.

$$f(x) = \begin{Bmatrix} \frac{1}{3}, & -1 < x < 2 \\ 0, & \text{elsewhere} \end{Bmatrix} \quad \text{is} \quad M_X(t) = \begin{cases} \frac{e^{2t} - e^{-t}}{3t}, & t \neq 0 \\ 1, & t = 0 \end{cases}$$

5. Find the moment generating function for triangular distribution defined by

$$f(x) = \begin{Bmatrix} x, & 0 \le x \le 1 \\ 2 - x, & 1 \le x \le 2 \end{Bmatrix}$$ (*M.T.U. 2013*)

6. If $P(X = x) = \frac{1}{2^x}$, $x = 1, 2, 3, \ldots$, find the moment generating function of x. Hence obtain the variance. (*U.P.T.U. 2014*)

Answers

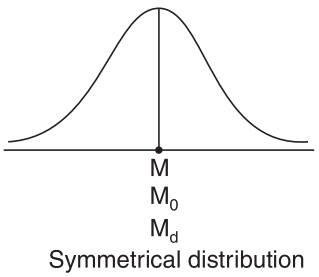
1. $\frac{p}{1 - e^t(1-p)}$ **3.** $\frac{e^{\theta t}}{1 - \theta^2 t^2}$ or $1 + \theta t + \frac{3\theta^2 t^2}{2!} + \ldots$; $E(X) = \theta$, $V(x) = 2\theta^2$

5. $M_x(t) = 1 + t + \frac{1}{18}t^2 + \ldots$ **6.** $\frac{e^t}{2 - e^t}$; $\frac{1}{2}$.

3.10 SKEWNESS

For a symmetrical distribution, the frequencies are symmetrically distributed about the mean *i.e.,* variates equidistant from the mean have equal frequencies. Also, the mean, mode and median coincide and median lies half-way between the two quartiles.

Thus, $M = M_0 = M_d$ and $Q_3 - M = M - Q_1$.

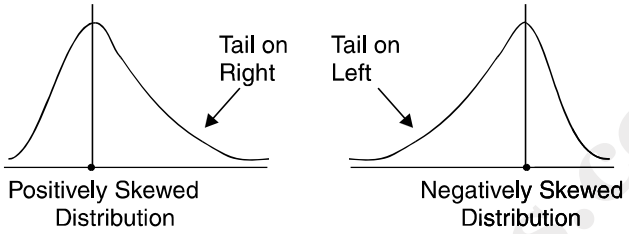


Symmetrical distribution

3.11 MEANING OF SKEWNESS (*U.P.T.U. 2015*)

If the curve of the distribution is not symmetrical, it may admit of tail on either side of the distribution. **Skewness means lack of symmetry or lopsidedness in a frequency distribution.**

The object of measuring skewness is to estimate the extent to which a distribution is distorted from a perfectly symmetrical distribution. Skewness indicates whether the curve is turned more to one side than to other *i.e.,* whether the curve has a longer tail on one side. Skewness can be positive as well as negative.



Skewness is positive if the longer tail of the distribution lies towards the right and negative if it lies towards the left.

3.12 TESTS OF SKEWNESS

1. If A.M. = Mode = Median, then there is no skewness in the distribution. In other words, the curve of the frequency distribution would be symmetrical, bell-shaped.

2. If A.M. is less than (greater than), the value of mode, the tail would be on left (right) side, *i.e.,* the distribution is negatively (positively) skewed.

3. If sum of frequencies of values less than mode is equal to the sum of frequencies of values greater than mode, then there would be no skewness.

4. If quartiles are equidistant from median, then there would be no skewness.

3.13 METHODS OF MEASURING SKEWNESS (*U.P.T.U. 2007*)

Relative measures of skewness are called the **coefficient of skewness**. They are independent of the units of measurement and as such, they are pure numbers.

Following are the methods of measuring skewness:

1. Karl Pearson's Method
2. Bowley's Method
3. Kelly's Method
4. Method of Moments.

Here, we will discuss Karl Pearson's method and the method of moments only.

3.13.1. Karl Pearson's Method

This method is based on the fact that in a symmetrical distribution, the value of A.M. is equal to that of mode. As we have already noted that the distribution is positively skewed if A.M. > Mode and negatively skewed if A.M. < Mode. The Karl Pearson's coefficient of skewness is given by:

$$\textbf{Karl Pearson's coefficient of skewness} = \frac{\textbf{A.M.} - \textbf{Mode}}{\textbf{S.D.}}$$

We have already studied the methods of calculating A.M., Mode and S.D. of frequency distributions. If mode is ill-defined in some frequency distribution, then the value of empirical mode is used in the formula.

$$\text{Empirical mode} = 3\ \text{Median} - 2\ \text{A.M.}$$

$$\therefore \quad \text{Coeff. of skewness} = \frac{\text{A.M.} - \text{Mode}}{\text{S.D.}}$$

$$= \frac{\text{A.M.} - (3\ \text{Median} - 2\ \text{A.M.})}{\text{S.D.}} = \frac{3\ \text{A.M.} - 3\ \text{Median}}{\text{S.D.}}$$

$$\therefore \quad \textbf{Karl Pearson's coefficient of skewness} = \frac{\textbf{3(A.M. – Median)}}{\textbf{S.D.}}$$

The coefficient of skewness as calculated by using this method gives magnitude as well as direction of skewness, present in the distribution. Practically, its value lies between – 1 and 1. For a symmetrical distribution, its value comes out to be zero.

The Karl Pearson's coefficient of skewness is generally denoted by 'SK_P'.

(*i*) If $SK_P = 0 \Leftrightarrow \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} = 0$

$\Leftrightarrow$ Mean = Mode

$\Leftrightarrow$ Distribution is symmetrical.

Thus a distribution is a symmetrical distribution iff $SK_P = 0$.

(*ii*) If $SK_P > 0 \Leftrightarrow \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} > 0$

$\Leftrightarrow$ Mean – Mode > 0

$\Leftrightarrow$ Mean > Mode

$\Leftrightarrow$ Distribution is positively skewed.

Thus a distribution is a positively skewed distribution iff $SK_P > 0$.

(*iii*) If $SK_P < 0 \Leftrightarrow \frac{\text{Mean} - \text{Mode}}{\text{S.D.}} < 0$

$\Leftrightarrow$ Mean – Mode < 0

$\Leftrightarrow$ Mean < Mode

$\Leftrightarrow$ Distribution is negatively skewed.

Thus a distribution is a negatively skewed distribution iff $SK_P < 0$.

EXAMPLES

Example 1. *Karl Pearson's coefficient of skewness of a distribution is 0.32, its standard deviation is 6.5 and mean is 29.6. Find the mode of the distribution.*

Sol. We have $SK_P = 0.32$, S.D. = 6.5, $\bar{x} = 29.6$.

Now $$SK_P = \frac{\bar{x} - \text{Mode}}{\text{S.D.}}$$

$$\therefore \quad 0.32 = \frac{29.6 - \text{Mode}}{6.5}$$

$$\Rightarrow \quad 29.6 - \text{Mode} = 0.32 \times 6.5 = 2.08$$

$$\Rightarrow \quad \text{Mode} = 29.6 - 2.08 = 27.52.$$

Example 2. *For a moderately skewed data, the arithmetic mean is 100, the variance is 35 and Karl Pearson's coefficient of skewness is 0.2. Find its mode and median.*

Sol. We have $\bar{x}$ = 100, Variance = 35, SK_P = 0.2.

Now $$SK_P = \frac{\bar{x} - \text{Mode}}{\sigma}$$

$\therefore$ $$0.2 = \frac{100 - \text{Mode}}{\sqrt{35}} \qquad (\because \text{ S.D.} = \sqrt{\text{variance}})$$

$\Rightarrow$ $100 - \text{Mode} = 0.2 \times 5.92 = 1.184$

$\Rightarrow$ $\text{Mode} = 100 - 1.184 = 98.816.$

Also, $\text{Mode} = 3 \text{ Median} - 2\bar{x} \Rightarrow 98.816 = 3 \text{ Median} - 2(100).$

$\therefore$ $3 \text{ Median} = 98.816 + 200 = 298.816$

$\therefore$ $$\text{Median} = \frac{298.816}{3} = 99.61.$$

Example 3. *In a certain distribution, the following results were obtained :*

A.M. = 45, Median = 48, Coefficient of skewness = – 0.4. The person who gave you this data, failed to give the value of S.D. You are required to estimate it with the help of available data.

Sol. We have

coeff. of skewness = – 0.4, A.M. = 45, median = 48.

Now, coeff. of skewness $= \frac{3(\bar{x} - \text{Median})}{\text{S.D.}}$

$\Rightarrow$ $$-\frac{4}{10} = \frac{3(45-48)}{\text{S.D.}} = \frac{-9}{\text{S.D.}}$$

$\Rightarrow$ $4 \text{ S.D.} = 90$

$$\text{S.D.} = \frac{90}{4} = 22.5.$$

Example 4. *The sum of 20 observations is 300 and sum of their squares is 5000. The median is 15. Find the Karl Pearson's coefficient of skewness.*

Sol. Let 'x' be the variable under consideration.

We have $n = 20$, $\Sigma x = 300$, $\Sigma x^2 = 5000$, median = 15.

Now, $$\bar{x} = \frac{\Sigma x}{n} = \frac{300}{20} = 15$$

$$\text{S.D.} = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2} = \sqrt{\frac{5000}{20} - (15)^2} = \sqrt{250 - 225} = \sqrt{25} = 5$$

Now, Karl Pearson's coeff. of skewness

$$= \frac{3(\bar{x} - \text{Median})}{\text{S.D.}} = \frac{3(15-15)}{5} = \frac{0}{5} = 0.$$

Example 5. *Find the coefficient of skewness by Karl Pearson's method for the following data:*

Value	*6*	*12*	*18*	*24*	*30*	*36*	*42*
Frequency	*4*	*7*	*9*	*18*	*15*	*10*	*3*

Sol. **Calculation of $\bar{x}$, S.D.**

Value x	f	$d = x - A$, $A = 24$	$u = d/h$, $h = 6$	fu	fu^2
6	4	– 18	– 3	– 12	36
12	7	– 12	– 2	– 14	28
18	9	– 6	– 1	– 9	9
24	18	0	0	0	0
30	15	6	1	15	15
36	10	12	2	20	40
42	3	18	3	9	27
	N = 66			$\Sigma fu = 9$	$\Sigma fu^2 = 155$

$$\textbf{A.M. } \bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h = 24 + \left(\frac{9}{66}\right)6 = 24.82$$

$$\textbf{S.D.} = \sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} \times h = 6 \times \sqrt{\frac{155}{66} - \left(\frac{9}{66}\right)^2} = 8.94$$

Mode. **Grouping Table**

x	*I* f	*II*	*III*	*IV*	*V*	*VI*
6	4					
12	7	11				
18	9		16	20		
24	**18**	**27**			**34**	
30	15		**33**	**43**		**42**
36	10	25			28	
42	**3**		13			

Analysis Table

Column	*24*	*18*	*30*	*36*	*12*
I	1				
II	1	1			
III	1		1		
IV	1		1	1	
V	1	1			
VI	1	1	1		1
Total	6	3	3	1	1

$\therefore$ Mode = 24

$\therefore$ $$SK_P = \frac{\bar{x} - \text{Mode}}{\text{S.D.}} = \frac{24.82 - 24}{8.94} = \frac{0.82}{8.94} = 0.092.$$

Example 6. *Calculate Karl Pearson's coefficient of skewness for the following data:*

Income (in ₹)	*500—600*	*600—700*	*700—800*	*800—900*	*900—1000*	*1000—1100*
No. of employees	*8*	*12*	*4*	*2*	*1*	*1*

Sol. **Calculation of $\bar{x}$, Mode, S.D.**

Income (in ₹)	*No. of employees* f	*Mid-points of classes* x	$d = x - A$, $A = 750$	$u = d/h$, $h = 100$	fu	fu^2
500—600	8	550	– 200	– 2	– 16	32
600—700	12	650	– 100	– 1	– 12	12
700—800	4	750	0	0	0	0
800—900	2	850	100	1	2	2
900—1000	1	950	200	2	2	4
1000—1100	1	1050	300	3	3	9
	N = 28				$\Sigma fu = -21$	$\Sigma fu^2 = 59$

A.M. $$\bar{x} = A + \left(\frac{\Sigma fu}{N}\right)h = 750 + \left(-\frac{21}{28}\right)(100) = 750 - 75 = ₹\ 675$$

Mode. By inspection, modal class is 600–700

$$\therefore \quad \text{Mode} = l + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)h$$

Here $l = 600$, $\Delta_1 = 12 - 8 = 4$, $\Delta_2 = 12 - 4 = 8$, $h = 100$

$$\therefore \quad \text{Mode} = 600 + \left(\frac{4}{4+8}\right)(100) = 600 + 33.33 = ₹\ 633.33$$

$$\text{S.D.} = \left[\sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2}\right] \times h = \left[\sqrt{\frac{59}{28} - \left(-\frac{21}{28}\right)^2}\right] \times 100$$

$$= \sqrt{2.1071 - 0.5625} \times 100 = \sqrt{1.5446} \times 100 = 1.2428 \times 100 = ₹\ 124.28$$

Now, Karl Pearson's coeff. of skewness $= \dfrac{\bar{x} - \text{Mode}}{\text{S.D.}} = \dfrac{675 - 633.33}{124.28} = 0.34.$

ASSIGNMENT

1. A frequency distribution gives the following results:

Coeff. of variation = 5

Karl Pearson's Coeff. of Skewness = 0.5

S.D. = 2

Find A.M. and Mode of the distribution.

2. Find Pearson's coeff. of skewness from the following frequency distribution:

Height (in inches)	60–62	63–65	66–68	69–71	72–74
Frequency	5	18	42	27	8

3. From the following data, calculate the coefficient of skewness based on mean, median and S.D.

Variable	100–110	110–120	120–130	130–140	140–150	150–160	160–170	170–180
Frequency	4	16	36	52	64	40	32	11

4. From the following data, find out the Karl Pearson's coefficient of skewness:

Measurement	10	11	12	13	14	15
Frequency	2	4	10	8	5	1

5. Calculate Karl Pearson's coefficient of skewness for the following frequency distribution:

Marks more than	0	10	20	30	40	50	60	70
No. of students	100	90	75	50	25	15	5	0

6. For the following frequency distribution, calculate the value of Karl Pearson's coeff. of skewness:

Temp. (°C)	– 40 to – 30	– 30 to – 20	– 20 to – 10	– 10 to 0	0 to 10	10 to 20	20 to 30
No. of days	10	28	30	42	65	180	10

7. From the following data, calculate Karl Pearson's coefficient of skewness:

Marks (above)	0	10	20	30	40	50	60	70	80
No. of students	150	140	100	80	80	70	30	14	0

8. Find out the mean wage and coefficient of skewness from the following data:

35 men gets at the rate of ₹ 4.5 per man
40 men gets at the rate of ₹ 5.5 per man
48 men gets at the rate of ₹ 6.5 per man
100 men gets at the rate of ₹ 7.5 per man
125 men gets at the rate of ₹ 8.5 per man
87 men gets at the rate of ₹ 9.5 per man
43 men gets at the rate of ₹ 10.5 per man
22 men gets at the rate of ₹ 11.5 per man.

9. Calculate Karl Pearson's coefficient of skewness:

Wages (in ₹)	70–80	80–90	90–100	100–110	110–120	120–130	130–140	140–150
No. of workers	12	18	35	42	50	45	20	8

10. Find the mean, mode, S.D. and Karl Pearson's coeff. of skewness for the following:

Yrs. under	10	20	30	40	50	60
No. of persons	15	32	51	78	97	109

11. Compute the coeff. of skewness from the following figures : 25, 15, 23, 40, 27, 25, 23, 25, 20.

12. In a discrete series of 20 terms, the sum of the terms is 200, the sum of the squares of the terms is 5000 and the median is 15. Find Karl Pearson's coefficient of skewness.

13. Calculate the coefficient of skewness based on mean, median and standard deviation from the following data:

C.I.	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
f	3	6	11	24	28	16	9	3

[*M.T.U. (MBA) 2012*]

14. Which of the following two series is symmetrical?

Series α: Mean = 22, Median = 24, S.D. = 10

Series β: Mean = 22, Median = 25, S.D. = 12

15. Following table gives the data relating to marks obtained by students who appeared for B. Tech. III Semester examination in Mathematics III at a centre. Calculate Karl Pearson's coefficient of skewness from the said data:

Marks	0 to 10	10 to 20	20 to 30	30 to 40	40 to 50	50 to 60	60 to 70	70 to 80
No. of students	10	40	20	0	10	40	16	14

[**Hint.** Since max. frequency corresponds to two classes very far from each other so mode is ill-defined and $SK_P = \dfrac{3(\text{Mean} - \text{Median})}{\text{S.D.}}$].

Answers

1. Mean = 40, Mode = 39 **2.** 0.0356 **3.** – 0.0087 **4.** 0.3604

5. – 0.0627 **6.** – 0.6617 **7.** – 0.7539

8. Mean wage = ₹ 8.07, Coeff. of skewness = – 0.2422 **9.** – 0.3314

10. Mean = 29.95, Mode = 35, S.D. = 15.49, Coeff. of skewness = – 0.32 **11.** – 0.03

12. – 1.22 **13.** – 0.08608 **14.** Series α **15.** 0.754.

3.13.2. Method of Moments

In this method, second and third central moments of the distribution are used. This measure of skewness is called the **Moment coefficient of skewness** and is given by:

$$\textbf{Moment coefficient of skewness} = \frac{\mu_3}{\sqrt{\mu_2^3}}.$$

[*G.B.T.U. (C.O.) 2009, 2011*]

For a symmetrical distribution, its value would come out to be zero. The coefficient of skewness as calculated by this method gives the magnitude as well as direction of the skewness present in the distribution.

In Statistics, we define $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3}$.

$\therefore$ Moment coefficient of skewness can also be written as $= \dfrac{\mu_3}{\sqrt{\mu_2^3}} = \pm\sqrt{\beta_1}$.

The sign with $\sqrt{\beta_1}$ is to be taken as that of μ_3. The moment coefficient of skewness is also denoted by γ_1. The moment coefficient of skewness is generally denoted by 'SK_M'.

EXAMPLES

Example 1. *The first three central moments of a distribution are 0, 15, – 31. Find the moment coefficient of skewness.*

Sol. We have $\mu_1 = 0$, $\mu_2 = 15$ and $\mu_3 = -31$

Moment coefficient of skewness $= \dfrac{\mu_3}{\sqrt{\mu_2^3}} = \dfrac{-31}{\sqrt{(15)^3}} = -\dfrac{31}{58.09} = -0.53.$

Example 2. *The first four moments of a distribution about the value 5 of the variable are 2, 20, 40 and 50. Calculate the moment coefficient of skewness.*

Sol. We have $A = 5$, $\mu_1' = 2$, $\mu_2' = 20$, $\mu_3' = 40$ and $\mu_4' = 50$

Now $\mu_2 = \mu_2' - (\mu_1')^2 = 20 - (2)^2 = 16$

$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = 40 - 3(2)(20) + 2(2)^3$

$= 40 - 120 + 16 = -64$

Moment coefficient of skewness $= \dfrac{\mu_3}{\sqrt{\mu_2^3}} = \dfrac{-64}{\sqrt{(16)^3}} = \dfrac{-64}{64} = -1.$

Example 3. *Calculate the moment coefficient of skewness for the following distribution :*

Classes	*2.5–7.5*	*7.5–12.5*	*12.5–17.5*	*17.5–22.5*	*22.5–27.5*	*27.5–32.5*	*32.5–37.5*
Frequency	*8*	*15*	*20*	*32*	*23*	*17*	*5*

Sol. **Calculation of Moment Coefficient of Skewness**

Classes	*f*	*Mid-pts.* *x*	*d = x – A* *A = 20*	*u = d / h* *h = 5*	*fu*	fu^2	fu^3
2.5–7.5	8	5	– 15	– 3	– 24	72	– 216
7.5–12.5	15	10	– 10	– 2	– 30	60	– 120
12.5–17.5	20	15	– 5	– 1	– 20	20	– 20
17.5–22.5	32	20	0	0	0	0	0
22.5–27.5	23	25	5	1	23	23	23
27.5–32.5	17	30	10	2	34	68	136
32.5–37.5	5	35	15	3	15	45	135
	N = 120				$\Sigma fu = -2$	$\Sigma fu^2 = 288$	$\Sigma fu^3 = -62$

Now, $\mu_1' = \left(\frac{\Sigma fu}{N}\right)h = \left(\frac{-2}{120}\right)5 = -0.083$

$\mu_2' = \left(\frac{\Sigma fu^2}{N}\right)h^2 = \left(\frac{288}{120}\right)5^2 = 60$

$\mu_3' = \left(\frac{\Sigma fu^3}{N}\right)h^3 = \left(\frac{-62}{162}\right)5^3 = -64.583$

Now, $\mu_2 = \mu_2' - \mu_1'^2 = 60 - (-0.083)^2 = 59.993$

$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2\mu_1'^3 = -64.583 - 3(-0.083)(60) + 2(-0.083)^3$

$= -49.644.$

$\therefore$ Moment coefficient of skewness $= \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{-49.644}{\sqrt{(59.993)^3}} = -0.1068.$

ASSIGNMENT

1. The first-three central moments of a distribution are 0, 2.5, 0.7. Find the value of the moment coefficient of skewness. (*U.P.T.U. 2015*)

2. In a certain distribution, the first four moments about the point 4 are – 1.5, 17, – 30 and 308. Calculate the moment coefficient of skewness. (*U.P.T.U. 2014*)

3. The first three moments of a frequency distribution about origin '5' are – 0.55, 4.46 and – 0.43. Find the moment coefficient of skewness.

4. Calculate the moment coefficient of skewness for the following data:

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

5. Calculate the moment coefficient of skewness from the following data:

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

6. For the following frequency distribution, find the first four moments about the mean. Also find the value of β_1. Is it a symmetrical distribution?

x	2	3	4	5	6
f	1	3	7	3	1

7. Compute the coefficient of skewness from the following data: [*G.B.T.U. (C.O.) 2009*]

x	6	7	8	9	10	11	12
f	3	6	9	13	8	5	4

8. In two frequency distributions, the second moments about mean are 36 and 49 respectively while third moments about mean are 43.2 and 85.75. Compare the skewness in the two frequency distributions. (*G.B.T.U. 2012*)

9. The first three moments about the origin are given by $\nu_1 = \frac{n+1}{2}$, $\nu_2 = \frac{(n+1)(2n+1)}{6}$, and $\nu_3 = \frac{n(n+1)^2}{4}$. Examine the skewness of the data.

10. (*i*) Define skewness of a distribution. (*U.P.T.U. 2015*)

(*ii*) Define the coefficients of skewness. (*M.T.U. 2012, 2013*)

Answers

1. 0.17708 **2.** 0.7017 **3.** 0.7781 **4.** 0.0726

5. 0 **6.** 0, 0.933, 0, 2.533, Yes **7.** 0.0903.

8. γ_1 (for I distribution) = 0.2, γ_2 (for II distribution) = 0.25

Second distribution is more positively skewed than the first.

9. data is symmetrical.

3.14 KURTOSIS

[*U.P.T.U. (C.O.) 2008; U.P.T.U. 2006, 2015*]

Given two frequency distributions which have the same variability as measured by the standard deviation, they may be relatively more or less flat topped than the normal curve. A frequency curve may be symmetrical but it may not be equally flat topped with the normal curve. The relative flatness of the top is called **kurtosis** and is measured by β_2. Kurtosis refers to the bulginess of the curve of a frequency distribution.

Curves which are neither flat nor sharply peaked are called normal curves or **mesokurtic curves.**

Curves which are flatter than the normal curve are called **platykurtic curves.**

Curves which are more sharply peaked than the normal curve are called **leptokurtic curves.**

3.15 MEASURE OF KURTOSIS

[*G.B.T.U. (C.O.) 2009, 2011; U.P.T.U. 2007*]

The measure of kurtosis is denoted by β_2 and is defined as

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

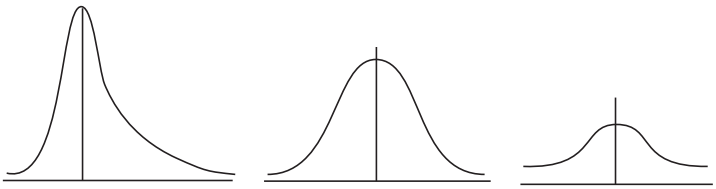
where μ_2 and μ_4 are respectively the second and fourth moments about mean of the distribution.

If $\beta_2 > 3$, the distribution is **leptokurtic.** If $\beta_2 = 3$, the distribution is ***mesokurtic***. If $\beta_2 < 3$, the distribution is **platykurtic.** The kurtosis of a distribution is also measured by using Greek letter 'γ_2' which is defined as $\gamma_2 = \beta_2 - 3$.

$\therefore \quad \gamma_2 > 0 \Rightarrow \beta_2 - 3 > 0 \Rightarrow \beta_2 > 3 \Rightarrow$ the distribution is **leptokurtic.**

Similarly, if $\gamma_2 = 0$, then $\beta_2 = 3 \Rightarrow$ The distribution is **mesokurtic.**

$\gamma_2 < 0 \Rightarrow \beta_2 < 3 \Rightarrow$ the distribution is **platykurtic.**



3.15.1. Steps for Computing β_2

I. *If the value of* μ_2 *and* μ_4 *are given, then find* β_2 *by using the formula:* $\beta_2 = \dfrac{\mu_4}{\mu_2^{\,2}}$.

II. *If raw moments* μ_1', μ_2', μ_3' *and* μ_4' *are given, then calculate:*

$\mu_2 = \mu_2' - \mu_1'^2$ *and* $\mu_4 = \mu_4' - 4\mu_3'\,\mu_1' + 6\mu_2'\,\mu_1'^2 - 3\mu_1'^4$

Now, find $\beta_2 = \dfrac{\mu_4}{\mu_2^2}$.

III. *If moments are not given, then first find* μ_2 *and* μ_4 *by using the given data and then use the formula:* $\beta_2 = \dfrac{\mu_4}{\mu_2^2}$.

IV. *The given distribution is leptokurtic, mesokurtic and platykurtic according as* $\beta_2 > 3$, $\beta_2 = 3$ *and* $\beta_2 < 3$ *respectively.*

EXAMPLES

Example 1. *The first four moments about mean of a frequency distribution are 0, 100, – 7 and 35000. Discuss the kurtosis of the distribution.*

Sol. We have, $\mu_1 = 0, \mu_2 = 100, \mu_3 = -7$ and $\mu_4 = 35000$

Now, $$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{35000}{(100)^2} = 3.5 > 3$$

$\therefore$ The distribution is **leptokurtic.**

Example 2. *The first four moments of a distribution about the value '4' of the variable are – 1.5, 17,– 30 and 108. State whether the distribution is leptokurtic or platykurtic.*

(U.P.T.U. 2007, 2014)

Sol. We have, $\mu_1' = -1.5, \mu_2' = 17, \mu_3' = -30, \mu_4' = 108$

Moments about mean:

$$\mu_2 = \mu_2' - \mu_1'^2 = 17 - (-1.5)^2 = 14.75$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$
$$= 108 - 4(-30)(-1.5) + 6(17)(-1.5)^2 - 3(-1.5)^4 = 142.3125$$

Kurtosis: $$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{142.3125}{(14.75)^2} = 0.6541$$

Since $\beta_2 < 3$, the distribution is **platykurtic.**

Example 3. *The first four moments of a distribution about x = 4 are 1, 4, 10 and 45. Obtain the various characteristics of the distribution on the basis of the given information. Comment upon the nature of the distribution.*

Sol. We have A = 4, $\mu_1' = 1$, $\mu_2' = 4$, $\mu_3' = 10$ and $\mu_4' = 45$

Moments about mean:

$$\mu_1 = 0 \text{ (always)}$$
$$\mu_2 = \mu_2' - \mu_1'^2 = 4 - (1)^2 = 3$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = 10 - 3(4)(1) + 2(1)^3 = 0$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$
$$= 45 - 4(10)(1) + 6(4)(1)^2 - 3(1)^4 = 26$$

Skewness: Moment coefficient of skewness, $\gamma_1 = \dfrac{\mu_3}{\sqrt{\mu_2^3}} = \dfrac{0}{\sqrt{(3)^3}} = \mathbf{0}$.

$\therefore$ The distribution is **symmetrical.**

Kurtosis: $\beta_2 = \dfrac{\mu_4}{\mu_2^{\,2}} = \dfrac{26}{(3)^2} = 2.89 < 3$ $\therefore$ The distribution is **platykurtic.**

Example 4. *The standard deviation of a symmetric distribution is 5. What must be the value of the fourth moment about the mean in order that the distribution be*

(i) leptokurtic (ii) mesokurtic (iii) platykurtic?

Sol. We have, $\sigma = 5 \Rightarrow \sigma^2 = 25 \Rightarrow \mu_2 = 25$ $\quad | \because \mu_2 = \sigma^2$

Now, $\beta_2 = \dfrac{\mu_4}{\mu_2^{\,2}} = \dfrac{\mu_4}{625}$

Thus, the distribution will be

(*i*) Leptokurtic if $\beta_2 > 3 \Rightarrow \dfrac{\mu_4}{625} > 3 \Rightarrow \mu_4 > 1875$

(*ii*) Mesokurtic if $\beta_2 = 3 \Rightarrow \dfrac{\mu_4}{625} = 3 \Rightarrow \mu_4 = 1875$

(*iii*) Platykurtic if $\beta_2 < 3 \Rightarrow \dfrac{\mu_4}{625} < 3 \Rightarrow \mu_4 < 1875$.

Example 5. *The first four moments about the working mean 28.5 of a distribution are 0.294, 7.144, 42.409 and 454.98. Calculate the moments about the mean. Also evaluate β_1, β_2 and comment upon the skewness and kurtosis of the distribution.* (U.P.T.U. 2006)

Sol. We have, $\mu_1' = 0.294$, $\mu_2' = 7.144$, $\mu_3' = 42.409$, $\mu_4' = 454.98$

Moments about mean

$$\mu_1 = 0$$
$$\mu_2 = \mu_2' - \mu_1'^2 = 7.144 - (.294)^2 = 7.0576$$
$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3$$
$$= 42.409 - 3(7.144)(.294) + 2(.294)^3 = 36.1588$$
$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'\mu_1'^2 - 3\mu_1'^4$$
$$= 454.98 - 4(42.409)(.294) + 6(7.144)(.294)^2 - 3(.294)^4$$
$$= 408.7896$$

Calculation of β_1 and β_2

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = 3.7193 \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} = 8.2070$$

Skewness

Since β_1 is positive, $\gamma_1 = 1.9285$ | μ_3 is positive

$\therefore$ The distribution is **positively skewed.**

Kurtosis

Since $\beta_2 = 8.2070 > 3$

$\therefore$ The distribution is **leptokurtic.**

Example 6. *The first four moments of a distribution about the value '0' are – 0.20, 1.76, – 2.36 and 10.88. Find the moments about the mean and measure the kurtosis.*

(U.P.T.U. 2009)

Sol. We have, $\mu'_1 = -0.20, \mu'_2 = 1.76, \mu'_3 = -2.36, \mu'_4 = 10.88$

Moments about the mean:

$$\mu_1 = 0$$

$$\mu_2 = \mu'_2 - \mu_1'^2 = 1.76 - (-0.20)^2 = 1.72$$

$$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu_1'^3$$
$$= -2.36 - 3(1.76)(-0.20) + 2(-0.20)^3 = -1.32$$

$$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu_1'^2 - 3\mu_1'^4$$
$$= 10.88 - 4(-2.36)(-0.20) + 6(1.76)(-0.20)^2 - 3(-0.20)^4$$
$$= 9.4096$$

Kurtosis:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3.180638$$

Since, $\beta_2 > 3$ hence the distribution is **leptokurtic.**

Example 7. *The following table represents the height of a batch of 100 students. Calculate kurtosis.*

Height (in cm)	*59*	*61*	*63*	*65*	*67*	*69*	*71*	*73*	*75*
No. of students	*0*	*2*	*6*	*20*	*40*	*20*	*8*	*2*	*2*

[U.P.T.U. (C.O.) 2008]

Sol. To calculate β_2, we will have to first find the values of μ_2 and μ_4.

Moments about 67

$$\mu_1' = \left(\frac{\Sigma fu}{N}\right)h = \left(\frac{12}{100}\right)(2) = 0.24$$

$$\mu_2' = \left(\frac{\Sigma fu^2}{N}\right)h^2 = \left(\frac{164}{100}\right)(4) = 6.56$$

Height (cm) x	*No. of students* f	$u = \frac{x - 67}{2}$	fu	fu^2	fu^3	fu^4
59	0	– 4	0	0	0	0
61	2	– 3	– 6	18	– 54	162
63	6	– 2	– 12	24	– 48	96
65	20	– 1	– 20	20	– 20	20
67	40	0	0	0	0	0
69	20	1	20	20	20	20
71	8	2	16	32	64	128
73	2	3	6	18	54	162
75	2	4	8	32	128	512
	$N = \Sigma f = 100$		$\Sigma fu = 12$	$\Sigma fu^2 = 164$	$\Sigma fu^3 = 144$	$\Sigma fu^4 = 1100$

$$\mu_3' = \left(\frac{\Sigma fu^3}{N}\right) h^3 = \frac{144}{100} \times 8 = 11.52$$

$$\mu_4' = \left(\frac{\Sigma fu^4}{N}\right) h^4 = \frac{1100}{100} \times 16 = 176$$

Moments about mean

$$\mu_2 = \mu_2' - \mu_1'^2 = 6.56 - (.24)^2 = 6.5024$$

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

$$= 176 - 4\,(11.52)\,(0.24) + 6\,(6.56)\,(0.24)^2 - 3\,(0.24)^4 = 167.19798$$

Kurtosis

Measure of kurtosis $\beta_2 = \frac{\mu_4}{\mu_2^2} = 3.9544 > 3$

Hence, the distribution is **leptokurtic.**

ASSIGNMENT

1. The first four moments about mean of a frequency distribution are 0, 60, – 50 and 8020 respectively. Discuss the kurtosis of the distribution.

2. The μ_2 and μ_4 for a distribution are found to be 2 and 12 respectively. Discuss the kurtosis of the distribution.

3. (*i*) The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Test the kurtosis of the distribution. (*M.T.U. 2013*)

(*ii*) Define skewness and kurtosis of a distribution. The first four moments of a distribution are 0, 2.5, 0.7 and 18.71. Find the coefficient of skewness and kurtosis. (*U.P.T.U. 2015*)

4. The standard deviation of symmetric distribution is 4. What must be the value of μ_4 so that the distribution may be mesokurtic?

5. (*i*) If the first four moments about the value '5' of the variable are – 4, 22, – 117 and 560, find the value of β_2 and discuss the kurtosis.

(*ii*) The first four moments of a distribution about the value 5 of the variable are 2, 20, 40 and 50. Calculate the moments about the mean and comment upon the skewness and kurtosis of the distribution. (*G.B.T.U. 2011*)

6. (*i*) Calculate the value of β_2 for the following distribution:

Class	2.5–7.5	7.5–12.5	12.5–17.5	17.5–22.5	22.5–27.5	27.5–32.5	32.5–37.5
Frequency	8	15	20	32	23	17	5

(*ii*) Compute the value of β_2 for the following distribution. Is the distribution platykurtic?

Class	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Frequency	1	20	69	108	78	22	2

7. (*i*) Calculate $\mu_1, \mu_2, \mu_3, \mu_4$ for the frequency distribution of heights of 100 students given in the following table and hence find coefficient of skewness and kurtosis.

Height (*cm.*) *Class interval*	144.5 – 149.5	149.5 – 154.5	154.5 – 159.5	159.5 – 164.5	164.5 – 169.5	169.5 – 174.5	174.5 – 179.5
Frequency	2	4	13	31	32	15	3

(*G.B.T.U. 2011*)

(*ii*) Find all four central moments and discuss skewness and kurtosis for the frequency distribution given in the following table:

Range of Expenditure (*in ₹ 100 per month*)	2–4	4–6	6–8	8–10	10–12
No. of families	38	292	389	212	69

[*G.B.T.U. 2013; M.T.U. 2012*]

8. (*i*) Find the measures of skewness and kurtosis on the basis of moments for the following distribution:

x	1	3	5	7	9
f	1	4	6	4	1

[*G.B.T.U.* (*C.O.*) *2011*]

(*ii*) Find the measure of skewness and kurtosis on the basis of moments for the following distribution:

Marks	5–15	15–25	25–35	35–45	45–55
No. of Students	1	3	5	7	4

(*M.T.U. (MBA) 2011*)

9. Calculate β_1 and β_2 from the following data:

Profit (in lakhs of ₹)	10–20	20–30	30–40	40–50	50–60
No. of companies	18	20	30	22	10

Indicate the nature of frequency curve.

10. Prove that the frequency distribution curve of the following frequency distribution is leptokurtic.

Class	10–15	15–20	20–25	25–30	30–35	35–40	40–45	45–50	50–55
Frequency	1	4	8	19	35	20	7	5	1

11. Calculate the first four moments about the mean of the following distribution:

x	2	2.5	3	3.5	4	4.5	5
f	5	38	65	92	70	40	10

Also find the measures of skewness and kurtosis. (*M.T.U. 2012*)

12. Calculate the first four moments about the mean for the following frequency distribution and hence find the coefficient of skewness and kurtosis and comment upon the nature of the distribution.

Class-interval	5–10	10–15	15–20	20–25	25–30	30–35	35–40
Frequency	6	8	17	21	15	11	2

(*G.B.T.U. 2013*)

13. Define the coefficients of kurtosis. [*M.T.U. 2014; G.B.T.U. (C.O.) 2009, 2011; U.P.T.U. 2007*]

14. (*i*) What do you mean by kurtosis? Explain in brief. [*U.P.T.U. (C.O.) 2008*]

(*ii*) Define kurtosis of a distribution. (*U.P.T.U. 2006*)

Answers

1. $\beta_2 = 2.2278$, Platykurtic **2.** $\beta_2 = 3$, Mesokurtic

3. (*i*) $\beta_2 = 3$, Mesokurtic (*ii*) 0.17708, 2.9936 **4.** $\mu_4 = 768$

5. (*i*) $\beta_2 = 0.8889$, Platykurtic

(*ii*) $\mu_1 = 0$, $\mu_2 = 16$, $\mu_3 = -64$, $\mu_4 = 162$

$\gamma_1 = -1$, $\beta_2 = 0.6328$; Negatively skewed and platykurtic

6. (*i*) $\beta_2 = 2.3216$, Platykurtic (*ii*) $\beta_2 = 2.7240$, Yes

7. (*i*) $\mu_1 = 0$, $\mu_2 = 36.66$, $\mu_3 = -85.104$, $\mu_4 = 4373.3832$, $\gamma_1 = -0.3834$, $\beta_2 = 3.2541$

(*ii*) $\mu_1 = 0$, $\mu_2 = 37267.04$, $\mu_3 = 1746530.688$, $\mu_4 = 3567851989$

$y_1 = 0.24275$, $\beta_2 = 2.5689$, positively skewed and platykurtic.

8. (*i*) $\gamma_1 = 0$, $\beta_2 = 2.5$

(*ii*) $\mu_1 = 0$, $\mu_2 = 125$, $\mu_3 = -600$, $\mu_4 = 37625$, $\gamma_1 = -0.4293$, $\beta_2 = 2.408$, negatively skewed and platykurtic.

9. $\beta_1 = 0.0001$, $\beta_2 = 2.047$, Platykurtic.

11. $\mu_1 = 0$, $\mu_2 = 0.45328125$, $\mu_3 = 0.009890625$, $\mu_4 = 0.502111743$, $\gamma_1 = 0.0324$, $\beta_2 = 2.44379$, positively skewed and platykurtic.

12. $\mu_1 = 0$, $\mu_2 = 56$, $\mu_3 = -176.5625$, $\mu_4 = 7502.9375$, $\gamma_1 = -0.4213$, $\beta_2 = 2.3925$; negatively skewed and platykurtic.

3.16 CURVE FITTING

Let there be two variables x and y which give us a set of n pairs of numerical values (x_1, y_1), (x_2, y_2).....(x_n, y_n). In order to have an approximate idea about the relationship of these two variables, we plot these n paired points on a graph thus, we get a diagram showing the simultaneous variation in values of both the variables called *scatter or dot diagram*. From scatter diagram, we get only an approximate non-mathematical relation between two variables. Curve fitting means an exact relationship between two variables by algebraic equations, infact this relationship is the equation of the curve. Therefore, curve fitting means to form an equation of the curve from the given data. Curve fitting is considered of immense importance both from the point of view of theoretical and practical statistics.

Theoretically, it is useful in the study of correlation and regression. Practically, it enables us to represent the relationship between two variables by simple algebraic expressions *e.g.*, polynomials, exponential or logarithmic functions.

It is also used to estimate the values of one variable corresponding to the specified values of the other variable.

The constants occurring in the equation of approximate curve can be found by following methods:

(*i*) Graphical method (*ii*) Method of group averages

(*iii*) Method of least squares (*iv*) Method of moments.

Out of the above four methods, we will only discuss and study here *method of least squares*.

3.17 METHOD OF LEAST SQUARES

[*U.P.T.U. MCA (C.O.) 2008; U.P.T.U. (C.O.) 2008 ; U.P.T.U. 2008*]

Method of least squares provides a unique set of values to the constants and hence suggests a curve of best fit to the given data.

Suppose we have m-paired observations (x_1, y_1), (x_2, y_2), ..., (x_m, y_m) of two variables x and y. It is required to fit a polynomial of degree n of the type

$$y = a + bx + cx^2 + ... + kx^n \qquad ...(1)$$

of these values. We have to determine the constants $a, b, c, ..., k$ such that it represents the curve of best fit of that degree.

In case $m = n$, we get in general a unique set of values satisfying the given system of equations.

But if $m > n$ then, we get m equations by putting different values of x and y in equation (1) and we want to find only the values of n constants. Thus, there may be no such solution to satisfy all m equations.

Therefore, we try to find out those values of $a, b, c, ..., k$ which satisfy all the equations as nearly as possible. We apply the method of least squares in such cases.

Putting $x_1, x_2, ..., x_m$ for x in (1), we get

$$\begin{aligned} y_1' &= a + bx_1 + cx_1^2 + ... + kx_1^n \\ y_2' &= a + bx_2 + cx_2^2 + ... + kx_2^n \\ &\vdots \qquad\qquad \vdots \\ y_m' &= a + bx_m + cx_m^2 + ... + kx_m^n \end{aligned}$$

where $y_1', y_2', ..., y_m'$ are the expected values of y for $x = x_1, x_2, ..., x_m$ respectively. The values $y_1, y_2, ..., y_m$ are called observed values of y corresponding to $x = x_1, x_2, ..., x_m$ respectively.

The expected values are different from observed values, the difference $y_r - y_r'$ for different values of r are called *residuals.*

Introduce a new quantity U such that

$$U = \Sigma(y_r - y_r')^2 = \Sigma(y_r - a - bx_r - cx_r^2 - ..... - kx_r^n)^2$$

The constants $a, b, c, ..., k$ are chosen in such a way that the sum of the squares of residuals is minimum.

Now the condition for U to be maximum or minimum is $\frac{\partial U}{\partial a} = 0 = \frac{\partial U}{\partial b} = \frac{\partial U}{\partial c} = ... = \frac{\partial U}{\partial k}$.

On simplifying these relations, we get

$$\Sigma y = ma + b\Sigma x + ... + k\Sigma x^n$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + ... + k\,\Sigma x^{n+1}$$
$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + ... + k\,\Sigma x^{n+2}$$
$$\vdots \qquad\qquad \vdots$$
$$\Sigma x^n y = a\Sigma x^n + b\Sigma x^{n+1} + ... + k\,\Sigma x^{2n}$$

These are known as *Normal equations* and can be solved as simultaneous equations to give the values of the constants $a, b, c, ..., k$. These equations are $(n + 1)$ in number.

If we calculate the second order partial derivatives and these values are put, they give a positive value of the function, so U is minimum.

This method does not help us to choose the degree of the curve to be fitted but helps us in finding the values of the constants when the form of the curve has already been chosen.

3.18 FITTING A STRAIGHT LINE

Let (x_i, y_i), $i = 1, 2, ..., n$ be n sets of observations of related data and

$$y = a + bx \qquad ...(1)$$

be the straight line to be fitted. The residual at $x = x_i$ is

$$E_i = y_i - f(x_i) = y_i - a - bx_i$$

Introduce a new quantity U such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

By the principle of Least squares, U is minimum

$$\therefore \qquad \frac{\partial U}{\partial a} = 0 \quad \text{and} \quad \frac{\partial U}{\partial b} = 0$$

$$\therefore \qquad 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0 \qquad \text{or} \quad \boxed{\Sigma y = na + b\Sigma x} \qquad ...(2)$$

and $$2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0 \qquad \text{or} \quad \boxed{\Sigma xy = a\Sigma x + b\Sigma x^2} \qquad ...(3)$$

Since x_i, y_i are known, equations (2) and (3) result two equations in a and b. Solving these, the best values for a and b can be known and hence equation (1).

Note. In case of change of origin,

if n is odd then, $\quad u = \dfrac{x - (\text{middle term})}{\text{interval } (h)}$

but if n is even then, $\quad u = \dfrac{x - (\text{mean of two middle terms})}{\frac{1}{2}(\text{interval})}$.

