7.6 MOMENTS, SKEWNESS AND KURTOSIS

So far we have studied two aspects of frequency distribution viz average and dispersion. In order to study few more aspects such as symmetry, shape of the frequency distribution (ie frequency curve) more general type of descriptive measure called moments are useful.

7.6.1 Moments

The r^{th} moment about the mean $\bar{x}$ (or central moments) of a frequency distribution is denoted by μ_r and is given by

$$\mu_r = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^r \quad \text{where } N = \sum_{i=1}^{n} f_i \text{ and } \bar{x} = \text{M.W. of the distribution}$$

putting $r = 0, 1, 2, 3, 4$ etc. we get

$\mu_0 = \frac{1}{N}\sum_{i=1}^{n} f_i = 1$

$\mu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x}) = \bar{x} - \bar{x} = 0$ gives first moment of distribution about the mean

$\mu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^2 = \sigma^2 =$ variance = (hence) gives second moment of the distribution about the mean

$\mu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^3$ gives 3^{rd} moment of the distribution about the mean

$\mu_4 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - \bar{x})^4$ gives 4^{th} moment of the distribution about the mean and so on

Since actual evaluation of r^{th} moment μ_r about the mean $\bar{x}$ is numerically complicated, we find r^{th} moments ν_r about arbitrary arbitrary convenient number A of the distribution with much less calculation. Refer to $\S$ 7.5. The r^{th} moment ν_r about arbitrary number (assumed mean) is given by

$\nu_r = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^r$ can be seen as putting $r = 0, 1, 2$ etc. that

$\nu_0 = 1$

$\nu_1 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A) = \frac{1}{N}\sum f_i x_i - \frac{\sum f_i}{N} A = \bar{x} - A$

$\nu_2 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^2 = \sigma^2$ the mean square deviation

$\nu_3 = \frac{1}{N}\sum_{i=1}^{n} f_i(x_i - A)^3$ etc.

7.6.2 Relation Between μ_r and ν_r

We know by definition of μ_r

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

$= \frac{1}{N}\sum f_i(x_i - A + A - \bar{x})^r$

$= \frac{1}{N}\sum f_i\left[(x_i - A) - (\bar{x} - A)\right]^r$

Let $x_i - A = d_i$ and hence $\bar{x} - A = \frac{\sum f_i x_i}{N} - \frac{\sum f_i A}{N} = \frac{\sum f_i x_i}{N} - A = \bar{x} - A$

$\mu_r = \frac{1}{N}\sum f_i(d_i - \bar{d})^r$

Thus

On expanding $(d_i - \bar{d})^r$ binomially we obtain

$\mu_r = \frac{1}{N}\sum f_i\left[d_i^r - {}^rC_1 d_i^{r-1}\bar{d} + {}^rC_2 d_i^{r-2}\bar{d}^2 - \ldots + (-1)^r \bar{d}^r\right]$

7.6 MOMENTS, SKEWNESS AND KURTOSIS

So far we have studied tow aspects of frequency distribution viz. average and dispersion. In order to study few more aspects such as symmetry, shape of the frequency distribution (or frequency curve), more general type of descriptive measure called moments are useful.

7.6.1 Moments

The r^{th} moment about the mean $\bar{x}$ (or central moments) of a frequency distribution is denoted by μ_r and is given by,

$$\mu_r = \frac{1}{N}\sum f\left(x-\bar{x}\right)^r, \quad \text{where, } N = \sum f \text{ and } \bar{x} = \text{A.M. of the distribution.}$$

Putting r = 0, 1, 2, 3, 4 etc., we get

$$\mu_0 = \frac{1}{N}\sum f = 1,$$

$$\mu_1 = \frac{1}{N}\sum f\left(x-\bar{x}\right) = \bar{x}-\bar{x} = 0 \text{ gives first moment of distribution about the mean.}$$

$$\mu_2 = \frac{1}{N}\sum f\left(x-\bar{x}\right)^2 = \sigma^2 = \text{Variance} = \text{Var}(x) \text{ gives second moment of the distribution about the mean.}$$

$$\mu_3 = \frac{1}{N}\sum f\left(x-\bar{x}\right)^3 \text{ gives } 3^{rd} \text{ moment of the distribution about the mean.}$$

$$\mu_4 = \frac{1}{N}\sum f\left(x-\bar{x}\right)^4 \text{ gives } 4^{th} \text{ moment of the distribution about the mean and so on}$$