EXAMPLES

Example 1. *By the method of least squares, find the straight line that best fits the following data:*

x:	*1*	*2*	*3*	*4*	*5*
y:	*14*	*27*	*40*	*55*	*68.*

(U.P.T.U. 2008)

Sol. Let the straight line of best fit be $y = a + bx$ ...(1)

Normal equations are $\Sigma y = ma + b\Sigma x$ ...(2)

and $\Sigma xy = a\Sigma x + b\Sigma x^2$ ...(3)

Here $m = 5$

Table is as below:

x	y	xy	x^2
1	14	14	1
2	27	54	4
3	40	120	9
4	55	220	16
5	68	340	25
$\Sigma x = 15$	$\Sigma y = 204$	$\Sigma xy = 748$	$\Sigma x^2 = 55$

Substituting in (2) and (3), we get

$$204 = 5a + 15b$$
$$748 = 15a + 55b$$

Solving, we get $a = 0$, $b = 13.6$

Hence required straight line is $y = 13.6x$

Example 2. *Fit a straight line to the following data by least square method:*

x:	*0*	*1*	*2*	*3*	*4*
y:	*1*	*1.8*	*3.3*	*4.5*	*6.3.*

(*U.K.T.U. 2011*)

Sol. Let the straight line obtained from the given data be $y = a + bx$ then the normal equations are

$$\Sigma y = ma + b\,\Sigma x \quad ...(1)$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad ...(2)$$

Here, $m = 5$

x	y	xy	x^2
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9
4	6.3	25.2	16
$\Sigma x = 10$	$\Sigma y = 16.9$	$\Sigma xy = 47.1$	$\Sigma x^2 = 30$

From (1) and (2), $16.9 = 5a + 10b$

and $47.1 = 10a + 30b$

Solving, we get $a = 0.72, b = 1.33$

$\therefore$ Required line is $y = 0.72 + 1.33\,x$.

Example 3. *Show that the best fitting linear function for the points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) may be expressed in the form*

$$\begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \end{vmatrix} = 0 \qquad (i = 1, 2, ..., n)$$

Show that the line passes through the mean point $(\bar{x}, \bar{y})$.

Sol. Let the best fitting linear function be $y = a + bx$ ...(1)

Then the normal equations are

$$\Sigma y_i = na + b\Sigma x_i \quad ...(2)$$

and $$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 \quad ...(3)$$

Equations (1), (2), (3) may be rewritten as

$$bx - y + a = 0$$

$$b\Sigma x_i - \Sigma y_i + na = 0$$

and $$b\Sigma x_i^2 - \Sigma x_i y_i + a\Sigma x_i = 0$$

Eliminating a and b between these equations

$$\begin{vmatrix} x & y & 1 \\ \Sigma x_i & \Sigma y_i & n \\ \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \end{vmatrix} = 0 \quad ...(4)$$

which is the required best fitting linear function for the mean point $(\bar{x}, \bar{y})$,

$$\bar{x} = \frac{1}{n}\Sigma x_i, \quad \bar{y} = \frac{1}{n}\Sigma y_i .$$

Clearly, the line (4) passes through point $(\bar{x}, \bar{y})$ as two rows of determinant being equal make it zero.

ASSIGNMENT

1. Fit a straight line to the following data regarding x as the independent variable:

(*i*)

x	1	2	3	4	5	6
y	1200	900	600	200	110	50

(*ii*)

x	71	68	73	69	67	65	66	67
y	69	72	70	70	68	67	68	64

(*iii*)

x	0	5	10	15	20	25
y	12	15	17	22	24	30

2. (*i*) Find the best values of a and b so that $y = a + bx$ fits the given data:

x	0	1	2	3	4
y	1.0	2.9	4.8	6.7	8.6

(*ii*) Fit a straight line of the form $y = a_0 + a_1x$ to the data: [*U.P.T.U. (C.O.) 2008*]

x	1	2	3	4	6	8
y	2.4	3.1	3.5	4.2	5.0	6.0

3. Fit a straight line approximate to the data:

x	1	2	3	4
y	3	7	13	21

4. A simply supported beam carries a concentrated load P(*lb*) at its mid-point. Corresponding to various values of P, the maximum deflection Y (*in*) is measured. The data are given below. Find a law of the type $Y = a + b$P

P	100	120	140	160	180	200
Y	0.45	0.55	0.60	0.70	0.80	0.85

5. What straight line best fits the following data in the least square sense?

x	1	2	3	4
y	0	1	1	2

[*G.B.T.U. (MCA) 2010*]

6. The weight of a calf taken at weekly intervals are given below. Fit a straight line using method of least squares and calculate the average rate of growth per week.

Age	1	2	3	4	5	6	7	8	9	10
Weight	52.5	58.7	65	70.2	75.4	81.1	87.2	95.5	102.2	108.4

7. Find the least square line for the data points

$(-1, 10)$, $(0, 9)$, $(1, 7)$, $(2, 5)$, $(3, 4)$, $(4, 3)$, $(5, 0)$ and $(6, -1)$.

8. Find the least square line $y = a + bx$ for the data:

x_i	-2	-1	0	1	2
y_i	1	2	3	3	4

9. If P is the pull required to lift a load W by means of a pulley block, find a linear law of the form $P = m$W $+ c$ connecting P and W, using the data:

P	12	15	21	25
W	50	70	100	120

where P and W are taken in kg-wt. (*U.P.T.U. 2007*)

10. (*i*) Using the method of least squares, fit a straight line to the following data:

x	1	2	3	4	5
y	2	4	6	8	10

(*ii*) Using the method of least squares, fit a straight line from the following data:

x	0	2	4	5	6
y	5.012	10	15	21	30

(*U.P.T.U. 2009*)

(*iii*) Find the least square line that fits the following data, assuming that x-values are free from error: [*U.P.T.U. MCA (SUM) 2008*]

x	1	2	3	4	5	6
y	5.04	8.12	10.64	13.18	16.20	20.04

Answers

1. (*i*) $y = 1361.97 - 243.42\,x$ (*ii*) $y = 39.5454 + 0.4242\,x$ (*iii*) $y = 11.285 + 0.7\,x$

2. (*i*) $y = 1 + 1.9x$ (*ii*) $y = 2.0253 + 0.502\,x$ **3.** $y = -4 + 6x$

4. $Y = 0.004P + 0.048$ **5.** $y = -0.5 + 0.6x$ **6.** $y = 45.74 + 6.16x$, 6.16

7. $y = -1.6071429x + 8.6428571$ **8.** $y = 2.6 + (0.7)\,x$ **9.** $P = 2.2759 + 0.1879\,W$

10. (*i*) $y = 2x$ (*ii*) $y = 3.07734 + 3.86031\,x$ (*iii*) $y = 2.0253 + 2.908\,x$.

3.19 FITTING OF AN EXPONENTIAL CURVE $y = ae^{bx}$

Taking logarithm on both sides, we get

$$\log_{10} y = \log_{10} a + bx \log_{10} e$$

i.e.,
$$\boxed{Y = A + BX} \qquad \ldots(1)$$

where $Y = \log_{10} y$, $A = \log_{10} a$, $B = b \log_{10} e$ and $X = x$

The normal equations for (1) are

$$\Sigma Y = nA + B\Sigma X \quad \text{and} \quad \Sigma XY = A\Sigma X + B\Sigma X^2$$

Solving these, we get A and B.

Then a = antilog A and $b = \dfrac{B}{\log_{10} e}$.

3.20 FITTING OF THE CURVE $y = ax^b$

Taking logarithm on both sides, we get

$$\log_{10} y = \log_{10} a + b \log_{10} x$$

i.e.,
$$\boxed{Y = A + BX} \qquad \ldots(1)$$

where $Y = \log_{10} y$, $A = \log_{10} a$, $B = b$ and $X = \log_{10} x$.

The normal equations to (1) are

$$\Sigma Y = nA + B\Sigma X$$

and $$\Sigma XY = A\Sigma X + B\Sigma X^2$$

which results A and B on solving and a = antilog A, b = B.

3.21 FITTING OF THE CURVE $y = ab^x$

Taking logarithm on both sides, we get

$$\log y = \log a + x \log b$$

$$\Rightarrow \quad \boxed{Y = A + BX} \qquad \ldots(1)$$

where $Y = \log y$, $A = \log a$, $B = \log b$, $X = x$.

This is a linear equation in Y and X.

For estimating A and B, normal equations are

$$\Sigma Y = nA + B\,\Sigma X$$

and $$\Sigma XY = A\,\Sigma X + B\,\Sigma X^2$$

where n is the number of pairs of values of x and y.

Ultimately, a = antilog (A) and b = antilog (B).

3.22 FITTING OF THE CURVE $pv^\gamma = k$

$$pv^\gamma = k \quad \Rightarrow \quad v = k^{1/\gamma} p^{-1/\gamma}$$

Taking logarithm on both sides, we get

$$\log v = \frac{1}{\gamma} \log k - \frac{1}{\gamma} \log p$$

$$\Rightarrow \quad \boxed{Y = A + BX}$$

where $Y = \log v$, $A = \frac{1}{\gamma} \log k$, $B = -\frac{1}{\gamma}$ and $X = \log p$

γ and k are determined by above equations. Normal equations are obtained as that of the straight line.

3.23 FITTING OF THE CURVE OF TYPE $xy = b + ax$

$$xy = b + ax \qquad \Rightarrow \quad y = \frac{b}{x} + a$$

$$\Rightarrow \quad Y = bX + a, \qquad \text{where } X = \frac{1}{x}.$$

Normal equations are $\Sigma Y = na + b\Sigma X$ and $\Sigma XY = a\Sigma X + b\Sigma X^2$.

3.24 FITTING OF THE CURVE $y = ax^2 + \frac{b}{x}$

Let the n points be (x_1, y_1), (x_2, y_2), ... , (x_n, y_n)

Error of estimate for i^{th} point (x_i, y_i) is

$$E_i = \left(y_i - ax_i^2 - \frac{b}{x_i} \right)$$

By principle of Least squares, the values of a and b are such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} \left(y_i - ax_i^2 - \frac{b}{x_i} \right)^2 \text{ is minimum.}$$

Normal equations are given by

$$\frac{\partial U}{\partial a} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} x_i^2 y_i = a \sum_{i=1}^{n} x_i^4 + b \sum_{i=1}^{n} x_i$$

and

$$\frac{\partial U}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \frac{y_i}{x_i} = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} \frac{1}{x_i^2}$$

or Dropping the suffix i, normal equations are

$$\boxed{\Sigma x^2 y = a\ \Sigma x^4 + b \Sigma x} \quad \text{and} \quad \boxed{\sum \frac{y}{x} = a\ \Sigma x + b \sum \frac{1}{x^2}.}$$

3.25 FITTING OF THE CURVE $y = ax + bx^2$ (*U.P.T.U. 2014*)

Error of estimate for i^{th} point (x_i, y_i) is $E_i = (y_i - ax_i - bx_i^2)$

By principle of Least squares, the values of a and b are such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - ax_i - bx_i^2)^2 \text{ is minimum.}$$

Normal equations are given by

$$\frac{\partial U}{\partial a} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i^3$$

and

$$\frac{\partial U}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} x_i^2 y_i = a \sum_{i=1}^{n} x_i^3 + b \sum_{i=1}^{n} x_i^4$$

or Dropping the suffix i, normal equations are

$$\boxed{\Sigma xy = a\ \Sigma x^2 + b \Sigma x^3} \quad \text{and} \quad \boxed{\Sigma x^2 y = a\ \Sigma x^3 + b \Sigma x^4.}$$

3.26 FITTING OF THE CURVE $y = ax + \frac{b}{x}$

Error of estimate for i^{th} point (x_i, y_i) is

$$E_i = y_i - ax_i - \frac{b}{x_i}$$

By principle of Least squares, the values of a and b are such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} \left(y_i - ax_i - \frac{b}{x_i} \right)^2 \text{ is minimum.}$$

Normal equations are given by

$$\frac{\partial U}{\partial a} = 0$$

$$\Rightarrow \quad 2 \sum_{i=1}^{n} \left(y_i - ax_i - \frac{b}{x_i} \right)(-x_i) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i^2 + nb \qquad \ldots(1)$$

and

$$\frac{\partial U}{\partial b} = 0$$

$$\Rightarrow \quad 2 \sum_{i=1}^{n} \left(y_i - ax_i - \frac{b}{x_i} \right)\left(-\frac{1}{x_i} \right) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} \frac{y_i}{x_i} = na + b \sum_{i=1}^{n} \frac{1}{x_i^2} \qquad \ldots(2)$$

Dropping the suffix i, normal equations are

$$\boxed{\Sigma xy = a\Sigma x^2 + nb} \text{ and } \boxed{\sum \frac{y}{x} = na + b \sum \frac{1}{x^2}}$$

where n is the no. of pairs of values of x and y.

3.27 FITTING OF THE CURVE $y = a + \frac{b}{x} + \frac{c}{x^2}$

Normal equations are

$$\Sigma y = ma + b \sum \frac{1}{x} + c \sum \frac{1}{x^2}$$

$$\sum \frac{y}{x} = a \sum \frac{1}{x} + b \sum \frac{1}{x^2} + c \sum \frac{1}{x^3}$$

$$\sum \frac{y}{x^2} = a \sum \frac{1}{x^2} + b \sum \frac{1}{x^3} + c \sum \frac{1}{x^4}$$

where m is number of pairs of values of x and y.

3.28 FITTING OF THE CURVE $y = \frac{c_0}{x} + c_1 \sqrt{x}$

Error of estimate for i^{th} point (x_i, y_i) is

$$E_i = y_i - \frac{c_0}{x_i} - c_1 \sqrt{x_i}$$

By principle of Least squares, the values of a and b are such that

$$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - \frac{c_0}{x_i} - c_i\sqrt{x_i})^2 \text{ is minimum.}$$

Normal equations are given by

$$\frac{\partial U}{\partial c_0} = 0 \quad \text{and} \quad \frac{\partial U}{\partial c_1} = 0$$

Now, $$\frac{\partial U}{\partial c_0} = 0$$

$$\Rightarrow \quad 2\sum_{i=1}^{n} \left(y_i - \frac{c_0}{x_i} - c_1\sqrt{x_i}\right)\left(-\frac{1}{x_i}\right) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} \frac{y_i}{x_i} = c_0 \sum_{i=1}^{n} \frac{1}{x_i^2} + c_1 \sum_{i=1}^{n} \frac{1}{\sqrt{x_i}} \qquad ...(1)$$

Also, $$\frac{\partial U}{\partial c_1} = 0$$

$$\Rightarrow \quad 2\sum_{i=1}^{n} \left(y_i - \frac{c_0}{x_i} - c_1\sqrt{x_i}\right)(-\sqrt{x_i}) = 0$$

$$\Rightarrow \quad \sum_{i=1}^{n} y_i\sqrt{x_i} = c_0 \sum_{i=1}^{n} \frac{1}{\sqrt{x_i}} + c_1 \sum_{i=1}^{n} x_i \qquad ...(2)$$

Dropping suffix i, normal equations (1) and (2) become

$$\sum \frac{y}{x} = c_0 \sum \frac{1}{x^2} + c_1 \sum \frac{1}{\sqrt{x}}$$

and $$\sum y\sqrt{x} = c_0 \sum \frac{1}{\sqrt{x}} + c_1 \sum x.$$

3.29 FITTING OF THE CURVE $2^x = ax^2 + bx + c$

Normal equations are

$$\Sigma\, 2^x x^2 = a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2$$

$$\Sigma\, 2^x x = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x$$

and $$\Sigma\, 2^x = a\Sigma x^2 + b\Sigma x + mc$$

where, m is number of points (x_i, y_i)

3.30 FITTING OF THE CURVE $y = ae^{-3x} + be^{-2x}$

Normal equations are

$$\Sigma y e^{-3x} = a\ \Sigma e^{-6x} + b\ \Sigma e^{-5x}$$

and $$\Sigma y e^{-2x} = a\ \Sigma e^{-5x} + b\ \Sigma e^{-4x}$$

EXAMPLES

Example 1. *Find the curve of best fit of the type* $y = ae^{bx}$ *to the following data by the method of Least squares:*

x:	*1*	*5*	*7*	*9*	*12*
y:	*10*	*15*	*12*	*15*	*21.*

Sol. The curve to be fitted is $y = ae^{bx}$

or $Y = A + BX$, where, $Y = \log_{10} y$, $A = \log_{10} a$, $X = x$ and $B = b \log_{10} e$

$\therefore$ The normal equations are $\Sigma Y = 5A + B\Sigma X$

and $\Sigma XY = A\Sigma X + B\Sigma X^2$

$X = x$	y	$Y = log_{10} y$	X^2	XY
1	10	1.0000	1	1
5	15	1.1761	25	5.8805
7	12	1.0792	49	7.5544
9	15	1.1761	81	10.5849
12	21	1.3222	144	15.8664
$\Sigma X = 34$		$\Sigma Y = 5.7536$	$\Sigma X^2 = 300$	$\Sigma XY = 40.8862$

Substituting the above values in the normal equations, we get

$$5.7536 = 5A + 34B$$

and $$40.8862 = 34A + 300B$$

On solving, $A = 0.9766$; $B = 0.02561$

$$\therefore \quad a = \text{antilog}_{10} A = 9.4754 \; ; \; b = \frac{B}{\log_{10} e} = 0.059$$

Hence the required curve is $y = 9.4754e^{0.059x}$.

Example 2. *Determine the constants a and b by the method of least squares such that* $y = ae^{bx}$ *fits the following data:*

x	2	4	6	8	10
y	4.077	11.084	30.128	81.897	222.62

Sol. $y = ae^{bx}$

Taking log on both sides

$$\log y = \log a + bx \log e$$

or $Y = A + BX$,

where, $Y = \log y$, $A = \log a$, $B = b \log_{10} e$, $X = x$.

Normal equations are

$$\Sigma Y = mA + B\Sigma X \quad ...(1)$$

and $$\Sigma XY = A\Sigma X + B\Sigma X^2. \quad ...(2)$$

Here, $m = 5$.

Table is as follows:

x	y	X	Y	XY	X^2
2	4.077	2	.61034	1.22068	4
4	11.084	4	1.04469	4.17876	16
6	30.128	6	1.47897	8.87382	36
8	81.897	8	1.91326	15.30608	64
10	222.62	10	2.347564	23.47564	100
		$\Sigma X = 30$	$\Sigma Y = 7.394824$	$\Sigma XY = 53.05498$	$\Sigma X^2 = 220$

Substituting these values in equations (1) and (2), we get

$$7.394824 = 5A + 30B$$

and $$53.05498 = 30A + 220B.$$

Solving, we get $A = 0.1760594$ and $B = 0.2171509$

$\therefore$ $a = \text{antilog}(A) = \text{antilog}(0.1760594) = 1.49989$

and $$b = \frac{B}{\log_{10} e} = \frac{0.2171509}{0.4342945} = 0.50001$$

Hence the required equation is

$$y = 1.49989\, e^{0.50001x}.$$

Example 3. *Obtain a relation of the form $y = ab^x$ for the following data by the method of least squares:* [G.B.T.U. MCA (SUM) 2010]

x	2	3	4	5	6
y	8.3	15.4	33.1	65.2	127.4

Sol. The curve to be fitted is $y = ab^x$

or $$Y = A + Bx,$$

where, $A = \log_{10} a$, $B = \log_{10} b$ and $Y = \log_{10} y$.

$\therefore$ The normal equations are $\Sigma Y = 5A + B\Sigma x$

and $$\Sigma xY = A\Sigma x + B\Sigma x^2.$$

x	y	$Y = \log_{10} y$	x^2	xY
2	8.3	0.9191	4	1.8382
3	15.4	1.1872	9	3.5616
4	33.1	1.5198	16	6.0792
5	65.2	1.8142	25	9.0710
6	127.4	2.1052	36	12.6312
$\Sigma x = 20$		$\Sigma Y = 7.5455$	$\Sigma x^2 = 90$	$\Sigma xY = 33.1812$

Substituting the above values, we get

$$7.5455 = 5A + 20B \quad \text{and} \quad 33.1812 = 20A + 90B.$$

On solving $A = 0.31$ and $B = 0.3$

$\therefore$ $a = \text{antilog } A = 2.04$ and $b = \text{antilog } B = 1.995$.

Hence the required curve is $y = 2.04(1.995)^x$.

Example 4. *Obtain the least squares fit of the form* $f(t) = a\,e^{-3t} + be^{-2t}$ *for the data:* (U.P.T.U. 2008)

t:	*0.1*	*0.2*	*0.3*	*0.4*
f(t):	*0.76*	*0.58*	*0.44*	*0.35*

Sol. Normal equations to the curve $f(t) = a\,e^{-3t} + be^{-2t}$ are:

$$\Sigma f(t)\,e^{-3t} = a\,\Sigma e^{-6t} + b\,\Sigma e^{-5t} \quad \ldots(1)$$

$$\Sigma f(t)\,e^{-2t} = a\,\Sigma e^{-5t} + b\,\Sigma e^{-4t} \quad \ldots(2)$$

See art. 3.30

Table of values is

t	$f(t)$	e^{-4t}	e^{-5t}	e^{-6t}	$f(t)\,e^{-2t}$	$f(t)\,e^{-3t}$
0.1	0.76	0.6703	0.6065	0.5488	0.6222	0.5630
0.2	0.58	0.4493	0.3679	0.3012	0.3888	0.3183
0.3	0.44	0.3012	0.2231	0.1653	0.2415	0.1789
0.4	0.35	0.2019	0.1353	0.0907	0.1573	0.1054
Total		Σe^{-4t} = 1.6227	Σe^{-5t} = 1.3328	Σe^{-6t} = 1.106	$\Sigma f(t)\,e^{-2t}$ = 1.4098	$\Sigma f(t)\,e^{-3t}$ = 1.1656

Substituting values in (1) and (2), we get

$$1.106\,a + 1.3328\,b = 1.1656$$

$$1.3328\,a + 1.6227\,b = 1.4098$$

On solving, we get $a = 0.6778$, $b = 0.3121$.

Hence the least squares fit is $f(t) = 0.6778\,e^{-3t} + 0.3121e^{-2t}$.

Example 5. *By the method of least squares, find the curve* $y = ax + bx^2$ *that best fits the following data:* (U.P.T.U. 2014)

x	*1*	*2*	*3*	*4*	*5*
y	*1.8*	*5.1*	*8.9*	*14.1*	*19.8*

Sol. Normal equations are

$$\Sigma xy = a\Sigma x^2 + b\Sigma x^3 \quad \ldots(1)$$

and

$$\Sigma x^2 y = a\Sigma x^3 + b\Sigma x^4 \quad \ldots(2)$$

Let us form a table as below:

x	y	x^2	x^3	x^4	xy	x^2y
1	1.8	1	1	1	1.8	1.8
2	5.1	4	8	16	10.2	20.4
3	8.9	9	27	81	26.7	80.1
4	14.1	16	64	256	56.4	225.6
5	19.8	25	125	625	99	495
Total		$\Sigma x^2 = 55$	$\Sigma x^3 = 225$	$\Sigma x^4 = 979$	$\Sigma xy = 194.1$	$\Sigma x^2y = 822.9$

Substituting these values in equations (1) and (2), we get

$$194.1 = 55\,a + 225\,b$$

and $$822.9 = 225\,a + 979\,b$$

$$\Rightarrow \qquad a = \frac{83.85}{55} \simeq 1.52 \quad \text{and} \quad b = \frac{317.4}{664} \simeq .49$$

Hence required parabolic curve is $y = 1.52\,x + 0.49\,x^2$.

Example 6. *Fit the curve $pv^{\gamma} = k$ to the following data:* [U.P.T.U. MCA (C.O.) 2007]

p (kg/cm^2)	0.5	1	1.5	2	2.5	3
v (litres)	1620	1000	750	620	520	460

Sol. $$pv^{\gamma} = k$$

$$\Rightarrow \qquad v = \left(\frac{k}{p}\right)^{1/\gamma} = k^{1/\gamma}\,p^{-1/\gamma}$$

Taking log, $$\log v = \frac{1}{\gamma}\log k - \frac{1}{\gamma}\log p$$

which is of the form $$Y = A + BX$$

where $Y = \log v$, $X = \log p$, $A = \frac{1}{\gamma}\log k$ and $B = -\frac{1}{\gamma}$

p	v	X	Y	XY	X^2
0.5	1620	– 0.30103	3.20952	– 0.96616	0.09062
1	1000	0	3	0	0
1.5	750	0.17609	2.87506	0.50627	0.03101
2	620	0.30103	2.79239	0.84059	0.09062
2.5	520	0.39794	2.716	1.08080	0.15836
3	460	0.47712	2.66276	1.27046	0.22764
Total		$\Sigma X = 1.05115$	$\Sigma Y = 17.25573$	$\Sigma XY = 2.73196$	$\Sigma X^2 = 0.59825$

Here, $m = 6$

Normal equations are

$$17.25573 = 6A + 1.05115\,B$$

and $$2.73196 = 1.05115\,A + 0.59825\,B$$

Solving these, we get

$$A = 2.99911 \quad \text{and} \quad B = -0.70298$$

$$\therefore \qquad \gamma = -\frac{1}{B} = \frac{1}{0.70298} = 1.42252$$

Again, $$\log k = \gamma A = 4.26629$$

$$\therefore \qquad k = \text{antilog}\,(4.26629) = 18462.48$$

Hence, required curve is

$$pv^{1.42252} = 18462.48.$$

Example 7. *The pressure of the gas corresponding to various volumes V is measured, given by the following data:*

V (cm^3):	*50*	*60*	*70*	*90*	*100*
P ($kg\ cm^{-2}$):	*64.7*	*51.3*	*40.5*	*25.9*	*78*

Fit the data to the equation $PV^{\gamma} = C$.

Sol. $PV^{\gamma} = C$

$\Rightarrow$ $P = CV^{-\gamma}$

Taking log on both sides, we get

$$\log P = \log C - \gamma \log V$$

$\Rightarrow$ $Y = A + BX$

where, $Y = \log P$, $A = \log C$, $B = -\gamma$, $X = \log V$

Normal equations are

$$\Sigma Y = mA + B\Sigma X \quad \ldots(1)$$

and $$\Sigma XY = A\Sigma X + B\Sigma X^2 \quad \ldots(2)$$

Here $m = 5$

The table is as below:

V	*P*	*X = log V*	*Y = log P*	*XY*	X^2
50	64.7	1.69897	1.81090	3.07666	2.88650
60	51.3	1.77815	1.71012	3.04085	3.16182
70	40.5	1.84510	1.60746	2.96592	3.40439
90	25.9	1.95424	1.41330	2.76193	3.81905
100	78	2	1.89209	3.78418	4
		$\Sigma X = 9.27646$	$\Sigma Y = 8.43387$	$\Sigma XY = 15.62954$	$\Sigma X^2 = 17.27176$

From Normal equations, we have

$$8.43387 = 5A + 9.27646\ B$$

and $$15.62954 = 9.27646\ A + 17.27176\ B$$

Solving these, we get

$$A = 2.22476,\ B = -0.28997$$

$\therefore$ $\gamma = -B = 0.28997$

$$C = \text{antilog}\ (A) = \text{antilog}\ (2.22476) = 167.78765$$

Hence, the required equation of curve is

$$PV^{0.28997} = 167.78765.$$

Example 8. (*i*) *Given the following experimental values:*

x: 0 1 2 3

y: 2 4 10 15

Fit by the method of Least squares a parabola of the type $y = a + bx^2$.

(*ii*) *Find the Least squares fit of the form* $y = a_0 + a_1 x^2$ *to the following data:*

x: – 1 0 1 2

y: 2 5 3 0 (U.P.T.U. 2008)

Sol. (*i*) Error of estimate for i^{th} point (x_i, y_i) is $E_i = (y_i - a - bx_i^2)$

By method of Least squares, the values of a, b are chosen such that

$$U = \sum_{i=1}^{4} E_i^2 = \sum_{i=1}^{4} (y_i - a - bx_i^2)^2 \text{ is minimum.}$$

Normal equation are given by

$$\frac{\partial U}{\partial a} = 0 \Rightarrow \Sigma y = ma + b\Sigma x^2 \quad ...(1)$$

and

$$\frac{\partial U}{\partial b} = 0 \Rightarrow \Sigma x^2 y = a\Sigma x^2 + b\Sigma x^4 \quad ...(2)$$

x	y	x^2	x^2y	x^4
0	2	0	0	0
1	4	1	4	1
2	10	4	40	16
3	15	9	135	81
Total	$\Sigma y = 31$	$\Sigma x^2 = 14$	$\Sigma x^2 y = 179$	$\Sigma x^4 = 98$

Here $m = 4$

From (1) and (2), $31 = 4a + 14b$ and $179 = 14a + 98b$

Solving for a and b, we get $a = 2.71$, $b = 1.44$

Hence the required curve is $y = 2.71 + 1.44\, x^2$.

(*ii*) Normal equations are

$$\Sigma y = ma_0 + a_1 \Sigma x^2 \quad ...(1)$$

and

$$\Sigma x^2 y = a_0 \Sigma x^2 + a_1 \Sigma x^4 \quad ...(2)$$

The table is as follows:

x	y	x^2	x^2y	x^4
– 1	2	1	2	1
0	5	0	0	0
1	3	1	3	1
2	0	4	0	16
	$\Sigma y = 10$	$\Sigma x^2 = 6$	$\Sigma x^2 y = 5$	$\Sigma x^4 = 18$

Here, $m = 4$

From (1) and (2), $10 = 4a_0 + 6a_1$

$5 = 6a_0 + 18a_1$

$\Rightarrow$ $a_0 = 4.1667$, $a_1 = -1.1111$

Hence, the required curve is $y = 4.1667 - 1.1111\, x^2$

Example 9. *Use the method of Least squares to fit the curve:* $y = \frac{c_0}{x} + c_1\sqrt{x}$ *to the following table of values:* [G.B.T.U. (MCA) 2007, 2011]

x:	*0.1*	*0.2*	*0.4*	*0.5*	*1*	*2*
y:	*21*	*11*	*7*	*6*	*5*	*6.*

Sol. As derived in art. 3.28, normal equations to the curve $y = \frac{c_0}{x} + c_1\sqrt{x}$ are

$$\sum \frac{y}{x} = c_0 \sum \frac{1}{x^2} + c_1 \sum \frac{1}{\sqrt{x}} \quad ...(1)$$

and

$$\sum y\sqrt{x} = c_0 \sum \frac{1}{\sqrt{x}} + c_1 \sum x \quad ...(2)$$

Table is as below:

x	y	y/x	$y\sqrt{x}$	$\frac{1}{\sqrt{x}}$	$\frac{1}{x^2}$
0.1	21	210	6.64078	3.16228	100
0.2	11	55	4.91935	2.23607	25
0.4	7	17.5	4.42719	1.58114	6.25
0.5	6	12	4.24264	1.41421	4
1	5	5	5	1	1
2	6	3	8.48528	0.70711	0.25
$\Sigma x = 4.2$		$\Sigma(y/x) = 302.5$	$\Sigma y\sqrt{x} = 33.71524$	$\sum \frac{1}{\sqrt{x}} = 10.10081$	$\sum \frac{1}{x^2} = 136.5$

From equations (1) and (2), we have

$$302.5 = 136.5\, c_0 + 10.10081\, c_1$$

and $$33.71524 = 10.10081\, c_0 + 4.2\, c_1$$

Solving these, we get

$$c_0 = 1.97327 \quad \text{and} \quad c_1 = 3.28182$$

Hence the required equation of curve is

$$y = \frac{1.97327}{x} + 3.28182\sqrt{x}.$$

ASSIGNMENT

1. (*i*) Using the method of least squares, fit the non-linear curve of the form $y = ae^{bx}$ to the following data:

x	0	2	4
y	5.012	10	31.62

(*ii*) The voltage V across a capacitor at time t seconds is given by the following table. Use the principle of least squares to fit a curve of the form $V = ae^{kt}$ to the data:

t	0	2	4	6	8
V	150	63	28	12	5.6

(*iii*) For the data given below, find the equation to the best fitting exponential curve of the form $y = ae^{bx}$.

x	1	2	3	4	5	6
y	1.6	4.5	13.8	40.2	125	300

2. (*i*) State some important curve-fitting procedures. (*U.P.T.U. 2008*)

(*ii*) Derive the least square equations for fitting a curve of the type $y = ax + \frac{b}{x}$ to a set of n points (x_i, y_i) ; $i = 1, 2, \ldots, n$.

3. (*i*) Fit a curve $y = ax^b$ to the following data:

x	1	2	3	4	5	6
y	2.98	4.26	5.21	6.1	6.8	7.5

(*ii*) Fit a least square geometric curve $y = ax^b$ to the following data:

x	1	2	3	4	5
y	0.5	2	4.5	8	12.5

(*iii*) Fit a curve of the form $y = ax^b$ to the data given below:

x	1	2	3	4	5
y	7.1	27.8	62.1	110	161

4. (*i*) Fit a curve of the form $y = ab^x$ in least square sence to the data given below:

x	2	3	4	5	6
y	144	172.8	207.4	248.8	298.5

(*ii*) Fit an exponential curve of the form $y = ab^x$ to the following data:

x	1	2	3	4	5	6	7	8
y	1	1.2	1.8	2.5	3.6	4.7	6.6	9.1

5. The pressure and volume of a gas are related by the equation $pv^a = b$ where a and b are constants. Fit this equation to the following set of data:

p (kg/cm^2)	0.5	1	1.5	2	2.5	3
v (litres)	1.62	1	0.75	0.62	0.52	0.46

6. (*i*) Determine the constants of the curve $y = ax + bx^2$ for the following data:

x	0	1	2	3	4
y	2.1	2.4	2.6	2.7	3.4

(*ii*) Using method of least squares, derive the normal equations to fit the curve $y = ax^2 + bx$. Hence fit this curve to the following data: (*G.B.T.U. 2011*)

x	1	2	3	4	5	6	7	8
y	1	1.2	1.8	2.5	3.6	4.7	6.6	9.1

7. Fit a curve of the type $xy = ax + b$ to the following data:

x	1	3	5	7	9	10
y	36	29	28	26	24	15

8. Fit a relation $y = ax + \dfrac{b}{x}$ which satisfies the following data, using method of least squares:

x	1	2	3	4	5	6	7	8
y	5.4	6.2	8.2	10.3	12.6	14.8	17.2	19.5

(*G.B.T.U. 2010*)

9. Derive the least square equations for fitting a curve of the type $y = ax^2 + \dfrac{b}{x}$ to a set of n points.

Hence fit a curve of this type to the data:

x	1	2	3	4
y	– 1.51	0.99	3.88	7.66

10. Derive the least squares approximations of the type $ax^2 + bx + c$ to the function 2^x at the points $x_i = 0, 1, 2, 3, 4$.

11. A person runs the same race track for 5 consecutive days and is timed as follows:

Day (x)	1	2	3	4	5
Time (y)	15.3	15.1	15	14.5	14

Make a least square fit to the above data using a function $a + \dfrac{b}{x} + \dfrac{c}{x^2}$.

12. Use the method of least squares to fit the curve $y = c_0 x + \frac{c_1}{\sqrt{x}}$ for the following data:

x	0.2	0.3	0.5	1	2
y	16	14	11	6	3

13. Experiments with a periodic process gave the following data:

$t°$	0	50	100	150	200	250	300	350
y	0.754	1.762	2.041	1.412	0.303	– 0.484	– 0.38	0.520

Estimate the parameters a and b in the model $y = b + a \sin t$ using the least squares approximation.

14. A physicist wants to approximate the following data:

x	0	0.5	1	2
$f(x)$	0	0.57	1.46	5.05

using a function $a\, e^{bx} + c$. He believes that $b \approx 1$. Compute the values of a and c that give the best least squares approximation assuming that $b = 1$.

15. Determine the normal equations if the cubic polynomial $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ is fitted to the data (x_i, y_i); $0 \le i \le m$.

16. Estimate y at $x = 5$ by fitting a least squares curve of the form $y = \frac{b}{x\,(x - a)}$ to the following data:

x	3.6	4.8	6	7.2	8.4	9.6	10.8
y	0.83	0.31	0.17	0.10	0.07	0.05	0.04

(**Hint:** Rewrite the equation as $\frac{1}{y} = -\frac{a}{b} x + \frac{1}{b} x^2$)

Answers

1. (*i*) $y = 4.642\, e^{0.46x}$ (*ii*) $V = 146.3\, e^{-0.4118t}$ (*iii*) $y = 0.5580\, e^{1.0631\,x}$

3. (*i*) $y = 2.978\, x^{0.5143}$ (*ii*) $y = 0.5012\, x^{1.9977}$ (*iii*) $y = 7.173 x^{1.952}$

4. (*i*) $y = 99.86\, (1.2)^x$ (*ii*) $y = 0.6823\, (1.384)^x$ **5.** $pv^{1.42} = 0.99$

6. (*i*) $a = 1.97$, $b = -0.298$ (*ii*) $y = 0.107798\, x^2 + 0.217125\, x$ **7.** $xy = 16.18x + 40.78$

8. $y = 2.39188\, x + \frac{2.98195}{x}$ **9.** $y = 0.509x^2 - \frac{2.04}{x}$

10. $y = 1.143x^2 - 0.971x + 1.286$ **11.** $y = 13.0065 + \frac{6.7512}{x} - \frac{4.4738}{x^2}$

12. $y = -1.1836\, x + \frac{7.5961}{\sqrt{x}}$ **13.** $a = 1.312810$, $b = 0.752575$

14. $a = 0.784976$, $b = -0.733298$

15. $\Sigma y = m\, a_0 + a_1 \Sigma x + a_2 \Sigma x^2 + a_3 \Sigma x^3$

$\Sigma xy = a_0 \Sigma x + a_1 \Sigma x^2 + a_2 \Sigma x^3 + a_3 \Sigma x^4$

$\Sigma x^2 y = a_0 \Sigma x^2 + a_1 \Sigma x^3 + a_2 \Sigma x^4 + a_3 \Sigma x^5$

$\Sigma x^3 y = a_0 \Sigma x^3 + a_1 \Sigma x^4 + a_2 \Sigma x^5 + a_3 \Sigma x^6$.

16. $y = \dfrac{3.774}{x\,(x-2)}$; $y\,(5) = 0.2516$

3.31 CORRELATION

In a bivariate distribution, if the change in one variable affects a change in the other variable, the variables are said to be *correlated.*

If the two variables deviate in the same direction *i.e.,* if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be *direct or positive.*

e.g., the correlation between income and expenditure is positive.

If the two variables deviate in opposite direction *i.e.,* if the increase (or decrease) in one results in a corresponding decrease (or increase) in the other, correlation is said to be *inverse or negative.*

e.g., the correlation between volume and the pressure of a perfect gas or the correlation between price and demand is negative.

Correlation is said to be *perfect* if the deviation in one variable is followed by a corresponding **proportional deviation** in the other.

3.32 REASONS RESPONSIBLE FOR THE EXISTENCE OF CORRELATION

1. Due to mere chance. The correlation between variables may be due to mere chance. Consider the data regarding six students selected at random from a college.

Students:	A	B	C	D	E	F
% of marks obtained in previous exam.:	43%	47%	60%	80%	55%	40%
Height (in inches):	60	62	65	70	64	59

Here the variables are moving in the same direction and a high degree of correlation is expected between the variables. We cannot expect this degree of correlation to hold good for any other sample drawn from the concerned population. In this case, the correlation has occurred just due to chance.

2. Due to the effect of some common cause. The correlation between variables may be due to the effect of some common cause. For example, positive correlation between the number of girls seeking admission in colleges A and B of a city may be due to the effect of increasing interest of girls towards higher education.

3. Due to the presence of cause-effect relationship between variables. For example, a high degree correlation between 'temperature' and 'sale of coffee' is due to the fact that people like taking coffee in winter season.

4. Due to the presence of interdependent relationship between the variables. For example, the presence of correlation between amount spent on entertainment of family and total expenditure of family is due to fact that both variables affect each other.

3.33 SCATTER OR DOT DIAGRAMS

It is the simplest method of the diagrammatic representation of bivariate data. Let (x_i, y_i), $i = 1, 2, 3, \ldots\ldots, n$ be a bivariate distribution. Let the values of the variables x and y be plotted along the x-axis and y-axis on a suitable scale. Then corresponding to every ordered pair, there corresponds a point or dot in the xy-plane. The diagram of dots so obtained is called a *dot or scatter diagram.*

If the dots are very close to each other and the number of observations is not very large, a fairly good correlation is expected. If the dots are widely scattered, a poor correlation is expected.

3.34 KARL PEARSON'S CO-EFFICIENT OF CORRELATION (OR PRODUCT MOMENT CORRELATION CO-EFFICIENT)

[*U.P.T.U. (C.O.) 2009; U.P.T.U. 2006, 2007, 2015*]

Correlation co-efficient between two variables x and y, usually denoted by $r(x, y)$ or r_{xy} is a numerical measure of linear relationship between them and is defined as

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2 \cdot \frac{1}{n}\Sigma(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}.$$

$$\therefore \quad \boxed{r_{xy} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y}}$$

Alternate form of r(x, y):

$$\boxed{r(x, y) = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}}$$

Here n is the no. of pairs of values of x and y.

Note. Correlation co-efficient is independent of change of origin and scale.

Let us define two new variables u and v as

$u = \frac{x - a}{h}, v = \frac{y - b}{k}$ where a, b, h, k are constants, then $r_{xy} = r_{uv}$.

Then, $$\boxed{r(u, v) = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}.}$$

EXAMPLES

Example 1. *Find the coefficient of correlation between the values of x and y:*

[U.P.T.U. (C.O.) 2008]

x	*1*	*3*	*5*	*7*	*8*	*10*
y	*8*	*12*	*15*	*17*	*18*	*20*

Sol. Here, $n = 6$. The table is as follows:

x	y	x^2	y^2	xy
1	8	1	64	8
3	12	9	144	36
5	15	25	225	75
7	17	49	289	119
8	18	64	324	144
10	20	100	400	200
$\Sigma x = 34$	$\Sigma y = 90$	$\Sigma x^2 = 248$	$\Sigma y^2 = 1446$	$\Sigma xy = 582$

Karl Pearson's coefficient of correlation is given by

$$r(x, y) = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{(6 \times 582) - (34 \times 90)}{\sqrt{(6 \times 248) - (34)^2}\sqrt{(6 \times 1446) - (90)^2}} = 0.9879$$

Example 2. *The following data regarding the heights (y) and weights (x) of 100 college students are given:*

$\Sigma x = 15000$, $\Sigma x^2 = 2272500$, $\Sigma y = 6800$, $\Sigma y^2 = 463025$ and $\Sigma xy = 1022250$.

Find the correlation coefficient between height and weight.

Sol. Here, $n = 100$

Correlation co-efficient $r(x, y)$ is given by

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{(100 \times 1022250) - (15000 \times 6800)}{\sqrt{(100 \times 2272500) - (15000)^2}\sqrt{(100 \times 463025) - (6800)^2}} = 0.6.$$

Example 3. *Find the co-efficient of correlation for the following table:* (U.K.T.U. 2011)

x:	*10*	*14*	*18*	*22*	*26*	*30*
y:	*18*	*12*	*24*	*6*	*30*	*36*

Sol. Let $u = \frac{x - 22}{4}$, $v = \frac{y - 24}{6}$

x	y	u	v	u^2	v^2	uv
10	18	– 3	– 1	9	1	3
14	12	– 2	– 2	4	4	4
18	24	– 1	0	1	0	0
22	6	0	– 3	0	9	0
26	30	1	1	1	1	1
30	36	2	2	4	4	4
Total		$\Sigma u = -3$	$\Sigma v = -3$	$\Sigma u^2 = 19$	$\Sigma v^2 = 19$	$\Sigma uv = 12$

Here, $n = 6$, $\bar{u} = \frac{1}{n}\Sigma u = \frac{1}{6}(-3) = -\frac{1}{2}$; $\bar{v} = \frac{1}{n}\Sigma v = \frac{1}{6}(-3) = -\frac{1}{2}$

$$r_{uv} = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{(6 \times 12) - (-3)(-3)}{\sqrt{(6 \times 19) - (-3)^2}\sqrt{(6 \times 19) - (-3)^2}} = \frac{63}{\sqrt{105}\sqrt{105}} = 0.6$$

Hence, $r_{xy} = r_{uv} = 0.6$.

Example 4. *Ten students got the following percentage of marks in Principles of Economics and Statistics:*

Roll Nos.:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Marks in Economics:	*78*	*36*	*98*	*25*	*75*	*82*	*90*	*62*	*65*	*39*
Marks in Statistics:	*84*	*51*	*91*	*60*	*68*	*62*	*86*	*58*	*53*	*47*

Calculate the co-efficient of correlation.