Since actual evaluation of r^{th} moment μ_r about the mean $\bar{x}$ is numerically complicated, we find r^{th} moments viz. μ_r' about arbitrary convenient number A of the distribution with much less calculation (Refer Art. 7.4). The r^{th} moment μ_r' about arbitrary number (assumed mean) is given by

$$\mu_r' = \frac{1}{N}\sum f(x-A)^r, \text{ it can be seen on putting } r = 0, 1, 2, \ldots \text{etc. that}$$

$$\mu_0' = 1$$

$$\mu_1' = \frac{1}{N}\sum f(x-A) = \frac{1}{N}\sum fx - \frac{\sum f}{N}A = \bar{x} - A$$

$$\mu_2' = \frac{1}{N}\sum f(x-A)^2 = S^2 \text{ the mean square deviation}$$

$$\mu_3' = \frac{1}{N}\sum f(x-A)^3 \text{ etc.}$$

7.6.2 Relation Between μ_r and μ_r'

We know by definition of μ_r,

$$\mu_r = \frac{1}{N}\sum f\left(x-\bar{x}\right)^r$$

$$= \frac{1}{N}\sum f\left(x-A+A-\bar{x}\right)^r$$

Let $d = x - A$ and hence $\bar{d} = \frac{\sum fd}{N} = \frac{\sum fx}{N} - \frac{A.\sum f}{N}$ or $\bar{d} = \bar{x} - A = \mu_1'$

Thus, $$\mu_r = \frac{1}{N}\sum f\left(d-\bar{d}\right)^r$$

On expanding $\left(d-\bar{d}\right)^r$ binomially, we obtain

$$\mu_r = \frac{1}{N}\sum f\left(d^r - {}^rC_1\, d^{r-1}\,\bar{d} + {}^rC_2\, d^{r-2}\,\bar{d}^2 + \ldots + (-1)^r\,\bar{d}^r\right)$$

where $^rC_1 = r,\ ^rC_2 = \frac{r(r-1)}{2!},\ ^rC_3 = \frac{r(r-1)(r-2)}{3!}$ etc.

$\therefore \quad \mu_r = \frac{1}{N}\sum fd^r - {}^rC_1 \frac{1}{N}\sum fd^{r-1}\,\bar{d} + {}^rC_2 \frac{1}{N}\sum fd^{r-2}\,\bar{d}^2 - {}^rC_3 \frac{1}{N}\sum fd^{r-3}\,\bar{d}^3 \ldots (-1)^r \frac{1}{N}\sum f\bar{d}^r$

Using, $\frac{1}{N}\sum f d^r = \mu_r'$ and $\bar{d} = \mu_1'$, relation between μ_r and μ_r' is

$\therefore \quad \mu_r = \mu_r' - {}^rC_1\,\mu_{r-1}'\,\mu_1' + {}^rC_2\,\mu_{r-2}'(\mu_1')^2 + \ldots + (-1)^r\,(\mu_r')^r$

We have already seen that $\mu_0 = 1,\ \mu_1 = 0$.

Putting r = 2, 3, 4 etc., we get

$$\mu_2 = \mu_2' - {}^2C_1\,\mu_1'\,\mu_1' + {}^2C_2\,\mu_0'\,\mu_1' \qquad \ldots (A)$$

$$= \mu_2' - 2\mu_1'^2 + \mu_1'^2 = \mu_1' - \mu_1'^2$$

$$\mu_3 = \mu_3' - {}^2C_1\,\mu_2'\,\mu_1' + {}^3C_2\,\mu_1'\,\mu_1'^2 - \mu_1'^3$$

$$= \mu_3' - {}^3C_1\,\mu_2' + {}^3C_2\,\mu_1'\,\mu_1'^2 - \mu_1'^3 \qquad \ldots (B)$$

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

$$\mu_4 = \mu_4' - {}^4C_1\,\mu_3'\,\mu_1' + {}^4C_2\,\mu_2'\,\mu_1'^2 - {}^4C_2\,\mu_1'\,\mu_1'^3 + {}^4C_4\,\mu_1'^4$$

$$= \mu_4' - 4\mu_3'\,\mu_1' + 6\mu_2'\,\mu_1'^2 - 4\mu_1'^4 + \mu_1'^4 \qquad \ldots (C)$$

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

The moments of higher order μ_5, μ_6 etc. can be similarly expressed.

Note : From the above relations (A), (B) and (C) we note the following :

While dealing with data presented in group frequency distribution, to reduce the calculation of μ_r' further, we use the following procedure.

(i) Sum of the coefficients on R.H.S. of each of the relations is zero.

(ii) First term is the expression is positive and alternative terms are negative.

(iii) The last term in the expression of μ_r is $(\mu_1')^r$.

Put $u = \frac{x-A}{h}$ where, h is taken generally width of class interval, then the expressions for the moments μ_r' about any arbitrary point A (assumed or convenient mean) are given by

$$\mu_r' = \frac{1}{N}\sum f(x-A)^r, \qquad N = \sum f$$

$$= \frac{1}{N}\sum f(hu)^r$$

$\therefore \quad \mu_r' = h^r \frac{1}{N}\sum f u^r,\ r = 1, 2, 3, \ldots$ ... (D)

We know that first moment about mean $\bar{x}$ is $\mu_1 = 0$. The second, third and fourth moments about the mean $\bar{x}$ are obtained using relations (A), (B) and (C).