Sol. Let the marks in the two subjects be denoted by x and y respectively.

x	y	$u = x - 65$	$v = y - 66$	u^2	v^2	uv
78	84	13	18	169	324	234
36	51	– 29	– 15	841	225	435
98	91	33	25	1089	625	825
25	60	– 40	– 6	1600	36	240
75	68	10	2	100	4	20
82	62	17	– 4	289	16	– 68
90	86	25	20	625	400	500
62	58	– 3	– 8	9	64	24
65	53	0	– 13	0	169	0
39	47	– 26	– 19	676	361	494
Total		$\Sigma u = 0$	$\Sigma v = 0$	$\Sigma u^2 = 5398$	$\Sigma v^2 = 2224$	$\Sigma uv = 2734$

Here, $n = 10$, $\bar{u} = \frac{1}{n}\Sigma u_i = 0$, $\bar{v} = \frac{1}{n}\Sigma v_i = 0$

$$r_{uv} = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$= \frac{(10 \times 2734) - (0 \times 0)}{\sqrt{(10 \times 5398) - (0)^2}\sqrt{(10 \times 2224) - (0)^2}} = 0.789$$

Hence, $r_{xy} = r_{uv} = 0.789$.

Example 5. *A computer while calculating correlation co-efficient between two variables X and Y from 25 pairs of observations obtained the following results :*

$n = 25$, $\Sigma X = 125$, $\Sigma X^2 = 650$,

$\Sigma Y = 100$, $\Sigma Y^2 = 460$, $\Sigma XY = 508$.

It was, however, later discovered at the time of checking that he had copied down two pairs as

X	*Y*
6	*14*
8	*6*

while the correct values were

X	*Y*
8	*12*
6	*8*

Obtain the correct value of correlation co-efficient.

Sol. Corrected $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$
Corrected $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$
Corrected $\Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$
Corrected $\Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$
Corrected $\Sigma XY = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$

(Subtract the incorrect values and add the corresponding correct values)

$$\overline{X} = \frac{1}{n}\Sigma X = \frac{1}{25} \times 125 = 5 \; ; \quad \overline{Y} = \frac{1}{n}\Sigma Y = \frac{1}{25} \times 100 = 4$$

Corrected $$r_{xy} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{(25 \times 520) - (125 \times 100)}{\sqrt{(25 \times 650) - (125)^2}\sqrt{(25 \times 436) - (100)^2}} = 0.67.$$

Example 6. *If $z = ax + by$ and r is the correlation co-efficient between x and y, show that*

$$\sigma_z^2 = a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\,\sigma_x\,\sigma_y.$$

Sol. $z = ax + by$

$\Rightarrow$ $\bar{z} = a\bar{x} + b\bar{y}$, $z_i = ax_i + by_i$

$$z_i - \bar{z} = a(x_i - \bar{x}) + b(y_i - \bar{y})$$

Now, $$\sigma_z^2 = \frac{1}{n}\Sigma(z_i - \bar{z})^2 = \frac{1}{n}\Sigma[a(x_i - \bar{x}) + b(y_i - \bar{y})]^2$$

$$= \frac{1}{n}\Sigma[a^2(x_i - \bar{x})^2 + b^2(y_i - \bar{y})^2 + 2ab(x_i - \bar{x})(y_i - \bar{y})]$$

$$= a^2 \cdot \frac{1}{n}\Sigma(x_i - \bar{x})^2 + b^2 \cdot \frac{1}{n}\Sigma(y_i - \bar{y})^2 + 2ab \cdot \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

$$= a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\,\sigma_x\sigma_y \qquad \because \; r = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x\sigma_y}$$

Example 7. *Establish the formula:* $\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r_{xy}\sigma_x\sigma_y$

where r_{xy} is the correlation coefficient between x and y. Using the above formula, calculate the correlation coefficient from the following data relating to the marks of 10 candidates in aptitude test (x) and Achievement rating (y).

	Marks									
Aptitude (x):	*22*	*53*	*46*	*67*	*43*	*35*	*88*	*11*	*95*	*13*
Achievement (y):	*18*	*39*	*31*	*42*	*55*	*64*	*82*	*10*	*96*	*14*

Sol. Let $z = x - y$

$\therefore$ $\bar{z} = \bar{x} - \bar{y}$

$\therefore$ $z - \bar{z} = (x - \bar{x}) - (y - \bar{y})$

or, $$(z - \bar{z})^2 = (x - \bar{x})^2 + (y - \bar{y})^2 - 2(x - \bar{x})(y - \bar{y})$$

Summing up for n terms,

$$\Sigma(z-\bar{z})^2 = \Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2 - 2\Sigma(x-\bar{x})(y-\bar{y})$$

or,
$$\frac{\Sigma(z-\bar{z})^2}{n} = \frac{\Sigma(x-\bar{x})^2}{n} + \frac{\Sigma(y-\bar{y})^2}{n} - \frac{2\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

$$\Rightarrow \quad \sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y, \qquad \text{where } r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}$$

$$\Rightarrow \quad \boxed{\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y} \qquad \ldots(1)$$

Now, $n = 10$

$$\bar{x} = \frac{22+53+46+67+43+35+88+11+95+13}{10} = \frac{473}{10} = 47.3$$

$$\bar{y} = \frac{18+39+31+42+55+64+82+10+96+14}{10} = \frac{451}{10} = 45.1$$

Now we form the table as

x	y	$z = x - y$	$x - \bar{x}$	$y - \bar{y}$	$z - \bar{z}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(z-\bar{z})^2$
22	18	4	– 25.3	– 27.1	1.8	640.09	734.41	3.24
53	39	14	5.7	– 6.1	11.8	32.49	37.21	139.24
46	31	15	– 1.3	– 14.1	12.8	1.69	198.81	163.84
67	42	25	19.7	– 3.1	22.8	38.09	9.61	519.84
43	55	– 12	– 4.3	9.9	– 14.2	18.49	98.01	201.64
35	64	– 29	– 12.3	18.9	– 31.2	151.29	357.21	973.44
88	82	6	40.7	36.9	3.8	1656.49	1361.61	14.44
11	10	1	– 36.3	– 35.1	– 1.2	1317.69	1232.01	1.44
95	96	– 1	47.7	50.9	– 3.2	2275.29	2590.81	10.24
13	14	– 1	– 34.3	– 31.1	– 3.2	1176.49	967.21	10.24
						$\Sigma(x-\bar{x})^2$ = 7658.1	$\Sigma(y-\bar{y})^2$ = 7586.9	$\Sigma(z-\bar{z})^2$ = 2037.6

where, $\bar{z} = \frac{\Sigma z}{n} = \frac{22}{10} = 2.2$

Now, $\sigma_x^2 = \frac{\Sigma(x-\bar{x})^2}{n} = \frac{7658.1}{10} = 765.81, \quad \sigma_y^2 = \frac{\Sigma(y-\bar{y})^2}{n} = \frac{7586.9}{10} = 758.69$

$$\sigma_{x-y}^2 = \sigma_z^2 = \frac{\Sigma(z-\bar{z})^2}{n} = \frac{2037.6}{10} = 203.76.$$

Substituting the values in the formula (1),

$$\sigma_{x-y}^2 = \sigma_x^2 + \sigma_y^2 - 2r\,\sigma_x\sigma_y$$

$$\Rightarrow \quad 203.76 = 765.81 + 758.69 - 2r\,(27.67)(27.54)$$

$$r = \frac{1524.5 - 203.76}{1524.06} = 0.866.$$

Example 8. *(i) Calculate coefficient of correlation from the following results:*

$n = 10,\ \Sigma X = 100,\ \Sigma Y = 150,\ \Sigma(X-10)^2 = 180,\ \Sigma(Y-15)^2 = 215,\ \Sigma(X-10)(Y-15) = 60.$

(ii) Calculate Karl Pearson's coefficient of correlation between X and Y for the following information:

$n = 12,\ \Sigma X = 120,\ \Sigma Y = 130,\ \Sigma(X-8)^2 = 150,\ \Sigma(Y-10)^2 = 200$ *and* $\Sigma(X-8)(Y-10) = 50$

Sol. (*i*) Mean of first series, $\overline{X} = \dfrac{\Sigma X}{n} = \dfrac{100}{10} = 10$

Mean of second series, $\overline{Y} = \dfrac{\Sigma Y}{n} = \dfrac{150}{10} = 15$

Now, $$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2 \,.\, \Sigma(Y-\overline{Y})^2}}$$

$$= \frac{\Sigma(X-10)(Y-15)}{\sqrt{\Sigma(X-10)^2 \,.\, \Sigma(Y-15)^2}} = \frac{60}{\sqrt{180 \times 215}} = 0.305$$

(*ii*) $\Sigma x = \Sigma(X-8) = \Sigma X - \Sigma 8 = 120 - (8 \times 12) = 24$

$\Sigma y = \Sigma(Y-10) = \Sigma Y - \Sigma 10 = 130 - (10 \times 12) = 10$

$\Sigma xy = \Sigma(X-8)(Y-10) = 50$ (given)

$\Sigma x^2 = \Sigma(X-8)^2 = 150$

$\Sigma y^2 = \Sigma(Y-10)^2 = 200$

Now, $$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\ \sqrt{n\Sigma y^2 - (\Sigma y)^2}} = \frac{(12 \times 50) - (24 \times 10)}{\sqrt{(12 \times 150) - (24)^2}\ \sqrt{(12 \times 200) - (10)^2}}$$

$$= \frac{360}{\sqrt{1224}\ \sqrt{2300}} = 0.2146.$$

3.35 CALCULATION OF CO-EFFICIENT OF CORRELATION FOR A BIVARIATE FREQUENCY DISTRIBUTION

If the bivariate data on x and y is presented on a two way correlation table and f is the frequency of a particular rectangle in the correlation table, then

$$r_{xy} = \frac{\Sigma fxy - \dfrac{1}{n}\Sigma fx\ \Sigma fy}{\sqrt{\left[\Sigma fx^2 - \dfrac{1}{n}(\Sigma fx)^2\right]\left[\Sigma fy^2 - \dfrac{1}{n}(\Sigma fy)^2\right]}}$$

Since change of origin and scale do not affect the co-efficient of correlation.

$\therefore$ $r_{xy} = r_{uv}$ where the new variables u, v are properly chosen.

Example 9. *The following table gives according to age the frequency of marks obtained by 100 students in an intelligence test:*

Age (in years) / *Marks*	*18*	*19*	*20*	*21*	*Total*
10–20	*4*	*2*	*2*		*8*
20–30	*5*	*4*	*6*	*4*	*19*
30–40	*6*	*8*	*10*	*11*	*35*
40–50	*4*	*4*	*6*	*8*	*22*
50–60		*2*	*4*	*4*	*10*
60–70		*2*	*3*	*1*	*6*
Total	*19*	*22*	*31*	*28*	*100*

Calculate the co-efficient of correlation between age and intelligence.

Sol. Let age and intelligence be denoted by x and y respectively.

Mid value	*x* / *y*	*18*	*19*	*20*	*21*	*f*	*u*	*fu*	fu^2	*fuv*
15	10–20	4	2	2		8	– 3	– 24	72	30
25	20–30	5	4	6	4	19	– 2	– 38	76	20
35	30–40	6	8	10	11	35	– 1	– 35	35	9
45	40–50	4	4	6	8	22	0	0	0	0
55	50–60		2	4	4	10	1	10	10	2
65	60–70		2	3	1	6	2	12	24	– 2
	f	19	22	31	28	100	Total	**– 75**	**217**	**59**
	v	– 2	– 1	0	1	Total				
	fv	– 38	– 22	0	28	**– 32**				
	fv^2	76	22	0	28	**126**				
	fuv	56	16	0	– 13	**59**				

Let us define two new variables u and v as $u = \frac{y - 45}{10}$, $v = x - 20$

$$r_{xy} = r_{uv} = \frac{\Sigma fuv - \frac{1}{n}\Sigma fu\,\Sigma fv}{\sqrt{\left[\Sigma fu^2 - \frac{1}{n}(\Sigma fu)^2\right]\left[\Sigma fv^2 - \frac{1}{n}(\Sigma fv)^2\right]}}$$

$$= \frac{59 - \frac{1}{100}(-75)(-32)}{\sqrt{\left[217 - \frac{1}{100}(-75)^2\right]\left[126 - \frac{1}{100}(-32)^2\right]}} = \frac{59 - 24}{\sqrt{\frac{643}{4} \times \frac{2894}{25}}} = 0.25.$$

3.36 RANK CORRELATION

Sometimes we have to deal with problems in which data cannot be quantitatively measured but qualitative assessment is possible.

Let a group of n individuals be arranged in order of merit or proficiency in possession of two characteristics A and B. The ranks in the two characteristics are, in general, different. For example, if A stands for intelligence and B for beauty, it is not necessary that the most intelligent individual may be the most beautiful and *vica versa.* Thus an individual who is ranked at the top for the characteristic A *may be* ranked at the bottom for the characteristic B. Let (x_i, y_i), $i = 1, 2, \ldots\ldots n$ be the ranks of the n individuals in the group for the characteristics A and B respectively. Pearsonian co-efficient of correlation between the ranks x_i's and y_i's is called the *rank correlation co-efficient* between the characteristics A and B for that group of individuals.

Thus rank correlation co-efficient

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \, \Sigma(y_i - \bar{y})^2}} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \qquad \ldots(1)$$

Now x_i's and y_i's are merely the permutations of n numbers from 1 to n. *Assuming that no two individuals are bracketed or tied in either classification i.e., $(x_i, y_i) \neq (x_j, y_j)$ for $i \neq j$, both x and y take all integral values from 1 to n.*

$$\therefore \qquad \bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + 3 + \ldots\ldots + n) = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

$$\Sigma x_i = 1 + 2 + 3 + \ldots\ldots + n = \frac{n(n+1)}{2} = \Sigma y_i$$

$$\Sigma x_i^2 = 1^2 + 2^2 + \ldots\ldots + n^2 = \frac{n(n+1)(2n+1)}{6} = \Sigma y_i^2$$

If D_i denotes the difference in ranks of the i^{th} individual, then

$$D_i = x_i - y_i = (x_i - \bar{x}) - (y_i - \bar{y}) \qquad [\because \ \bar{x} = \bar{y}]$$

$$\frac{1}{n}\Sigma D_i^2 = \frac{1}{n}\Sigma[(x_i - \bar{x}) - (y_i - \bar{y})]^2$$

$$= \frac{1}{n}\Sigma(x_i - \bar{x})^2 + \frac{1}{n}\Sigma(y_i - \bar{y})^2 - 2 \cdot \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

$$= \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y \qquad \ldots(2) \quad | \text{ using (1)}$$

But $$\sigma_x^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2 = \sigma_y^2$$

$$\therefore \quad \text{From (2), } \frac{1}{n}\Sigma D_i^2 = 2\sigma_x^2 - 2r\sigma_x^2 = 2(1-r)\,\sigma_x^2 = 2(1-r)\left[\frac{1}{n}\Sigma x_i^2 - \bar{x}^2\right]$$

$$= 2(1-r)\left[\frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}\right]$$

$$= (1-r)(n+1)\left[\frac{4n+2-3n-3}{6}\right] = \frac{(1-r)(n^2-1)}{6} \quad \text{or} \quad 1 - r = \frac{6\Sigma D_i^2}{n(n^2-1)}$$

Hence, $$\boxed{r = 1 - \left[\frac{6\Sigma D_i^2}{n(n^2-1)}\right]}$$

Note. This is called *Spearman's Formula for* Rank Correlation.

$\Sigma d_i = \Sigma(x_i - y_i) = \Sigma x_i - \Sigma y_i = 0$ always.

This serves as a check on calculations.

EXAMPLES

Example 1. *Compute the rank correlation coefficient for the following data:*

Person:	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
Rank in Maths:	*9*	*10*	*6*	*5*	*7*	*2*	*4*	*8*	*1*	*3*
Rank in Physics:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*

Sol. Here the ranks are given and $n = 10$.

Person	R_1	R_2	$D = R_1 - R_2$	D^2
A	9	1	8	64
B	10	2	8	64
C	6	3	3	9
D	5	4	1	1
E	7	5	2	4
F	2	6	– 4	16
G	4	7	– 3	9
H	8	8	0	0
I	1	9	– 8	64
J	3	10	– 7	49
				$\Sigma D^2 = 280$

$$\therefore \quad r = 1 - \left\{\frac{6\Sigma D^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6 \times 280}{10(100-1)}\right\} = 1 - 1.697 = -0.697.$$

Example 2. *The marks secured by recruits in the selection test (X) and in the proficiency test (Y) are given below:*

Serial No.:	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
X:	*10*	*15*	*12*	*17*	*13*	*16*	*24*	*14*	*22*
Y:	*30*	*42*	*45*	*46*	*33*	*34*	*40*	*35*	*39*

Calculate the rank correlation co-efficient.

Sol. Here the marks are given. Therefore, first of all, write down ranks. In each series, the item with the largest size is ranked 1, next largest 2 and so on. Here $n = 9$.

X	10	15	12	17	13	16	24	14	22	Total
Y	30	42	45	46	33	34	40	35	39	
Ranks in X (x)	9	5	8	3	7	4	1	6	2	
Ranks in Y (y)	9	3	2	1	8	7	4	6	5	
$D = x - y$	0	2	6	2	– 1	– 3	– 3	0	– 3	0
D^2	0	4	36	4	1	9	9	0	9	72

$$\therefore \qquad r = 1 - \left\{\frac{6\Sigma D^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6 \times 72}{9 \times 80}\right\} = 1 - .6 = 0.4$$

Example 3. *Rank correlation co-efficient of marks obtained by 10 students in Mathematics and English was found to be 0.5. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct rank correlation co-efficient.*

Sol. Incorrect $r_k = 0.5$, $n = 10$

$$\therefore \quad \text{Incorrect } \Sigma D^2 = \frac{n(n^2-1)(1-r_k)}{6} = \frac{10 \times 99 \times 0.5}{6} = 82.5$$

Now, correct ΣD^2 = Incorrect $\Sigma D^2 - (3)^2 + (7)^2 = 82.5 - 9 + 49 = 122.5$

$$\therefore \qquad \text{Correct } r_k = 1 - \left\{\frac{6\Sigma D^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6 \times 122.5}{10(100-1)}\right\} = 0.2575.$$

Example 4. *Ten competitors in a beauty contest were ranked by three judges in the following orders:*

First Judge:	1	6	5	10	3	2	4	9	7	8
Second Judge:	3	5	8	4	7	10	2	1	6	9
Third Judge:	6	4	9	8	1	2	3	10	5	7

Use the method of rank correlation to determine which pair of judges has the nearest approach to common taste in beauty?

Sol. Let R_1, R_2, R_3 be the ranks given by three judges.

Calculation of rank correlation coefficient

Competitor	R_1	R_2	R_3	$D_{12} = R_1 - R_2$	$D_{13} = R_1 - R_3$	$D_{23} = R_2 - R_3$	D_{12}^2	D_{13}^2	D_{23}^2
A	1	3	6	– 2	– 5	– 3	4	25	9
B	6	5	4	1	2	1	1	4	1
C	5	8	9	– 3	– 4	– 1	9	16	1
D	10	4	8	6	2	– 4	36	4	16
E	3	7	1	– 4	2	6	16	4	36
F	2	10	2	– 8	0	8	64	0	64
G	4	2	3	2	1	– 1	4	1	1
H	9	1	10	8	– 1	– 9	64	1	81
I	7	6	5	1	2	1	1	4	1
J	8	9	7	– 1	1	2	1	1	4
$n = 10$							$\Sigma D_{12}^2 = 200$	$\Sigma D_{13}^2 = 60$	$\Sigma D_{23}^2 = 214$

Rank correlation coefficient between first and second judges,

$$r_{k_{12}} = 1 - \left\{\frac{6\Sigma D_{12}^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6 \times 200}{10\,(99)}\right\} = -\,0.212$$

Rank correlation coefficient between first and third judges,

$$r_{k_{13}} = 1 - \left\{\frac{6\Sigma D_{13}{}^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6\times 60}{10\times 99}\right\} = 0.636$$

Rank correlation coefficient between second and third judges,

$$r_{k_{23}} = 1 - \left\{\frac{6\Sigma D_{23}{}^2}{n(n^2-1)}\right\} = 1 - \left\{\frac{6\times 214}{10\times 99}\right\} = -0.297$$

Correlation between first and second judges is negative *i.e.,* their opinions regarding beauty test are opposite to each other. Similarly, opinions of second and third judges are opposite to each other, but the opinions of first and third judges are of similar type as their correlation is positive. It means that their likings and dislikings are very much common.

3.37 TIED RANKS

If any two or more individuals have same rank or the same value in the series of marks, then the above formula fails and requires an adjustment. In such cases, each individual is given an average rank. This common average rank is the average of the ranks which these individuals would have assumed if they were slightly different from each other. Thus, if two individual are ranked equal at the sixth place, they would have assumed the 6th and 7th ranks if they were ranked slightly different. Their common rank = $\frac{6+7}{2}$ = 6.5. If three individuals are ranked equal at fourth place, they would have assumed the 4th, 5th and 6th ranks if they were ranked slightly different. Their common rank = $\frac{4+5+6}{3}$ = 5.

Adjustment. Add $\frac{1}{12}\, m(m^2-1)$ to ΣD^2 where m stands for the number of times an item is repeated.

This adjustment factor is to be added for each repeated item.

Thus $$r = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}\, m(m^2-1) + \frac{1}{12}\, m(m^2-1) + \ldots\right\}}{n(n^2-1)}.$$

Example 5. *Obtain the rank correlation co-efficient for the following data:*

X:	*68*	*64*	*75*	*50*	*64*	*80*	*75*	*40*	*55*	*64*
Y:	*62*	*58*	*68*	*45*	*81*	*60*	*68*	*48*	*50*	*70.*

Sol. Here, marks are given, so write down the ranks.

X	68	64	75	50	64	80	75	40	55	64	Total
Y	62	58	68	45	81	60	68	48	50	70	
Ranks in X (x)	4	6	2.5	9	6	1	2.5	10	8	6	
Ranks in Y (y)	5	7	3.5	10	1	6	3.5	9	8	2	
$D = x - y$	– 1	– 1	– 1	– 1	5	– 5	– 1	1	0	4	0
D^2	1	1	1	1	25	25	1	1	0	16	72

In the X-series, the value 75 occurs twice. Had these values been slightly different, they would have been given the ranks 2 and 3. Therefore, the common rank given to them is $\frac{2+3}{2}$ = 2.5. The value 64 occurs thrice. Had these values been slightly different, they would have been given the ranks 5, 6, and 7. Therefore the common rank given to them is $\frac{5+6+7}{3}$ = 6. Similarly, in the Y-series, the value 68 occurs twice. Had these values been slightly different, they would have been given the ranks 3 and 4? Therefore, the common rank given to them is $\frac{3+4}{2}$ = 3.5.

Thus, m has the values 2, 3, 2.

$$\therefore \quad r = 1 - \frac{6\left\{\Sigma D^2 + \frac{1}{12}m_1(m_1^2 - 1) + \frac{1}{12}m_2(m_2^2 - 1) + \frac{1}{12}m_3(m_3^2 - 1)\right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6\left[72 + \frac{1}{12}.2(2^2 - 1) + \frac{1}{12}.3(3^2 - 1) + \frac{1}{12}.2(2^2 - 1)\right]}{10(10^2 - 1)}$$

$$= 1 - \left\{\frac{6 \times 75}{990}\right\} = \frac{6}{11} = 0.545.$$

ASSIGNMENT

1. Calculate the coefficient of correlation for the following data: (*U.P.T.U. 2006*)

Husband's age (in yrs.) x	23	27	28	28	29	30	31	33	35	36
Wife's age (in yrs.) y	18	20	22	27	21	29	27	29	28	29

2. Calculate the coefficient of correlation for the following data:

Height of father (in inches)	65	66	67	67	68	69	70	72
Height of son (in inches)	67	68	65	68	72	72	69	71

3. Define Karl Pearson's coefficient of correlation. How would you interpret the sign and magnitude of a correlation coefficient?

4. Calculate the coefficient of correlation between the marks obtained by 8 students in Mathematics and Statistics:

Students	A	B	C	D	E	F	G	H
Mathematics	25	30	32	35	37	40	42	45
Statistics	08	10	15	17	20	23	24	25

[*U.P.T.U. (C.O.) 2009*]

5. Find the correlation coefficient between x and y for the following data:

x	60	34	40	50	45	41	22	43
y	75	32	34	40	45	33	12	30

[*U.K.T.U. 2010*]

6. The marks of the same 15 students in two subjects A and B are analysed. The two numbers within the brackets denote the ranks of the same student in A and B respectively:

(1, 10) (2, 7) (3, 2) (4, 6) (5, 4) (6, 8) (7, 3) (8, 1) (9, 11) (10, 15) (11, 9) (12, 5) (13, 14) (14, 12) (15, 13)

Use Spearman's formula to find the rank correlation coefficient.

7. Ten students got the following percentage of marks in Chemistry and Physics:

Students:	1	2	3	4	5	6	7	8	9	10
Marks in Chemistry:	78	36	98	25	75	82	90	62	65	39
Marks in Physics:	84	51	91	60	68	62	86	58	63	47

Calculate the rank correlation co-efficient.

8. A firm not sure of the response to its product in ten different colour shades decides to produce them in those colour shades, if the ranking of these colour shades by two typical consumer judges is highly correlated.

The two judges rank the ten colours in the following order:

Colour :	*Red*	*Green*	*Blue*	*Yellow*	*White*	*Black*	*Pink*	*Purple*	*Orange*	*Ivory*
Ranking by I Judge :	6	4	3	1	2	7	9	8	10	5
Ranking by II Judge :	4	1	6	7	5	8	10	9	3	2

Is there any agreement between the two judges, to allow the introduction of the product by the firm in the market?

9. Two judges in a music competition rank the 12 entries as follows:

x	1	2	3	4	5	6	7	8	9	10	11	12
y	12	9	6	10	3	5	4	7	8	2	11	1

What degree of agreement is there between the judgement of the two judges?

10. Calculate the coefficient of correlation between the following ages of husband (x) and wife (y) by taking 30 and 28 as assumed mean incase of x and y respectively:

x :	24	27	28	28	29	30	32	33	35	35	40
y :	18	20	22	25	22	28	28	30	27	30	32

Answers

1. 0.82 **2.** 0.603 **4.** 0.9804 **5.** 0.9158
6. 0.51 **7.** 0.84 **8.** 0.22, no **9.** – 0.454
10. 0.8926.

3.38 REGRESSION ANALYSIS (*U.P.T.U. 2015*)

The term 'regression' was first used by Sir Francis Galton (1822–1911), a British Biometrician in connection with the height of parents and their offsprings. He found that the offspring of tall or short parents tend to regress to the average height. In other words, though tall fathers do tend to have tall sons yet the average height of tall fathers is more than the average height of their sons and the average height of short fathers is less than the average height of their sons.

The term 'regression' stands for some sort of functional relationship between two or more related variables. The only fundamental difference, if any, between problems of curve-fitting and regression is that in regression, any of the variables may be considered as independent or dependent while in curve-fitting, one variable cannot be dependent.

Regression measures the nature and extent of correlation. Regression is the estimation or prediction of unknown values of one variable from known values of another variable.

3.39 CURVE OF REGRESSION AND REGRESSION EQUATION

If two variates x and y are correlated *i.e.,* there exists an association or relationship between them, then the scatter diagram will be more or less concentrated round a curve. This curve is called the *curve of regression* and the relationship is said to be expressed by means of *curvilinear regression*.

The mathematical equation of the regression curve is called regression equation.

3.40 LINEAR REGRESSION

When the points of the scatter diagram concentrate round a straight line, the regression is called linear and this straight line is known as the line of regression.

Regression will be called non-linear if there exists a relationship other than a straight line between the variables under consideration.

3.41 LINES OF REGRESSION (*U.P.T.U. 2006, 2007*)

A line of regression is the straight line which gives the best fit in the least square sense to the given frequency.

In case of n pairs (x_i, y_i) ; $i = 1, 2, ..., n$ from a bivariate data, we have no reason or justification to assume y as dependent variable and x as independent variable. Either of the two may be estimated for the given values of the other. Thus if we wish to estimate y for given values of x, we shall have the regression equation of the form $y = a + bx$, called the regression line of y on x. If we wish to estimate x for given values of y, we shall have the regression line of the form $x = A + By$, called the regression line of x on y.

Thus it implies, in general, *we always have two lines of regression*.

If the line of regression is so chosen that the sum of squares of deviation parallel to the axis of y is minimised [*See Fig. (a)*], it is called *the line of regression of y on x* and it gives *the best estimate of y for any given value of x.*

If the line of regression is so chosen that the sum of squares of deviations parallel to the axis of x is minimised [*See Fig. (b)*], it is called *the line of regression of x on y* and it gives *the best estimate of x for any given value of y.*

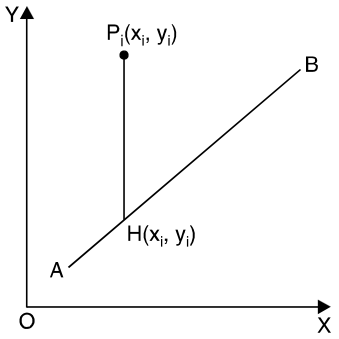


Fig. (*a*)

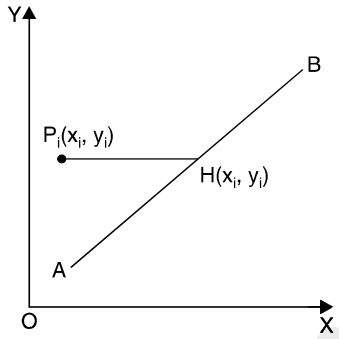


Fig. (*b*)

The independent variable is called *predictor* or regresser or explanator and the dependent variable is called the *predictant* or regressed or explained variable.

3.42 DERIVATION OF LINES OF REGRESSION

3.42.1. Line of Regression of y on x

To obtain the line of regression of y on x, we shall assume y as dependent variable and x as independent variable. Let $y = a + bx$ be the equation of regression line of y on x.

The residual for i^{th} point is $E_i = y_i - a - bx_i$.

Introduce a new quantity U such that

$$U = \sum_{i=1}^{n} E_i^{\,2} = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad ...(1)$$

According to the *principle of Least squares*, the constants a and b are chosen in such a way that the sum of the squares of residuals is minimum.

Now, the condition for U to be maximum or minimum is

$$\frac{\partial U}{\partial a} = 0 \quad \text{and} \quad \frac{\partial U}{\partial b} = 0$$

From (1), $$\frac{\partial U}{\partial a} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1)$$

$$\frac{\partial U}{\partial a} = 0 \text{ gives } \; 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$

$\Rightarrow$ $$\boxed{\Sigma y = na + b\,\Sigma x} \qquad ..(2)$$

Also, $$\frac{\partial U}{\partial b} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i)$$

$$\frac{\partial U}{\partial b} = 0 \text{ gives } \; 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0$$

$\Rightarrow$ $$\boxed{\Sigma xy = a\,\Sigma x + b\,\Sigma x^2} \qquad ...(3)$$

Equations (2) and (3) are called *normal equations.*

Solving (2) and (3) for '*a*' and '*b*', we get

$$b = \frac{\Sigma xy - \frac{1}{n}\Sigma x\,\Sigma y}{\Sigma x^2 - \frac{1}{n}(\Sigma x)^2} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} \quad ...(4)$$

and

$$a = \frac{\Sigma y}{n} - b\frac{\Sigma x}{n} = \bar{y} - b\bar{x} \quad ...(5)$$

Eqn. (5) gives $\bar{y} = a + b\bar{x}$

Hence $y = a + bx$ line passes through point $(\bar{x}, \bar{y})$.

Putting $a = \bar{y} - b\bar{x}$ in equation of line $y = a + bx$, we get

$$\boxed{y - \bar{y} = b(x - \bar{x})} \quad ...(6)$$

Equation (6) is called regression line of y on x. 'b' is called the regression coefficient of y on x and is usually denoted by b_{yx}.

Hence eqn. (6) can be rewritten as

$$\boxed{y - \bar{y} = b_{yx}(x - \bar{x})}$$

where $\bar{x}$ and $\bar{y}$ are mean values while

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2}$$

In equation (3), shifting the origin to $(\bar{x}, \bar{y})$, we get

$$\Sigma(x - \bar{x})(y - \bar{y}) = a\,\Sigma(x - \bar{x}) + b\,\Sigma(x - \bar{x})^2$$

$$\Rightarrow \quad nr\,\sigma_x\sigma_y = a(0) + bn\sigma_x^2$$

$$\Rightarrow \quad b = r\,\frac{\sigma_y}{\sigma_x}$$

$\because \quad \Sigma(x - \bar{x}) = 0,$

$\frac{1}{n}\Sigma(x - \bar{x})^2 = \sigma_x^2$

and $\frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\sigma_x\sigma_y} = r$

Hence, regression coefficient b_{yx} can also be defined as

$$\boxed{b_{yx} = r\,\frac{\sigma_y}{\sigma_x}}$$

where r is the coefficient of correlation, σ_x and σ_y are the standard deviations of x and y series respectively.

3.42.2. Line of Regression of x on y

Proceeding in the same way as 3.42.1, we can derive the regression line of x on y as

$$\boxed{x - \bar{x} = b_{xy}(y - \bar{y})}$$

where b_{xy} is the regression coefficient of x on y and is given by

$$\boxed{b_{xy} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma y^2 - (\Sigma y)^2}}$$

or

$$\boxed{b_{xy} = r\,\frac{\sigma_x}{\sigma_y}}$$

where the terms have their usual meanings.

Note. If $r = 0$, the two lines of regression become $y = \bar{y}$ and $x = \bar{x}$ which are two straight lines parallel to x and y axes respectively and passing through their means $\bar{y}$ and $\bar{x}$. They are mutually perpendicular. If $r = \pm 1$, the two lines of regression will coincide.

3.43 USE OF REGRESSION ANALYSIS (*U.P.T.U. 2008*)

(*i*) In the field of Business, this tool of statistical analysis is widely used. Businessmen are interested in predicting future production, consumption, investment, prices, profits and sales etc.

(*ii*) In the field of economic planning and sociological studies, projections of population, birth rates, death rates and other similar variables are of great use.

3.44 COMPARISON OF CORRELATION AND REGRESSION ANALYSIS

[*G.B.T.U. M.B.A. (C.O.) 2011*]

Both the correlation and regression analysis helps us in studying the relationship between two variables yet they differ in their approach and objectives.

(*i*) Correlation studies are meant for studying the covariation of the two variables. They tell us whether the variables under study move in the same direction or in reverse directions. The degree of their covariation is also reflected in the correlation co-efficient but the correlation study does not provide the nature of relationship. It does not tell us about the relative movement in the variables and we cannot predict the value of one variable corresponding to the value of other variable. This is possible through regression analysis.

(*ii*) Regression presumes one variable as a cause and the other as its effect. The independent variable is supposed to be affecting the dependent variable and as such we can estimate the values of the dependent variable by projecting the relationship between them. However, correlation between two series is not necessarily a cause-effect relationship.

(*iii*) Coefficient of correlation cannot exceed unity but one of the regression coefficients can have a value higher than unity but the product of two regression coefficients can never exceed unity.

3.45 PROPERTIES OF REGRESSION CO-EFFICIENTS

Property I. *Correlation co-efficient is the geometric mean between the regression co-efficients.*

Proof. The co-efficients of regression are $\frac{r\sigma_y}{\sigma_x}$ and $\frac{r\sigma_x}{\sigma_y}$.

G.M. between them $= \sqrt{\frac{r\sigma_y}{\sigma_x} \times \frac{r\sigma_x}{\sigma_y}} = \sqrt{r^2} = r$ = co-efficient of correlation.

Property II. *If one of the regression co-efficients is greater than unity, the other must be less than unity.*

Proof. The two regression co-efficients are $b_{yx} = \frac{r\sigma_y}{\sigma_x}$ and $b_{xy} = \frac{r\sigma_x}{\sigma_y}$.

Let $b_{yx} > 1$, then $\frac{1}{b_{yx}} < 1$ ...(1)

Since $b_{yx}.\ b_{xy} = r^2 \le 1$ ($\because\ -1 \le r \le 1$)

$\therefore\quad b_{xy} \le \frac{1}{b_{yx}} < 1.$ | Using (1)

Similarly, if $b_{xy} > 1$, then $b_{yx} < 1$.

Property III. *Arithmetic mean of regression co-efficients is greater than the correlation co-efficient.*

Proof. We have to prove that

$$\frac{b_{yx} + b_{xy}}{2} > r$$

or $$r\frac{\sigma_y}{\sigma_x} + r\frac{\sigma_x}{\sigma_y} > 2r$$

or $$\sigma_x^2 + \sigma_y^2 > 2\sigma_x\sigma_y$$

or $$(\sigma_x - \sigma_y)^2 > 0, \quad \text{which is true.}$$

Property IV. *Regression co-efficients are independent of the origin but not of scale.*

Proof. Let $u = \frac{x-a}{h}, v = \frac{y-b}{k}$, where a, b, h and k are constants

$$b_{yx} = \frac{r\sigma_y}{\sigma_x} = r \cdot \frac{k\sigma_v}{h\sigma_u} = \frac{k}{h}\left(\frac{r\sigma_v}{\sigma_u}\right) = \frac{k}{h} b_{vu}$$

Similarly, $b_{xy} = \frac{h}{k} b_{uv}$.

Thus, b_{yx} and b_{xy} are both independent of a and b but not of h and k.

Property V. *The correlation co-efficient and the two regression co-efficients have same sign.*

Proof. Regression co-efficient of y on $x = b_{yx} = r\frac{\sigma_y}{\sigma_x}$

Regression co-efficient of x on $y = b_{xy} = r\frac{\sigma_x}{\sigma_y}$

Since σ_x and σ_y are both positive; b_{yx}, b_{xy} and r have same sign.

3.46 ANGLE BETWEEN TWO LINES OF REGRESSION

If θ is the acute angle between the two regression lines in the case of two variables x and y, show that

$$\tan\theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}, \quad \text{where } r, \sigma_x, \sigma_y \text{ have their usual meanings.}$$

Explain the significance of the formula when $r = 0$ and $r = \pm 1$.

[U.P.T.U. 2007, 2015; G.B.T.U. (C.O.) 2011]

Proof. Equations to the lines of regression of y on x and x on y are

$$y - \bar{y} = \frac{r\sigma_y}{\sigma_x}(x - \bar{x}) \quad \text{and} \quad x - \bar{x} = \frac{r\sigma_x}{\sigma_y}(y - \bar{y})$$

Their slopes are $m_1 = \frac{r\sigma_y}{\sigma_x}$ and $m_2 = \frac{\sigma_y}{r\sigma_x}$.

$$\therefore \quad \tan\theta = \pm\frac{m_2 - m_1}{1 + m_2 m_1} = \pm\frac{\frac{\sigma_y}{r\sigma_x} - \frac{r\sigma_y}{\sigma_x}}{1 + \frac{\sigma_y^2}{\sigma_x^2}}$$

$$= \pm \frac{1-r^2}{r} \cdot \frac{\sigma_y}{\sigma_x} \cdot \frac{\sigma_x^{\,2}}{\sigma_x^{\,2} + \sigma_y^{\,2}} = \pm \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^{\,2} + \sigma_y^{\,2}}$$

Since $r^2 \leq 1$ and σ_x, σ_y are positive.

$\therefore$ +ve sign gives the acute angle between the lines.

Hence $$\boxed{\tan \theta = \frac{1-r^2}{r} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^{\,2} + \sigma_y^{\,2}}}$$

when $r = 0$, $\theta = \frac{\pi}{2}$ $\therefore$ The two lines of regression are perpendicular to each other.

Hence the estimated value of y is the same for all values of x and vice-versa.

When $r = \pm 1$, $\tan \theta = 0$ so that $\theta = 0$ or π

Hence the lines of regression coincide and there is perfect correlation between the two variates x and y.

EXAMPLES

Example 1. *If the regression coefficients are 0.8 and 0.2, what would be the value of coefficient of correlation?*

Sol. We know that,

$$r^2 = b_{yx} \cdot b_{xy} = 0.8 \times 0.2 = 0.16$$

Since r has the same sign as both the regression coefficients b_{yx} and b_{xy}

Hence $r = \sqrt{0.16} = 0.4$.

Example 2. *Calculate linear regression coefficients from the following:*

x	$\rightarrow$	1	2	3	4	5	6	7	8
y	$\rightarrow$	3	7	10	12	14	17	20	24

Sol. Linear regression coefficients are given by

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} \quad \text{and} \quad b_{xy} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma y^2 - (\Sigma y)^2}$$

Let us prepare the following table:

x	y	x^2	y^2	xy
1	3	1	9	3
2	7	4	49	14
3	10	9	100	30
4	12	16	144	48
5	14	25	196	70
6	17	36	289	102
7	20	49	400	140
8	24	64	576	192
$\Sigma x = 36$	$\Sigma y = 107$	$\Sigma x^2 = 204$	$\Sigma y^2 = 1763$	$\Sigma xy = 599$

Here, $n = 8$

$$\therefore \quad b_{yx} = \frac{(8 \times 599) - (36 \times 107)}{(8 \times 204) - (36)^2} = \frac{940}{336} = 2.7976$$

and $$b_{xy} = \frac{(8 \times 599) - (36 \times 107)}{(8 \times 1763) - (107)^2} = \frac{940}{2655} = 0.3540$$

Example 3. *The following table gives age (x) in years of cars and annual maintenance cost (y) in hundred rupees:*

X:	*1*	*3*	*5*	*7*	*9*
Y:	*15*	*18*	*21*	*23*	*22*

Estimate the maintenance cost for a 4 year old car after finding the regression equation.

Sol.

x	y	xy	x^2
1	15	15	1
3	18	54	9
5	21	105	25
7	23	161	49
9	22	198	81
$\Sigma x = 25$	$\Sigma y = 99$	$\Sigma xy = 533$	$\Sigma x^2 = 165$

Here, $n = 5$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{25}{5} = 5, \ \bar{y} = \frac{\Sigma y}{n} = \frac{99}{5} = 19.8$$

$$\therefore \quad b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{(5 \times 533) - (25 \times 99)}{(5 \times 165) - (25)^2} = 0.95$$

Regression line of y on x is given by

$$y - \bar{y} = b_{yx}\,(x - \bar{x})$$

$$\Rightarrow \quad y - 19.8 = 0.95\,(x - 5)$$

$$\Rightarrow \quad y = 0.95x + 15.05$$

When $x = 4$ years, $y = (0.95 \times 4) + 15.05 = 18.85$ hundred rupees = ₹ 1885.

Example 4. *In a partially destroyed laboratory record of an analysis of a correlation data, the following results only are legible:*

Variance of x = 9

Regression equations: 8x – 10y + 66 = 0, 40x – 18y = 214.

What were (a) the mean values of x and y (b) the standard deviation of y and the co-efficient of correlation between x and y? [U.P.T.U. 2008, 2009; U.K.T.U. 2010]

Sol. (*a*) **Since both the lines of regression pass through the point** $(\bar{x}, \bar{y})$ therefore, we have

$$8\bar{x} - 10\bar{y} + 66 = 0 \quad \ldots(1)$$

$$40\bar{x} - 18\bar{y} - 214 = 0 \quad \ldots(2)$$

Multiplying (1) by 5, $40\bar{x} - 50\bar{y} + 330 = 0$ ...(3)

Subtracting (3) from (2), $32\bar{y} - 544 = 0$ $\therefore$ $\bar{y} = 17$

$\therefore$ From (1), $8\bar{x} - 170 + 66 = 0$ or $8\bar{x} = 104$ $\therefore$ $\bar{x} = 13$

Hence, $\bar{x} = 13, \quad \bar{y} = 17$

(*b*) Variance of $x = \sigma_x^2 = 9$ (given)

$\therefore$ $\sigma_x = 3$

The equations of lines of regression can be written as

$$y = 0.8x + 6.6 \quad \text{and} \quad x = 0.45y + 5.35$$

$\therefore$ The regression co-efficient of y on x is $\dfrac{r\sigma_y}{\sigma_x} = 0.8$ ...(4)

The regression co-efficient of x on y is $\dfrac{r\sigma_x}{\sigma_y} = 0.45$ ...(5)

Multiplying (4) and (5), $r^2 = 0.8 \times 0.45 = 0.36$ $\therefore$ $r = 0.6$

(+ve sign with square root is taken because regression co-efficients are +ve).

From (4), $\sigma_y = \dfrac{0.8\sigma_x}{r} = \dfrac{0.8 \times 3}{0.6} = 4.$

Example 5. *The regression lines of y on x and x on y are respectively* $y = ax + b$, $x = cy + d$. *Show that*

$$\frac{\sigma_y}{\sigma_x} = \sqrt{\frac{a}{c}}, \ \bar{x} = \frac{bc + d}{1 - ac} \ \text{and} \ \bar{y} = \frac{ad + b}{1 - ac}.$$

[U.P.T.U. (C.O.) 2009, U.P.T.U. (MCA) 2008]

Sol. The regression line of y on x is

$$y = ax + b \quad ...(1)$$

$\therefore$ $b_{yx} = a$

The regression line of x on y is

$$x = cy + d \quad ...(2)$$

$\therefore$ $b_{xy} = c$

We know that, $b_{yx} = r\,\dfrac{\sigma_y}{\sigma_x}$ ...(3)

and $b_{xy} = r\,\dfrac{\sigma_x}{\sigma_y}$ ...(4)

Dividing eqn. (3) by (4), we get

$$\frac{b_{yx}}{b_{xy}} = \frac{\sigma_y^{\,2}}{\sigma_x^{\,2}} \quad \Rightarrow \quad \frac{a}{c} = \frac{\sigma_y^{\,2}}{\sigma_x^{\,2}} \quad \Rightarrow \quad \frac{\sigma_y}{\sigma_x} = \sqrt{\frac{a}{c}}$$

Since both the regression lines pass through the point $(\bar{x}, \bar{y})$ therefore,

$$\bar{y} = a\bar{x} + b \qquad \text{and} \qquad \bar{x} = c\bar{y} + d$$

$$\Rightarrow \quad a\bar{x} - \bar{y} = -b \qquad \ldots(5)$$

$$\bar{x} - c\bar{y} = d \qquad \ldots(6)$$

Multiplying equation (6) by a and then subtracting from (5), we get

$$(ac - 1)\,\bar{y} = -ad - b \quad \Rightarrow \quad \bar{y} = \frac{ad + b}{1 - ac}$$

Similarly, we get $\bar{x} = \dfrac{bc + d}{1 - ac}$.

Example 6. *For two random variables, x and y with the same mean, the two regression equations are* $y = ax + b$ *and* $x = \alpha y + \beta$. *Show that* $\dfrac{b}{\beta} = \dfrac{1 - a}{1 - \alpha}$. *Find also the common mean.*

(G.B.T.U. 2010)

Sol. Here, $b_{yx} = a,\ b_{xy} = \alpha$

Let the common mean be m, then regression lines are

$$y - m = a\,(x - m)$$

$$\Rightarrow \quad y = ax + m\,(1 - a) \qquad \ldots(1)$$

and $$x - m = \alpha(y - m)$$

$$\Rightarrow \quad x = \alpha y + m\,(1 - \alpha) \qquad \ldots(2)$$

Comparing (1) and (2) with the given equations.

$$b = m\,(1 - a),\ \beta = m\,(1 - \alpha)$$

$$\therefore \quad \frac{b}{\beta} = \frac{1 - a}{1 - \alpha}$$

Since regression lines pass through $(\bar{x}, \bar{y})$

$\therefore$ $\bar{x} = \alpha\bar{y} + \beta$ and $\bar{y} = a\bar{x} + b$ will hold.

$\Rightarrow$ $m = am + b,\ m = \alpha m + \beta$

$\Rightarrow$ $am + b = \alpha m + \beta$

$\Rightarrow$ $m = \dfrac{\beta - b}{a - \alpha}$.

Example 7. (*i*) *Obtain the line of regression of y on x for the data given below:*

x:	*1.53*	*1.78*	*2.60*	*2.95*	*3.42*
y:	*33.50*	*36.30*	*40.00*	*45.80*	*53.50.*

(*ii*) *The following data regarding the heights (y) and weights (x) of 100 college students are given:*

$\Sigma x = 15000,\quad \Sigma x^2 = 2272500,\quad \Sigma y = 6800,\quad \Sigma y^2 = 463025$ *and* $\Sigma xy = 1022250.$

Find the equation of regression line of height on weight.

Sol. (*i*) The line of regression of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x}) \qquad \ldots(1)$$

where b_{yx} is the coefficient of regression given by

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} \qquad ...(2)$$

Now we form the table as,

x	y	x^2	xy
1.53	33.50	2.3409	51.255
1.78	36.30	2.1684	64.614
2.60	40.00	6.76	104
2.95	45.80	8.7025	135.11
3.42	53.50	11.6964	182.97
$\Sigma x = 12.28$	$\Sigma y = 209.1$	$\Sigma x^2 = 32.6682$	$\Sigma xy = 537.949$

Here, $n = 5$

$$b_{yx} = \frac{(5 \times 537.949) - (12.28 \times 209.1)}{(5 \times 32.6682) - (12.28)^2} = 9.726$$

Also, mean $\bar{x} = \frac{\Sigma x}{n} = \frac{12.28}{5} = 2.456$ and $\bar{y} = \frac{\Sigma y}{n} = \frac{209.1}{5} = 41.82$

$\therefore$ From (1), we get

$$y - 41.82 = 9.726(x - 2.456) = 9.726x - 23.887$$

$$y = 17.932 + 9.726x$$

(*ii*) $$\bar{x} = \frac{\Sigma x}{n} = \frac{15000}{100} = 150, \quad \bar{y} = \frac{\Sigma y}{n} = \frac{6800}{100} = 68$$

Regression coefficient of y on x,

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{(100 \times 1022250) - (15000 \times 6800)}{(100 \times 2272500) - (15000)^2} = 0.1$$

Regression line of height (y) on weight (x) is given by

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

$\Rightarrow$ $$y - 68 = 0.1(x - 150)$$

$\Rightarrow$ $$y = 0.1x + 53.$$

Example 8. *For 10 observations on price (x) and supply (y), the following data were obtained (in appropriate units):*

$$\Sigma x = 130, \quad \Sigma y = 220, \quad \Sigma x^2 = 2288, \quad \Sigma y^2 = 5506 \text{ and } \Sigma xy = 3467$$

Obtain the two lines of regression and estimate the supply when the price is 16 units.

Sol. Here, $n = 10$, $\bar{x} = \frac{\Sigma x}{n} = 13$ and $\bar{y} = \frac{\Sigma y}{n} = 22$

Regression coefficient of y on x is

$$b_{yx} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma x^2 - (\Sigma x)^2} = \frac{(10 \times 3467) - (130 \times 220)}{(10 \times 2288) - (130)^2} = 1.015$$

$\therefore$ Regression line of y on x is

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

$$y - 22 = 1.015(x - 13)$$

$$\Rightarrow \quad y = 1.015x + 8.805 \qquad ...(1)$$

Regression coefficient of x on y is

$$b_{xy} = \frac{n\,\Sigma xy - \Sigma x\,\Sigma y}{n\,\Sigma y^2 - (\Sigma y)^2} = \frac{(10 \times 3467) - (130 \times 220)}{(10 \times 5506) - (220)^2} = 0.9114$$

Regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 13 = 0.9114(y - 22)$$

$$x = 0.9114y - 7.0508 \qquad ...(2)$$

Since we are to estimate supply (y) when price (x) is given therefore we are to use regression line of y on x here.

When $x = 16$ units,

$$y = 1.015(16) + 8.805 = 25.045 \text{ units.}$$

Example 9. *The following results were obtained from records of age (x) and systolic blood pressure (y) of a group of 10 men:*

	x	y
Mean	*53*	*142*
Variance	*130*	

and $\Sigma(x - \bar{x})(y - \bar{y}) = 1220$

Find the appropriate regression equation and use it to estimate the blood pressure of a man whose age is 45.

Sol. Given

Mean $\bar{x} = 53$ and $\bar{y} = 142$; Variance $\sigma_x^2 = 130$

$n = 10$ and $\Sigma(x - \bar{x})(y - \bar{y}) = 1220$

Since we are to estimate blood pressure (y) of a 45 years old man, we will find regression line of y on x.

Regression coefficient

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\,\sigma_x\,\sigma_y}\left(\frac{\sigma_y}{\sigma_x}\right)$$

$$= \frac{\Sigma(x - \bar{x})(y - \bar{y})}{n\,\sigma_x^2} = \frac{1220}{(10)(130)} = 0.93846$$

Regression line of y on x is given by

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

$$\Rightarrow \quad y - 142 = 0.93846\,(x - 53)$$

$$\Rightarrow \quad y = 0.93846\,x + 92.26162$$

When $x = 45$, $y = 134.49$

Hence the required blood pressure = 134.49.

Example 10. *The following results were obtained from marks in Applied Mechanics and Engineering Mathematics in an examination:*

	Applied Mechanics (x)	*Engineering Mathematics (y)*
Mean	*47.5*	*39.5*
Standard Deviation	*16.8*	*10.8*

r = 0.95.

Find both the regression equations. Also estimate the value of y for x = 30.

Sol. $\bar{x} = 47.5, \quad \bar{y} = 39.5$

$\sigma_x = 16.8, \quad \sigma_y = 10.8$ and $r = 0.95.$

Regression coefficients are

$$b_{yx} = r\,\frac{\sigma_y}{\sigma_x} = 0.95 \times \frac{10.8}{16.8} = 0.6107$$

and

$$b_{xy} = r\,\frac{\sigma_x}{\sigma_y} = 0.95 \times \frac{16.8}{10.8} = 1.477.$$

Regression line of y on x is

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

$$\Rightarrow \quad y - 39.5 = 0.6107\,(x - 47.5) = 0.6107x - 29.008$$

$$y = 0.6107x + 10.49 \qquad ...(1)$$

Regression line of x on y is

$$x - \bar{x} = b_{xy}\,(y - \bar{y})$$

$$\Rightarrow \quad x - 47.5 = 1.477\,(y - 39.5)$$

$$x = 1.477y - 10.8415 \qquad ...(2)$$

Putting $x = 30$ in equation (1), we get

$$y = (0.6107)(30) + 10.49 = 28.81.$$

Example 11. *The equations of two regression lines, obtained in a correlation analysis of 60 observations are:*

$$5x = 6y + 24 \text{ and } 1000y = 768x - 3608.$$

What is the correlation coefficient? Show that the ratio of coefficient of variability of x to that of y is $\frac{5}{24}$. *What is the ratio of variances of x and y?*

Sol. Regression line of x on y is

$$5x = 6y + 24$$

$$\Rightarrow \quad x = \frac{6}{5}y + \frac{24}{5}$$

$$\therefore \quad b_{xy} = \frac{6}{5} \qquad ...(1)$$

Regression line of y on x is

$$1000y = 768x - 3608$$

$$\Rightarrow \quad y = 0.768x - 3.608$$

$$\therefore \quad b_{yx} = 0.768 \qquad ...(2)$$

From (1), $r\,\frac{\sigma_x}{\sigma_y} = \frac{6}{5}$ ...(3)

From (2), $r\,\frac{\sigma_y}{\sigma_x} = 0.768$ ...(4)

Multiplying equations (3) and (4), we get

$$r^2 = 0.9216 \quad \Rightarrow \quad r = 0.96 \qquad ...(5)$$

Dividing (4) by (3), we get

$$\frac{\sigma_x^{\,2}}{\sigma_y^{\,2}} = \frac{6}{5 \times 0.768} = 1.5625.$$

Taking square root, we get

$$\frac{\sigma_x}{\sigma_y} = 1.25 = \frac{5}{4} \qquad ...(6)$$

Since the regression lines pass through the point $(\bar{x}, \bar{y})$, we have

$$5\bar{x} = 6\bar{y} + 24$$

$$1000\,\bar{y} = 768\bar{x} - 3608.$$

Solving the above equations for $\bar{x}$ and $\bar{y}$, we get $\bar{x} = 6$, $\bar{y} = 1$.

Co-efficient of variability of $x = \frac{\sigma_x}{\bar{x}}$

Co-efficient of variability of $y = \frac{\sigma_y}{\bar{y}}$.

$\therefore$ Required ratio $= \frac{\sigma_x}{\bar{x}} \times \frac{\bar{y}}{\sigma_y} = \frac{\bar{y}}{\bar{x}}\left(\frac{\sigma_x}{\sigma_y}\right) = \frac{1}{6} \times \frac{5}{4} = \frac{5}{24}$. | Using (6)

Example 12. *A panel of two judges, A and B, graded seven TV serial performances by awarding marks independently as shown in the following table:*

Performance	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Marks by A	*46*	*42*	*44*	*40*	*43*	*41*	*45*
Marks by B	*40*	*38*	*36*	*35*	*39*	*37*	*41*

The eighth TV performance which judge B could not attend, was awarded 37 marks by judge A. If the judge B had also been present, how many marks would be expected to have been awarded by him to the eighth TV performance?

Use regression analysis to answer this question.

Sol. Let the marks awarded by judge A be denoted by x and the marks awarded by judge B be denoted by y.

Here, $n = 7$; $\bar{x} = \frac{\Sigma x}{n} = \frac{46 + 42 + 44 + 40 + 43 + 41 + 45}{7} = 43$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{40 + 38 + 36 + 35 + 39 + 37 + 41}{7} = 38$$

Let us form the table as

x	y	xy	x^2
46	40	1840	2116
42	38	1596	1764
44	36	1584	1936
40	35	1400	1600
43	39	1677	1849
41	37	1517	1681
45	41	1845	2025
$\Sigma x = 301$	$\Sigma y = 266$	$\Sigma xy = 11459$	$\Sigma x^2 = 12971$

Regression coefficient,

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{(7 \times 11459) - (301 \times 266)}{(7 \times 12971) - (301)^2} = 0.75$$

Regression line of y on x is given by

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

$$y - 38 = 0.75(x - 43)$$

$\Rightarrow$ $y = 0.75x + 5.75$

when $x = 37$, $y = 0.75(37) + 5.75 = 33.5$ marks

Hence, if judge B had also been present, 33.5 marks would be expected to have been awarded to the eighth TV performance.

Example 13. *Two variables x and y have zero means, the same variance* σ^2 *and zero correlation, show that:*

$$u = x \cos \alpha + y \sin \alpha \quad \text{and} \quad v = x \sin \alpha - y \cos \alpha$$

have the same variance σ^2 *and zero correlation.* (U.P.T.U. 2007)

Sol. We are given that

$$r(x, y) = 0 \Rightarrow \text{Cov}(x, y) = 0, \quad \sigma_x^2 = \sigma_y^2 = \sigma^2$$

We have, $\sigma_u^2 = V(x \cos \alpha + y \sin \alpha)$

$= \cos^2 \alpha \, V(x) + \sin^2 \alpha \, V(y) + 2 \sin \alpha \cos \alpha \, \text{Cov}(x, y)$

$= (\cos^2 \alpha + \sin^2 \alpha)\, \sigma^2 \quad | \because \text{Cov}(x, y) = 0$

$= \sigma^2$

Similarly, $\sigma_v^2 = \sigma^2$

$\text{Cov}(u, v) = E[(u - \bar{u})(v - \bar{v})]$

$= E[(x \cos \alpha + y \sin \alpha - \bar{x} \cos \alpha - \bar{y} \sin \alpha)$

$(x \sin \alpha - y \cos \alpha - \bar{x} \sin \alpha + \bar{y} \cos \alpha)]$

$= E[(x \cos \alpha + y \sin \alpha)(x \sin \alpha - y \cos \alpha)] \quad | \because \bar{x} = 0 = \bar{y}$

$= [E(x^2) - E(y^2)] \sin \alpha \cos \alpha + E(xy)(\sin^2 \alpha - \cos^2 \alpha)$

$= 0 \quad | \because \sigma_x^2 = \sigma_y^2 = \sigma^2$ and $E(xy) = 0$

$\therefore$ $r = \dfrac{\text{Cov}(u, v)}{\sigma_u \, \sigma_v} = 0.$

ASSIGNMENT

1. (*i*) Discuss regression and its importance. Given the following data:

x:	1	5	3	2	1	1	7	3
y:	6	1	0	0	1	2	1	5

Find a regression line of x on y. (*U.P.T.U. 2008*)

(*ii*) In a study between the amount of rainfall and the quantity of air pollution removed the following data were collected:

Daily rainfall: (in .01 cm)	4.3	4.5	5.9	5.6	6.1	5.2	3.8	2.1
Pollution removed: (mg/m³)	12.6	12.1	11.6	11.8	11.4	11.8	13.2	14.1

Find the regression line of y on x.

(*iii*) Find the two lines of regression and coefficient of correlation for the data given below:

$n = 18, \Sigma x = 12, \Sigma y = 18, \Sigma x^2 = 60, \Sigma y^2 = 96, \Sigma xy = 48$ [*U.P.T.U. (MCA) 2009*]

(*iv*) From the data given, find the equation of lines of regression of x on y and y on x. Also calculate the correlation co-efficient.

x:	2	4	6	8	10
y:	5	7	9	8	11

(*U.P.T.U. 2011*)

2. (*i*) Can Y = 5 + 2.8 X and X = 3 – 0.5 Y be the estimated regression equations of Y on X and X on Y respectively? Explain your answer with suitable theoretical arguments.

(*ii*) Find the co-efficient of correlation when the two regression equations are

$$X = -0.2\,Y + 4.2, \quad Y = -0.8\,X + 8.4$$

3. (*i*) If F is the pull required to lift a load W by means of a pulley block, fit a linear law of the form $F = mW + c$ connecting F and W, using the data

W:	50	70	100	120
F:	12	15	21	25

where F and W are in kg wt. Compute F when W = 150 kg wt. (*U.P.T.U. 2007*)

(*ii*) A simply supported beam carries a concentrated load P (kg) at its mid-point. The following table gives maximum deflection y (cm) corresponding to various values of P:

P:	100	120	140	160	180	200
y:	0.45	0.55	0.60	0.70	0.80	0.85

Find a law of the form $y = a + bP$. Also find the value of maximum deflection when P = 150 kg.

4. (*i*) Find both the lines of regression of following data:

x:	5.60	5.65	5.70	5.81	5.85
y:	5.80	5.70	5.80	5.79	6.01

(*ii*) Obtain regression line of x on y for the given data:

x:	1	2	3	4	5	6
y:	5.0	8.1	10.6	13.1	16.2	20.0

[*U.P.T.U. (MCA) 2007*]

(*iii*) Given that:

x:	1	3	5	7	8	10
y:	8	12	15	17	18	20

Find the equations of both lines of regression. [*U.P.T.U. (C.O.) 2008*]

5. (*i*) The two regression equations of the variables x and y are $x = 19.13 - 0.87y$ and $y = 11.64 - 0.50x$. Find (*a*) mean of x's (*b*) mean of y's and (*c*) correlation coefficient between x and y.

(*ii*) Two random variables have the regression lines with equations $3x + 2y = 26$ and $6x + y = 31$. Find the mean values and the correlation coefficient between x and y.

[*G.B.T.U. (MBA) 2011*]

(*iii*) In a partially destroyed laboratory data, only the equations giving the two lines of regression of y on x and x on y are available and are respectively

$7x - 16y + 9 = 0, \quad 5y - 4x - 3 = 0.$

Calculate the coefficient of correlation, $\bar{x}$ and $\bar{y}$.

(*iv*) The regression equations calculated from a given set of observations for two random variables are

$x = -0.4y + 6.4$ and $y = -0.6x + 4.6$

Calculate (*i*) $\bar{x}$ (*ii*) $\bar{y}$ (*iii*) r.

(*v*) Two lines of regression are given by

$x + 2y - 5 = 0$ and $2x + 3y - 8 = 0$ and $\sigma_x^2 = 12$,

Calculate:

(*a*) the mean values of x and y (*b*) variance of y

(*c*) the coefficient of correlation between x and y. [*U.P.T.U. (MCA) 2008*, *G.B.T.U. (C.O.) 2011*]

6. An analyst for a company was studying travelling expenses (y) in ₹ and duration (x) of these trips for 102 sales trip. He has found relation between x and y linear and data as follows:

$\Sigma x = 510, \quad \Sigma y = 7140, \quad \Sigma x^2 = 4150, \quad \Sigma xy = 54900, \quad \Sigma y^2 = 740200$

Calculate (*i*) Two regression lines

(*ii*) A given trip has to take 7 days. How much money should be allowed so that they will not run short of money?

7. Assuming that we conduct an experiment with 8 fields planted with corn, four fields having no nitrogen fertiliser and four fields having 80 kgs of nitrogen fertilizer. The resulting corn yields are shown in table in bushels per acre :

Field:	1	2	3	4	5	6	7	8
Nitrogen (kgs) x:	0	0	0	0	80	80	80	80
Corn yield y: (acre)	120	360	60	180	1280	1120	1120	760

(*a*) Compute a linear regression equation of y on x.

(*b*) Predict corn yield for a field treated with 60 kgs of fertilizer.

8. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between their lines of regression is $\tan^{-1}\left(\frac{3}{5}\right)$, show that $\sigma_x = \frac{1}{2}\sigma_y$. (*U.P.T.U. 2009*)

9. Given N = 50, Mean of y = 44, Variance of x is $\frac{9}{16}$ of the variance of y.

Regression equation of x on y is $3y - 5x = -180$

Find (*i*) Mean of x (*ii*) Coeff. of correlation between x and y.

10. The means of a bivariate frequency distribution are at (3, 4) and $r = 0.4$. The line of regression of y on x is parallel to the line $y = x$. Find the two lines of regression and estimate value of x when $y = 1$.

11. The following results were obtained in the analysis of data on yield of dry bark in ounces (y) and age in years (x) of 200 cinchona plants:

	x	y
Average:	9.2	16.5
Standard deviation:	2.1	4.2

Correlation coefficient = 0.84

Construct the two lines of regression and estimate the yield of dry bark of a plant of age 8 years.

12. A panel of judges A and B graded 7 debators and independently awarded the following marks:

Debator:	1	2	3	4	5	6	7
Marks by A:	40	34	28	30	44	38	31
Marks by B:	32	39	26	30	38	34	28

An eighth debator was awarded 36 marks by judge A while judge B was not present. If judge B were also present, how many marks would you expect him to award to the eighth debator assuming that the same degree of relationship exists in their judgement?

Answers

1. (*i*) $72x = -20y + 247$ (*ii*) $y = -0.6842x + 15.5324$

(*iii*) $y = 0.6923x + 0.53846$; $x = 0.4615y + 0.2051$

(*iv*) $x = 1.3y - 4.4$, $y = .65x + 4.1$; $r = .9192$

2. (*i*) No (*ii*) $r = -0.4$

3. (*i*) F = 0.18793W + 2.27595; F = 30.4654 kg wt.

(*ii*) $y = 0.04765 + 0.004071$ P; $y = 0.6583$ cm

4. (*i*)Regression line of y on x: $y = 0.74306x + 1.56821$

Regression line of x on y: $x = 0.63602y + 2.0204$

(*ii*) $x = 0.34195y - 0.660355$ (*iii*) $y = 1.3012x + 7.6265$; $x = 0.75y - 5.5833$.

5. (*i*) (*a*) 15.935 (*b*) 3.67 (*c*) – 0.659

(*ii*) $\bar{x} = 4$, $\bar{y} = 7$, $r = -0.5$ (*iii*) $r = 0.7395$, $\bar{x} = -0.1034$, $\bar{y} = 0.5172$

(*iv*) $\bar{x} = 6$, $\bar{y} = 1$, $r = -0.48989$

(*v*) (*a*) $\bar{x} = 1$, $\bar{y} = 2$ (*b*) 4 (*c*) $-\frac{\sqrt{3}}{2}$

6. (*i*) $y = 12x + 10$, $x = 0.07986y - 0.59068$ (*ii*) ₹ 94

7. (*a*) $y = 11.125x + 180$ (*b*) 847.5 acre

9. (*i*) 62.4 (*ii*) 0.8

10. $y = x + 1$; $x = 0.16y + 2.36$; $x = 2.52$

11. $y = 1.68x + 1.044$, $x = 0.42y + 2.27$; $y = 14.484$

12. 33 marks.

3.47 POLYNOMIAL FIT: NON-LINEAR REGRESSION

Let $$y = a + bx + cx^2 \quad ...(1)$$

be a second degree parabolic curve of regression of y on x to be fitted for the data (x_i, y_i), $i = 1, 2, ..., n$.

Residual at $x = x_i$ is

$$E_i = y_i - f(x_i) = y_i - a - bx_i - cx_i^2$$

Now, let $$U = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2$$

By principle of Least squares, U should be minimum for the best values of a, b and c.

For this, $$\frac{\partial U}{\partial a} = 0, \frac{\partial U}{\partial b} = 0 \text{ and } \frac{\partial U}{\partial c} = 0$$

$$\frac{\partial U}{\partial a} = 0 \quad \Rightarrow \quad 2\sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)(-1) = 0$$

$$\Rightarrow \boxed{\Sigma y = na + b\Sigma x + c\Sigma x^2} \qquad ...(1)$$

$$\frac{\partial U}{\partial b} = 0 \quad \Rightarrow \quad 2\sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)(-x_i) = 0$$

$$\Rightarrow \boxed{\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3} \qquad ...(2)$$

$$\frac{\partial U}{\partial c} = 0 \quad \Rightarrow \quad 2\sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)(-x_i^2) = 0$$

$$\Rightarrow \boxed{\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4} \qquad ...(3)$$

Equations (1), (2) and (3) are the normal equations for fitting a second degree parabolic curve of regression of y on x. Here n is the no. of pairs of values of x and y.

EXAMPLES

Example 1. (*a*) *Fit a second degree parabola to the following data:*

x	*0.0*	*1.0*	*2.0*
y	*1.0*	*6.0*	*17.0*

(*b*) *Fit a second degree curve of regression of y on x to the following data:*

x	*1.0*	*2.0*	*3.0*	*4.0*
y	*6.0*	*11.0*	*18.0*	*27*

(*c*) *Fit a second degree parabola in the following data:*

x	*0.0*	*1.0*	*2.0*	*3.0*	*4.0*
y	*1.0*	*4.0*	*10.0*	*17.0*	*30.0*

Sol. The equation of second degree parabola is given by

$$y = a + bx + cx^2 \quad ...(1)$$

Normal equations are

$$\Sigma y = ma + b\Sigma x + c\Sigma x^2 \quad ...(2)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad ...(3)$$

and

$$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \quad ...(4)$$

(a) Here, $m = 3$. Table is as follows:

x	y	x^2	x^3	x^4	xy	x^2y
0	1	0	0	0	0	0
1	6	1	1	1	6	6
2	17	4	8	16	34	68
Total 3	24	5	9	17	40	74

Substituting in eqns. (2), (3) and (4), we get

$$24 = 3a + 3b + 5c \quad ...(5)$$

$$40 = 3a + 5b + 9c \quad ...(6)$$

$$74 = 5a + 9b + 17c \quad ...(7)$$

Solving eqns. (5), (6) and (7), we get $a = 1, b = 2, c = 3$

Hence the required second degree parabola is $y = 1 + 2x + 3x^2$

(b) Here, $m = 4$. Table is as follows:

x	y	x^2	x^3	x^4	xy	x^2y
1	6	1	1	1	6	6
2	11	4	8	16	22	44
3	18	9	27	81	54	162
4	27	16	64	256	108	432
$\Sigma x = 10$	$\Sigma y = 62$	$\Sigma x^2 = 30$	$\Sigma x^3 = 100$	$\Sigma x^4 = 354$	$\Sigma xy = 190$	$\Sigma x^2y = 644$

Substituting values in eqns. (2), (3) and (4), we get

$$62 = 4a + 10b + 30c \quad ...(8)$$

$$190 = 10a + 30b + 100c \quad ...(9)$$

$$644 = 30a + 100b + 354c \quad ...(10)$$

Solving equations (8), (9) and (10), we get $a = 3, b = 2, c = 1$

Hence the required second degree parabola is $y = 3 + 2x + x^2$

(c) Here, $m = 5$. Table is as follows:

x	y	x^2	x^3	x^4	xy	x^2y
0.0	1.0	0	0	0	0	0
1.0	4.0	1	1	1	4	4
2.0	10.0	4	8	16	20	40
3.0	17.0	9	27	81	51	153
4.0	30.0	16	64	256	120	480
$\Sigma x = 10$	$\Sigma y = 62$	$\Sigma x^2 = 30$	$\Sigma x^3 = 100$	$\Sigma x^4 = 354$	$\Sigma xy = 195$	$\Sigma x^2y = 677$

Substituting values in eqns. (2), (3) and (4), we get

$$62 = 5a + 10b + 30c \quad ...(11)$$

$$195 = 10a + 30b + 100c \quad ...(12)$$

$$677 = 30a + 100b + 354c \quad ...(13)$$

Solving eqns. (11), (12) and (13), we get $a = 1.2$, $b = 1.1$ and $c = 1.5$

Hence the required second degree parabola is $y = 1.2 + 1.1x + 1.5x^2$.

Example 2. *Fit a second degree parabola to the following data taking y as dependent variable:*

x	1	2	3	4	5	6	7	8	9
y	2	6	7	8	10	11	11	10	9

Sol. Normal equations to fit a second degree parabola of the form $y = a + bx + cx^2$ are

$$\left.\begin{aligned} \Sigma y &= ma + b\Sigma x + c\Sigma x^2 \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \\ \Sigma x^2 y &= a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \end{aligned}\right\} \quad ...(1)$$

and

Here, $m = 9$

x	y	x^2	x^3	x^4	xy	x^2y
1	2	1	1	1	2	2
2	6	4	8	16	12	24
3	7	9	27	81	21	63
4	8	16	64	256	32	128
5	10	25	125	625	50	250
6	11	36	216	1296	66	396
7	11	49	343	2401	77	539
8	10	64	512	4096	80	640
9	9	81	729	6561	81	729
$\Sigma x = 45$	$\Sigma y = 74$	$\Sigma x^2 = 285$	$\Sigma x^3 = 2025$	$\Sigma x^4 = 15333$	$\Sigma xy = 421$	$\Sigma x^2y = 2771$

Putting in (1), we get

$$74 = 9a + 45b + 285c$$

$$421 = 45a + 285b + 2025c$$

$$2771 = 285a + 2025b + 15333c$$

Solving the above equations, we get $a = -1$, $b = 3.55$, $c = -0.27$

Hence the required equation of second degree parabola is $y = -1 + 3.55x - 0.27x^2$.

Example 3. *Employ the method of least squares to fit a parabola $y = a + bx + cx^2$ in the data:* (x, y): $(-1, 2)$, $(0, 0)$, $(0, 1)$, $(1, 2)$

Sol. Normal equations to the parabola $y = a + bx + cx^2$ are

$$\Sigma y = ma + b\Sigma x + c\Sigma x^2 \quad ...(1)$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma x^3 \quad ...(2)$$

and $$\Sigma x^2 y = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 \quad ...(3)$$

Here, $m = 4$. The table is as follows:

x	y	x^2	x^3	x^4	xy	x^2y
– 1	2	1	– 1	1	– 2	2
0	0	0	0	0	0	0
0	1	0	0	0	0	0
1	2	1	1	1	2	2
$\Sigma x = 0$	$\Sigma y = 5$	$\Sigma x^2 = 2$	$\Sigma x^3 = 0$	$\Sigma x^4 = 2$	$\Sigma xy = 0$	$\Sigma x^2y = 4$

Substituting these values in equations (1), (2) and (3); we get

$$5 = 4a + 2c \quad \ldots(4)$$

$$0 = 2b \quad \ldots(5)$$

and $$4 = 2a + 2c \quad \ldots(6)$$

Solving (4), (5) and (6), we get $a = 0.5$, $b = 0$ and $c = 1.5$

Hence the required second degree parabola is $y = 0.5 + 1.5x^2$

Example 4. *Fit a second degree parabola to the following data by Least Squares method:*

x	*1*	*2*	*3*	*4*	*5*
y	*1090*	*1220*	*1390*	*1625*	*1915*

[U.P.T.U. (MCA) 2009, U.P.T.U. 2007; U.K.T.U. 2010]

Sol. Here $m = 5$ (odd)

Let $u = x - 3$, $v = y - 1220$

x	y	u	v	u^2	u^2v	uv	u^3	u^4
1	1090	– 2	– 130	4	– 520	260	– 8	16
2	1220	– 1	0	1	0	0	– 1	1
3	1390	0	170	0	0	0	0	0
4	1625	1	405	1	405	405	1	1
5	1915	2	695	4	2780	1390	8	16
Total		$\Sigma u = 0$	$\Sigma v = 1140$	$\Sigma u^2 = 10$	$\Sigma u^2v = 2665$	$\Sigma uv = 2055$	$\Sigma u^3 = 0$	$\Sigma u^4 = 34$

Putting these values in normal equations, we get

$$1140 = 5a' + 10c', \quad 2055 = 10b', \quad 2655 = 10a' + 34c'$$

$$\Rightarrow \quad a' = 173, \quad b' = 205.5, \quad c' = 27.5$$

$$\therefore \quad v = 173 + 205.5u + 27.5u^2 \quad \ldots(1)$$

Put $u = x - 3$ and $v = y - 1220$

From (1), $y - 1220 = 173 + 205.5\,(x - 3) + 27.5\,(x - 3)^2$

$$\Rightarrow \quad y = 27.5x^2 + 40.5x + 1024.$$

3.48 MULTIPLE LINEAR REGRESSION

Now we proceed to discuss the case where the dependent variable is a function of two or more linear or non-linear independent variables. Consider such a linear function as

$$y = a + bx + cz \quad \ldots(1)$$

The sum of the squares of residual is

$$U = \sum_{i=1}^{n} (y_i - a - bx_i - cz_i)^2 \qquad \ldots(2)$$

Differentiating U partially w.r.t. a, b, c ; we get

$$\frac{\partial U}{\partial a} = 0 \quad \Rightarrow \quad 2\sum_{i=1}^{n} (y_i - a - bx_i - cz_i)\ (-1) = 0$$

$$\frac{\partial U}{\partial b} = 0 \quad \Rightarrow \quad 2\sum_{i=1}^{n} (y_i - a - bx_i - cz_i)\ (-x_i) = 0$$

and

$$\frac{\partial U}{\partial c} = 0 \quad \Rightarrow \quad 2\sum_{i=1}^{n} (y_i - a - bx_i - cz_i)\ (-z_i) = 0$$

which on simplification and omitting the suffix i, yields.

$$\Sigma y = ma + b\Sigma x + c\Sigma z$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2 + c\Sigma xz$$
$$\Sigma yz = a\Sigma z + b\Sigma xz + c\Sigma z^2$$

Solving the above three equations, we get values of a, b and c. Consequently, we get the linear function $y = a + bx + cz$ called **regression plane.**

EXAMPLES

Example 1. *Obtain a regression plane by using multiple linear regression to fit the data given below:*

x	1	2	3	4
z	0	1	2	3
y	12	18	24	30

[U.P.T.U. MCA (C.O.) 2008]

Sol. Let $y = a + bx + cz$ be the required regression plane where a, b, c are the constants to be determined by following equations:

$$\Sigma y = ma + b\Sigma x + c\Sigma z$$
$$\Sigma yx = a\Sigma x + b\Sigma x^2 + c\Sigma zx$$

and

$$\Sigma yz = a\Sigma z + b\Sigma zx + c\Sigma z^2$$

Here, $m = 4$

x	z	y	x^2	z^2	yx	zx	yz
1	0	12	1	0	12	0	0
2	1	18	4	1	36	2	18
3	2	24	9	4	72	6	48
4	3	30	16	9	120	12	90
$\Sigma x = 10$	$\Sigma z = 6$	$\Sigma y = 84$	$\Sigma x^2 = 30$	$\Sigma z^2 = 14$	$\Sigma yx = 240$	$\Sigma zx = 20$	$\Sigma yz = 156$

Substitution yields, $84 = 4a + 10b + 6c$

$240 = 10a + 30b + 20c$

and $156 = 6a + 20b + 14c$

Solving, we get $a = 10, b = 2, c = 4$

Hence the required regression plane is $y = 10 + 2x + 4z$

Example 2. *Find the multiple linear regression of X_1 on X_2 and X_3 from the data relating to three variables:*

X_1	4	6	7	9	13	15
X_2	15	12	8	6	4	3
X_3	30	24	20	14	10	4

Sol. Let $X_1 = a + bX_2 + cX_3$ be the required regression plane where a, b, c are the constants, determined by following normal equations

$$\Sigma X_1 = ma + b\Sigma X_2 + c\Sigma X_3$$
$$\Sigma X_1X_2 = a\Sigma X_2 + b\Sigma X_2^2 + c\Sigma X_2X_3$$
$$\Sigma X_1X_3 = a\Sigma X_3 + b\Sigma X_2X_3 + c\Sigma X_3^2$$

Here, $m = 6$

X_1	X_2	X_3	X_1X_2	X_2^2	X_2X_3	X_1X_3	X_3^2
4	15	30	60	225	450	120	900
6	12	24	72	144	288	144	576
7	8	20	56	64	160	140	400
9	6	14	54	36	84	126	196
13	4	10	52	16	40	130	100
15	3	4	45	9	12	60	16
Total 54	48	102	339	494	1034	720	2188

Substituting the values, we get

$54 = 6a + 48b + 102c$

$339 = 48a + 102b + 1034c$

$720 = 102a + 1034b + 2188c$

On solving, we get, $a = 16.413$, $b = -0.00536$, $c = -0.4335$

Hence $X_1 = 16.413 - 0.00536X_2 - 0.4335X_3$

ASSIGNMENT

1. Fit a parabola of the form $y = a + bx + cx^2$ to the data:

x	1	2	3	4
y	1.7	1.8	2.3	3.2

by the method of least squares. (*U.P.T.U. 2009*)

2. Find the best values of a_0, a_1, a_2 so that the parabola $y = a_0 + a_1x + a_2x^2$ fits the data:

x	1	1.5	2	2.5	3	3.5	4
y	1.1	1.2	1.5	2.6	2.8	3.3	4.1

[*U.P.T.U. (C.O.) 2008*]

3. (*i*) Fit a second degree parabola to the following data:

x	1	2	3	4	5
y	25	28	33	39	46

[*U.P.T.U. (C.O.) 2011*]

(*ii*) Fit a second degree parabola to the following data:

x	1	2	3	4	5	6	7	8	9	10
y	124	129	140	159	228	289	315	302	263	210

(*U.P.T.U. 2009*)

4. Fit a second degree parabola to the following data taking x as the independent variable:

(*i*)

x	0	1	2	3	4
y	1	5	10	22	38

(*ii*)

x	1	2	3	4	5	6	7	8	9
y	3	7	8	9	11	12	13	14	15

(*U.P.T.U. 2007*)

5. The profit of a certain company in Xth year of its life are given by:

x	1	2	3	4	5
y	1250	1400	1650	1950	2300

Taking $u = x - 3$ and $v = \dfrac{y - 1650}{50}$, show that the parabola of second degree of v on u is $v + 0.086 = 5.3\,u + 0.643u^2$ and deduce that the parabola of second degree of y on x is

$$y = 1144 + 72x + 32.15x^2$$

6. (*i*) The corresponding values of x and y are given below:

x	87	84	79	64	47	37
y	292	283	270	235	197	181

Fit a parabola of the form $y = ax^2 + bx + c$. Also find the value of y for $x = 80$ correct upto third place of decimal. (*U.P.T.U. 2006*)

(*ii*) Determine the constants a, b and c by the method of least squares such that $y = ax^2 + bx + c$ fits the following data:

x	2	4	6	8	10
y	4.01	11.08	30.12	81.89	222.62

7. The velocity V of a liquid is known to vary with temperature T, according to a quadratic law $V = a + bT + cT^2$. Find the best values of a, b and c for the following table:

T	1	2	3	4	5	6	7
V	2.31	2.01	1.80	1.66	1.55	1.47	1.41

[*G.B.T.U. (MCA) 2010*]

8. The following table gives the results of the measurements of train resistances, V is the velocity in miles per hour, R is the resistance in pounds per ton:

V	20	40	60	80	100	120
R	5.5	9.1	14.9	22.8	33.3	46

If R is related to V by the relation $R = a + bV + cV^2$; find a, b and c by using the method of least squares.

9. Find the multiple linear regression of X_1 on X_2 and X_3 from the data relating to three variables:

X_1	7	12	17	20
X_2	4	7	9	12
X_3	1	2	5	8

(*U.P.T.U. 2009*)

10. Fit a second degree parabola to the following data:

x	1	2	3	4	5	6	7	8	9
y	2	6	7	8	10	11	8	13	5

(*U.P.T.U. 2015*)

Answers

1. $y = 2 - 0.5x + 0.2x^2$ **2.** $y = 0.45714 + 0.39286x + 0.12857x^2$

3. (*i*) $y = 22.8 + 1.44x + 0.64x^2$ (*ii*) $y = 18.866 + 66.1576x - 4.3333x^2$

4. (*i*) $y = 1.43 + 0.24x + 2.21x^2$ (*ii*) $y = 1.5238 + 2.38398x - 0.10173x^2$

6. (*i*) $y = 0.010626822x^2 + 0.908257322x + 132.2040143$; 272.876

(*ii*) $a = 5.358035714$, $b = -38.89492857$, $c = 67.56$

7. $V = 2.5928 - 0.3258T + 0.02274T^2$ **8.** $R = 4.35 + 0.00241V + 0.0028705V^2$

9. $X_1 = 0.6441 + 1.661X_2 + 0.0169X_3$ **10.** $y = -1.619 + 4.031x - 0.339x^2$.

3.49 THEORETICAL PROBABILITY DISTRIBUTIONS

Generally, frequency distribution are formed from the observed or experimental data. However, frequency distribution of certain populations can be deduced mathematically by fitting theoretical probability distribution under certain assumptions.

Frequency distributions can be classified under two heads:

(*i*) Observed Frequency Distributions.

(*ii*) Theoretical or Expected Frequency Distributions.

Observed frequency distributions are based on actual observation and experimentation. If certain hypothesis is assumed, it is sometimes possible to derive mathematically what

the frequency distribution of certain universe should be. Such distributions are called **Theoretical Distributions**.

Theoretical probability distributions are of two types:

(*i*) Discrete probability distribution. Binomial, poisson, geometric, negative binomial, hypergeometric, multinomial, multivariate hypergeometric distributions.

(*ii*) Continuous probability distributions.

Uniform, normal Gamma, exponential, χ^2, Beta, bivariate normal, t, F-distributions.

Here, we will study three important theoretical probability distributions:

1. Binomial Distribution (or Bernoulli's Distribution)
2. Poisson's Distribution
3. Normal Distribution.

3.50 BINOMIAL PROBABILITY DISTRIBUTION [*G.B.T.U. 2010, 2013*]

It was discovered by a Swiss Mathematician Jacob James Bernoulli in the year 1700.

This distribution is concerned with trials of a repretitive nature in which only the occurrence or non-occurrence, success or failure, acceptance or rejection, yes or no of a particular event is of interest.

For convenience, we shall call the occurrence of the event 'a success' and its non-occurrence 'a failure'.

Let there be n independent trials in an experiment. Let a random variable X denote the number of successes in these n trials. Let p be the probability of a success and q that of a failure in a single trial so that $p + q = 1$. Let the trials be independent and p be constant for every trial.

Let us find the probability of r successes in n trials.

r successes can be obtained in n trials in ${}^nC_r$ ways.

$$\therefore \quad P(X = r) = {}^nC_r \underbrace{P(S\,S\,S\,.....\,S}_{r \text{ times}} \quad \underbrace{F\,F\,F\,.....\,F}_{(n-r) \text{ times}}$$

$$= {}^nC_r \underbrace{P(S)\,P(S)\,.....\,P(S)}_{r \text{ factors}} \quad \underbrace{P(F)\,P(F)\,.....\,P(F)}_{(n-r) \text{ factors}}$$

$$= {}^nC_r \underbrace{p\,p\,p\,.....\,p}_{r \text{ factors}} \quad \underbrace{q\,q\,q\,.....\,q}_{(n-r) \text{ factors}}$$

$$= {}^nC_r p^r q^{n-r} \qquad ...(1)$$

Hence $\boxed{\mathbf{P(X = r) = {}^nC_r\, p^r\, q^{n-r}}}$ where **p + q = 1 and r = 0, 1, 2, ..., n.**

The distribution (1) is called the *binomial probability distribution* and X is called the *binomial variate*.

Note 1. P(X = r) is usually written as P(r).

Note 2. The successive probabilities P(r) in (1) for r = 0, 1, 2, ..., n are

$${}^nC_0 q^n,\ {}^nC_1 q^{n-1} p,\ {}^nC_2 q^{n-2} p^2,\ .....,\ {}^nC_n p^n$$

which are the successive terms of the binomial expansion of $(q + p)^n$. That is why this distribution is called "binomial" distribution.

Note 3. n and p occurring in the binomial distribution are called the *parameters* of the distribution.

Note 4. In a binomial distribution:

(*i*) n, the number of trials is finite.

(*ii*) each trial has only two possible outcomes usually called success and failure.

(*iii*) all the trials are independent.

(*iv*) p (and hence q) is constant for all the trials.

3.51 RECURRENCE OR RECURSION FORMULA FOR THE BINOMIAL DISTRIBUTION

In a binomial distribution,

$$P(r) = {}^nC_r\, q^{n-r}\, p^r = \frac{n!}{(n-r)!\,r!}\, q^{n-r}\, p^r$$

$$P(r+1) = {}^nC_{r+1}\, q^{n-r-1}\, p^{r+1} = \frac{n!}{(n-r-1)!\,(r+1)!}\, q^{n-r-1}\, p^{r+1}$$

$$\therefore \quad \frac{P(r+1)}{P(r)} = \frac{(n-r)!}{(n-r-1)!} \times \frac{r!}{(r+1)!} \times \frac{p}{q}$$

$$= \frac{(n-r)\times(n-r-1)!}{(n-r-1)!} \times \frac{r!}{(r+1)\times r!} \times \frac{p}{q} \times = \left(\frac{n-r}{r+1}\right) \cdot \frac{p}{q}$$

$$\Rightarrow \quad \boxed{P(r+1) = \frac{n-r}{r+1} \cdot \frac{p}{q}\, P(r)}$$

which is the required recurrence formula. Applying this formula successively, we can find P(1), P(2), P(3), ..., if P(0) is known.

3.52 MEAN AND VARIANCE OF THE BINOMIAL DISTRIBUTION

[*U.P.T.U. 2008, G.B.T.U. 2012*]

For the binomial distribution, $P(r) = {}^nC_r\, q^{n-r}\, p^r$

$$\text{Mean } \mu = \sum_{r=0}^{n} rP(r) = \sum_{r=0}^{n} r \,.\, {}^nC_r q^{n-r}\, p^r$$

$$= 0 + 1 \,.\, {}^nC_1\, q^{n-1}\, p + 2 \,.\, {}^nC_2\, q^{n-2}\, p^2 + 3 \,.\, {}^nC_3\, q^{n-3}\, p^3 + \ldots\ldots + n \,.\, {}^nC_n\, p^n$$

$$= nq^{n-1}\, p + 2 \,.\, \frac{n(n-1)}{2.1}\, q^{n-2} p^2 + 3 \,.\, \frac{n(n-1)(n-2)}{3.2.1}\, q^{n-3}\, p^3 + \ldots\ldots + n.p^n$$

$$= nq^{n-1}p + n(n-1)\, q^{n-2}p^2 + \frac{n(n-1)(n-2)}{2.1}\, q^{n-3}p^3 + \ldots\ldots + np^n$$

$$= np[{}^{n-1}C_0 q^{n-1} + {}^{n-1}C_1 q^{n-2}p + {}^{n-1}C_2 q^{n-3}p^2 + \ldots + {}^{n-1}C_{n-1} p^{n-1}]$$

$$= np(q+p)^{n-1} = np \qquad (\because \quad p+q=1)$$

Hence the mean of the binomial distribution in np.

$$\text{Variance } \sigma^2 = \sum_{r=0}^{n} r^2 P(r) - \mu^2 = \sum_{r=0}^{n} [r + r(r-1)]\, P(r) - \mu^2$$

$$= \sum_{r=0}^{n} rP(r) + \sum_{r=0}^{n} r(r-1)\, P(r) - \mu^2 = \mu + \sum_{r=2}^{n} r(r-1)\, {}^nC_r q^{n-r} p^r - \mu^2$$

(since the contribution due to $r = 0$ and $r = 1$ is zero)

$$= \mu + [2 \,.1\,.\, {}^{n}C_2 q^{n-2}p^2 + 3.2\,.\, {}^{n}C_3 q^{n-3}p^3 + \ldots + n(n-1)\, {}^{n}C_n p^n] - \mu^2$$

$$= \mu + \left[2\,.1\,.\frac{n(n-1)}{2\,.1}\,q^{n-2}p^2 + 3\,.2\,.\frac{n(n-1)(n-2)}{3\,.2\,.1}\,q^{n-3}p^3 + \ldots + n(n-1)p^n\right] - \mu^2$$

$$= \mu + [n(n-1)q^{n-2}p^2 + n(n-1)(n-2)q^{n-3}p^3 + \ldots + n(n-1)p^n] - \mu^2$$

$$= \mu + n(n-1)\,p^2[q^{n-2} + (n-2)q^{n-3}p + \ldots + p^{n-2}] - \mu^2$$

$$= \mu + n\,(n-1)p^2\,[{}^{n-2}C_0 q^{n-2} + {}^{n-2}C_1 q^{n-3}p + \ldots + {}^{n-2}C_{n-2}p^{n-2}] - \mu^2$$

$$= \mu + n(n-1)p^2(q+p)^{n-2} - \mu^2 = \mu + n(n-1)p^2 - \mu^2 \qquad [\because \quad q + p = 1]$$

$$= np + n(n-1)p^2 - n^2p^2 = np\,[1-p] = npq. \qquad [\because \quad \mu = np]$$

Hence the variance of the binomial distribution is npq.

Standard deviation of the binomial distribution is $\sqrt{npq}$.

3.53 MOMENT GENERATING FUNCTION OF BINOMIAL DISTRIBUTION

1. About origin

$$M_x(t) = E(e^{tx}) = \sum_{x=0}^{n} e^{tx}\ {}^{n}C_x\ p^x\ q^{n-x} = \sum_{x=0}^{n} {}^{n}C_x\ (pe^t)^x\ q^{n-x} = (q + pe^t)^n$$

2. About mean [*G.B.T.U. 2012, U.P.T.U. 2008, 2015*]

$$M_{x-np}(t) = E[e^{t(x-np)}]$$

$$= e^{-npt}\,E(e^{tx}) = e^{-npt}\,M_x(t) = e^{-npt}\,(q + pe^t)^n$$

$$= (qe^{-pt} + pe^{t-pt})^n = (qe^{-pt} + pe^{qt})^n \qquad |\ \because \quad 1 - p = q$$

3.54 MOMENTS ABOUT MEAN OF BINOMIAL DISTRIBUTION

$$M_{x-np}(t) = (qe^{-pt} + pe^{qt})^n$$

$$= \left[q\left(1 - pt + \frac{p^2t^2}{2!} - \frac{p^3t^3}{3!} + \ldots\right) + p\left(1 + qt + \frac{q^2t^2}{2!} + \frac{q^3t^3}{3!} + \ldots\right)\right]^n$$

$$= \left[(q+p) + \frac{t^2}{2!}\,pq\,(q+p) + \frac{t^3}{3!}\,pq\,(q^2 - p^2) + \frac{t^4}{4!}\,pq\,(q^3 + p^3) + \ldots\right]^n$$

$$= \left[1 + \left\{\frac{t^2}{2!}\,.\,pq + \frac{t^3}{3!}\,pq\,(q-p) + \frac{t^4}{4!}\,qp\,(1 - 3pq) + \ldots\right\}\right]^n$$

$$= \left[1 + {}^{n}C_1\left\{\frac{t^2}{2!}\,.\,pq + \frac{t^3}{3!}\,pq\,(q-p) + \frac{t^4}{4!}\,pq\,(1-3pq) + \ldots\right\}\right.$$

$$\left. + {}^{n}C_2\left\{\frac{t^2}{2!}\,.\,pq + \frac{t^3}{3!}\,pq\,(q-p) + \ldots\right\}^2 + \ldots\right]$$

Now,

$$\mu_2 = \text{coefficient of } \frac{t^2}{2!} = npq$$

$$\mu_3 = \text{coefficient of } \frac{t^3}{3!} = npq\,(q-p)$$

$$\mu_4 = \text{coefficient of } \frac{t^4}{4!} = npq\,(1-3pq) + 3n(n-1)\,p^2q^2$$

$$= 3n^2p^2q^2 + npq\,(1-6pq)$$

Hence,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(q-p)^2}{npq} = \frac{(1-2p)^2}{npq}$$

$$\therefore \quad \gamma_1 = \frac{1-2p}{\sqrt{npq}}$$

and

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 3 + \frac{1-6pq}{npq}$$

$$\therefore \quad \gamma_2 = \frac{1-6pq}{npq}$$

Note 1. $\gamma_1 = \dfrac{1-2p}{\sqrt{npq}}$ gives a **measure of skewness** of the binomial distribution. If $p < \frac{1}{2}$, skewness is positive, if $p > \frac{1}{2}$, skewness is negative and if $p = \frac{1}{2}$, it is zero.

$\beta_2 = 3 + \dfrac{1-6pq}{npq}$ gives a **measure of the kurtosis** of the binomial distribution.

Note 2. If n independent trials constitute one experiment and this experiment is repeated N times then the frequency of r successes is $N \cdot {}^nC_r\, p^r\, q^{n-r}$.

3.55 APPLICATIONS OF BINOMIAL DISTRIBUTION

1. In problems concerning no. of defectives in a sample production line.
2. In estimation of reliability of systems.
3. No. of rounds fired from a gun hitting a target.
4. In Radar detection.

EXAMPLES

Example 1. (*i*) *Comment on the following statement:*
For a Binomial distribution, mean is 6 and variance is 9.
(*ii*) *A die is tossed thrice. A success is getting 1 or 6 on a toss. Find the mean and variance of the number of success.*

Sol. (*i*)
$$\mu = np = 6 \qquad \ldots(1)$$
$$\sigma^2 = npq = 9 \qquad \ldots(2)$$

Dividing (2) by (1), we get

$$q = \frac{9}{6} = 1.5$$

which is impossible as $0 \le q \le 1$

$\therefore$ The above statement is **False.**

(*ii*) Prob. of getting success (1 or 6) on a toss = $\frac{2}{6} = \frac{1}{3} = p$

$\therefore \quad q = 1 - \frac{1}{3} = \frac{2}{3}$

No. of tosses of a die, $n = 3$

(*i*) Mean = $np = 3\left(\frac{1}{3}\right) = 1.$ (*ii*) Variance = $npq = (3)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right) = \frac{2}{3}.$

Example 2. *If 10% of the bolts produced by a machine are defective, determine the probability that out of 10 bolts choosen at random*

(*i*) *1* (*ii*) *None* (*iii*) *at most 2 bolts will be defective.*

Sol. Here, p(defective) = $\frac{10}{100} = \frac{1}{10}$ (given)

$\therefore \quad q$(non-defective) = $1 - \frac{1}{10} = \frac{9}{10}$

Also, $n = 10$, (n is no. of bolts chosen). (given)

The probability of r defective bolts out of n bolts chosen at random is given by

$$P(r) = {}^nC_r p^r q^{n-r} \quad ...(1)$$

(*i*) Here $r = 1$,

$\therefore \quad P(1) = {}^{10}C_1 \left(\frac{1}{10}\right)^1 \left(\frac{9}{10}\right)^{10-1}$ | Using (1)

$$= 10\left(\frac{1}{10}\right)\left(\frac{9}{10}\right)^9 = (.9)^9 = 0.3874 \quad ...(2)$$

(*ii*) Here $r = 0$

$\therefore \quad P(0) = {}^{10}C_0 \left(\frac{1}{10}\right)^0 \left(\frac{9}{10}\right)^{10-0} = \left(\frac{9}{10}\right)^{10} = 0.3486$ ...(3) | Using (1)

(*iii*) Prob. that at most 2 bolts will be defective = $P(r \le 2) = P(0) + P(1) + P(2)$ ...(4)

Now, $P(2) = {}^{10}C_2 \left(\frac{1}{10}\right)^2 \left(\frac{9}{10}\right)^{10-2}$ | Using (1)

$$= 45\left(\frac{1}{100}\right)(0.43046) = 0.1937$$

$\therefore$ From (4), Required Probability $= P(0) + P(1) + P(2)$

$= 0.3486 + 0.3874 + 0.1937 = 0.9297.$

Example 3. *A binomial variable X satisfies the relation 9P (X = 4) = P(X = 2) when n = 6. Find the value of the parameter p and P(X = 1).*

Sol. We know that

$$P(X = r) = {}^nC_r p^r q^{n-r} \quad ...(1)$$

$\therefore \quad P(X = 4) = {}^6C_4 p^4 q^2 = 15p^4q^2$

and $P(X = 2) = {}^6C_2 p^2 q^4 = 15 p^2 q^4$ | Since $n = 6$

The given relation is

$$9\,P(X = 4) = P(X = 2) \quad \Rightarrow \quad 9(15p^4 q^2) = 15p^2q^4$$

$\Rightarrow \quad 9p^2 = q^2 = (1-p)^2 \quad | \because \quad p+q=1$

$\Rightarrow \quad 9p^2 = 1 + p^2 - 2p$

$\Rightarrow \quad 8p^2 + 2p - 1 = 0 \quad \Rightarrow \quad (4p-1)(2p+1) = 0$

$\therefore \quad \boxed{p = \frac{1}{4}} \quad | \because \quad p$ cannot be negative

Now, $P(X=1) = {}^6C_1 \left(\frac{1}{4}\right)^1 \left(\frac{3}{4}\right)^5 = .3559.$ $\quad \left| \because \quad q = 1-p = \frac{3}{4} \right.$

Example 4. *Fit a binomial distribution to the following frequency data:*

x :	*0*	*1*	*2*	*3*	*4*
f :	*30*	*62*	*46*	*10*	*2*

Sol. The table is as follows:

x	f	fx
0	30	0
1	62	62
2	46	92
3	10	30
4	2	8
	$\Sigma f = 150$	$\Sigma fx = 192$

Mean of observations $= \frac{\Sigma fx}{\Sigma f} = \frac{192}{150} = 1.28$

$\Rightarrow \quad np = 1.28$

$\Rightarrow \quad 4p = 1.28 \quad (n$ is no. of trials)

$\Rightarrow \quad p = 0.32$

$\therefore \quad q = 1 - p = 1 - 0.32 = 0.68$

Also, $N = 150 \quad | \because \quad N = \Sigma f$

Hence the binomial distribution is $= N(q+p)^n = 150\,(0.68 + 0.32)^4$.

Example 5. *A student is given a true-false examination with 8 questions. If he corrects at least 7 questions, he passes the examination. Find the probability that he will pass given that he guesses all questions.*

Sol. Here, n = no. of questions asked = 8

$p = \frac{1}{2}, q = \frac{1}{2}$ | Since the question can either be true or false

Probability that he will pass

$$= P(r \geq 7) = P(7) + P(8)$$

$$= {}^8C_7 \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^{8-7} + {}^8C_8 \left(\frac{1}{2}\right)^8 \left(\frac{1}{2}\right)^{8-8} = 8\left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^1 + 1.\left(\frac{1}{2}\right)^8$$

$$= \left(\frac{1}{2}\right)^8 (8+1) = \frac{9}{256} = .03516.$$

Example 6. *During war, 1 ship out of 9 was sunk on an average in making a certain voyage. What was the probability that exactly 3 out of a convoy of 6 ships would arrive safely?*

Sol. p, the probability of a ship arriving safely $= 1 - \frac{1}{9} = \frac{8}{9}$; $q = \frac{1}{9}$, $n = 6$

The probability that exactly 3 ships arrive safely $= P(r = 3) = {}^6C_3 \left(\frac{1}{9}\right)^3 \left(\frac{8}{9}\right)^3 = \frac{10240}{9^6}$.

Example 7. *A policeman fires 6 bullets on a dacoit. The probability that the dacoit will be killed by a bullet is 0.6. What is the probability that dacoit is still alive?*

Sol. Here n = no. of bullets fired = 6, $p = 0.6$, $q = 1 - p = 0.4$

Probability that dacoit is still alive (not killed)

$$= P(r = 0) = {}^nC_0\, p^0\, q^{n-0} = {}^6C_0\, (.6)^0\, (.4)^6 = (.4)^6 = .004096.$$

Example 8. *If the probability of hitting a target is 10% and 10 shots are fired independently. What is the probability that the target will be hit at least once?*

Sol. Here, $p = \frac{10}{100} = \frac{1}{10}$, $q = 1 - p = 1 - \frac{1}{10} = \frac{9}{10}$, $n = 10$

Probability that the target will be hit at least once

$$= P(r \geq 1) = 1 - P(r = 0)$$

$$= 1 - [{}^nC_0\, p^0\, q^n] = 1 - \left[{}^{10}C_0 \left(\frac{1}{10}\right)^0 \left(\frac{9}{10}\right)^{10}\right] = 0.6513.$$

Example 9. *Out of 800 families with 4 children each, how many families would be expected to have (i) 2 boys and 2 girls (ii) at least one boy (iii) no girl (iv) atmost two girls? Assume equal probabilities for boys and girls.* (U.P.T.U. 2014)

Sol. Since probabilities for boys and girls are equal,

p = probability of having a boy $= \frac{1}{2}$; q = probability of having a girl $= \frac{1}{2}$

$n = 4$ N = 800

(*i*) The expected number of families having 2 boys and 2 girls

$$= 800\, {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = 800 \times 6 \times \frac{1}{16} = 300.$$

(*ii*) The expected number of families having *at least one boy*

$$= 800 \left[{}^4C_1 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right) + {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 + {}^4C_3 \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 + {}^4C_4 \left(\frac{1}{2}\right)^4\right]$$

$$= 800 \times \frac{1}{16} [4 + 6 + 4 + 1] = 750$$

(*iii*) The expected number of families having no girl *i.e.*, having 4 boys

$$= 800 \,.\, {}^4C_4 \left(\frac{1}{2}\right)^4 = 50.$$

(*iv*) The expected number of families having *atmost two girls i.e.*, having *at least 2 boy*

$$= 800 \left[{}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 + {}^4C_3 \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)^3 + {}^4C_4 \left(\frac{1}{2}\right)^4\right] = 800 \times \frac{1}{16} [6 + 4 + 1] = 550.$$

Example 10. *Six dice are thrown 729 times. How many times do you expect at least three dice to show a five or six?*

Sol. p = the chance of getting 5 or 6 with one die = $\frac{2}{6} = \frac{1}{3}$

$$q = 1 - \frac{1}{3} = \frac{2}{3},\ n = 6,\ \text{N} = 729$$

since dice are in sets of 6 and there are 729 sets.

The expected number of times *at least three* dice showing five or six

$$= \text{N} \cdot \text{P}(r \geq 3)$$

$$= 729\,[\text{P}(3) + \text{P}(4) + \text{P}(5) + \text{P}(6)]$$

$$= 729\left[{}^6\text{C}_3\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right)^3 + {}^6\text{C}_4\left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right)^4 + {}^6\text{C}_5\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^5 + {}^6\text{C}_6\left(\frac{1}{3}\right)^6\right]$$

$$= \frac{729}{3^6}\,[160 + 60 + 12 + 1] = 233.$$

Example 11. *The probability of a man hitting a target is* $\frac{1}{3}$. *How many times must he fire so that the probability of his hitting the target at least once is more than 90%?*

Sol. $p = \frac{1}{3}$

The probability of not hitting the target in n trials is q^n.

Therefore, to find the smallest n for which the probability of hitting at least once is more than 90%, we have

$$1 - q^n > 0.9$$

$$\Rightarrow \quad 1 - \left(\frac{2}{3}\right)^n > 0.9$$

$$\Rightarrow \quad \left(\frac{2}{3}\right)^n < 0.1$$

The smallest n for which the above inequality holds true is 6 hence he must fire 6 times.

Example 12. *In a bombing action, there is 50% chance that any bomb will strike the target. Two direct hits are needed to destroy the target completely. How many bombs are required to be dropped to give a 99% chance or better of completely destroying the target?*

Sol. We have, $p = \frac{50}{100} = \frac{1}{2}$

Since the probability must be greater than 0.99, if n bombs are dropped, we have

$${}^n\text{C}_2\left(\frac{1}{2}\right)^n + {}^n\text{C}_3\left(\frac{1}{2}\right)^n + {}^n\text{C}_4\left(\frac{1}{2}\right)^n + \ldots\ldots + {}^n\text{C}_n\left(\frac{1}{2}\right)^n \geq 0.99$$

$$\Rightarrow \quad \left(\frac{1}{2}\right)^n [{}^n\text{C}_2 + {}^n\text{C}_3 + {}^n\text{C}_4 + \ldots\ldots + {}^n\text{C}_n] \geq 0.99$$

$$\frac{2^n - n - 1}{2^n} \geq 0.99$$

$$\Rightarrow \quad 1 - \frac{1+n}{2^n} \geq 0.99$$

$$\Rightarrow \quad \frac{1+n}{2^n} \leq 0.01$$

$$\Rightarrow \quad 2^n \geq 100n + 100$$

By trial, $n = 11$ satisfies the inequality.

Hence 11 bombs are required to be dropped.

ASSIGNMENT

1. (*i*) Ten coins are tossed simultaneously. Find the probability of getting at least seven heads.

(*ii*) A die is thrown five times. If getting an odd number is a success, find the probability of getting at least four successes. (*M.T.U. 2012*)

2. (*a*) The probability of any ship of a company being destroyed on a certain voyage is 0.02. The company owns 6 ships for the voyage. What is the probability of :

(*i*) losing one ship (*ii*) losing atmost two ships (*iii*) losing none?

(*b*) Assume that on the average one telephone number out of fifteen called between 2 P.M. and 3 P.M. on week-days is busy. What is the probability that if 6 randomly selected telephone numbers are called (*i*) not more than 3 (*ii*) at least 3 of them will be busy?

3. (*i*) The incidence of occupational disease in an industry is such that the workers have a 20% chance of suffering from it. What is the probability that out of six workers chosen at random, four or more will suffer from the disease?

(*ii*) The probability that a man aged 60 will live to be 70 is 0.65. What is the probability that out of 10 men, now 60, at least 7 will live to be 70?

4. (*i*) If the mean of a binomial distribution is 3 and the variance is $\frac{3}{2}$, find the probability of obtaining at least 4 successes.

(*ii*) In a binomial distribution, for $n = 5$ if $P(x = 1) = 0.4096$ and $P(x = 2) = 0.2048$, then find the value of p.

(*iii*) The sum and product of the mean and variance of a binomial distribution are $\frac{25}{3}$ and $\frac{50}{3}$ respectively. Find the distribution. (*U.P.T.U. 2007*)

(*iv*) If the probability of a defective bolt is 0.1, find (*a*) The mean (*b*) The standard deviation for the distribution in a total of 400 bolts.

(*v*) If the moment generating function of a random variable X is $\left(\frac{1}{3} + \frac{2}{3}e^t\right)^5$, find P(X = 2).

5. (*a*) The probability that a bomb dropped from a plane will strike the target is $\frac{1}{5}$. If six bombs are dropped, find the probability that (*i*) exactly two will strike the target, (*ii*) at least two will strike the target.

(*b*) Four persons in a group of 20 are graduates. If 4 persons are selected at random from 20, find the probability that

(*i*) all are graduates (*ii*) at least one is a graduate.

6. A bag contains 5 white, 7 red and 8 black balls. If four balls are drawn, one by one, with replacement, what is the probability that

(*i*) none is white (*ii*) all are white

(*iii*) at least one is white (*iv*) only 2 are white?

7. (*i*) In a hurdle race, a player has to cross 10 hurdles. The probability that he will clear each hurdle is $\frac{5}{6}$. What is the probability that he will knock down fewer than 2 hurdles ?

(*ii*) If on an average one ship in every ten is wrecked, find the probability that out of 5 ships expected to arrive, 4 at least will arrive safely.

8. The prob. that a bulb produced by a factory will fuse after a use of 150 days is 0.05. Find the prob. that out of 5 such bulbs [*M.T.U.* (*B. Pharma*) *2011*]

(*i*) None (*ii*) Atmost one

(*iii*) More than one (*iv*) At least one will fuse after 150 days of use.

9. Five cards are drawn successively with replacement from a well-shuffled deck of 52 cards. What is the probability that

(*i*) all the five cards are spades (*ii*) only three are spades (*iii*) none is spade?

10. Manish takes a step forward with probability 0.4 and backward with probability 0.6. Find the probability that at the end of 11 steps, he is one step away from the starting point.

[**Hint.** Manish will take either 6 steps forward and 5 backward or 5 steps forward and 6 backward].

11. (*a*) In 800 families with 5 children each, how many families would be expected to have (*i*) 3 boys and 2 girls, (*ii*) 2 boys and 3 girls, (*iii*) no girl (*iv*) at the most two girls. (Assume probabilities for boys and girls to be equal.)

(*b*) Out of 800 families with 5 children each, how many would you expect to have (*i*) 3 boys (*ii*) 5 girls (*iii*) either 2 or 3 boys? Assume equal probabilities for boys and girls.

(*M.T.U. 2012*)

(*c*) Out of 320 families with 5 children each, what percentage would be expected to have (*i*) 2 boys and 3 girls (*ii*) at least one boy? Assume equal probability for boys and girls.

(*G.B.T.U. 2011*)

12. In 100 sets of ten tosses of an unbiased coin, in how many cases do you expect to get

(*i*) 7 heads and 3 tails (*ii*) at least 7 heads?

13. The following data are the number of seeds germinating out of 10 on damp filter for 80 sets of seeds. Fit a binomial distribution to this data:

x:	0	1	2	3	4	5	6	7	8	9	10	Total
f:	6	20	28	12	8	6	0	0	0	0	0	80

[**Hint.** Here $n = 10$, $N = 80$, Mean $= \frac{\Sigma fx}{\Sigma f} = 2.175 \quad \therefore \quad np = 2.175$ etc.]

14. Fit a binomial distribution for the following data and compare the theoretical frequencies with the actual ones :

x:	0	1	2	3	4	5
f:	2	14	20	34	22	8

15. Fit a binomial distribution to the data given in the following table:

(*i*)

x:	0	1	2	3	4
f:	24	41	28	5	2

(*M.T.U. 2012*)

(*ii*)

x:	0	1	3	4
f:	28	62	10	4

(*U.K.T.U. 2011*)

16. (*i*) Assuming half the population of a town consumes chocolates so that the chance of an individual being consumer is $\frac{1}{2}$ and that 100 investigators each take 10 individuals to see whether they are consumers, how many investigators would you expect to report that 3 people or less were consumers? (*U.P.T.U. 2006*)

(*ii*) Assuming that 20% of the population of a city are literate, so that the chance of an individual being literate is $\frac{1}{5}$ and assuming that 100 investigators each take 10 individuals to see whether they are literate, how many investigators would you expect to report 3 or less were literate?

17. Following results were obtained when 100 batches of seeds were allowed to germinate on damp filter paper in a laboratory : $\beta_1 = \frac{1}{15}, \beta_2 = \frac{89}{30}$. Determine the Binomial distribution. Calculate the expected frequency for $x = 8$ assuming $p > q$.

18. A coffee connoisseur claims that he can distinguish between a cup of instant coffee and a cup of percolator coffee 75% of the time. It is agreed that his claim will be accepted if he correctly identifies at least 5 of the 6 cups. Find his chances of having the claim (*i*) accepted (*ii*) rejected when he does have the ability he claims.

19. A multiple-choice test consists of 8 questions with 3 answers to each question of which only one is correct. A student answers each question by rolling a balanced die and checking the first answer if he gets 1 or 2, the second answer if he gets 3 or 4 and the third answer if he gets 5 or 6. To get a distinction, the student must secure at least 75% correct answers. If there is no negative marking, what is the probability that the student secures a distinction?

20. An irregular six-faced die is thrown and the expectation that in 10 throws, it will give five even numbers is twice the expectation that it will give four even numbers. How many times in 10,000 sets of 10 throws each, would you expect it to give no even number?

Answers

1. (*i*) $\frac{11}{64}$ (*ii*) $\frac{3}{16}$

2. (*a*) (*i*) 0.1085, (*ii*) 0.9997, (*iii*) 0.8858 (*b*) (*i*) 0.9997 (*ii*) 0.005

3. (*i*) $\frac{53}{3125}$ (*ii*) 0.5137

4. (*i*) $\frac{11}{32}$, (*ii*) $\frac{1}{5}$, (*iii*) $\left(\frac{2}{3}+\frac{1}{3}\right)^{15}$
(*iv*) (*a*) 40 (*b*) 6 (*v*) 0.1646

5. (*a*) (*i*) 0.246 (*ii*) 0.345 (*b*) (*i*) 0.0016 (*ii*) 0.5904

6. (*i*) $\frac{81}{256}$, (*ii*) $\frac{1}{256}$, (*iii*) $\frac{175}{256}$, (*iv*) $\frac{27}{128}$ **7.** (*i*) $\frac{5}{2}\left(\frac{5}{6}\right)^9$, (*ii*) 0.91854

8. (*i*) $\left(\frac{19}{20}\right)^5$ (*ii*) $\frac{6}{5}\left(\frac{19}{20}\right)^4$ (*iii*) $1-\frac{6}{5}\left(\frac{19}{20}\right)^4$ (*iv*) $1-\left(\frac{19}{20}\right)^5$

9. (*i*) $\left(\frac{1}{4}\right)^5$ (*ii*) $90\left(\frac{1}{4}\right)^5$ (*iii*) $\left(\frac{3}{4}\right)^5$ **10.** 0.36787

11. (*a*) (*i*) 250 (*ii*) 250 (*iii*) 25 (*iv*) 400
(*b*) (*i*) 250 (*ii*) 25 (*iii*) 500
(*c*) (*i*) 31.25% (*ii*) 96.875%

12. (*i*) 12 nearly (*ii*) 17 nearly **13.** $80\,(0.7825 + 0.2175)^{10}$

14. $100\,(0.432 + 0.568)^5$ **15.** (*i*) $100\,(0.7 + 0.3)^4$, (*ii*) $104\,(0.7404 + 0.2596)^4$

16. (*i*) 17 (*ii*) 88 **17.** 100 $\left(\frac{1}{4}+\frac{3}{4}\right)^{20}$, 0.075168752

18. (*i*) 0.534 (*ii*) 0.466

19. 0.0197 **20.** 1.

POISSON DISTRIBUTION

3.56 POISSON DISTRIBUTION AS A LIMITING CASE OF BINOMIAL DISTRIBUTION

[*U.P.T.U. 2007 ; G.B.T.U. 2010, 2013 ; M.T.U. 2013, 2014*]

Poisson distribution was discovered by S.D. Poisson in the year 1837.

If the parameters n and p of a binomial distribution are known, we can find the distribution. But in situations where n is very large and p is very small, application of binomial distribution is very labourious. However, if we assume that as $n \to \infty$ and $p \to 0$ such that np always remains finite, say λ, we get the Poisson approximation to the binomial distribution.

Now, for a binomial distribution

$$P(X = r) = {}^nC_r q^{n-r} p^r$$

$$= \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!} \times (1-p)^{n-r} \times p^r$$

$$= \frac{n(n-1)(n-2)\ldots(n-r+1)}{r!} \times \left(1-\frac{\lambda}{n}\right)^{n-r} \times \left(\frac{\lambda}{n}\right)^r \qquad \left| \text{Since } np = \lambda \quad \therefore \quad p = \frac{\lambda}{n}\right.$$

$$= \frac{\lambda^r}{r!} \times \frac{n(n-1)(n-2)\ldots(n-r+1)}{n^r} \times \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^r}$$

$$= \frac{\lambda^r}{r!}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\ldots\ldots\left(\frac{n-r+1}{n}\right) \times \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^r}$$

$$= \frac{\lambda^r}{r!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\ldots\left(1-\frac{r-1}{n}\right) \times \frac{\left[\left(1-\frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right]^{-\lambda}}{\left(1-\frac{\lambda}{n}\right)^r}$$

As $n \to \infty$, each of the $(r-1)$ factors

$$\left(1-\frac{1}{n}\right), \left(1-\frac{2}{n}\right), \ldots\ldots, \left(1-\frac{r-1}{n}\right)$$ tends to 1. Also $\left(1-\frac{\lambda}{n}\right)^r$ tends to 1.

Since $\underset{x \to \pm\infty}{\text{Lt}} \left(1+\frac{1}{x}\right)^x = e$, the Naperian base. $\therefore$ $\left[\left(1-\frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right]^{-\lambda} \longrightarrow e^{-\lambda}$ as $n \to \infty$

Hence in the limiting case when $n \to \infty$, we have

$$\boxed{P(X = r) = \frac{e^{-\lambda}.\lambda^r}{r!} \ (r = 0, 1, 2, 3, ...)} \qquad ...(1)$$

where λ is a finite number = np.

(1) represents a probability distribution which is called the *Poisson probability distribution*.

Note 1. λ is called the parameter of the distribution.

Note 2. The sum of the probabilities $P(r)$ for $r = 0, 1, 2, 3, ...$ is 1, since

$$P(0) + P(1) + P(2) + P(3) + ... = e^{-\lambda} + \frac{\lambda e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^3 e^{-\lambda}}{3!} + ...$$

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

3.57 RECURRENCE FORMULA FOR THE POISSON DISTRIBUTION

[*U.P.T.U. (B. Pharma.) 2009*]

For Poisson distribution, $P(r) = \frac{e^{-\lambda}\,\lambda^r}{r!}$ and $P(r+1) = \frac{e^{-\lambda}\,\lambda^{r+1}}{(r+1)!}$

$$\therefore \qquad \frac{P(r+1)}{P(r)} = \frac{\lambda r!}{(r+1)!} = \frac{\lambda}{r+1}$$

or $$P(r+1) = \frac{\lambda}{r+1} P(r), \ r = 0, 1, 2, 3, ...$$

This is called the *recurrence* or *recursion formula* for the Poisson distribution.

3.58 MEAN AND VARIANCE OF THE POISSON DISTRIBUTION [*U.P.T.U. 2006*]

For the Poisson distribution, $P(r) = \frac{e^{-\lambda}\,\lambda^r}{r!}$

$$\text{Mean } \mu = \sum_{r=0}^{\infty} rP(r) = \sum_{r=0}^{\infty} r \cdot \frac{e^{-\lambda}\,\lambda^r}{r!}$$

$$= e^{-\lambda} \sum_{r=1}^{\infty} \frac{\lambda^r}{(r-1)!} = e^{-\lambda}\left(\lambda + \frac{\lambda^2}{1!} + \frac{\lambda^3}{2!} + ...\right)$$

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

Thus, the mean of the Poisson distribution is equal to the parameter λ.

$$\text{Variance } \sigma^2 = \sum_{r=0}^{\infty} r^2 P(r) - \mu^2 = \sum_{r=0}^{\infty} r^2 \cdot \frac{\lambda^r e^{-\lambda}}{r!} - \lambda^2 = e^{-\lambda} \sum_{r=1}^{\infty} \frac{r^2 \lambda^r}{r!} - \lambda^2$$

$$= e^{-\lambda}\left[\frac{1^2 \cdot \lambda}{1!} + \frac{2^2 \cdot \lambda^2}{2!} + \frac{3^2 \cdot \lambda^3}{3!} + \frac{4^2 \cdot \lambda^4}{4!} + ...\right] - \lambda^2$$

$$= \lambda e^{-\lambda}\left[1 + \frac{2\lambda^2}{1!} + \frac{3\lambda^2}{2!} + \frac{4\lambda^3}{3!} + ...\right] - \lambda^2$$

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

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

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

$= \lambda e^{-\lambda}\,[e^{\lambda} + \lambda e^{\lambda}] - \lambda^2 = \lambda e^{-\lambda} \cdot e^{\lambda}\,(1 + \lambda) - \lambda^2 = \lambda(1 + \lambda) - \lambda^2 = \lambda.$

Hence, the variance of the Poisson distribution is also λ.

Thus, the mean and the variance of the Poisson distribution are each equal to the parameter λ.

Note 1. The mean and the variance of the Poisson distribution can also be derived from those of the binomial distribution in the limiting case when $n \to \infty$, $p \to 0$ and $np = \lambda$.

Mean of Binomial distribution is np.

$\therefore$ Mean of Poisson distribution $= \underset{n \to \infty}{\text{Lt}}\ np = \underset{n \to \infty}{\text{Lt}}\ \lambda = \lambda$

Variance of Binomial distribution is $npq = np\,(1 - p)$

$\therefore$ Variance of Poisson distribution $= \underset{n \to \infty}{\text{Lt}}\ np\,(1 - p) = \underset{n \to \infty}{\text{Lt}}\ \lambda\left(1 - \frac{\lambda}{n}\right) = \lambda.$

Note 2. For Poisson distribution, $\mu_3 = \lambda$ and $\mu_4 = 3\lambda^2 + \lambda$.

Coefficients of skewness and kurtosis are given by

$$\beta_1 = \frac{1}{\lambda} \text{ and } \gamma_1 = \frac{1}{\sqrt{\lambda}}. \text{ Also, } \beta_2 = 3 + \frac{1}{\lambda} \text{ and } \gamma_2 = \frac{1}{\lambda}$$

Hence Poisson distribution is always a skewed distribution.

Remark. While fitting the Poisson distribution to a given data, we round the figures to the nearest integer but it should be kept in mind that the total of the observed and the expected frequencies should be same.

3.59 MODE OF POISSON DISTRIBUTION

Let $\quad P(x = r) = e^{-\lambda}\frac{\lambda^r}{r!},\ r = 0, 1, 2, ..., \infty$

The value of r which has a greater probability than any other value is the mode of the Poisson distribution. Thus r is mode if

$$P(X = r) \geq P(X = r + 1) \text{ and } P(X = r) \geq P(X = r - 1)$$

$$\Rightarrow \quad \frac{e^{-\lambda} \cdot \lambda^r}{r!} \geq \frac{e^{-\lambda} \cdot \lambda^{r+1}}{(r+1)!} \quad \text{and} \quad \frac{e^{-\lambda} \cdot \lambda^r}{r!} \geq \frac{e^{-\lambda} \cdot \lambda^{r-1}}{(r-1)!}$$

$$\Rightarrow \quad 1 \geq \frac{\lambda}{r+1} \quad \text{and} \quad \frac{\lambda}{r} \geq 1$$

$$\Rightarrow \quad r \geq \lambda - 1 \quad \text{and} \quad r \leq \lambda \quad i.e., \quad \lambda - 1 \leq r \leq \lambda$$

Case I. If λ is a positive integer, there are two modes $\lambda - 1$ and λ.

Case II. If λ is not a positive integer, there is one mode and is the integral value between $\lambda - 1$ and λ.

3.60 APPLICATIONS OF POISSON DISTRIBUTION

This distribution is applied to problems concerning :

(*i*) Arrival pattern of defective vehicles in a workshop.

(*ii*) Patients in a hospitals.

(*iii*) Telephone calls.

(*iv*) Demand pattern for certain spare parts.

(*v*) Number of fragments from a shell hitting a target.

(*vi*) Emission of radioactive (α) particles.

EXAMPLES

Example 1. *If the variance of the Poisson distribution is 2, find the probabilities for r = 1, 2, 3, 4 from the recurrence relation of the Poisson distribution. Also find P(r ≥ 4).* (M.T.U. 2013)

Sol. λ, the parameter of Poisson distribution = Variance = 2

Recurrence relation for the Poisson distribution is

$$P(r+1) = \frac{\lambda}{r+1} P(r) = \frac{2}{r+1} P(r) \qquad ...(1)$$

Now $$P(r) = \frac{e^{-\lambda} \lambda^r}{r!} \Rightarrow P(0) = \frac{e^{-2} (2)^0}{0!} = e^{-2} = 0.1353$$

Putting $r = 0, 1, 2, 3$ in (1), we get

$$P(1) = 2P(0) = 2 \times 0.1353 = 0.2706 ; \quad P(2) = \frac{2}{2} P(1) = 0.2706$$

$$P(3) = \frac{2}{3} P(2) = \frac{2}{3} \times 0.2706 = 0.1804 ; \quad P(4) = \frac{2}{4} P(3) = \frac{1}{2} \times 0.1804 = 0.0902.$$

Now, $P(r \geq 4) = 1 - [P(0) + P(1) + P(2) + P(3)]$

$= 1 - [0.1353 + 0.2706 + 0.2706 + 0.1804] = 0.1431.$

Example 2. *Using Poisson distribution, find the probability that the ace of spades will be drawn from a pack of well-shuffled cards at least once in 104 consecutive trials.* (U.P.T.U. 2015)

Sol. $$p = \frac{1}{52}, n = 104$$

$\therefore$ $$\lambda = np = 104 \times \frac{1}{52} = 2$$

Prob. (at least once) $= P(r \geq 1) = 1 - P(0)$

$$= 1 - \frac{e^{-\lambda} . \lambda^0}{0!} = 1 - e^{-2} = 1 - 0.135335 \approx 0.8647.$$

Example 3. (*i*) *Fit a Poisson distribution to the following data and calculate theoretical frequencies.*

Deaths:	*0*	*1*	*2*	*3*	*4*
Frequencies:	*122*	*260*	*15*	*2*	*1*

[U.P.T.U. 2014 ; U.K.T.U. 2010]

(*ii*) *The frequency of accidents per shift in a factory is shown in the following table:*

Accident per shift	*Frequency*
0	*192*
1	*100*
2	*24*
3	*3*
4	*1*
Total	320

Calculate the mean number of accidents per shift. Fit a Poisson distribution and calculate theoretical frequencies.

Sol. (*i*) Mean of given distribution = $\frac{\Sigma fx}{\Sigma f}$

$$\Rightarrow \qquad \lambda = \frac{60 + 30 + 6 + 4}{200} = 0.5$$

Required Poisson distribution = $N \cdot \frac{e^{-\lambda} \cdot \lambda^r}{r!} = 200 \cdot \frac{e^{-.5} (.5)^r}{r!} = (121.306) \frac{(0.5)^r}{r!}$

r	*N. P(r)*	*Theoretical frequency*
0	$121.306 \times \frac{(0.5)^0}{0!} = 121.306$	121
1	$121.306 \times \frac{(0.5)^1}{1!} = 60.653$	61
2	$121.306 \times \frac{(0.5)^2}{2!} = 15.163$	15
3	$121.306 \times \frac{(0.5)^3}{3!} = 2.527$	3
4	$121.306 \times \frac{(0.5)^4}{4!} = 0.3159$	0
		Total = 200

(*ii*) Mean number of accidents per shift

$$\lambda = \frac{\Sigma fx}{\Sigma f} = \frac{100 + 48 + 9 + 4}{320} = 0.5031$$

∴ Required Poisson distribution

$$= N \cdot \frac{e^{-\lambda} \cdot \lambda^r}{r!} = 320 \cdot \frac{e^{-.5031} (.5031)^r}{r!} = \frac{(193.48)(.5031)^r}{r!}$$

r	*N. P(r)*	*Theoretical frequency*
0	193.48	194
1	97.34	97
2	24.38	24
3	4.10	4
4	0.51	1
		Total = 320

Example 4. (*i*) *Suppose that a book of 600 pages contains 40 printing mistakes. Assume that these errors are randomly distributed throughout the book and r, the number of errors per page has a Poisson distribution. What is the probability that 10 pages selected at random will be free from errors?*

(*ii*) *Wireless sets are manufactured with 25 solders joints each, on the average 1 joint in 500 is defective. How many sets can be expected to be free from defective joints in a consignment of 10000 sets?*

Sol. (*i*) $p = \frac{40}{600} = \frac{1}{15}, \quad n = 10$

∴ $\lambda = np = 10\left(\frac{1}{15}\right) = \frac{2}{3}$

$$P(r) = \frac{e^{-\lambda} \lambda^r}{r!} = \frac{e^{-2/3} (2/3)^r}{r!}$$

∴ $P(0) = \frac{e^{-2/3} (2/3)^0}{0!} = e^{-2/3} = 0.51.$

(*ii*) $p = \frac{1}{500}, n = 25$

∴ $\lambda = np = 25 \times \frac{1}{500} = \frac{1}{20} = 0.05$

No. of sets in a consignment, N= 10000

Probability of having no defective joint = $P(r = 0) = \frac{e^{-.05} (0.05)^0}{0!} = 0.9512.$

∴ The expected no. of sets free from defective joints = 0.9512 × 10000 = 9512.

Example 5. *A manufacturer knows that the condensors he makes contain on an average 1% of defectives. He packes them in boxes of 100. What is the probability that a box picked at random will contain 4 or more faulty condensors?*

Sol. $p = 0.01, n = 100$

∴ $\lambda = np = 1$

$$P(r) = \frac{e^{-\lambda} \lambda^r}{r!} = \frac{e^{-1}}{r!}$$

P(4 or more faulty condensors) = P(4) + P(5) + ... + P(100)

$$= 1 - [P(0) + P(1) + P(2) + P(3)]$$

$$= 1 - \left[\frac{e^{-1}}{0!} + \frac{e^{-1}}{1!} + \frac{e^{-1}}{2!} + \frac{e^{-1}}{3!}\right]$$

$$= 1 - e^{-1}\left[1 + 1 + \frac{1}{2} + \frac{1}{6}\right] = 1 - \frac{8}{3e} = 1 - 0.981 = 0.019.$$

Example 6. (*i*) *If the probabilities of a bad reaction from a certain injection is 0.0002, determine the chance that out of 1000 individuals more than two will get a bad reaction.*

(*ii*) *The probability that a man aged 50 years will die within a year is 0.01125. What is the probability that of 12 such men, at least 11 will reach their 51st birthday?*

(*Given:* $e^{-.135} = 0.87366$)

Sol. (*i*) Here, $p = 0.0002, \quad n = 1000$

$\therefore \quad \lambda = np = 1000 \times 0.0002 = 0.2.$

Since the prob. of bad reaction is very small and no. of trials is very high, we use Poisson distribution here.

The prob. that out of 100 individuals, more than 2 will get a bad reaction is

$$= P(r > 2) = 1 - P(r \le 2) = 1 - [P(0) + P(1) + P(2)] \qquad ...(1)$$

Now, $P(0) = \dfrac{e^{-0.2}\,(0.2)^0}{0} = 0.8187$ (Here $r = 0$)

$$P(1) = \frac{e^{-0.2}\,(0.2)^1}{1} = 0.1637 \qquad \text{(Here } r = 1\text{)}$$

and $P(2) = \dfrac{e^{-0.2}\,(0.2)^2}{2} = 0.0164.$ (Here $r = 2$)

$\therefore$ From (1), Reqd. probability = 1 – [0.8187 + 0.1637 + 0.0164] = 0.0012.

(*ii*) $p = 0.01125, n = 12$

$\therefore \quad \lambda = np = 12 \times 0.01125 = 0.135$

P (at least 11 survive) = P(atmost 1 dies)

$$= P(0) + P(1) = \frac{e^{-\lambda}.\lambda^0}{0!} + \frac{e^{-\lambda}.\lambda^1}{1!}$$

$$= e^{-.135}\,(1 + 0.135) = 1.135 \times 0.87366 = 0.9916.$$

Example 7. *A car-hire firm has two cars, which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with mean 1.5. Calculate the proportion of days on which neither car is used and the proportion of days on which some demand is refused* ($e^{-1.5} = 0.2231$).

Sol. Since the number of demands for a car is distributed as a Poisson distribution with mean $\lambda = 1.5$.

$\therefore$ Proportion of days on which neither car is used

= Probability of there being no demand for the car

$$= \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-1.5} = 0.2231$$

Proportion of days on which some demand is refused

= probability for the number of demands to be more than two

$$= 1 - P(x \le 2) = 1 - \left(e^{-\lambda} + \frac{\lambda e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} \right)$$

$$= 1 - e^{-1.5} \left(1 + 1.5 + \frac{(1.5)^2}{2} \right) = 0.1912625.$$

Example 8. *Suppose the number of telephone calls on an operator received from 9 : 00 to 9 : 05 follow a Poisson distribution with a mean 3. Find the probability that*

(*i*) *The operator will receive no calls in that time interval tomorrow.*

(*ii*) *In the next three days, the operator will receive a total of 1 call in that time interval.* (*Given:* $e^{-3} = 0.04978$)

Sol. Here, $\lambda = 3$

(*i*) $$P(0) = \frac{e^{-\lambda} . \lambda^0}{0!} = e^{-3} = 0.04978$$

(*ii*) Reqd. probability = P(0) P(0) P(1) + P(0) P(1) P(0) + P(1) P(0) P(0)

$$= 3 \left\{ \frac{e^{-\lambda} . \lambda^0}{0!} \right\}^2 \frac{e^{-\lambda} . \lambda^1}{1!} = 9(e^{-3})^3 = 0.00111.$$

Example 9. *The no. of arrivals of customers during any day follows Poisson distribution with a mean of 5. What is the probability that the total no. of customers on two days selected at random is less than 2?* (*Given:* $e^{-10} = 4.54 \times 10^{-5}$)

Sol. $\lambda = 5$

Arrival of Customers

I day	*II day*	*Total*
0	0	0
0	1	1
1	0	1

Reqd. probability = P(0) P(0) + P(0) P(1) + P(1) P(0)

$$= \frac{e^{-5} .5^0}{0!} \cdot \frac{e^{-5} .5^0}{0!} + \frac{e^{-5} .5^0}{0!} \cdot \frac{e^{-5} .5^1}{1!} + \frac{e^{-5} .5^1}{1!} \cdot \frac{e^{-5} .5^0}{0!}$$

$$= e^{-10} + 2 . 5 . e^{-10} = 11 e^{-10} = 11 \times 4.54 \times 10^{-5}$$

$$= 4.994 \times 10^{-4}.$$

Example 10. *An insurance company finds that 0.005% of the population dies from a certain kind of accident each year. What is the probability that the company must pay off no more than 3 of 10,000 insured risks against such incident in a given year?*

Sol. $$p = \frac{0.005}{100} = 0.00005, \quad n = 10000$$

$\therefore$ $\lambda = np = 10000 \times 0.00005 = 0.5$

Reqd. Probability = $1 - P(r \le 3) = 1 - [P(0) + P(1) + P(2) + P(3)]$

$$= 1 - \left[\frac{e^{-0.5}(0.5)^0}{0!} + \frac{e^{-0.5}(0.5)^1}{1!} + \frac{e^{-0.5}(0.5)^2}{2!} + \frac{e^{-0.5}(0.5)^3}{3!}\right]$$

$$= 1 - e^{-5}\,[1 + 0.5 + 0.125 + 0.021] = 0.0016.$$

Example 11. *In a certain factory turning out razor blades, there is a small chance of 0.002 for any blade to be defective. The blades are supplied in packets of 10. Calculate the approximate number of packets containing no defective, one defective and two defective blades in a consignment of 10,000 packets.* (*Given:* $e^{-0.02} = 0.9802$) [U.P.T.U. 2009]

Sol. p(defective) = 0.002

$n = 10$ (no. of blades in a packet)

$\therefore \quad \lambda = np = 10 \times 0.002 = 0.02$

No. of packets in the consignment, N = 10,000.

(*i*) Probability of having no defective = $P(0) = \dfrac{e^{-0.02}\,(0.02)^0}{0!} = 0.9802$ | Here $r = 0$

Approximate no. of packets having zero defective in the consignment = 0.9802×10000

$= 9802$

(*ii*) Probability of having one defective = $P(1) = \dfrac{e^{-0.02}\,(0.02)^1}{1!} = 0.9802 \times 0.02 = 0.019604$

Approximate no. of packets having one defective in the consignment

$= 0.019604 \times 10000 \approx 196.$

(*iii*) Probability of having two defective blades

$$P(2) = \frac{e^{-0.02}\,(0.02)^2}{2!} = \frac{(0.980198) \times (0.0004)}{2} = 0.000196.$$

$\therefore$ Approximate no. of packet having two defectives in the consignment

$= 0.000196 \times 10000 = 1.96 \approx 2.$

Example 12. (*i*) *Six coins are tossed 6400 times. Using the Poisson distribution, determine the approximate probability of getting six heads x times.* [U.P.T.U. (C.O.) 2008]

(*ii*) *A Poisson distribution has a double mode at x = 3 and x = 4. What is the probability that x will have one or the other of these two values?* [U.P.T.U. (C.O.) 2008]

Sol. (*i*) Probability of getting one head with one coin = $\dfrac{1}{2}$.

$\therefore$ The probability of getting six heads with six coins = $\left(\dfrac{1}{2}\right)^6 = \dfrac{1}{64}$

$\therefore$ Average number of six heads with six coins in 6400 throws = $np = 6400 \times \dfrac{1}{64} = 100$

$\therefore$ The mean of the Poisson distribution = 100.

Approximate probability of getting six heads x times when the distribution is Poisson

$$= \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(100)^x \cdot e^{-100}}{x!}.$$

(*ii*) Since 2 modes are given when λ is an integer, modes are $\lambda - 1$ and λ.

$\therefore \quad \lambda - 1 = 3 \quad \Rightarrow \quad \lambda = 4$

Probability (when $r = 3$) $= \dfrac{e^{-4}\,(4)^3}{3\,!}$

Probability (when $r = 4$) $= \dfrac{e^{-4}\,(4)^4}{4\,!}$

$\therefore$ Required Probability = P($r = 3$ or 4) = P($r = 3$) + P($r = 4$)

$$= \frac{e^{-4}\,(4)^3}{3\,!} + \frac{e^{-4}\,(4)^4}{4\,!} = \frac{64}{3}\,e^{-4} = 0.39073.$$

Example 13. *For a Poisson distribution with mean m, show that*

$$\mu_{r+1} = mr\,\mu_{r-1} + m\,\frac{d\mu_r}{dm} \text{ where, } \mu_r = \sum_{x=0}^{\infty} (x-m)^r\,\frac{e^{-m}\,.\,m^x}{x\,!}. \quad \text{(U.P.T.U. 2007)}$$

Sol. $$\mu_r = \sum_{x=0}^{\infty} (x-m)^r\,.\frac{e^{-m}\,.\,m^x}{x\,!}$$

$$\frac{d\mu_r}{dm} = \sum_{x=0}^{\infty}\left[\frac{-\,e^{-m}}{x\,!}\,.\,m^x\,(x-m)^r + \frac{e^{-m}}{x\,!}\{xm^{x-1}\,(x-m)^r - r(x-m)^{r-1}\,.\,m^x\}\right]$$

$$\Rightarrow \quad m\frac{d\mu_r}{dm} = \sum_{x=0}^{\infty}\frac{e^{-m}}{x\,!}\,m^x\,(x-m)^{r+1} - rm\sum_{x=0}^{\infty}\frac{e^{-m}}{x\,!}\,m^x\,(x-m)^{r-1} = \mu_{r+1} - mr\,\mu_{r-1}$$

$$\Rightarrow \quad \mu_{r+1} = m\,\frac{d\mu_r}{dm} + mr\,\mu_{r-1}.$$

Example 14. *Show that in a Poisson distribution with unit mean, mean deviation about mean is* $\left(\dfrac{2}{e}\right)$ *times the standard deviation.* (G.B.T.U. 2012)

Sol. Here, $\lambda = 1$

$\therefore \quad P(X = x) = \dfrac{e^{-1}\,.\,(1)^x}{x\,!} = \dfrac{e^{-1}}{x\,!}$; $x = 0, 1, 2, \ldots$

Mean deviation about mean 1 is

$$= \sum_{x=0}^{\infty} |\,x-1\,|\,p(x) = e^{-1}\sum_{x=0}^{\infty}\frac{|\,x-1\,|}{x\,!} = e^{-1}\left[1 + \frac{1}{2!} + \frac{2}{3!} + \frac{3}{4!} + \ldots\right] \quad \ldots(1)$$

we have, $$\frac{n}{(n+1)\,!} = \frac{\overline{n+1} - 1}{(n+1)\,!} = \frac{1}{n\,!} - \frac{1}{(n+1)\,!}$$

$\therefore$ From (1), Mean deviation about mean $= e^{-1}\left[1 + \left(1 - \dfrac{1}{2!}\right) + \left(\dfrac{1}{2!} - \dfrac{1}{3!}\right) + \ldots\right]$

$$= e^{-1}\,(1+1) = \frac{2}{3} \times 1 = \frac{2}{e} \times \text{S.D.} \qquad | \text{ Since Variance = mean = 1}$$

ASSIGNMENT

1. If X is a Poisson variate such that P(X = 2) = 9P(X = 4) + 90P(X = 6), find the standard deviation.

2. If a random variable has a Poisson distribution such that P(1) = P(2), find
(*i*) mean of the distribution (*ii*) P(4).

3. Suppose that X has a Poisson distribution. If P(X = 2) = $\frac{2}{3}$ P(X = 1) find, (*i*) P(X = 0) (*ii*) P(X = 3).

4. A certain screw making machine produces on average 2 defective screws out of 100, and packs them in boxes of 500. Find the probability that a box contains 15 defective screws.

5. (*i*) The incidence of occupational disease in an industry is such that the workmen have a 10% chance of suffering from it. What is the probability that in a group of 7, five or more will suffer from it?
(*ii*) The experience shows that 4 industrial accidents occur in a plant on an average per month. Calculate the probabilities of less than 3 accidents in a certain month. Use Poisson distribution. (Given : $e^{-4} = 0.01832$). [*M.T.U.* (*MBA*) *2011*]

6. (*i*) Suppose a book of 585 pages contains 43 typographical errors. If these errors are randomly distributed throughout the book, what is the probability that 10 pages, selected at random, will be free from errors ?
(*ii*) Assume that the probability of an individual coalminer being killed in a mine accident during a year is $\frac{1}{2400}$. Use Poisson's distribution to calculate the probability that in a mine employing 200 miners there will be at least one fatal accident in a year.

7. (*i*) A manufacturer of cotter pins knows that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantee that not more than 10 pins will be defective, what is the approximate prob. that a box will fail to meet the guaranteed quality?
(*ii*) An insurance company insures 4000 people against loss of both eyes in a car accident. Based on previous data it was assumed 10 persons out of 1,00,000 will have such type of injury in car accident. What is probability that more than 2 of the insured will collect on their policy in a given year? (*M.T.U. 2013*)

8. Records show that the probability is 0.00002 that a car will have a flat tyre while driving over a certain bridge. Use Poisson distribution to find the probability that among 20,000 cars driven over the bridge, not more than one will have a flat tyre.

9. Between the hours of 2 and 4 P.M., the average no. of phone calls per minute coming into the switch board of a company is 2.5. Find the probability that during a particular minute, there will be no phone call at all. [Given : $e^{-2} = 0.13534$ and $e^{-0.5} = 0.60650$.]

10. (*i*) Fit a Poisson distribution to the following data given the number of yeast cells per square for 400 squares :

No. of cells per sq. :	0	1	2	3	4	5	6	7	8	9	10
No. of squares :	103	143	98	42	8	4	2	0	0	0	0

It is given that $e^{-1.3225} = 0.2665$.

(*ii*) Data was collected over a period of 10 years, showing number of deaths from horse kicks in each of the 200 army corps. The distribution of deaths was as follows:

No. of deaths :	0	1	2	3	4	Total
Frequency :	109	65	22	3	1	200

Fit a Poisson distribution to the data and calculate the theoretical frequencies.
[*M.T.U.* (*B. Pharma*) *2011 ; M.T.U.* (*MBA*) *2011*]

(*iii*) The following table gives the no. of days in a 50 day period during which automobile accidents occured in a city.

No. of accidents :	0	1	2	3	4
No. of days :	21	18	7	3	1

Fit a Poisson distribution to the data. (*G.B.T.U. 2011*)

11. The number of accidents in a year involving taxi drivers in a city follows a Poisson distribution with mean equal to 3. Out of 1000 taxi drivers, find approximately the number of drivers with
(*i*) no accidents (*ii*) more than 3 accidents in a year.

12. The distribution of the number of road accidents per day in a city is Poisson with mean 4. Find the number of days out of 100 days when there will be
(*i*) no accident (*ii*) at least 2 accidents
(*iii*) atmost 3 accidents (*iv*) between 2 and 5 accidents.

13. It is given that 2% of the electric bulbs manufactured by a company are defective. Using Poisson distribution, find the probability that a sample of 200 bulbs will contain (*i*) no defective bulb (*ii*) two defective bulbs (*iii*) at the most three defective bulbs. (*G.B.T.U. 2011*)

14. A manager accepts the work submitted by his typist only when there is no mistake in the work. The typist has to type on an average 20 letters per day of about 200 words each. Find the chance of her making a mistake if
(*i*) less than 1% of the letters submitted by her are rejected
(*ii*) on 90% days all the letters submitted by her are accepted.
As the probability of making a mistake is small, you may use Poisson distribution. (Take $e = 2.72$)
[**Hint:** (*i*) $e^{-200p} \geq 0.99$ (*ii*) $e^{-4000p} = 0.90$]

15. In a certain factory manufacturing razor blades, there is a small chance of 0.002 for any blade to be defective. The blades are supplied in packets of 10. Use suitable distribution to calculate the approximate number of packets containing no defective, one defective and two defective blades respectively in a consignment of 20,000 packets. (*M.T.U. 2014*)

Answers

1. 1 **2.** (*i*) 2 (*ii*) $\frac{2}{3e^2}$ **3.** (*i*) $e^{-(4/3)}$ (*ii*) $\frac{32}{81}e^{-(4/3)}$ **4.** 0.035

5. (*i*) 0.0008 (*ii*) 0.23816 **6.** (*i*) 0.4795 (*ii*) 0.08 **7.** (*i*) 0.0136875 (*ii*) 0.007926

8. 0.938 **9.** 0.08208

10. (*i*) Theoretical frequencies are 107, 141, 93, 41, 13, 4, 1, 0, 0, 0, 0.
(*ii*) $108.67 \times \frac{(0.61)^r}{r!}$; 109, 66, 20, 4 and 1 (*iii*) $(50)\frac{e^{-0.9}(0.9)^r}{r!}$; 20, 18, 8, 3 and 1.

11. (*i*) 50 (*ii*) 353

12. (*i*) 2 (*ii*) 91 (*iii*) 43 (*iv*) 39

13. (*i*) 0.018315 (*ii*) 0.146525 (*iii*) 0.43347

14. (*i*) $p \leq 0.0000506$ (*ii*) $p = 0.0000263$ **15.** 19604, 392, 4

3.61 NORMAL DISTRIBUTION [*U.P.T.U. 2007*]

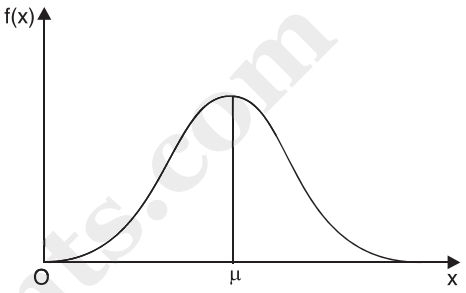
The normal distribution is a continuous distribution. It can be derived from the binomial distribution in the limiting case when n, the number of trials is very large and p, the probability of a success, is close to $\frac{1}{2}$. The general equation of the normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where the variable x can assume all values from $-\infty$ to $+\infty$. μ and σ, called the parameters of the distribution, are respectively the mean and the standard deviation of the distribution and $-\infty < \mu < \infty$, $\sigma > 0$. x is called the normal variate and $f(x)$ is called probability density function of the normal distribution.

If a variable x has the normal distribution with mean μ and standard deviation σ, we briefly write $x : N(\mu, \sigma^2)$.

The graph of the normal distribution is called the *normal curve*. It is bell-shaped and symmetrical about the mean μ. The two tails of the curve extend to $+\infty$ and $-\infty$ towards the positive and negative directions of the x-axis respectively and gradually approach the x-axis without ever metting it. The curve is unimodal and the mode of the normal distribution coincides with its mean μ. The line $x = \mu$ divides the area under the normal curve above x-axis into two equal parts. Thus, the median of the distribution also coincides with its mean and mode. The area under the normal curve between any two given ordinates $x = x_1$ and $x = x_2$ represents the probability of values falling into the given interval. The total area under the normal curve above the x-axis is 1.



3.62 BASIC PROPERTIES OF THE NORMAL DISTRIBUTION

The probability density function of the normal distribution is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

(*i*) $f(x) \geq 0$ (*ii*) $\int_{-\infty}^{\infty} f(x)\, dx = 1$,

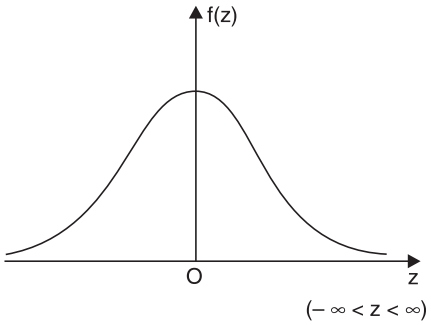
i.e., the total area under the normal curve above the x-axis is 1.

(*iii*) The normal distribution is symmetrical about its mean.

(*iv*) It is a unimodal distribution. The mean, mode and median of this distribution coincide.

3.63 STANDARD FORM OF THE NORMAL DISTRIBUTION

If X is a normal random variable with mean μ and standard deviation σ, then the random variable $Z = \frac{X - \mu}{\sigma}$ has the normal distribution with mean 0 and standard deviation 1. The random variable Z is called the *standardized* (or *standard*) normal random variable.



The probability density function for the normal distribution in standard form is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

It is free from any parameter. This helps us to compute areas under the normal probability curve by making use of standard tables.

Note 1. If $f(z)$ is the probability density function for the normal distribution, then

$$P(z_1 \le Z \le z_2) = \int_{z_1}^{z_2} f(z)\,dz = F(z_2) - F(z_1), \quad \text{where} \quad F(z) = \int_{-\infty}^{z} f(z)\,dz = P(Z \le z)$$

The function $F(z)$ defined above is called the *distribution function* for the normal distribution.

Note 2. The probabilities $P(z_1 \le Z \le z_2)$, $P(z_1 < Z \le z_2)$, $P(z_1 \le Z < z_2)$ and $P(z_1 < Z < z_2)$ are all regarded to be the same.

Note 3. $F(-z_1) = 1 - F(z_1)$.

3.64 NORMAL DISTRIBUTION AS A LIMITING FORM OF BINOMIAL DISTRIBUTION (when $p = q$)

(U.P.T.U. 2007)

Let N $(q + p)^n$ be the binomial distribution. If $p = q$ then $p = q = \frac{1}{2}$ (since $p + q = 1$) and consequently the binomial distribution is symmetrical. Let n be an even integer say $2k$, k being an integer. Since $n \to \infty$, the frequencies of r and $r + 1$ successes can be written in following forms:

$$f(r) = N.\ {}^{2k}C_r \left(\frac{1}{2}\right)^{2k}$$

$$f(r+1) = N.\ {}^{2k}C_{r+1} \left(\frac{1}{2}\right)^{2k}$$

$$\therefore \quad \frac{f(r+1)}{f(r)} = \frac{{}^{2k}C_{r+1}}{{}^{2k}C_r} = \frac{2k-r}{r+1}$$

The frequency of r successes will be greater than the frequency of $(r + 1)$ successes if

$$f(r) > f(r+1)$$

$$\Rightarrow \quad \frac{f(r+1)}{f(r)} < 1$$

$$\Rightarrow \quad 2k - r < r + 1$$

$$\Rightarrow \quad r > k - \frac{1}{2} \qquad \ldots(1)$$

In a similar way, the frequency of r successes will be greater than the frequencies of $(r - 1)$ successes if $r < k + \frac{1}{2}$ ...(2)

In view of (1) and (2), we observe that if $k - \frac{1}{2} < r < k + \frac{1}{2}$ the frequency corresponding to r successes will be the greatest. Clearly, $r = k$ is the value of the success corresponding to which the frequency is maximum. Suppose it is y_0. Then, we have

$$y_0 = N.\ {}^{2k}C_k \left(\frac{1}{2}\right)^{2k} = N.\ \frac{2k!}{k!\,k!}\left(\frac{1}{2}\right)^{2k}$$

Let y_x be the frequency of $k + x$ successes then, we have

$$y_x = \mathrm{N} \cdot {}^{2k}\mathrm{C}_{k+x} \left(\frac{1}{2}\right)^{2k} = \mathrm{N} \cdot \left(\frac{1}{2}\right)^{2k} \cdot \frac{2k!}{(k+x)!\,(k-x)!}$$

Now,
$$\frac{y_x}{y_0} = \frac{k!\,k!}{(k+x)!\,(k-x)!} = \frac{k\,(k-1)\,(k-2)\ldots(k-x+1)}{(k+x)\,(k+x-1)\ldots(k+1)}$$

$$= \frac{\left(1-\frac{1}{k}\right)\left(1-\frac{2}{k}\right)\ldots\left\{1-\frac{x-1}{k}\right\}}{\left(1+\frac{1}{k}\right)\left(1+\frac{2}{k}\right)\ldots\left(1+\frac{x}{k}\right)}$$

Taking log on both sides,

$$\log \frac{y_x}{y_0} = \left[\log\left(1-\frac{1}{k}\right) + \log\left(1-\frac{2}{k}\right) + \ldots + \log\left(1-\frac{x-1}{k}\right)\right]$$

$$-\left[\log\left(1+\frac{1}{k}\right) + \log\left(1+\frac{2}{k}\right) + \ldots + \log\left(1+\frac{x}{k}\right)\right] \quad \ldots(3)$$

Now, writing expression for each term and neglecting higher powers of $\frac{x}{k}$ (very small quantity), we get from (3),

$$\log \frac{y_x}{y_0} = -\frac{1}{k}\{1+2+3+\ldots+(x-1)\} - \frac{1}{k}\{1+2+3+\ldots+(x-1)+x\}$$

$$= -\frac{2}{k}\{1+2+3+\ldots+(x-1)\} - \frac{x}{k}$$

$$= -\frac{2}{k}\frac{(x-1)\,x}{2} - \frac{x}{k} = -\frac{x^2}{k}$$

$\therefore$ $$y_x = y_0\, e^{-x^2/k}$$

$\Rightarrow$ $$\boxed{y_x = y_0\, e^{-x^2/2\sigma^2}} \quad | \because \quad \sigma^2 = npq = \frac{n}{4} = \frac{k}{2}$$

which is **normal distribution.**

3.65 MEAN AND VARIANCE OF NORMAL DISTRIBUTION *(U.P.T.U. 2015)*

1. The A.M. of a continuous distribution $f(x)$ is given by

$$\text{A.M. } (\bar{x}) = \frac{\int_{-\infty}^{\infty} x\, f(x)\, dx}{\int_{-\infty}^{\infty} f(x)\, dx} \quad | \text{ By definition}$$

Consider the normal distribution with μ, σ as the parameters then

$$\bar{x} = \int_{-\infty}^{\infty} x.\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Since $\int_{-\infty}^{\infty} f(x)\,dx$ = area under normal curve = 1

Put $\frac{x-\mu}{\sigma} = z$ so that $x = \mu + \sigma z \quad \therefore \quad dx = \sigma dz$

so,
$$\bar{x} = \int_{-\infty}^{\infty} (\mu + \sigma z)\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} (\sigma\, dz)$$

$$= \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\, e^{-\frac{1}{2}z^2}\, dz$$

$$= \mu + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}\, d\left(\frac{z^2}{2}\right) \qquad \left| \because \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz = 1 \right.$$

$$= \mu + \frac{\sigma}{\sqrt{2\pi}} \left(\frac{e^{-z^2/2}}{-1}\right)_{-\infty}^{\infty}$$

$$\boxed{\bar{x} = \mu}$$

2. By definition,

$$\text{Variance} = \int_{-\infty}^{\infty} (x - \bar{x})^2 f(x)\, dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\, dx + \bar{x}^2 \int_{-\infty}^{\infty} f(x)\, dx - 2\bar{x} \int_{-\infty}^{\infty} x f(x)\, dx$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\, dx + \bar{x}^2 - 2\bar{x}\,\bar{x} \qquad \left| \because \int_{-\infty}^{\infty} f(x)\, dx = 1 \text{ and } \int_{-\infty}^{\infty} x f(x)\, dx = \bar{x} \right.$$

$$= \int_{-\infty}^{\infty} x^2 f(x)\, dx - \bar{x}^2 \qquad \ldots(1)$$

Now, Let $I = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2} dx$

Put $\frac{x-\bar{x}}{\sigma} = z$ so that $x = \bar{x} + \sigma z \;\therefore \quad dx = \sigma\, dz$

Hence,
$$I = \int_{-\infty}^{\infty} (\bar{x} + \sigma z)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2}\, \sigma\, dz$$

$$= \frac{1}{\sqrt{2\pi}} \left[\sigma^2 \int_{-\infty}^{\infty} z^2 e^{-z^2/2}\, dz + \bar{x}^2 \int_{-\infty}^{\infty} e^{-z^2/2}\, dz + 2\sigma\bar{x} \int_{-\infty}^{\infty} z\, e^{-z^2/2}\, dz\right]$$

$$= \frac{-\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\, d(e^{-z^2/2}) + \bar{x}^2 \,.\, 1 + 2\sigma\bar{x} \,.\, 0$$

$$= -\frac{\sigma^2}{\sqrt{2\pi}} \left(z\, e^{-z^2/2} \right)_{-\infty}^{\infty} + \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}\, dz + \bar{x}^2$$

$$= 0 + \sigma^2 \,.\, 1 + \bar{x}^2 = \sigma^2 + \bar{x}^2$$

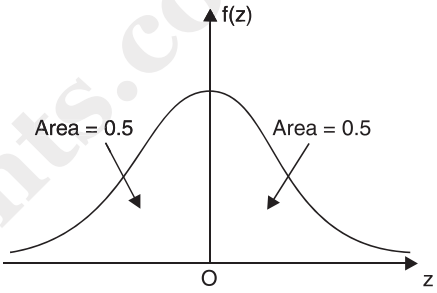
$\therefore$ From (1), $\boxed{\text{Variance} = \sigma^2 + \bar{x}^2 - \bar{x}^2 = \sigma^2}$

$\therefore$ The standard deviation of the normal distribution is σ.

3.66 AREA UNDER THE NORMAL CURVE

By taking $z = \dfrac{x - \mu}{\sigma}$, standard normal curve is formed. The total area under this curve is 1.

The area under the curve is divided into two equal parts by $z = 0$. The area between the ordinate $z = 0$ and any other ordinate can be noted from the supplied table. It should be noted that mean for the normal distribution is 0.



3.67 APPLICATIONS OF NORMAL DISTRIBUTION

De Moivre made the discovery of this distribution in 1733.

This distribution has an important application in the theory of errors made by chance in experimental measurements. Its more applications are in computation of hit probability of a shot and in statistical inference in almost every branch of science.

EXAMPLES

Example 1. *A sample of 100 dry battery cells tested to find the length of life produced the following results:*

$$\bar{x} = 12 \text{ hours}, \sigma = 3 \text{ hours}.$$

Assuming the data to be normally distributed, what percentage of battery cells are expected to have life

(*i*) *more than 15 hours* (*ii*) *less than 6 hours* (*iii*) *between 10 and 14 hours?*

Sol. Here x denotes the length of life of dry battery cells.

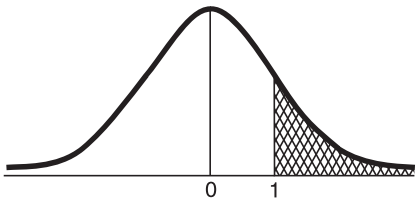
Also $$z = \frac{x - \bar{x}}{\sigma} = \frac{x - 12}{3}.$$

(*i*) When $x = 15$, $z = 1$

$\therefore$ $P(x > 15) = P(z > 1)$

$= P(0 < z < \infty) - P(0 < z < 1)$

$= .5 - 0.3413 = 0.1587 = 15.87\%$.

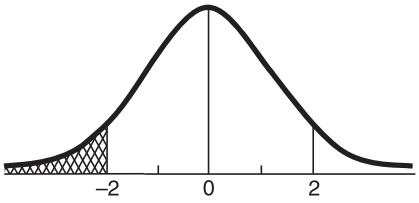


(*ii*) When $x = 6$, $z = -2$

$\therefore$ $P(x < 6) = P(z < -2)$

$= P(z > 2) = P(0 < z < \infty) - P(0 < z < 2)$

$= 0.5 - 0.4772 = 0.0228 = 2.28\%$.



(*iii*) When $x = 10$, $z = -\frac{2}{3} = -0.67$

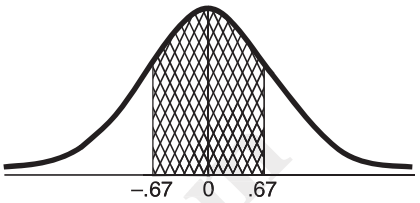
When $x = 14$, $z = \frac{2}{3} = 0.67$

$P(10 < x < 14)$

$= P(-0.67 < z < 0.67)$

$= 2P(0 < z < 0.67) = 2 \times 0.2485$

$= 0.4970 = 49.70\%$.



Example 2. *In a sample of 1000 cases, the mean of a certain test is 14 and S.D. is 2.5. Assuming the distribution to be normal, find*

(*i*) *how many students score between 12 and 15?*

(*ii*) *how many score above 18?*

(*iii*) *how many score below 8?*

(*iv*) *how many score 16?*

Sol. (*i*) $z_1 = \frac{x_1 - \mu}{\sigma} = \frac{12 - 14}{2.5} = -0.8$

$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{15 - 14}{2.5} = 0.4$

Area lying between – 0.8 and 0.4

= Area between 0 to 0.8 + Area between 0 to 0.4

= 0.2881 + 0.1554 = 0.4435

Reqd. no. of students = 1000 × 0.4435 = 444 (app.)

(*ii*) $z = \frac{18 - 14}{2.5} = 1.6$

Area right to 1.6 = 0.5 – (Area between 0 and 1.6) = 0.5 – 0.4452 = 0.0548

Reqd. no. of students = 1000 × 0.0548 = 54.8 ≈ 55 (app.)

(*iii*) $z = \frac{8 - 14}{2.5} = -2.4$

Area left to – 2.4 = 0.5 – (Area between 0 and 2.4) = 0.5 – 0.4918 = 0.0082

$\therefore$ Reqd. no. of students = 1000 × 0.0082 = 8.2 ≈ 8 (app.).

(*iv*) $z_1 = \frac{15.5 - 14}{2.5} = 0.6$

$z_2 = \frac{16.5 - 14}{2.5} = 1$

Area between 0.6 and 1 = 0.3413 – 0.2257 = 0.1156

$\therefore$ Reqd. no. of students = 1000 × 0.1156 = 115.6 ≈ 116 (app.).

Example 3. *Assume mean height of soldiers to be 68.22 inches with a variance of 10.8 inches square. How many soldiers in a regiment of 1,000 would you expect to be over 6 feet tall, given that the area under the standard normal curve between z = 0 and z = 0.35 is 0.1368 and between z = 0 and z = 1.15 is 0.3746.* [G.B.T.U. (C.O.) 2011]

Sol. x = 6 feets = 72 inches

$$\therefore \quad z = \frac{x - \mu}{\sigma} = \frac{72 - 68.22}{\sqrt{10.8}} = 1.15$$

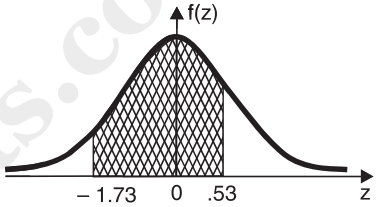
$$P(x > 72) = P(z > 1.15) = 0.5 - P(0 \le z \le 1.15)$$
$$= 0.5 - 0.3746 = 0.1254$$

$\therefore$ Expected no. of soldiers = $1000 \times 0.1254 = 125.4 \approx 125$ (app.).

Example 4. *A large number of measurement is normally distributed with a mean 65.5″ and S.D. of 6.2″. Find the percentage of measurements that fall between 54.8″ and 68.8″.*

Sol. Mean μ = 65.5 inches, S.D. σ = 6.2 inches

x_1 = 54.8 inches, x_2 = 68.8 inches



$$\therefore \quad z_1 = \frac{x_1 - \mu}{\sigma} = \frac{54.8 - 65.5}{6.2} = -1.73$$

and $$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{68.8 - 65.5}{6.2} = 0.53$$

Now, $P(-1.73 \le z \le 0.53) = P(-1.73 \le z \le 0) + P(0 \le z \le 0.53)$

$= P(0 \le z \le 1.73) + P(0 \le z \le 0.53)$

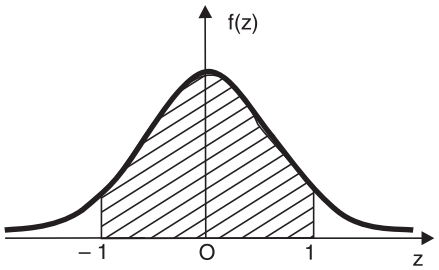
$= 0.4582 + 0.2019 = 0.6601$ | By table

$\therefore$ Reqd. percentage of measurements = 66.01%.

Example 5. *A manufacturer knows from experience that the resistance of resistors he produces is normal with mean μ = 100 ohms and standard deviation σ = 2 ohms. What percentage of resistors will have resistance between 98 ohms and 102 ohms?*

Sol. $\mu = 100\ \Omega$, $\sigma = 2\ \Omega$, $x_1 = 98\ \Omega$, $x_2 = 102\ \Omega$

$$\therefore \quad z_1 = \frac{x_1 - \mu}{\sigma} = \frac{98 - 100}{2} = -1$$



and $$z_2 = \frac{x_2 - \mu}{\sigma} = \frac{102 - 100}{2} = 1.$$

Now, $P(98 < x < 102) = P(-1 < z < 1)$

$= P(-1 \le z \le 0) + P(0 \le z \le 1)$

$= P(0 \le z \le 1) + P(0 \le z \le 1)$

$= 0.3413 + 0.3413 = 0.6826.$

$\therefore$ Percentage of resistors having resistance between 98 Ω and 102 Ω = 68.26%.

Example 6. *In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and standard deviation of the distribution. It is given that if* $f(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-\frac{1}{2}x^2}\, dx$ *then f(0.5) = 0.19 and f(1.4) = 0.42.* (M.T.U. 2013)

Sol. Let μ and σ be the mean and S.D. respectively.

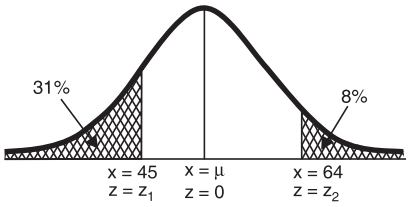
31% of the items are under 45.

$\Rightarrow$ Area to the left of the ordinate $x = 45$ is 0.31

When $x = 45$, let $z = z_1$

$P(z_1 < z < 0) = 0.5 - 0.31 = 0.19$

From the tables, the value of z corresponding to this area is 0.5



$\therefore \quad z_1 = -0.5 [z_1 < 0]$

When $x = 64$, let $z = z_2$

$P(0 < z < z_2) = 0.5 - 0.08 = 0.42$

From the tables, the value of z corresponding to this area is 1.4.

$\therefore \quad z_2 = 1.4$

Since $z = \dfrac{x - \mu}{\sigma}$

$\therefore \quad -0.5 = \dfrac{45 - \mu}{\sigma}$ and $1.4 = \dfrac{64 - \mu}{\sigma}$

$\Rightarrow \quad 45 - \mu = -0.5\sigma$ ...(1)

and $64 - \mu = 1.4\sigma$ ...(2)

Subtracting $-19 = -1.9\sigma \quad \therefore \quad \sigma = 10$

From (1), $45 - \mu = -0.5 \times 10 = -5 \quad \therefore \quad \mu = 50.$

Example 7. *The life of army shoes is normally distributed with mean 8 months and standard deviation 2 months. If 5000 pairs are insured, how many pairs would be expected to need replacement after 12 months?* $\left[\text{Given that } P(z \geq 2) = 0.0228 \text{ and } z = \dfrac{x - \mu}{\sigma}\right]$.

Sol. Mean $(\mu) = 8$, Standard Deviation $(\sigma) = 2$

Number of pairs of shoes = 5000, Total months $(x) = 12$

when $x = 12$, $z = \dfrac{x - \mu}{\sigma} = \dfrac{12 - 8}{2} = 2$

Area $(z \geq 2) = 0.0228$

Number of pairs whose life is more than 12 months = $5000 \times 0.0228 = 114$

Pair of shoes needing replacement after 12 months = $5000 - 114 = 4886$.

Example 8. *The mean inside diameter of a sample of 200 washers produced by a machine is 0.502 cm and the standard deviation is 0.005 cm. The purpose for which these washers are intended allows a minimum tolerance in the diameter of 0.496 to 0.508 cm, otherwise the washers are considered defective. Determine the percentage of defective washers produced by the machine. Assume the diameters are normally distributed.*

Sol. Given: Mean $\mu = 0.502$ cm, S.D. $\sigma = 0.005$ cm, $x_1 = 0.496$ cm, $x_2 = 0.508$ cm.

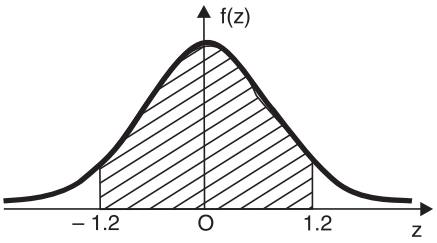
Now, $z_1 = \dfrac{x_1 - \mu}{\sigma} = \dfrac{0.496 - 0.502}{0.005} = -1.2$

$z_2 = \dfrac{x_2 - \mu}{\sigma} = \dfrac{0.508 - 0.502}{0.005} = 1.2$

Area for non-defective washers

$$= P(-1.2 \le z \le 1.2)$$
$$= P(-1.2 \le z \le 0) + P(0 \le z \le 1.2)$$
$$= P(0 \le z \le 1.2) + P(0 \le z \le 1.2)$$
$$= 0.3849 + 0.3849 = 0.7698$$
$$= 76.98\%.$$

$\therefore$ Percentage of defective washers $= 100 - 76.98$
$= 23.02\%$.



Example 9. *Assuming that the diameters of 1000 brass plugs taken consecutively from a machine, form a normal distribution with mean 0.7515 cm and standard deviation 0.002 cm, how many of the plugs are likely to be rejected if the approved diameter is 0.752 ± 0.004 cm.*

Sol. Tolerance limits of the diameter of non-defective plugs are

$0.752 - 0.004 = 0.748$ cm. and $0.752 + 0.004 = 0.756$ cm.

Standard normal variable, $z = \dfrac{x - \mu}{\sigma}$

If $x_1 = 0.748$, $z_1 = \dfrac{0.748 - 0.7515}{0.002} = -1.75$

If $x_2 = 0.756$, $z_2 = \dfrac{0.756 - 0.7515}{0.02} = 2.25$

Area from $(z_1 = -1.75)$ to $(z_2 = 2.25)$

$$= P(-1.75 \le z \le 2.25) = P(-1.75 \le z \le 0) + P(0 \le z \le 2.25)$$
$$= P(0 \le z \le 1.75) + P(0 \le z \le 2.25) = 0.4599 + 0.4878 = 0.9477$$

Number of plugs which are likely to be rejected $= 1000 \times (1 - 0.9477) = 1000 \times .0523 = 52.3$

Hence approximately 52 plugs are likely to be rejected.

Example 10. *If the heights of 300 students are normally distributed with mean 64.5 inches and standard deviation 3.3 inches, find the height below which 99% of the students lie.*

Sol. Mean $\mu = 64.5$ inches, S.D. $\sigma = 3.3$ inches

Area between 0 and $\dfrac{x - 64.5}{3.3} = 0.99 - 0.5 = 0.49$

From the table, for the area 0.49, $z = 2.327$

The corresponding value of x is given by

$$\frac{x - 64.5}{3.3} = 2.327$$

$\Rightarrow$ $x - 64.5 = 7.68$

$\Rightarrow$ $x = 7.68 + 64.5 = 72.18$ inches.

Hence 99% students are of height less than 6 ft. 0.18 inches.

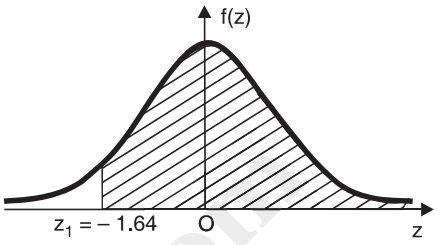
Example 11. *The income of a group of 10,000 persons was found to be normally distributed with mean ₹ 750 p.m. and standard deviation of ₹ 50. Show that, of this group, about 95% had income exceeding ₹ 668 and only 5% had income exceeding ₹ 832. Also find the lowest income among the richest 100.*

Sol. Given: $\mu = 750$, $\sigma = 50$

Standard normal variable, $z = \dfrac{x - \mu}{\sigma}$

(*i*) If $x_1 = 668$, $z_1 = \dfrac{x_1 - \mu}{\sigma} = \dfrac{668 - 750}{50} = -1.64$

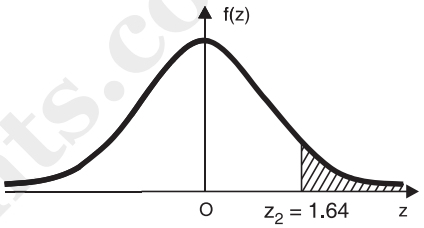
$$P(x_1 > 668) = P(z_1 > -1.64)$$
$$= 0.5 + P(-1.64 \le z \le 0)$$
$$= 0.5 + P(0 \le z \le 1.64)$$
$$= 0.5 + 0.4495$$
$$= 0.9495$$



∴ Required percentage of persons having income exceeding ₹ 668 = 94.95% ≈ 95% (approx.)

(*ii*) If $x_2 = 832$, $z_2 = \dfrac{x_2 - \mu}{\sigma} = \dfrac{832 - 750}{50} = 1.64$

$$P(x_2 > 832) = P(z_2 > 1.64)$$
$$= 0.5 - P(0 \le z \le 1.64)$$
$$= 0.5 - 0.4495 = 0.0505$$



∴ Required percentage of persons having income exceeding ₹ 832 = 5.05% ≈ 5% (approx.)

(*iii*) Let x be the lowest income among the richest 100 persons *i.e.,* 1% of 10,000.

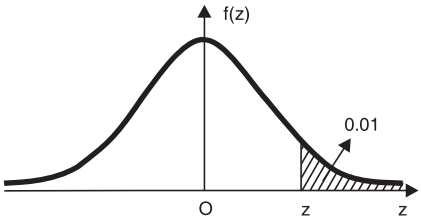
Thus, area between O and $z = 0.49$ (see figure) by Normal distribution table,

$$z = 2.33$$

Thus, $\dfrac{x - \mu}{\sigma} = 2.33$

$\Rightarrow \dfrac{x - 750}{50} = 2.33$

$\Rightarrow x = 866.5$



Hence ₹ 866.5 is the minimum income among the richest 100 persons.

Example 12. *255 metal rods were cut roughly 6 inches over size. Finally the lengths of the over size amount, were measured exactly and grouped with 1 inch intervals, there being in all 12 groups* $\dfrac{1''}{2} - 1\dfrac{1''}{2}$, $1\dfrac{1''}{2} - 2\dfrac{1''}{2}$, ..., $11\dfrac{1''}{2} - 12\dfrac{1''}{2}$.

The frequency distribution for the 255 lengths was as follows:

Length (inches) Central value	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Frequency	*2*	*10*	*19*	*25*	*40*	*44*	*41*	*28*	*25*	*15*	*5*	*1*

Fit a normal curve to this data.

Sol. The equation of the normal curve for N observations is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2} \quad ...(1)$$

x	f	$u = x - 6$	fu	fu^2
1	2	– 5	– 10	50
2	10	– 4	– 40	160
3	19	– 3	– 57	171
4	25	– 2	– 50	100
5	40	– 1	– 40	40
6	44	0	0	0
7	41	1	41	41
8	28	2	56	112
9	25	3	75	225
10	15	4	60	240
11	5	5	25	125
12	1	6	6	36
Total	255		66	1300

Mean, $$\mu = a + \frac{\Sigma fu}{\Sigma f} = 6 + \frac{66}{255} = 6.259$$

Variance, $$\sigma^2 = \frac{\Sigma fu^2}{\Sigma f} - \left(\frac{\Sigma fu}{\Sigma f}\right)^2 = \frac{1300}{225} - \left(\frac{66}{255}\right)^2 = 5.031$$

$\therefore$ $\sigma = 2.243$

Thus, we have $N = 255$, Mean, $\mu = 6.259''$, S.D. $\sigma = 2.243''$

Hence the fitted curve is

$$y = \frac{255}{2.243\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x - 6.259}{2.243}\right)^2} \quad | \text{ From (1)}$$

$$= \frac{113.68}{\sqrt{2\pi}} e^{-0.099(x-6.259)^2}.$$

Example 13. *Show that the area under the normal curve is unity.*

Sol. Area under the normal curve is given by

$$A = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Put $\frac{x - \mu}{\sigma} = z$ so $dx = \sigma\, dz$

$\therefore$ $$A = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2} (\sigma dz) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-z^2/2}\, dz$$

Now, $$A \,.\, A = A^2 = \left(\sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-x^2/2}\, dx\right)\left(\sqrt{\frac{2}{\pi}} \int_0^{\infty} e^{-y^2/2}\, dy\right)$$

$$= \frac{2}{\pi} \int_0^{\infty}\int_0^{\infty} e^{-\left(\frac{x^2 + y^2}{2}\right)} dx\, dy \quad | \text{ where } x \text{ and } y \text{ are dummy variables}$$

Put $x = r\cos\theta$, $y = r\sin\theta$ so that $J = r$ changing to polar coordinates,

$$A^2 = \frac{2}{\pi}\int_0^{\pi/2}\int_0^{\infty} e^{-r^2/2}\, r\, dr\, d\theta = \int_0^{\infty} e^{-r^2/2}\, d\left(\frac{r^2}{2}\right) = 1$$

$\therefore$ A = Area under the normal curve = 1

Example 14. *Prove that for normal distribution, the mean deviation from the mean equals to* $\frac{4}{5}$ *of the standard deviation approximately.* (U.P.T.U. 2009)

Sol. Let μ and σ be the mean and standard deviation of the normal distribution. Then by definition,

Mean deviation from the mean

$$= \int_{-\infty}^{\infty} |x-\mu|\, f(x)\, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} |x-\mu|\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \sigma|z|\, e^{-\frac{1}{2}z^2}\, \sigma\, dz \qquad \left|\begin{array}{l}\text{when } \dfrac{x-\mu}{\sigma} = z \\ \Rightarrow\ dx = \sigma\, dz\end{array}\right.$$

$$= \sigma\sqrt{\frac{2}{\pi}}\int_0^{\infty} z\, e^{-z^2/2}\, dz$$

$$= \sigma\sqrt{\frac{2}{\pi}}\left[-e^{-z^2/2}\right]_0^{\infty} = \sqrt{\frac{2}{\pi}}\,\sigma = 0.7979\,\sigma \approx 0.8\sigma \approx \frac{4}{5}\sigma$$

ASSIGNMENT

1. In a test on 2000 electric bulbs, it was found that the life of a particular make, was normally distributed with an average life of 2040 hours and S.D. of 60 hours, estimate the number of bulbs likely to burn for

(*i*) more than 2150 hours (*ii*) less than 1950 hours

(*iii*) more than 1920 hours but less than 2160 hours. (*U.P.T.U. 2008*)

2. An aptitude test for selecting officers in a bank is conducted on 1000 candidates. The average score is 42 and the standard deviation of score is 24. Assuming normal distribution for the scores, find

(*i*) the number of candidates whose scores exceed 60

(*ii*) the number of candidates whose scores lie between 30 and 60.

3. (*i*) In a normal distribution exactly normal, 7% of the items are under 35 and 89% are under 63. What are the mean and standard deviation of the distribution? (*G.B.T.U. 2010*)

(*ii*) In a normal distribution, 0.0107 of the items lie below 42 and 0.0446 of the items lie above 82. What is the mean and standard deviation of the normal distribution?

[*U.P.T.U. (MBA) 2009*]

4. If Z is a standard normal variable, find the following probabilities: [*G.B.T.U. (MBA) 2010*]

(*i*) $P(Z < 1.2)$ (*ii*) $P(Z > -1.2)$ (*iii*) $P(-1.2 < Z < 1.3)$.

5. An aptitude test was conducted on 900 employees of the Metro Tyres Limited in which the mean score was found to be 50 units and standard deviation was 20. On the basis of this information, you are required to answer the following questions:

(*i*) What was the number of employees whose mean score was less than 30?

(*ii*) What was the number of employees whose mean score exceeded 70?

(*iii*) What was the number of employees whose mean score were between 30 and 70?

$\frac{x-\mu}{\sigma}$	0.25	0.50	0.70	1.00	1.25	1.50
Area	0.0987	0.1915	0.2734	0.3413	0.3944	0.4332

[*U.P.T.U. (MBA) 2009*]

6. (*a*) Students of a class were given a mechanical aptitude test. Their marks were found to be normally distributed with mean 60 and standard deviation 5. What percent of students scored?

(*i*) more than 60 marks? (*ii*) less than 56 marks? (*iii*) between 45 and 65 marks?

(*b*) 2000 students appeared in an examination. Distribution of marks is assumed to be normal with mean $\mu = 30$ and $\sigma = 6.25$. How many students are expected to get marks?

(*i*) between 20 and 40 (*ii*) less than 35 and (*iii*) above 50.

[*U.P.T.U. (MBA) 2012*]

(*c*) Suppose the weight W of 600 male students are normally distributed with mean $\mu = 70$ kg and standard deviation $\sigma = 5$ kg. Find number of students with weight

(*i*) between 69 and 74 kg (*ii*) more than 76 kg. (*G.B.T.U. 2013*)

7. (*a*) In an intelligence test administered to 1000 students, the average score was 42 and standard deviation 24. Find:

(*i*) the expected number of students scoring more than 50.

(*ii*) the number of students scoring between 30 and 54.

(*iii*) the value of score exceeded by top 100 students. [*G.B.T.U. (MBA) 2010*]

(*b*) The average monthly sales of 5000 firms are normally distributed. Its mean and standard deviation are ₹ 36000 and ₹ 10000 respectively. Find:

(*i*) the no. of firms having sales over ₹ 40000.

(*ii*) the no. of firms having sales between ₹ 30000 and ₹ 40000.

[Given area under normal curve from 0 to z for Z (0.4) = 0.1554 and Z (0.6) = 0.2257]

[*G.B.T.U. (MBA) 2010*]

(*c*) The daily wages of 1000 workers are distributed around a mean of ₹ 140 and with a standard deviation of ₹ 10. Estimate the number of workers whose daily wages will be

(*i*) between ₹ 140 and ₹ 144 (*ii*) less than ₹ 126

(*iii*) more than ₹ 160. (*G.B.T.U. 2012*)

8. (*a*) Records kept by the goods inwards department of a large factory show that the average no. of lorries arriving each week is 248. It is known that the distribution approximates to be normal with a standard deviation of 26.

If this pattern of arrival continues, what percentage of weeks can be expected to have number of arrivals of:

(*i*) less than 229 per week? (*ii*) more than 280 per week?

(*b*) Pipes for tobacco are being packed in fancy plastic boxes. The length of the pipe is normally distributed with $\mu = 5''$ and $\sigma = 0.1''$. The internal length of the boxes is 5.2″. What is the probability that the box would be small for the pipe?

[Given that : $\phi(1.8) = 0.9641$, $\phi(2) = 0.9772$, $\phi(2.5) = 0.9938$]

(*c*) A manufacturer of envelopes knows that the weight of the envelopes is normally distributed with mean 1.9 gm and variance 0.01 square gm. Find how many envelopes weighing

(*i*) 2 gm or more

(*ii*) 2.1 gm or more, can be expected in a given packet of 1000 envelopes? (*U.P.T.U. 2015*)

[Given: if t is the normal variable, then $\phi(0 \le t \le 1) = 0.3413$ and $\phi(0 \le t \le 2) = 0.4772$]

9. (*a*) East-East Airlines has the policy of employing only Indian woman whose height is between 62 inches and 69 inches. If the height of Indian women is approximately normally distributed with a mean of 64 inches and a standard deviation of 3 inches, Out of the 1000 applications received find the number of applicants that would be

(*i*) too tall (*ii*) too short (*iii*) of acceptable height.

(*b*) The mean height of 500 students is 151 cm and the standard deviation is 15 cm. Assuming that the heights are normally distributed, find how many students have heights between 120 and 155 cm?

(*c*) The monthly mess bill of a student who is staying in the hostel follows a normal distribution with a mean of ₹ 2000 and a standard deviation of ₹ 185. What is the probability that in the next month, his bill will go above ₹ 2400? [*U.P.T.U. (MBA) 2009*]

10. In an examination taken by 500 candidates, the average and the standard deviation of marks obtained (normally distributed) are 40% and 10%. Find approximately

(*i*) how many will pass, if 50% is fixed as a minimum?

(*ii*) what should be the minimum if 350 candidates are to pass?

(*iii*) how many have scored marks above 60%?

11. The income distribution of workers in a certain factory was found to be normal with mean ₹ 500 and standard deviation equal to ₹ 50. There were 228 persons getting above ₹ 600. How many workers were there in all?

(Area under the standard normal curve between height at 0 and 2 is 0.4772).

12. In a certain examination, the percentage of passes and distinction were 46 and 9 respectively. Estimate the average marks obtained by the candidates, the minimum pass and distinction marks being 40 and 75 respectively. Assume the distribution of marks to be normal.

$$\left[\textbf{Hint. } \frac{40-\mu}{\sigma} = 0.1 \quad \text{and} \quad \frac{75-\mu}{\sigma} = 1.34.\right]$$

13. The marks obtained by a no. of students for a certain subject are assumed to be approximately normally distributed with mean value 65 and S.D. of 5. If 3 students are taken at random from this set, what is the probability that exactly 2 of them will have marks over 70?

14. In an examination, it is laid down that a student passes if he secures 30% or more marks. He is placed in the first, second or third division according as he secures 60% or more marks, between 45% and 60% marks and marks between 30% and 45% respectively. He gets distinction incase he secures 80% or more marks. It is noticed from the result that 10% of the students failed in the examination, whereas 5% of them obtained distinction.

Calculate the percentage of students placed in the second division. (Assume normal distribution of marks).

[**Hint.** $P(X < 30) = 0.10$, $P(X \ge 80) = 0.05$]

15. Determine the minimum marks a student must get in order to receive an A grade if the top 10% of the students are awarded A grades in an examination where the mean mark is 72 and standard deviation is 9.

16. (*i*) When the mean of marks was 50% and S.D. 5% then 60% of the students failed in an examination. Determine the 'grace' marks to be awarded in order to show that 70% of the students passed. Assume that the marks are normally distributed.

$$\left[\textbf{Hint. } \frac{x_1 - 0.5}{0.05} = 0.25, \quad \frac{x_2 - 0.5}{0.05} = -0.52\right]$$

(*ii*) In a university examination of a particular year, 60% of the students failed when mean of the marks was 50% and S.D. 5%. University decided to relax the conditions of passing by lowering the pass marks to show its result 70%. Find the minimum marks for a student to pass supposing the marks to be normally distributed and no change in the performance of students takes place.

17. How does a normal distribution differ from a binomial distribution? What are the important properties of normal distribution? [*M.T.U. (MBA) 2012*]

18. If the skulls are classified as A, B and C according as the length-breadth index is under 75, between 75 and 80 or over 80, find approximately (assuming that the distribution is normal) the mean and standard deviation of a series in which A are 58%, B are 38% and C are 4%, being given that if $f(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-(x^2/2)}\, dx$ then $f(0.20) = 0.08$ and $f(1.75) = 0.46$.

[**Hint:** P(X < 75) = 0.58, P(X > 80) = 0.04]

19. The following table gives frequencies of occurrence of a variable X between certain limits:

Variable X	*Frequency*
Less than 40	30
40 or more but less than 50	33
50 and more	37

The distribution is exactly normal. Find the distribution and also obtain the fequency between X = 50 and X = 60.

20. The marks X obtained in Mathematics by 1000 students are normally distributed with mean 78% and standard deviation 11%.

Determine:

(*i*) how many students got marks above 90%?

(*ii*) What was the highest marks obtained by the lowest 10% of students?

(*iii*) Within what limits did the middle 90% of the students lie?

Answers

1. (*i*) 67 (*ii*) 134 (*iii*) 1909 **2.** (*i*) 227 (*ii*) 465

3. (*i*) $\bar{x} = 50.3$, $\sigma = 10.33$ (*ii*) $\mu = 65$, $\sigma = 10$

4. (*i*) 0.8849 (*ii*) 0.8849 (*iii*) 0.7881

5. (*i*) 143 (*ii*) 143 (*iii*) 614

6. (*a*) (*i*) 50%, (*ii*) 21.2%, (*iii*) 84% (*b*) (*i*) 1781, (*ii*) 1576, (*iii*) 1 (*c*) (*i*) 220, (*ii*) 69

7. (*a*) (*i*) 371, (*ii*) 383, (*iii*) 72.72 (*b*) (*i*) 1723 (*ii*) 1906 (*c*) (*i*) 155, (*ii*) 81, (*iii*) 23

8. (*a*) (*i*) 23% (*ii*) 11% (*b*) 0.0228 (*c*) (*i*) 159 (*ii*) 23

9. (*a*) (*i*) 48 (*ii*) 251 (*iii*) 701 (*b*) 294 (*c*) 0.0154

10. (*i*) 79 (*ii*) 35% (*iii*) 11 **11.** 10,000 **12.** 37.2

13. 0.06357 **14.** 34% **15.** 84 marks

16. (*i*) 3.85 (*ii*) 47.4. **18.** $\mu = 74.35$, $\sigma = 3.23$ **19.** $\mu = 46.12$, $\sigma = 11.76$, 25

20. (*i*) 138 (*ii*) 63.92% (*iii*) between 60 and 96.

3.68 POPULATION OR UNIVERSE

An aggregate of objects (animate or inanimate) under study is called **population or universe.** It is thus a collection of individuals or of their attributes (qualities) or of results of operations which can be numerically specified.

A universe containing a finite number of individuals or members is called a **finite inverse.** For example, the universe of the weights of students in a particular class.

A universe with infinite number of members is known as an **infinite universe.** For example, the universe of pressures at various points in the atmosphere.

In some cases, we may be even ignorant whether or not a particular universe is infinite, *e.g.*, the universe of stars.

The universe of concrete objects is an **existent universe.** The collection of all possible ways in which a specified event can happen is called a **hypothetical universe.** The universe of heads and tails obtained by tossing a coin an infinite number of times (provided that it does not wear out) is a hypothetical one.

3.69 SAMPLING

The statistician is often confronted with the problem of discussing universe of which he cannot examine every member *i.e.*, of which complete enumeration is impracticable. For example, if we want to have an idea of the average per capita income of the people of India, enumeration of every earning individual in the country is a very difficult task. Naturally, the question arises : What can be said about a universe of which we can examine only a limited number of members ? This question is the origin of the Theory of Sampling.

A finite subset of a universe is called a **sample.** A sample is thus a small portion of the universe. The number of individuals in a sample is called the **sample size.** The process of selecting a sample from a universe is called **sampling.**

The theory of sampling is a study of relationship existing between a population and samples drawn from the population. The fundamental object of sampling is to get as much information as possible of the whole universe by examining only a part of it. An attempt is thus made through sampling to give the maximum information about the parent universe with the minimum effort.

Sampling is quite often used in our day-to-day practical life. For example, in a shop we assess the quality of sugar, rice or any other commodity by taking only a handful of it from the bag and then decide whether to purchase it or not. A housewife normally tests the cooked products to find if they are properly cooked and contain the proper quantity of salt or sugar, by taking a spoonful of it.

3.70 SAMPLING METHODOLOGIES

Sampling methodologies are classified under two general categories:

1. Probability sampling and 2. Non-probability sampling

In the former, the researcher knows the exact possibility of selecting each member of the population while in the latter, the chance of being included in the sample is not known. A probability sample tends to be more difficult and costly to conduct. However, probability samples are the only type of samples where the results can be generalized from the sample to the population. In addition, probability samples allow the researcher to calculate the precision of the estimates obtained from the sample and to specify the sampling error.

Non-probability samples, incontrast, donot allow the study's findings to be generalized from the sample to the population. When discussing the results of a non-probability sample, the researchers must limit his/her findings to the persons or elements sampled.

This procedure also does not allow the researcher to calculate sampling statistics that provide information about the precision of the results. The advantage of non-probability sampling is the case in which it can be administered.

Non-probability samples tend to be less complicated and less time consuming than probability samples. If the researcher has no intention of generalizing beyond the sample, one of the non-probability sampling methodologies will provide the desired information.

3.71 NON-PROBABILITY SAMPLING

The three common types of non-probability samples are:

(*i*) Convenience Sampling. As the name implies, convenience sampling involves choosing respondents at the convenience of the researcher. Examples of convenience sampling include people-in-the street interviews—the sampling of people to which the researcher has easy access, such as a class of students and studies that use people who have volunteered to be questioned as a result of an advertisement or another type of promotion. A drawback to this methodology is the lack of sampling accuracy. Because the probability of inclusion in the sample is unknown for each respondent, none of the reliability or sampling precision statistics can be calculated. Convenience samples, however, are employed by researchers because the time and cost of collecting information can be reduced.

(*ii*) Quota Sampling [*G.B.T.U. (B. Pharm.) 2010*]

Quota sampling is often confused with stratified and cluster sampling—two probability sampling methodologies. All of these methodologies sample a population that has been subdivided into classes or categories.

The primary differences between the methodologies is that with stratified and cluster sampling, the classes are mutually exclusive and are isolated prior to sampling. Thus, the probability of being selected is known and members of the population selected to be sampled are not arbitrarily disqualified from being included in the results. In quota sampling, the classes cannot be isolated prior to sampling and respondents are categorized into the classes as the survey proceeds. As each class fills or reaches its quota, additional respondents that would have fallen into these classes are rejected or excluded from the results.

An example of a quota sample would be a survey in which the researcher desires to obtain a certain number of respondents from various income categories. Generally, researchers donot know the income of the persons they are sampling until they ask about income. Therefore, the researcher is unable to subdivide the population from which the sample is drawn into mutually exclusive income categories prior to drawing the sample.

(*iii*) Judgemental Sampling. In judgemental or purposive sampling, the researcher employs his or her own expert judgement about who to include in the sample frame. Prior knowledge and research skill are used in selecting the respondents or elements to be sampled.

An example of this type of sample would be a study of potential users of a new recreational facility that is limited to those persons who live within two miles of the new facility. Expert judgement based on past experience indicates that most of the use of this type of facility comes from persons living within two miles. However, by limiting the sample to only this group, usage projections may not be reliable if the usage characteristics of the new facility vary from those previously experienced. As with all non-probability sampling methods, the degree and direction of error introduced by the researcher cannot be measured and statistics that measure the precision of the estimates cannot be calculated.

3.72 PROBABILITY SAMPLING

Five methodologies are most commonly used for conducting probability sampling.

(*i*) Simple Random Sampling. Simple random sampling provides the base from which the other more complex sampling methodologies are derived.

To conduct a simple random sampling, the researcher must first prepare an exhaustive list (sampling frame) of all members of the population of interest. From this list, the sample is drawn so that each person or item has an equal chance of being drawn during each selection

round. Samples may be drawn with or without replacement. In practice, however, most simple random sampling for survey research is done without replacement ; that is, a person or item selected for sampling is removed from the population for all subsequent selections. At any draw, the process for a simple random sample without replacement must provide an equal chance of inclusion to any member of the population not already drawn. To draw a simple random sample without introducing researcher bias, computerized sampling programs and random numbers tables are used to impartially select the members of the population to be sampled.

An example of a simple random sample would be a survey of County employees. An exhaustive list of all County employees as of a certain date could be obtained from the Department of Human Resources. If 100 names were selected from this list using a random number table or a computerized sampling program, then a simple random sample would be created. Such a random sampling procedure has the advantage of reducing bias and enables the researcher to estimate sampling errors and the precision of the estimates derived through statistical calculations.

(*ii*) Stratified Random Sampling [*G.B.T.U. (B. Pharm.) 2010*]

Stratified random sampling involves categorizing the members of the population into mutually exclusive and collectively exhaustive groups. An independent simple random sample is then drawn from each group. Stratified sampling techniques can provide more precise estimates if the population being surveyed is more heterogeneous than the categorized groups, can enable the researcher to determine desired levels of sampling precision for each group, and can provide administrative efficiency.

An example of a stratified sample would be a sample conducted to determine the average income earned by families in the United States. To obtain more precise estimates of income, the researcher may want to stratify the sample by geographic region (northeast, mid-Atlantic, etc.) and/or stratify the sample by urban, suburban, and rural groupings. If the differences in income among the regions or groupings are greater than the income differences within the regions or groupings, precision of the estimates is improved. In addition, if the resarch organization has branch offices located in these regions, the administration of the survey can be decentralized and perhaps conducted in a more cost-efficient manner.

(*iii*) Cluster Sampling. Cluster sampling is similar to stratified sampling because the population to be sampled is subdivided into mutually exclusive groups. However, in cluster sampling, the groups are defined so as to maintain the heterogeneity of the population. It is the researcher's goal to establish clusters that are representative of the population as a whole, although in practice this may be difficult to achieve. After the clusters are established, a simple random sample of the clusters is drawn and the members of the chosen clusters are sampled. If all of the elements (members) of the clusters selected are sampled, then the sampling procedure is defined as **one-stage cluster sampling.** If a random sample of the elements of each selected cluster is drawn, then the sampling procedure is defined as **two-stage cluster sampling.**

Cluster sampling is frequently employed when the researcher is unable to compile a comprehensive list of all the elements in the population of interest. A cluster sample might be used by a researcher attempting to measure the age distribution of persons residing in Mumbai. It would be much more difficult for the researcher to compile a list of every person residing in Mumbai than to compile a list of residential addresses. In this example, each address would represent a cluster of elements (persons) to be sampled. If the elements contained in the clusters are as heterogeneous as the population, then estimates derived from cluster sampling are as precise as those from simple random sampling. However, if the heterogeneity of the clusters is less than that of the population, the estimates will be less precise.

(*iv*) Systematic Sampling. Systematic sampling, a form of one-stage cluster sampling, is often used in place of simple random sampling. In systematic sampling, the researcher selects every n^{th} member after randomly selecting the first through n^{th} element as the starting point. For example, if the researcher decides to sample every 20^{th} member of the population, a 5 percent sample, the starting point for the sample is randomly selected from the first 20 members. A systematic sample is a type of cluster sample because each of the first 20 members of the sampling frame defines a cluster that contains 5 percent of the population.

A researcher may choose to conduct a systematic sample instead of a simple random sample for several reasons. Systematic samples tend to be easier to draw and execute. The researcher does not have to jump backward and forward through the sampling frame to draw the members to be sampled. A systematic sample may spread the members selected for measurement more evenly across the entire population than simple random sampling. Therefore, in some cases, systematic sampling may be more representative of the population and more precise.

One of the most attractive aspects of systematic sampling is that this method can allow the researcher to draw a probability sample without complete prior knowledge of the sampling frame. For example, a survey of visitors to the County's publications desk could be conducted by sampling every 10^{th} visitor after randomly selecting the first through 10^{th} visitor as the starting point. By conducting the sample in this manner, it would not be necessary for the researcher to obtain a comprehensive list of visitors prior to drawing the sample.

As with other types of cluster sampling, systematic sampling is as precise as simple random sampling if the members contained in the clusters are as heterogeneous as the population. If this assumption is not valid, then systematic sampling will be less precise than simple random sampling. In conducting systematic sampling, it is also essential that the researcher does not introduce bias into the sample by selecting an inappropriate sampling interval. For instance, when conducting a sample of financial records, or other items that follow a calendar schedule, the researcher would not want to select "7" as the sampling interval because the sample would then be comprised of observations that were all on the same day of the week. Day-of-the-week influences may cause contamination of the sample, giving the researcher biased results.

(*v*) Multi-Stage Sampling. Multi-stage sampling is like cluster sampling, but involves selecting a sample within each chosen cluster, rather than including all units in the cluster. Thus, multi-stage sampling involves selecting a sample in at least two stages. In the first stage, large groups or clusters are selected. These clusters are designed to contain more population units than are required for the final sample.

In the second stage, population units are chosen from selected clusters to derive a final sample. If more than two stages are used, the process of choosing population units within clusters continues until the final sample is achieved.

An example of multi-stage sampling is where, firstly, electoral sub-divisions (clusters) are sampled from a city or state. Secondly, blocks of houses are selected from within the electoral sub-divisions and, thirdly, individual houses are selected from within the selected blocks of houses.

The advantages of multi-stage sampling are convenience, economy and efficiency. Multi-stage sampling does not require a complete list of members in the target population, which greatly reduces sample preparation cost. The list of members is required only for those clusters used in the final stage. The main disadvantage of multi-stage sampling is the same as for cluster sampling : lower accuracy due to higher sampling error.

3.73 PARAMETERS OF STATISTICS

The statistical constants of the population such as mean, the variance etc. are known as the parameters. The statistical concepts of the sample from the members of the sample to estimate the parameters of the population from which the sample has been drawn is known as *statistic*.

Population mean and variance are denoted by μ and σ^2, while those of the samples are given by $\bar{x}$, s^2.

3.74 STANDARD ERROR

The standard deviation of the sampling distribution of a statistic is known as the **standard error (S.E.).** It plays an important role in the theory of large samples and it forms a basis of the testing of hypothesis. If t is any statistic, for large sample. $z = \dfrac{t - \mathrm{E}(t)}{\mathrm{S.E.}(t)}$ is normally distributed with mean 0 and variance unity.

3.75 TEST OF SIGNIFICANCE

An important aspect of the sampling theory is to study the test of significance which will enable us to decide, on the basis of the results of the sample, whether

(*i*) *the deviation between the observed sample statistic and the hypothetical parameter value or*

(*ii*) *the deviation between two sample statistics*

is significant or might be attributed due to chance or the fluctuations of the sampling.

3.76 TESTING OF STATISTICAL HYPOTHESIS

Step 1. Null hypothesis: [*U.P.T.U. (MCA) 2007*]

For applying the tests of significance, we first set up a hypothesis which is a definite statement about the population parameter called Null Hypothesis. It is denoted by H_0.

Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true. First, we set up H_0 in clear terms.

Step 2. Alternative hypothesis:

Any hypothesis which is complementary to the null hypothesis (H_0) is called an alternative hypothesis. It is denoted by H_1.

For example, if we want to test the null hypothesis that the population has a specified mean μ_0 then we have

$$H_0 : \mu = \mu_0$$

then the alternative hypothesis will be

(*i*) $H_1 : \mu \neq \mu_0$ (Two tailed alternative hypothesis)

(*ii*) $H_1 : \mu > \mu_0$ (right tailed alternative hypothesis (or) single tailed)

(*iii*) $H_1 : \mu < \mu_0$ (left tailed alternative hypothesis (or) single tailed)

Hence alternative hypothesis helps to know whether the test is two tailed or one tailed. Therefore, we set up H_1 for this decision.

Step 3. Level of significance: [*U.P.T.U. (MCA) 2008, 2007*]

The probability of the value of the variate falling in the critical region is known as level of significance. A region corresponding to a statistic t in the sample space S which amounts to rejection of the null hypothesis H_0 is called as **critical region** or region of rejection while which amounts to acceptance of H_0 is called **acceptance region.** The probability α that a random value of the statistic t belongs to the critical region is known as the **level of significance.**

$$P(t \in w/H_0) = \alpha$$

i.e., the level of significance is the size of the type I error (refer art. 3.77) or the maximum producer's risk.

We select the appropriate level of significance in advance depending on the reliability of the estimates.

Step 4. Test statistic (or test criterion): We compute the test statistic z under the null hypothesis. For larger samples corresponding to the statistic t, the variable $z = \dfrac{t - E(t)}{S.E.(t)}$ is normally distributed with mean 0 and variance 1. The value of z given above under the null hypothesis is known as **test statistic.**

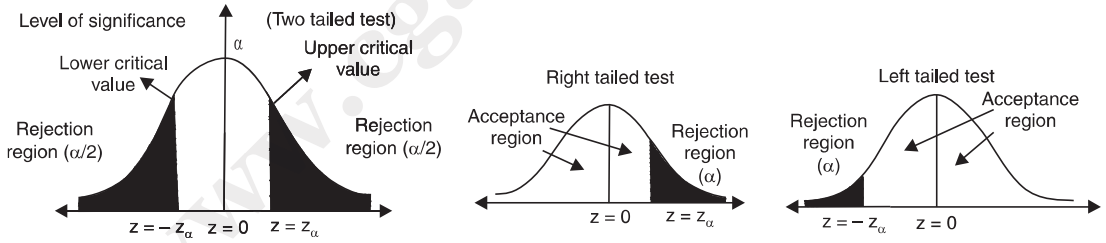
Step 5. Conclusion: We compare the computed value of z with the critical value z_α at level of significance (α). The critical value of z_α of the test statistic at level of significance α for a two tailed test is given by

$$p(|z| > z_\alpha) = \alpha \qquad \ldots(1)$$

i.e., z_α is the value of z so that the total area of the critical region on both tails is α. Since the normal curve is symmetrical, from equation (1), we get

$$p(z > z_\alpha) + p(z < -z_\alpha) = \alpha \text{ ; } i.e., \; 2p(z > z_\alpha) = \alpha \text{ ; } i.e., \; p(z > z_\alpha) = \alpha/2$$

i.e., the area of each tail is $\alpha/2$.



The critical value z_α is that value such that the area to the right of z_α is $\alpha/2$ and the area to the left of $-z_\alpha$ is $\alpha/2$.

In the case of one tailed test,

$p(z > z_\alpha) = \alpha$ if it is right tailed ; $p(z < -z_\alpha) = \alpha$ if it is left tailed.

The critical value of z for a single tailed test (right or left) at level of significance α is same as the critical value of z for two tailed test at level of significance 2α.

Using the equation, also using the normal tables, the critical value of z at different levels of significance (α) for both single tailed and two tailed test are calculated and listed below. The equations are

$$p(|z| > z_\alpha) = \alpha \text{ ; } p(z > z_\alpha) = \alpha \text{ ; } p(z < -z_\alpha) = \alpha$$

Level of significance			
	1% (0.01)	5% (0.05)	10% (0.1)
Two tailed test	$\mid z_\alpha \mid = 2.58$	$\mid z_\alpha \mid = 1.96$	$\mid z_\alpha \mid = 1.645$
Right tailed test	$z_\alpha = 2.33$	$z_\alpha = 1.645$	$z_\alpha = 1.28$
Left tailed test	$z_\alpha = -2.33$	$z_\alpha = -1.645$	$z_\alpha = -1.28$

If $|z| > z_\alpha$, we reject H_0 and conclude that there is significant difference. If $|z| < z_\alpha$, we accept H_0 and conclude that there is no significant difference.

3.77 ERRORS IN SAMPLING

The main aim of the sampling theory is to draw a valid conclusion about the population parameters on the basis of the sample results. In doing this we may commit the following two types of errors:

Type I. Error. [*U.P.T.U. (MCA) 2008, 2007*]

When H_0 is true, we may reject it.

P(Reject H_0 when it is true) = P(Reject H_0/H_0) = α

α is called the size of the type I error also referred to as **producer's risk.**

Type II. Error. When H_0 is wrong, we may accept it P(Accept H_0 when it is wrong) = P(Accept H_0/H_1) = β . β is called the size of the type II error, also referred to as **consumer's risk.**

Note. The values of the test statistic which separates the critical region and acceptance region are called the **critical values** or **significant values.** This value is dependent on (*i*) the level of significance used and (*ii*) the alternative hypothesis, whether it is one tailed or two tailed.

3.78 TEST OF SIGNIFICANCE OF SMALL SAMPLES

When the size of the sample is less than 30, then the sample is called small sample. For such sample it will not be possible for us to assume that the random sampling distribution of a statistic is approximately normal and the values given by the sample data are sufficiently close to the population values and can be used in their place for the calculation of the standard error of the estimate.

3.79 STUDENT'S t-DISTRIBUTION (t-Test)

[*G.B.T.U. (MBA) 2011 ; G.B.T.U. (MCA) 2010*]

This *t*-distribution is used when sample size is ≤ 30 and the population standard deviation is unknown.

t-statistic is defined as

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \quad \text{where,} \quad S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

$\bar{x}$ is the mean of sample, μ is population mean. S is the standard deviation of population and n is sample size.

If the standard deviation of the sample 's' is given then t-statistic is defined as

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

Note. The relation between s and S is $ns^2 = (n-1)S^2$.

3.79.1. The t-Table

The t-table given at the end is the probability integral of t-distribution. The t-distribution has different values for each degrees of freedom and when the degrees of freedom are infinitely large, the t-distribution is equivalent to normal distribution and the probabilities shown in the normal distribution tables are applicable.

3.79.2. Applications of t-Distribution

[*G.B.T.U. (MBA) 2011*]

Some of the applications of t-distribution are given below:

1. *To test if the sample mean* $(\bar{x})$ *differs significantly from the hypothetical value* μ *of the population mean.*
2. *To test the significance between two sample means.*
3. *To test the significance of observed partial and multiple correlation coefficients.*

3.79.3. Critical Value of t

The critical value or significant value of t at level of significance α, degrees of freedom γ for two tailed test is given by

$$P[|\ t\ | > t_\gamma(\alpha)] = \alpha$$
$$P[|\ t\ | \leq t_\gamma(\alpha)] = 1 - \alpha$$

The significant value of t at level of significance α, for a single tailed test can be got from those of two tailed test by referring to the values at 2α.

3.80 TEST I : t-TEST OF SIGNIFICANCE OF THE MEAN OF A RANDOM SAMPLE

To test whether the mean of a sample drawn from a normal population deviates significantly from a stated value when variance of the population is unknown.

H_0 : There is no significant difference between the sample mean $\bar{x}$ and the population mean μ *i.e.,* we use the statistic

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} \qquad \text{where} \quad S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$$

with degree of freedom $n - 1$.

At given level of significance α and degrees of freedom $(n-1)$, we refer to t-table t_α (two tailed or one tailed). If calculated t value is such that $|\ t\ | < t_\alpha$, the null hypothesis is accepted. If $|\ t\ | > t_\alpha$, H_0 is rejected.

3.80.1. Fiducial Limits of Population Mean

If t_α is the value of t at level of significance α at $(n-1)$ degrees of freedom then,

$$\left| \frac{\bar{x} - \mu}{S/\sqrt{n}} \right| < t_\alpha \text{ for acceptance of } H_0.$$

$$\bar{x} - t_\alpha \, S/\sqrt{n} < \mu < \bar{x} + t_\alpha \, S/\sqrt{n}$$

95% confidence limits (level of significance 5%) are $\bar{x} \pm t_{0.05} S/\sqrt{n}$.

99% confidence limits (level of significance 1%) are $\bar{x} \pm t_{0.01} S/\sqrt{n}$.

EXAMPLES

Example 1. *A random sample of size 16 has 53 as mean. The sum of squares of the deviation from mean is 135. Can this sample be regarded as taken from the population having 56 as mean? Obtain 95% and 99% confidence limits of the mean of the population.*

Sol. Null hypothesis, H_0 : There is no significant difference between the sample mean and hypothetical population mean *i.e.,* $\mu = 56$.

Alternative hypothesis, $H_1 : \mu \neq 56$ (**Two tailed test**)

Test statistic. Under H_0, test statistic is $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$

Given: $\bar{x} = 53$, $\mu = 56$, $n = 16$, $\Sigma(x - \bar{x})^2 = 135$

$$S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{135}{15}} = 3$$

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{53 - 56}{3/\sqrt{16}} = -4$$

$| t | = 4$

d.fv. $= 16 - 1 = 15$.

Conclusion. Since $| t | = 4 > t_{0.05} = 2.13$ *i.e.,* the calculated value of t is more than the tabulated value, the null hypothesis is rejected. Hence, the sample mean has not come from a population having 56 as mean.

95% confidence limits of the population mean

$$= \bar{x} \pm \frac{S}{\sqrt{n}} t_{0.05} = 53 \pm \frac{3}{\sqrt{16}} (2.13) = 51.4025, 54.5975$$

99% confidence limits of the population mean

$$= \bar{x} \pm \frac{S}{\sqrt{n}} t_{0.01,} = 53 \pm \frac{3}{\sqrt{16}} (2.95) = 50.7875, 55.2125.$$

Example 2. *The lifetime of electric bulbs for a random sample of 10 from a large consignment gave the following data:*

Item	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Life in '000 hrs.	*4.2*	*4.6*	*3.9*	*4.1*	*5.2*	*3.8*	*3.9*	*4.3*	*4.4*	*5.6*

Can we accept the hypothesis that the average lifetime of bulb is 4000 hrs?

Sol. Null hypothesis: H_0: There is no significant difference in the sample mean and population mean. *i.e.,* $\mu = 4000$ hrs.

Alternative hypothesis: $\mu \neq 4000$ hrs (**Two tailed test**)

Test statistic: Under H_0, the test statistic is $t = \dfrac{\bar{x} - \mu}{S/\sqrt{n}}$

x	4.2	4.6	3.9	4.1	5.2	3.8	3.9	4.3	4.4	5.6
$x - \bar{x}$	– 0.2	0.2	– 0.5	– 0.3	0.8	– 0.6	– 0.5	– 0.1	0	1.2
$(x - \bar{x})^2$	0.04	0.04	0.25	0.09	0.64	0.36	0.25	0.01	0	1.44

$$\bar{x} = \frac{\Sigma x}{n} = \frac{44}{10} = 4.4, \qquad \Sigma(x - \bar{x})^2 = 3.12$$

$$S = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}} = 0.589$$

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{4.4 - 4}{\left(\frac{0.589}{\sqrt{10}}\right)} = 2.123$$

For $\gamma = 9$, $t_{0.05} = 2.26$.

Conclusion. Since the calculated value of t is less than the tabulated value of t at 5% level of significance. $\therefore$ The null hypothesis $\mu = 4000$ hrs is accepted *i.e.*, the average lifetime of bulbs could be 4000 hrs.

Example 3. *A sample of 20 items has mean 42 units and S.D. 5 units. Test the hypothesis that it is a random sample from a normal population with mean 45 units.*

Sol. Null hypothesis: H_0 : There is no significant difference between the sample mean and the population mean. *i.e.,* $\mu = 45$ units

Alternative hypothesis, H_1: $\mu \neq 45$ **(Two tailed test)**

Given: $n = 20$, $\bar{x} = 42$, $s = 5$; $\gamma = 19$ d.f.

Test statistic: Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n - 1}} = \frac{42 - 45}{5/\sqrt{19}} = - 2.615$$

$\therefore \qquad | t | = 2.615$

The tabulated value of t at 5% level for 19 d.f. is $t_{0.05} = 2.09$.

Conclusion. Since the calculated value $| t |$ is greater than the tabulated value of t at 5% level of significance, the null hypothesis H_0 is rejected. *i.e.,* there is significant difference between the sample mean and population mean.

i.e., the sample could not have come from this population.

Example 4. *The 9 items of a sample have the following values*

45, 47, 50, 52, 48, 47, 49, 53, 51.

Does the mean of these values differ significantly from the assumed mean 47.5?

Sol. Here, $n = 9$, $\mu = 47.5$, $\bar{x} = \frac{\Sigma x}{n} = 49.1$

x	45	47	50	52	48	47	49	53	51
$x - \bar{x}$	– 4.1	– 2.1	0.9	2.9	– 1.1	– 2.1	– .1	3.9	1.9
$(x - \bar{x})^2$	16.81	4.41	.81	8.41	1.21	4.41	.01	15.21	3.61

$$\Sigma(x - \bar{x})^2 = 54.89,$$

$$S^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1} = 6.86$$

$\therefore \qquad S = 2.619$

Null hypothesis:

$H_0 : \mu = 47.5$

i.e., there is no significant difference between the sample and population means.

Alternative hypothesis:

$H_1 : \mu \neq 47.5$

Hence we apply **two-tailed test**.

Test statistic: Under H_0, the test statistic is

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{49.1 - 47.5}{(2.619/\sqrt{9})} = 1.8327$$

$t_{0.05} = 2.31$ for $\gamma = 8$

Conclusion: Since $|t|_{\text{calculated}} < t_{\text{tabulated}}$ at 5% level of significance, the null hypothesis H_0 is accepted *i.e.,* there is no significant difference between their means.

ASSIGNMENT

1. Ten individuals are chosen at random from a normal population of students and their marks are found to be 63, 63, 66, 67, 68, 69, 70, 70, 71, 71. In the light of these data, discuss the suggestion that mean mark of the population of students is 66.
2. The following values gives the lengths of 12 samples of Egyptian cotton taken from a consignment : 48, 46, 49, 46, 52, 45, 43, 47, 47, 46, 45, 50. Test if the mean length of the consignment can be taken as 46.
3. A sample of 18 items has a mean 24 units and standard deviation 3 units. Test the hypothesis that it is a random sample from a normal population with mean 27 units.
4. A random sample of 10 boys had the I.Q.'s 70, 120, 110, 101, 88, 83, 95, 98, 107 and 100. Do these data support the assumption of a population mean I.Q. of 160?

3.81 TEST II : t-TEST FOR DIFFERENCE OF MEANS OF TWO SMALL SAMPLES (from a Normal Population)

This test is used to test whether the two samples $x_1, x_2, \ldots\ldots, x_{n_1}, y_1, y_2, \ldots\ldots, y_{n_2}$ of sizes n_1, n_2 have been drawn from two normal populations with mean μ_1 and μ_2 respectively under the assumption that the population variance are equal. ($\sigma_1 = \sigma_2 = \sigma$).

H_0 : The samples have been drawn from the normal population with means μ_1 and μ_2 *i.e.,* $H_0 : \mu_1 \neq \mu_2$.

Let $\bar{x}$, $\bar{y}$ be their means of the two samples.

Under this H_0 the test statistic t is given by $t = \dfrac{(\bar{x} - \bar{y})}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

Degree of freedom is $n_1 + n_2 - 2$.

Note 1. If the two sample's standard deviations s_1, s_2 are given then we have $S^2 = \dfrac{n_1 {s_1}^2 + n_2 {s_2}^2}{n_1 + n_2 - 2}$.

Note 2. If s_1, s_2 are not given then $S^2 = \dfrac{\Sigma(x_1 - \bar{x}_1)^2 + \Sigma(x_2 - \bar{x}_2)^2}{n_1 + n_2 - 2}$.

EXAMPLES

Example 1. *Two samples of sodium vapour bulbs were tested for length of life and the following results were got*:

	Size	*Sample mean*	*Sample S.D.*
Type I	*8*	*1234 hrs*	*36 hrs*
Type II	*7*	*1036 hrs*	*40 hrs*

Is the difference in the means significant to generalise that Type I is superior to Type II regarding length of life?

Sol. Null hypothesis,

$H_0 : \mu_1 = \mu_2$ *i.e.,* two types of bulbs have same lifetime.

Alternative hypothesis,

$H_1 : \mu_1 > \mu_2$ *i.e.,* type I is superior to Type II.

Hence we use **right tailed test.**

$$S^2 = \frac{n_1 s_1^{\,2} + n_2 s_2^{\,2}}{n_1 + n_2 - 2} = \frac{8(36)^2 + 7(40)^2}{8 + 7 - 2} = 1659.076$$

$\therefore \quad S = 40.7317$

Test statistic: Under H_0, the test statistic t is given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1234 - 1036}{40.7317\sqrt{\frac{1}{8} + \frac{1}{7}}} = 18.1480$$

$t_{0.05}$ at d.f. $\gamma = n_1 + n_2 - 2 = 13$ is 1.77.

Conclusion. Since calculated $| t | > t_{\text{tabulated}}$ at 5% level of significance, H_0 is rejected.

$\therefore$ Type I is definitely superior to Type II.

Example 2. *Samples of sizes 10 and 14 were taken from two normal populations with S.D. 3.5 and 5.2. The sample means were found to be 20.3 and 18.6. Test whether the means of the two populations are the same at 5% level.*

Sol. We have, $\bar{x}_1 = 20.3$, $\bar{x}_2 = 18.6$, $n_1 = 10$, $n_2 = 14$, $s_1 = 3.5$, $s_2 = 5.2$

$$S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = 22.775$$

$\therefore \quad S = 4.772$

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ *i.e.,* the means of the two populations are the same.

Alternative hypothesis:

$H_1 : \mu_1 \neq \mu_2$

Test statistic: Under H_0, the test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{20.3 - 18.6}{4.772\sqrt{\frac{1}{10} + \frac{1}{14}}} = 0.8604$$

The tabulated value of t at 5% level of significance for 22 *d.f.* is $t_{0.05} = 2.0739$

Conclusion:

Since $t = 0.8604 < t_{0.05}$, the null hypothesis H_0 is accepted; *i.e.,* there is no significant difference between their means.

Example 3. *The height of 6 randomly chosen sailors in inches are 63, 65, 68, 69, 71 and 72. Those of 9 randomly chosen soldiers are 61, 62, 65, 66, 69, 70, 71, 72 and 73. Test whether the sailors are on the average taller than soldiers.*

Sol. Let X_1 and X_2 be the two samples denoting the heights of sailors and soldiers.

$$n_1 = 6, n_2 = 9$$

Null hypothesis, $H_0 : \mu_1 = \mu_2$.

i.e., the mean of both the population are the same.

Alternative hypothesis, $H_1 : \mu_1 > \mu_2$ (**one tailed test**)

Calculation of two sample means:

X_1	63	65	68	69	71	72
$X_1 - \overline{X}_1$	– 5	– 3	0	1	3	4
$(X_1 - \overline{X}_1)^2$	25	9	0	1	9	16

$$\overline{X}_1 = \frac{\Sigma X_1}{n_1} = 68 \; ; \; \Sigma(X_1 - \overline{X}_1)^2 = 60$$

X_2	61	62	65	66	69	70	71	72	73
$X_2 - \overline{X}_2$	– 6.66	– 5.66	– 2.66	1.66	1.34	2.34	3.34	4.34	5.34
$(X_2 - \overline{X}_2)^2$	44.36	32.035	7.0756	2.7556	1.7956	5.4756	11.1556	18.8356	28.5156

$$\overline{X}_2 = \frac{\Sigma X_2}{n_2} = 67.66 \; ; \; \Sigma(X_2 - \overline{X}_2)^2 = 152.0002$$

$$S^2 = \frac{1}{n_1 + n_2 - 2}[\Sigma(X_1 - \overline{X}_1)^2 + \Sigma(X_2 - \overline{X}_2)^2] = 16.3077$$

$\therefore$ $S = 4.038$

Test statistic:

Under H_0, $$t = \frac{\overline{X}_1 - \overline{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{68 - 67.666}{4.038\sqrt{\frac{1}{6} + \frac{1}{9}}} = 0.1569$$

The value of t at 5% level of significance for 13 d.f. is 1.77. (d.f. $= n_1 + n_2 - 2$)

Conclusion. Since $t_{\text{calculated}} < t_{0.05} = 1.77$, the null hypothesis H_0 is accepted.

i.e., there is no significant difference between their average.

i.e., the sailors are not on the average taller than the soldiers.

ASSIGNMENT

1. The mean life of 10 electric motors was found to be 1450 hrs with S.D. of 423 hrs. A second sample of 17 motors chosen from a different batch showed a mean life of 1280 hrs with a S.D. of 398 hrs. Is there a significant difference between means of the two samples?
2. The marks obtained by a group of 9 regular course students and another group of 11 part time course students in a test are given below:

   Regular: 56 62 63 54 60 51 67 69 58

   Part time: 62 70 71 62 60 56 75 64 72 68 66

   Examine whether the marks obtained by regular students and part time students differ significantly at 5% and 1% level of significance.
3. A group of 5 patients treated with the medicine A weigh 42, 39, 48, 60 and 41 kgs. A second group of 7 patients from the same hospital treated with medicine B weigh 38, 42, 56, 64, 68, 69 and 62 kgs. Do you agree with the claim that medicine B increases the weight significantly? It is given that the value of t at 10% level of significance for 10 degree of freedom is 1.81.

   [*G.B.T.U. (B. Pharm.) 2010*]
4. Two independent samples of sizes 7 and 9 have the following values:

   Sample A: 10 12 10 13 14 11 10

   Sample B: 10 13 15 12 10 14 11 12 11

   Test whether the difference between the mean is significant.
5. The average no. of articles produced by two machines per day are 200 and 250 with standard deviations 20 and 25 respectively on the basis of records of 25 days production. Can you regard both the machines equally efficient at 5% level of significance?

3.82 CHI-SQUARE (χ^2) TEST

[*G.B.T.U. 2010 ; G.B.T.U. MCA (SUM) 2010*]

When a coin is tossed 200 times, the theoretical considerations lead us to expect 100 heads and 100 tails. But in practice, these results are rarely achieved. The quantity χ^2 (a Greek letter, pronounced as chi-square) describes the magnitude of discrepancy between theory and observation. If $\chi^2 = 0$, the observed and expected frequencies completely coincide. The greater the discrepancy between the observed and expected frequencies, the greater is the value of χ^2. Thus χ^2 **affords a measure of the correspondence between theory and observation.**

If O_i ($i = 1, 2, ..., n$) is a set of observed (experimental) frequencies and E_i ($i = 1, 2, ..., n$) is the corresponding set of expected (theoretical or hypothetical) frequencies, then, χ^2 **is defined as**

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

where $\Sigma O_i = \Sigma E_i = N$ (total frequency) and degrees of freedom (d.f.) = $(n - 1)$.

Note. (*i*) If $\chi^2 = 0$, the observed and theoretical frequencies agree exactly.

(*ii*) If $\chi^2 > 0$ they do not agree exactly.

3.83 DEGREES OF FREEDOM

While comparing the calculated value of χ^2 with the tabular value, we have to determine the degrees of freedom.

If we have to choose any four numbers whose sum is 50, we can exercise our independent choice for any three numbers only, the fourth being 50 minus the total of the three numbers selected. Thus, though we were to choose any four numbers, our choice was reduced to three because of one condition imposed. There was only one restraint on our freedom and our degrees of freedom were $4 - 1 = 3$. If two restrictions are imposed, our freedom to choose will be further curtailed and degrees of freedom will be $4 - 2 = 2$.

In general, the number of degrees of freedom is the total number of observations less the number of independent constraints imposed on the observations. Degrees of freedom (d.f.) are usually denoted by ν.

Thus, $\nu = n - k$, where k is the number of independent constraints in a set of data of n observations.

Note. (*i*) For a $p \times q$ contingency table (p columns and q rows), $\nu = (p - 1)(q - 1)$

(*ii*) In the case of a contingency table, the expected frequency of any class

$$= \frac{\text{Total of rows in which it occurs} \times \text{Total of columns in which it occurs}}{\text{Total number of observations}}$$

3.84 APPLICATIONS OF CHI-SQUARE TEST

χ^2 test is one of the simplest and the most general test known. It is applicable to a very large number of problems in practice which can be summed up under the following heads:

(*i*) as a test of goodness of fit.

(*ii*) as a test of independence of attributes.

(*iii*) as a test of homogeneity of independent estimates of the population variance.

(*iv*) as a test of the hypothetical value of the population variance σ^2.

(*v*) as a list to the homogeneity of independent estimates of the population correlation coefficient.

3.85 CONDITIONS FOR APPLYING χ^2 TEST

χ^2-test is an approximate test for large values of n. For the validity of chi-square test of goodness of fit between theory and experiment, the following conditions must be satisfied.

(*a*) The sample observations should be independent.

(*b*) The constraints on the cell frequencies, if any, should be linear *e.g.,* $\Sigma n_i = \Sigma \lambda_i$ or $\Sigma O_i = \Sigma E_i$.

(*c*) N, the total number of frequencies should be reasonably large. It is difficult to say what constitutes largeness, but as an arbitrary figure, we may say that **N should be atleast 50,** however, few the cells.

(*d*) No theoretical cell-frequency should be small. Here again, it is difficult to say what constitutes smallness, but 5 should be regarded as the very minimum and **10 is better.** If small theoretical frequencies occur (*i.e.*, < 10), the difficulty is overcome by grouping two or more classes together before calculating (O – E). **It is important to remember that the number of degrees of freedom is determined with the number of classes after regrouping.**

Note 1. If any one of the theoretical frequency is less than 5, then we apply a corrected given by F Yates, which is usually known as 'Yates correction for continuity', we add 0.5 to the cell frequency which is less than 5 and adjust the remaining cell frequency suitably so that the marginal total is not changed.

Note 2. It may be noted that the χ^2-test depends only on the set of observed and expected frequencies and on degrees of freedom (d.f.). It does not make any assumption regarding the parent population from which the observations are taken. Since χ^2 does not involve any population parameters, it is termed as a statistic and the test is known as Non-parametric test or Distribution-free test.

3.86 THE χ^2 DISTRIBUTION

For large sample sizes, the sampling distribution of χ^2 can be closely approximated by a continuous curve known as the chi-square distribution. The probability function of χ^2 distribution is given by

$$f(\chi^2) = c(\chi^2)^{(\nu/2-1)}\, e^{-x^2/2}$$

where $e = 2.71828$, ν = number of degrees of freedom ; c = a constant depending only on ν.

Symbolically, the degrees of freedom are denoted by the symbol ν or by d.f. and are obtained by the rule $\nu = n - k$, where k refers to the number of independent constraints.

In general, when we fit a binomial distribution the number of degrees of freedom is one less than the number of classes ; when we fit a Poisson distribution the degrees of freedom are 2 less than the number of classes, because we use the total frequency and the arithmetic mean to get the parameter of the Poisson distribution. When we fit a normal curve the number of degrees of freedom are 3 less than the number of classes, because in this fitting we use the total frequency, mean and standard deviation.

If the data is given in a series of "n" numbers then degrees of freedom = $n - 1$.

In the case of Binomial distribution d.f. = $n - 1$

In the case of Poisson distribution d.f. = $n - 2$

In the case of Normal distribution d.f. = $n - 3$.

3.87 χ^2 TEST AS A TEST OF GOODNESS OF FIT

χ^2 test enables us to ascertain how well the theoretical distributions such as Binomial, Poisson or Normal etc. fit empirical distributions, *i.e.*, distributions obtained from sample data. If the **calculated value of χ^2 is less than the tabular value** at a specified level (generally 5%) of significance, the **fit is considered to be good** *i.e.*, the divergence between actual and expected frequencies is attributed to fluctuations of simple sampling. If the calculated value of χ^2 is greater than the tabular value, the fit is considered to be poor.

EXAMPLES

Example 1. *In experiments on pea breeding, the following frequencies of seeds were obtained:*

Round and yellow	*Wrinkled and yellow*	*Round and green*	*Wrinkled and green*	*Total*
315	*101*	*108*	*32*	*556*

Theory predicts that the frequencies should be in proportions 9 : 3 : 3 : 1. Examine the correspondence between theory and experiment.

Sol. Null hypothesis:

H_0 : The experimental result support the theory *i.e.,* there is no significant difference between the observed and theoretical frequency.

Under H_0, The theoretical (expected) frequencies can be calculated as follows:

$$E_1 = \frac{556 \times 9}{16} = 312.75 \qquad E_2 = \frac{556 \times 3}{16} = 104.25$$

$$E_3 = \frac{556 \times 3}{16} = 104.25 \qquad E_4 = \frac{556 \times 1}{16} = 34.75$$

To calculate the value of χ^2:

Observed frequency O_i	315	101	108	32
Expected Frequency E_i	312.75	104.25	104.25	34.75
$\frac{(O_i - E_i)^2}{E_i}$	0.016187	0.101319	0.134892	0.217626

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] = 0.470024$$

Tabular value of χ^2 at 5% level of significance for $n - 1 = 3$ *d.f.* is 7.815 *i.e.,* $\chi^2_{0.05} = 7.815$.

Conclusion: Since the calculated value of χ^2 is less than that of the tabulated value, hence H_0 is accepted. Therefore, the experimental results support the theory.

Example 2. *The following table gives the number of accidents that took place in an industry during various days of the week. Test if accidents are uniformly distributed over the week.*

Day	*Mon*	*Tue*	*Wed*	*Thu*	*Fri*	*Sat*
No. of accidents	*14*	*18*	*12*	*11*	*15*	*14*

Sol. Null hypothesis H_0: The accidents are uniformly distributed over the week.

Under this H_0, the expected frequencies of the accidents on each of these days $= \frac{84}{6} = 14$

Observed frequency O_i	14	18	12	11	15	14
Expected frequency E_i	14	14	14	14	14	14
$(O_i - E_i)^2$	0	16	4	9	1	0

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] = \frac{\Sigma(O_i - E_i)^2}{E_i} = \frac{30}{14} = 2.1428.$$

Tabular value of χ^2 at 5% level for (6 – 1 = 5 d.f.) is 11.09.

Conclusion: Since the calculated value of χ^2 is less than the tabulated value, H_0 is accepted *i.e.,* the accidents are uniformly distributed over the week.

Example 3. *A die is thrown 276 times and the results of these throws are given below*:

No. appeared on the die	*1*	*2*	*3*	*4*	*5*	*6*
Frequency	*40*	*32*	*29*	*59*	*57*	*59*

Test whether the die is biased or not.

Sol. Null hypothesis H_0: Die is unbiased.

Under this H_0, the expected frequencies for each digit is $\frac{276}{6} = 46$.

To find the value of χ^2

O_i	40	32	29	59	57	59
E_i	46	46	46	46	46	46
$(O_i - E_i)^2$	36	196	289	169	121	169

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right] = \frac{\Sigma(O_i - E_i)^2}{E_i} = \frac{980}{46} = 21.30.$$

Tabulated value of χ^2 at 5% level of significance for $(6 - 1 = 5)$ d.f. is 11.09.

Conclusion. Since the calculated value of $\chi^2 = 21.30 > 11.07$ the tabulated value, H_0 is rejected. *i.e.,* die is not unbiased or die is biased.

Example 4. *Records taken of the number of male and female births in 800 families having four children are as follows:* [U.P.T.U. (MCA) 2009]

No. of male births	*0*	*1*	*2*	*3*	*4*
No. of female births	*4*	*3*	*2*	*1*	*0*
No. of families	*32*	*178*	*290*	*236*	*64*

Test whether the data are consistent with the hypothesis that the Binomial law holds and the chance of male birth is equal to that of female birth, namely p = q = 1/2.

Sol. Null hypothesis H_0: The data are consistent with the hypothesis of equal probability for male and female births. *i.e.,* $p = q = 1/2$.

We use Binomial distribution to calculate theoretical frequency given by:

$$N(r) = N \times P(X = r) = N \times {}^nC_r\, p^r q^{n-r}$$

where N is the total frequency, $N(r)$ is the number of families with r male children, p and q are probabilities of male and female births respectively, n is the number of children.

$$N(0) = 800 \times {}^4C_0 \left(\frac{1}{2}\right)^4 = 50, \quad N(1) = 200, N(2) = 300, N(3) = 200 \text{ and } N(4) = 50$$

Observed frequency O_i	32	178	290	236	64
Expected frequency E_i	50	200	300	200	50
$(O_i - E_i)^2$	324	484	100	1296	196
$\frac{(O_i - E_i)^2}{E_i}$	6.48	2.42	0.333	6.48	3.92

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right] = 19.633.$$

Tabulated value of χ^2 at 5% level of significance for $5 - 1 = 4$ d.f. is 9.49.

Conclusion. Since the calculated value of χ^2 is greater than the tabulated value, H_0 is rejected. *i.e.,* the data are not consistent with the hypothesis that the Binomial law holds and that the chance of a male birth is not equal to that of a female birth.

Example 5. *The theory predicts the proportion of beans in the four groups, G_1, G_2, G_3, G_4 should be in the ratio 9 : 3 : 3 : 1. In an experiment with 1600 beans the numbers in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory?*

Sol. Null hypothesis H_0: The experimental result support the theory. *i.e.,* there is no significant difference between the observed and theoretical frequency.

Under H_0, the theoretical frequency can be calculated as follows:

$$E(G_1) = \frac{1600 \times 9}{16} = 900; \quad E(G_2) = \frac{1600 \times 3}{16} = 300;$$

$$E(G_3) = \frac{1600 \times 3}{16} = 300; \quad E(G_4) = \frac{1600 \times 1}{16} = 100$$

To calculate the value of χ^2.

Observed frequency O_i	882	313	287	118
Expected frequency E_i	900	300	300	100
$\frac{(O_i - E_i)^2}{E_i}$	0.36	0.5633	0.5633	3.24

$$\chi^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right] = 4.7266.$$

Tablular value of χ^2 at 5% level of significance for 3 d.f. is 7.815.

Conclusion: Since the calculated value of χ^2 is less than that of the tabulated value, hence H_0 is accepted. *i.e.,* the experimental results support the theory.

Example 6. *The following table shows the distribution of digits in numbers chosen at random from a telephone directory:*

Digits	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Frequency	*1026*	*1107*	*997*	*966*	*1075*	*933*	*1107*	*972*	*964*	*853*

Test whether the digits may be taken to occur equally frequently in the directory.

[G.B.T.U. (MCA) 2011]

Sol. Null hypothesis:

H_0 : The digits taken in the directory occur equally frequently *i.e.,* there is no significant difference between the observed and expected frequency.

Under H_0, the expected frequency = $\frac{10000}{10} = 1000$

Calculation of χ^2

O_i	1026	1107	997	966	1075	933	1107	972	964	853
E_i	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
$(O_i - E_i)^2$	676	11449	9	1156	5625	4489	11449	784	1296	21609

$$\chi^2 = \sum \left[\frac{(O_i - E_i)^2}{E_i}\right] = \frac{58542}{1000} = 58.542$$

The tabulated value of χ^2 at 5% level of significance for 9 d.f. is 16.919.

Conclusion: Since $\chi^2_{\text{calculated}} > \chi^2_{\text{tabulated}}$, H_0 is rejected *i.e.,* there is significant difference between the observed and theoretical frequencies. Therefore, the digits taken in the directory donot occur equally frequently.

Example 7. *When the first proof of 392 pages of a book of 1200 pages were read, the distribution of printing mistakes were found to be as follows:*

No. of mistakes in a page (x):	*0*	*1*	*2*	*3*	*4*	*5*	*6*
No. of pages (f):	*275*	*72*	*30*	*7*	*5*	*2*	*1*

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

Sol. Null Hypothesis, H_0 : Poisson distribution is a good fit to the data.

$$\text{Mean } (\lambda) = \frac{\sum fx}{\sum f} = \frac{189}{392} = 0.4821$$

The frequency of x mistakes per page is given by the poisson law as follows:

$$N(x) = N \cdot P(x)$$

$$= 392\left[\frac{e^{-0.4821}(0.4821)^x}{x!}\right] = \frac{242.05(0.4821)^x}{x!}\ ;\ 0 \le x \le 6$$

Under H_0, expected frequencies are,

N(0) = 242.05, N(1) = 116.69, N(2) = 28.13, N(3) = 4.52

N(4) = 0.54, N(5) = 0.052, N(6) = 0.0042

The χ^2-table is as follows:

Mistakes per page (x)	*Observed frequency (O_i)*	*Expected frequency (E_i) (correct to one place of decimal)*	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	275	242.1	1082.41	4.471
1	72	116.7	1998.09	17.121
2	30	28.1	3.61	0.128
3	7	4.5		
4	5 } 15	0.5 } 5.1		
5	2	0.1	98.01	19.217
6	1	0		
Total	392	392		40.937

$$\chi^2_{\text{cal.}} = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] = 40.937$$

d.f. = 7 – 1 – 1 – 3 = 2

One d.f. is lost because of linear constraint $\Sigma O_i = \Sigma E_i$. One d.f. is lost because the parameter λ has been estimated from the given data and is then used for computing the expected frequencies. 3 d.f. are lost because of grouping the last four expected cell frequencies which were less than 5.

Tabulated value of χ^2 for 2 d.f. at 5% level of significance is 5.991.

Conclusion : Since $\chi^2_{\text{cal.}} > \chi^2_{\text{tab.}}$, the null hypothesis is rejected at 5% level of significance. Hence, we conclude that poisson distribution is not a good fit to the given data.

Example 8. *Fit a Poisson distribution to the following data and best the goodness of fit :*

x :	0	1	2	3	4
f :	109	65	22	3	1

Sol. Null hypothesis, H_0: Poisson distribution is a good fit to the data.

$$\text{Mean } (\lambda) = \frac{\Sigma fx}{\Sigma f} = \frac{122}{200} = 0.61$$

$$N(x) = N \cdot P(x) = (200) \frac{e^{-.61} (.61)^x}{x!} = \frac{(108.67)(0.61)^x}{x!}$$

Under H_0, expected frequencies are

$$N(0) = 108.67 \approx 109, \quad N(1) = 66.29 \approx 66, \quad N(2) = 20.22 \approx 20$$

$$N(3) = 4.11 \approx 4, \quad N(4) = 0.63 \approx 1$$

The χ^2-table is as follows:

x	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	109	109	0	0
1	65	66	1	0.01515
2	22	20	4	0.2
3	3 } 4	4 } 5	1	0.2
4	1 }	1 }		
Total	200	200		0.41515

$$\chi^2_{\text{cal.}} = \sum \left[\frac{(O_i - E_i)^2}{E_i} \right] = 0.41515$$

d.f. = 5 – 1 – 1 – 1 = 2

Tabulated value of χ^2 for 2 d.f. at 5% level of significance is 5.991.

Conclusion: Since $\chi^2_{\text{cal.}} < \chi^2_{\text{tab.}}$, the null hypothesis H_0 is accepted at 5% level of significance. Hence we conclude that Poisson distribution is a good fit to the given data.

ASSIGNMENT

1. A sample analysis of examination results of 500 students, it was found that 220 students have failed, 170 have secured a third class, 90 have secured a second class and the rest, a first class. Do these figures support the general belief that above categories are in the ratio 4 : 3 : 2 : 1 respectively ? (The tabular value of χ^2 for d.f. 3 at 5% level of significance is 7.81).

[*U.P.T.U. (MBA) 2009*]

2. What is χ^2–test? [*G.B.T.U. 2010 ; G.B.T.U. MCA (C.O.) 2010*]

A die is thrown 90 times with the following results:

Face:	1	2	3	4	5	6	Total
Frequency:	10	12	16	14	18	20	90

Use χ^2-test to test whether these data are consistent with the hypothesis that die is unbiased.

Given $\chi^2_{0.05} = 11.07$ for 5 degrees of freedom. [*U.P.T.U. (MCA) 2007*]

3. A survey of 320 families with 5 children shows the following distribution:

No. of boys & girls:	5 boys & 0 girl	4 boys & 1 girl	3 boys & 2 girls	2 boys & 3 girls	1 boy & 4 girls	0 boy & 5 girls	Total
No. of families:	18	56	110	88	40	8	320

Given that values of χ^2 for 5 degrees of freedom are 11.1 and 15.1 at 0.05 and 0.01 significance level respectively, test the hypothesis that male and female births are equally probable.

(*G.B.T.U. 2010*)

4. A chemical extraction plant processes sea water to collect sodium chloride and magnesium. It is known that sea water contains sodium chloride, magnesium and other elements in the ratio 62 : 4 : 34. A sample of 200 tonnes of sea water has resulted in 130 tonnes of sodium chloride and 6 tonnes of magnesium. Are these data consistent with the known composition of sea water at 5% level of significance? (Given that the tabular value of χ^2 is 5.991 for 2 degree of freedom).

[*U.P.T.U. MCA (C.O.) 2008*]

5. The demand for a particular spare part in a factory was found to vary from day-to-day. In a sample study, the following information was obtained:

Days:	Mon	Tue	Wed	Thurs	Fri	Sat
No. of parts demanded:	1124	1125	1110	1120	1126	1115

Test the hypothesis that the number of parts demanded does not depend on the day of the week.

[**Given.** The values of chi-square significance at 5, 6, 7 d.f. are respectively 11.07, 12.59, 14.07 at 5% level of significance] (*G.B.T.U. 2011*)

6. The sales in a supermarket during a week are given below. Test the hypothesis that the sales donot depend on the day of the week using a significant level of 0.05.

Days	*Mon*	*Tue*	*Wed*	*Thurs*	*Fri*	*Sat*
Sales (in 1000 ₹)	65	54	60	56	71	84

7. 4 coins were tossed at a time and this operation is repeated 160 times. It is found that 4 heads occur 6 times, 3 heads occur 43 times, 2 heads occur 69 times, one head occur 34 times. Discuss whether the coin may be regarded as unbiased?

8. 200 digits are chosen at random from a set of tables. The frequencies of the digits were :

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

Use χ^2-test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the table, given that the value of χ^2 are respectively 16.9, 18.3 and 19.7 for 9, 10 and 11 degrees of freedom at 5% level of significance.

9. A genetical law says that children having one parent of blood group M and the other parent of blood group N will always be one of the three blood groups M, MN, N and that the average no. of children in these groups will be in the ratio 1 : 2 : 1. The report on an experiment states as follows:

"Of 162 children having one M parent and one N parent, 28.4% were found to be of group M, 42% of group MN and the rest of the group N." Do the data in the report conform to the expected genetic ratio 1 : 2 : 1?

10. Every clinical thermometer is classified into one of the four categories A, B, C and D on the basis of inspection and test. From past experience, it is knwon that thermometers produced by a certain manufacturer are distributed among the four categories in the following proportions:

Category:	A	B	C	D
Proportion:	0.87	0.09	0.03	0.01

A new lot of 1336 thermometers is submitted by the manufacturer for inspection and test and the following distribution into four categories results :

Category:	A	B	C	D
No. of the rmometers reported:	1188	91	47	10

Does this new lot of thermometers differ from the previous experience with regards to proportion of thermometers in each category?

11. Test for goodness of fit of a poisson distribution at 5% level of significance to the following frequency distribution:

(*i*)

x:	0	1	2	3	4	5	6	7	8
f:	52	151	130	102	45	12	5	1	2

[**Hint.** Group the last three frequencies]

(*ii*)

x:	0	1	2	3	4	5	6	7	8	9	10	11	12	13
f:	3	15	47	76	68	74	46	39	15	9	5	2	0	1

[**Hint.** Group the first two and last four frequencies]

(*iii*)

x:	0	1	2	3	4	5
f:	275	138	75	7	4	1

[**Hint.** Club the last three frequencies]

(*iv*)

x:	0	1	2	3	4
f:	419	352	154	56	19

12. (*i*) Fit a binomial distribution to the data and test for goodness of fit at 5% level of significance.

x:	0	1	2	3	4	5
y:	38	144	342	287	164	25

(*ii*) A random number table of 250 digits showed the following distribution of digits 0, 1, 2, ..., 9.

Digit:	0	1	2	3	4	5	6	7	8	9
Observed:	17	31	29	18	14	20	35	30	20	36
Frequency										
Expected:	25	25	25	25	25	25	25	25	25	25
Frequency										

Does the observed distribution differ significantly from expected distributions using a significance level of 0.01? Given that $\chi^2_{0.99}$ for 9 degrees of freedom is 21.7. [*G.B.T.U. MCA (SUM) 2010*]

Answers

1. No **2.** Yes

3. H_0 accepted at 1% level of significance and rejected at 5% level of significance

4. $\chi^2 = 1.025$, Yes **5.** H_0 accepted at 5% level **6.** H_0 accepted

7. Coin is unbiased **8.** H_0 accepted at 5% level **9.** H_0 accepted at 5% level

10. H_0 rejected at 5% level

11. (*i*) H_0 accepted at 5% level, (*ii*) H_0 accepted at 5% level, (*iii*) H_0 rejected at 5% level, not a good fit. (*iv*) H_0 accepted at 5% level

12. (*i*) H_0 accepted at 5% level, provides a good fit (*ii*) Yes.

3.88 χ^2 TEST AS A TEST OF INDEPENDENCE

With the help of χ^2 test, we can find whether or not, two attributes are associated. We take the null hypothesis that there is no association between the attributes under study, *i.e.*, **we assume that the two attributes are independent. If the calculated value of χ^2 is less than the table value** at a specified level (generally 5%) of significance, the hypothesis holds good, *i.e.*, **the attributes are independent** and do not bear any association. On the other hand, if the calculated value of χ^2 is greater than the table value at a specified level of significance, we say that the results of the experiment do not support the hypothesis. In other words, the attributes are associated. Thus a very useful application of χ^2 test is to investigate the relationship between trials or attributes which can be classified into two or more categories.

The sample data set out into two-way table, called **contingency table.**

Let us consider two attributes A and B divided into r classes $A_1, A_2, A_3, \ldots\ldots, A_r$ and B divided into s classes $B_1, B_2, B_3, \ldots\ldots, B_s$. If (A_i), (B_j) represents the number of persons possessing the attributes A_i, B_j respectively, ($i = 1, 2, \ldots\ldots, r, j = 1, 2, \ldots\ldots, s$) and $(A_i\ B_j)$ represent the number of persons possessing attributes A_i and B_j. Also we have $\sum_{i=1}^{r} A_i = \sum_{j=1}^{s} B_j = N$ where N is the total frequency. The contingency table for $r \times s$ is given as follows:

A / *B*	A_1	A_2	A_3	$\ldots A_r$	*Total*
B_1	(A_1B_1)	(A_2B_1)	(A_3B_1)	$\ldots\ldots(A_rB_1)$	B_1
B_2	(A_1B_2)	(A_2B_2)	(A_3B_2)	$\ldots\ldots(A_rB_2)$	B_2
B_3	(A_1B_3)	(A_2B_3)	(A_3B_3)	$\ldots\ldots(A_rB_3)$	B_3
.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....
B_s	(A_1B_s)	(A_2B_s)	(A_3B_s)	$\ldots\ldots(A_rB_s)$	(B_s)
Total	(A_1)	(A_2)	(A_3)	$\ldots\ldots(A_r)$	N

H_0 : Both the attributes are independent. *i.e.,* A and B are independent under the null hypothesis, we calculate the expected frequency as follows:

$$P(A_i) = \text{Probability that a person possesses the attribute } A_i = \frac{(A_i)}{N} \quad i = 1, 2, \ldots\ldots, r$$

$P(B_j)$ = Probability that a person possesses the attribute $B_j = \dfrac{(B_j)}{N}$

$P(A_iB_j)$ = Probability that a person possesses both attributes A_i and $B_j = \dfrac{(A_iB_j)}{N}$

If $(A_iB_j)_0$ is the expected number of persons possessing both the attributes A_i and B_j

$$(A_iB_j)_0 = NP(A_iB_j) = NP(A_i)(B_j)$$

$$= N\,\frac{(A_i)}{N}\,\frac{(B_j)}{N} = \frac{(A_i)(B_j)}{N} \qquad (\because \text{ A and B are independent})$$

Hence $$\boxed{\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\left[\frac{[(A_iB_j) - (A_iB_j)_0]^2}{(A_iB_j)_0}\right]}$$

which is distributed as a χ^2 variate with $(r-1)(s-1)$ degrees of freedom.

Note 1. For a 2×2 contingency table where the frequencies are $\dfrac{a/b}{c/d}$, χ^2 can be calculated from independent frequencies as $\chi^2 = \dfrac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}$.

Note 2. If the contingency table is not 2×2, then the above formula for calculating χ^2 can't be used. Hence, we have another formula for calculating the expected frequency $(A_iB_j)_0 = \dfrac{(A_i)(B_j)}{N}$

i.e., expected frequency in each cell is $= \dfrac{\text{Product of column total and row total}}{\text{whole total}}$.

Note 3. If $\dfrac{a|b}{c|d}$ is the 2×2 contingency table with two attributes, $Q = \dfrac{ad-bc}{ad+bc}$ is called the coefficient of association. If the attributes are independent then $\dfrac{a}{b} = \dfrac{c}{d}$.

Note 4. Yates's Correction. In a 2×2 table, if the frequencies of a cell is small, we make Yates's correction to make χ^2 continuous. Decrease by $\dfrac{1}{2}$ those cell frequencies which are greater than expected frequencies, and increase by $\dfrac{1}{2}$ those which are less than expectation. This will not affect the marginal columns. This correction is known as Yates's correction to continuity. After Yates's correction

$$\chi^2 = \frac{N\left(bc - ad - \frac{1}{2}N\right)^2}{(a+c)(b+d)(c+d)(a+b)} \quad \text{when} \quad ad - bc < 0$$

and

$$\chi^2 = \frac{N\left(ad - bc - \frac{1}{2}N\right)^2}{(a+c)(b+d)(c+d)(a+b)} \quad \text{when} \quad ad - bc > 0.$$

EXAMPLES

Example 1. *What are the expected frequencies of 2 × 2 contingency tables given below:*

(*i*)

a	b
c	d

(*ii*)

2	10
6	6

Sol. Observed frequencies → Expected frequencies

(*i*)

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$a+b+c+d=\text{N}$

→

$\frac{(a+c)(a+b)}{a+b+c+d}$	$\frac{(b+d)(a+b)}{a+b+c+d}$
$\frac{(a+c)(c+d)}{a+b+c+d}$	$\frac{(b+d)(c+d)}{a+b+c+d}$

Observed frequencies → Expected frequencies

(*ii*)

2	10	12
6	6	12
8	16	24

→

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

Example 2. *From the following table regarding the colour of eyes of father and son, test if the colour of son's eye is associated with that of the father.*

		Eye colour of son	
		Light	*Not light*
Eye colour of father	*Light*	*471*	*51*
	Not light	*148*	*230*

Sol. Null hypothesis H_0: The colour of son's eye is not associated with that of the father. *i.e.,* they are independent.

Under H_0, we calculate

$$\text{the expected frequency in each cell} = \frac{\text{Product of column total and row total}}{\text{whole total}}$$

Expected frequencies are:

Eye colour of son / *Eye colour of father*	*Light*	*Not light*	*Total*
Light	$\frac{619 \times 522}{900} = 359.02$	$\frac{289 \times 522}{900} = 167.62$	522
Not light	$\frac{619 \times 378}{900} = 259.98$	$\frac{289 \times 378}{900} = 121.38$	378
Total	619	289	900

$$\chi^2 = \frac{(471 - 359.02)^2}{359.02} + \frac{(51 - 167.62)^2}{167.62} + \frac{(148 - 259.98)^2}{259.98} + \frac{(230 - 121.38)^2}{121.38} = 261.498.$$

Tabulated value of χ^2 at 5% level for 1 d.f. is 3.841.

Conclusion. Since the calculated value of χ^2 > tabulated value of χ^2, H_0 is rejected. They are dependent *i.e.,* the colour of son's eye is associated with that of the father.

Example 3. *The following table gives the number of good and bad parts produced by each of the three shifts in a factory:*

	Good parts	*Bad parts*	*Total*
Day shift	*960*	*40*	*1000*
Evening shift	*940*	*50*	*990*
Night shift	*950*	*45*	*995*
Total	*2850*	*135*	*2985*

Test whether or not the production of bad parts is independent of the shift on which they were produced.

Sol. Null hypothesis H_0. The production of bad parts is independent of the shift on which they were produced. *i.e.,* the two attributes, production and shifts are independent.

Under H_0, $$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \left[\frac{[(A_i B_j)_0 - (A_i B_j)]^2}{(A_i B_j)_0} \right]$$

Calculation of expected frequencies

Let A and B be the two attributes namely production and shifts. A is divided into two classes A_1, A_2 and B is divided into three classes B_1, B_2, B_3.

$$(A_1B_1)_0 = \frac{(A_1)(B_1)}{N} = \frac{(2850) \times (1000)}{2985} = 954.77$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{(2850) \times (990)}{2985} = 945.226$$

$$(A_1B_3)_0 = \frac{(A_1)(B_3)}{N} = \frac{(2850) \times (995)}{2985} = 950$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{(135) \times (1000)}{2985} = 45.27$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{(135)\times(990)}{2985} = 44.773$$

$$(A_2B_3)_0 = \frac{(A_2)(B_3)}{N} = \frac{(135)\times(995)}{2985} = 45.$$

To calculate the value of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
(A_1B_1)	960	954.77	27.3529	0.02864
(A_1B_2)	940	945.226	27.3110	0.02889
(A_1B_3)	950	950	0	0
(A_2B_1)	40	45.27	27.7729	0.61349
(A_2B_2)	50	44.773	27.3215	0.61022
(A_2B_3)	45	45	0	0
				1.28126

The tabulated value of χ^2 at 5% level of significance for 2 degrees of freedom $(r-1)(s-1)$ is 5.991.

Conclusion: Since the calculated value of χ^2 is less than the tabulated value, we accept H_0. *i.e.,* the production of bad parts is independent of the shift on which they were produced.

Example 4. *From the following data, find whether hair colour and sex are associated.*

Colour / *Sex*	*Fair*	*Red*	*Medium*	*Dark*	*Black*	*Total*
Boys	*592*	*849*	*504*	*119*	*36*	*2100*
Girls	*544*	*677*	*451*	*97*	*14*	*1783*
Total	*1136*	*1526*	*955*	*216*	*50*	*3883*

Sol. Null hypothesis H_0. The two attributes hair colour and sex are not associated. *i.e.,* they are independent.

Let A and B be the attributes hair colour and sex respectively. A is divides into 5 classes ($r = 5$). B is divided into 2 classes ($s = 2$).

$\therefore$ Degrees of freedom $= (r-1)(s-1) = (5-1)(2-1) = 4$

Under H_0, we calculate $\chi^2 = \sum_{i=1}^{5}\sum_{j=1}^{2} \frac{[(A_iB_j)_0 - (A_iB_j)]^2}{(A_iB_j)_0}$

To calculate the expected frequency $(A_i B_j)_0$ as follows:

$$(A_1B_1)_0 = \frac{(A_1)(B_1)}{N} = \frac{1136\times 2100}{3883} = 614.37$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{1136 \times 1783}{3883} = 521.629$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{1526 \times 2100}{3883} = 852.289$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{1526 \times 1783}{3883} = 700.71$$

$$(A_3B_1)_0 = \frac{(A_3)(B_1)}{N} = \frac{955 \times 2100}{3883} = 516.482$$

$$(A_3B_2)_0 = \frac{(A_3)(B_2)}{N} = \frac{955 \times 1783}{3883} = 483.517$$

$$(A_4B_1)_0 = \frac{(A_4)(B_1)}{N} = \frac{216 \times 2100}{3883} = 116.816$$

$$(A_4B_2)_0 = \frac{(A_4)(B_2)}{N} = \frac{216 \times 1783}{3883} = 99.183$$

$$(A_5B_1)_0 = \frac{(A_5)(B_1)}{N} = \frac{50 \times 2100}{3883} = 27.04$$

$$(A_5B_2)_0 = \frac{(A_5)(B_2)}{N} = \frac{50 \times 1783}{3883} = 22.959$$

Calculation of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
A_1B_1	592	614.37	500.416	0.8145
A_1B_2	544	521.629	500.462	0.959
A_2B_1	849	852.289	10.8175	0.0127
A_2B_2	677	700.71	562.1641	0.8023
A_3B_1	504	516.482	155.800	0.3016
A_3B_2	451	438.517	155.825	0.3553
A_4B_1	119	116.816	4.7698	0.0408
A_4B_2	97	99.183	4.7654	0.0480
A_5B_1	36	27.04	80.2816	2.9689
A_5B_2	14	22.959	80.2636	3.495
				9.79975

$$\chi^2_{\text{cal.}} = 9.799.$$

Tabular value of χ^2 at 5% level of significance for 4 d.f. is 9.488.

Conclusion: Since the calculated value of χ^2 < tabulated value, H_0 is rejected. *i.e.,* the two attributes are not independent. *i.e.,* the hair colour and sex are associated.

Example 5. *Can vaccination be regarded as preventive measure of small pox as evidenced by the following data of 1482 persons exposed to small pox in a locality. 368 in all were attacked of these 1482 persons and 343 were vaccinated and of these only 35 were attacked.*

Sol. For the given data we form the contingency table. Let the two attributes be vaccination and exposed to small pox. Each attributes is divided into two classes.

Vaccination A / *Disease small pox B*	*Vaccinated*	*Not*	*Total*
Attacked	35	333	368
Not	308	806	1114
Total	343	1139	1482

Null hypothesis H_0. The two attributes are independent *i.e.,* vaccination can't be regarded as preventive measure of small pox.

Degrees of freedom $\nu = (r-1)(s-1) = (2-1)(2-1) = 1$

Under H_0, $\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{[(A_iB_j)_0 - (A_iB_j)]^2}{(A_iB_j)_0}$

Calculation of expected frequency

$$(A_1B_1)_0 = \frac{(A_1)(B_1)}{N} = \frac{343 \times 368}{1482} = 85.1713$$

$$(A_1B_2)_0 = \frac{(A_1)(B_2)}{N} = \frac{343 \times 1114}{1482} = 257.828$$

$$(A_2B_1)_0 = \frac{(A_2)(B_1)}{N} = \frac{1139 \times 368}{1482} = 282.828$$

$$(A_2B_2)_0 = \frac{(A_2)(B_2)}{N} = \frac{1139 \times 1114}{1482} = 856.171$$

Calculation of χ^2

Class	O_i	E_i	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
(A_1B_1)	35	85.1713	2517.159	29.554
(A_1B_2)	308	257.828	2517.229	8.1728
(A_2B_1)	333	282.828	2517.2295	7.5592
(A_2B_2)	806	856.171	2517.1292	2.9399
				48.2261

Calculated value of $\chi^2 = 48.2261$.

Tabulated value of χ^2 at 5% level of significance for 1 d.f. is 3.841.

Conclusion. Since the calculated value of χ^2 > tabulated value, H_0 is rejected.

i.e., the two attributes are not independent. *i.e.,* the vaccination can be regarded as preventive measure of small pox.

Example 6. *To test the effectiveness of inoculation against cholera, the following table was obtained:*

	Attacked	*Not attacked*	*Total*
Inoculated	*30*	*160*	*190*
Not inoculated	*140*	*460*	*600*
Total	*170*	*620*	*790*

(The figures represent the number of persons.)

Use χ^2-test to defend or refute the statement that the inoculation prevents attack from cholera. (U.P.T.U. 2009)

Sol. Null hypothesis H_0: The inoculation does not prevent attack from cholera.

Under H_0, we calculate the expected frequencies as:

	Attacked	*Not attacked*
Inoculated	$\frac{190 \times 170}{790} = 40.886$	$\frac{190 \times 620}{790} = 149.11$
Not inoculated	$\frac{600 \times 170}{790} = 129.11$	$\frac{600 \times 620}{790} = 470.89$

Calculation of χ^2

O_i	30	160	140	460
E_i	40.886	149.11	129.11	470.89
$\frac{(O_i - E_i)^2}{E_i}$	2.898	0.795	0.918	0.252

$$\chi^2_{\text{cal.}} = \sum \frac{(O_i - E_i)^2}{E_i} = 4.863$$

Tabulated value of χ^2 at 5% level of significance for 1 d.f. is 3.841.

Conclusion: Since $\chi^2_{\text{cal.}} > \chi^2_{\text{tab.}}$ at 5% level of significance, null hypothesis H_0 is rejected. Hence we defend the statement that inoculation prevents attack from cholera.

ASSIGNMENT

1. In a locality 100 persons were randomly selected and asked about their educational achievements. The results are given below:

		Education		
		Middle	*High school*	*College*
Sex	Male	10	15	25
	Female	25	10	15

Based on this information can you say that the education depends on sex.

2. The following data is collected on two characters:

	Smokers	*Non smokers*
Literate	83	57
Illiterate	45	68

Based on this information can you say that there is no relation between habit of smoking and literacy.

3. 500 students at school were graded according to their intelligences and economic conditions of their homes. Examine whether there is any association between economic condition and intelligence, from the following data:

Economic conditions	*Intelligence*	
	Good	*Bad*
Rich	85	75
Poor	165	175

4. In an experiment on the immunisation of goats from anthrox, the following results were obtained. Derive your inferences on the efficiency of the vaccine.

	Died anthrox	*Survived*
Inoculated with vaccine	2	10
Not inoculated	6	6

5. By using χ^2-test, find out whether there is any association between income level and type of schooling:

Income	*Public School*	*Govt. School*
Low	200	400
High	1000	400

(Given for degree of freedom 1, $\chi^2_{0.05} = 3.84$) [*U.P.T.U. 2008, G.B.T.U. (MBA) 2011*]

6. Examine by any suitable method, whether the nature of area is related to voting preference in the election for which the data are tabulated below:

Votes for Area	*A*	*B*	*Total*
Rural	620	480	1100
Urban	380	520	900
Total	1000	1000	2000

(*U.P.T.U. 2006*)

7. The groups of 100 people each were taken for testing the use of a vaccine. 15 persons contracted the disease out of the inoculated persons, while 25 contracted the disease in the other group. Test the efficiency of the vaccine using Chi-square test. (The value of χ^2 for one degree of freedom at 5% level of significance is 3.84).

Answers

1. Yes **2.** No **3.** No **4.** Not effective
5. Yes **6.** Yes **7.** Not associated.

TEST YOUR KNOWLEDGE

1. The fourth moment about the mean of a frequency distribution is 24. What must be the value of standard deviation in order that the distribution be platykurtic? (*M.T.U. 2012*)

2. Two events A and B have probabilities 0.25 and 0.50 respectively. The probability that both events A and B occurs in 0.14. Find the probability that neither A nor B occurs. (*M.T.U. 2013*)

[**Hint.** $P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B)$]

3. (*i*) Define the coefficients of skewness and kurtosis. (*U.P.T.U. 2014*)
(*ii*) Define skewness, coefficient of skewness, kurtosis and coefficient of kurtosis. (*M.T.U. 2013*)
(*iii*) Write a short note on skewness. [(*M.T.U.* (*B.Pharma*) *2011*]
(*iv*) What is meant by skewness? How is it measured? [(*M.T.U.* (*MBA*) *2012*]

4. What is the total probability theorem? (*U.P.T.U. 2014*)

5. The first four central moments of a distribution are 0, 2.5, 0.7 and 18.75. Comment on the kurtosis of the distribution. (*M.T.U. 2013*)

6. Find the moment generating function of Poisson distribution. (*M.T.U. 2013*)

7. Find the parameters p and q of the binomial distribution whose mean is 9 and variance is $\frac{9}{4}$.

(*M.T.U. 2012*)

8. If the sum of the mean and variance of a binomial distribution of 5 trials is $\frac{9}{5}$, find $P(X \geq 1)$.

(*M.T.U. 2013*)

9. It has been found that 2% of the tools produced by a certain machine are defective. What is the probability that in a shipment of 400 such tools, 3 or more will be defective? (*M.T.U. 2013*)

10. If $P(X = 0) = P(X = 1) = k$ in a Poisson distribution, then what is k?

11. For a Poisson variate X if $P(X = 1) = P(X = 2)$, then find $P(X = 4)$.

12. Find the total area under the curve of p.d.f. of a normal curve.

13. If for a Poisson distribution, $P(2) = P(3)$, then what is its probability function?

14. Find the parameters of a binomial distribution with mean = 8 and variance = 4.

15. If X is a normal variate with mean 30 and S.D. 5, find the probabilities that
(*i*) $26 \le X \le 40$ (*ii*) $X \ge 45$ and (*iii*) $|X - 30| > 5$

Answers

2. 0.39 **5.** $\beta_2 = 3$, Mesokurtic **6.** $M_x(t) = e^{\lambda(e^t - 1)}$ **7.** $q = \frac{1}{4}, p = \frac{3}{4}$

8. 0.67232 **9.** 0.9862 **10.** $\frac{1}{e}$ **11.** $\frac{2}{3e^2}$

12. 1 **13.** $\frac{e^{-3}(3)^x}{x!}$ **14.** $n = 16, p = \frac{1}{2}$

15. (*i*) 0.7653 (*ii*) 0.00135 (*iii*) 0.3174.