Note : r^{th} moment about the mean $\bar{x}$ (or central moment) of individual observation $(x_1, x_2, \ldots, x_n)$ denoted by μ_r and is given by

$$\mu_r = \frac{1}{n}\sum (x-\bar{x})^r, \qquad r = 0, 1, 2, 3, 4$$

r^{th} moment about the arbitrary mean A of individual observation $(x_1, x_2, \ldots, x_n)$ is denoted by μ'_r and is given by

$$\mu_r' = \frac{1}{n}\sum (x-A)^r \qquad \ldots (E)$$

Remark :

(i) Change of Origin Property : The rth moment about the mean $\bar{x}$ (central moments) are invariant to the change of origin. If $u = x - A$ then (μ_r of u) = (μ_r of x).

(ii) Change of Origin and Scale : If $u = \frac{x - A}{h}$ then $(\mu_r \text{ of } u) = \frac{1}{h^r} (\mu_r \text{ of } x)$.

Sheppard's Correction for Moments : In case of grouped frequency distribution, we take mid-values of class intervals to represent the class interval. This involves some error in calculation of moments. W.F. Sheppard suggested some corrective formulae :

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{1}{12} h^2$$

$$\mu_3 = \mu_3$$

$$\mu_4 \text{ (corrected)} = \mu_4 - \frac{1}{2} h^2 \mu_2 + \frac{7}{240} h^4, \quad \text{where, h is the width of class interval.}$$

7.6.3 Skewness

A frequency distribution is symmetric above a value 'a' (say); if the corresponding frequency curve is symmetric about a. For symmetric frequency curve, the point 'a' turns out to be arithmetic mean, mode as well as median refer Fig. 7.3(a).

Skewness signifies departure from symmetry (or a lacks of symmetry). We study skewness to have an idea about the shape of the curve which we draw with the given data.

If the frequency curve stretches to the right as in Fig. 7.3 (b) i.e. the mean is to the right of the mode then the distribution is right skewed or is said to have positive skewness. If the curve stretches to left or mode is to the right of the mean then the distribution is said to have negative skewness refer Fig. 7.3 (c).

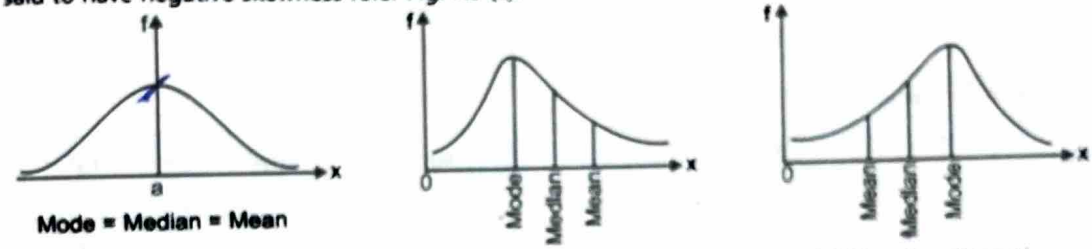


Fig. 7.3

The different measures of skewness are :

(i) $\text{Skewness} = \frac{3 (\text{Mean} - \text{Median})}{\text{Standard deviation}}$

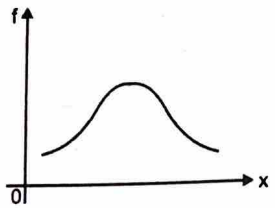
(ii) Coefficient of skewness, $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$

The distribution is positively skewed if skewness or coefficient of skewness β_1 is positive. If coefficient of skewness is negative the distribution is negatively skewed. It is also clear from (i). Now to decide the sign of β_1, We introduce the parameter $\gamma_1 = \pm \sqrt{\beta_1}$. Now $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$ where both numerator and denominator are positive. To decide the sign of γ_1 we associate with sign of μ_3. If μ_3 is negative, we take γ_1 as negative and μ_3 is positive, we take γ_1 as positive. In short distribution is positively skewed if μ_3 is positive and it is negatively skewed if μ_3 is negative.

7.6.4 Kurtosis

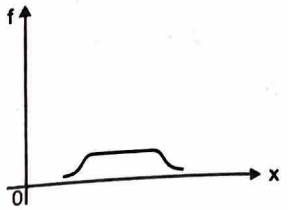
To get complete idea of the distribution in addition to the knowledge of mean, dispersion and skewness, we should have an idea of the flatness or peakedness of the curve. It is measured by the coefficient β_2 given by,

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \text{ and } \gamma_2 = \beta_2 - 3$$



(a) Mesokurtic

(b) Leptokurtic



(c) Platykurtic

Fig. 7.4

The curve of Fig. 7.4 (a) which is neither flat nor peaked is called the normal curve or Mesokurtic curve. $\gamma_2 = \beta_2 - 3$ gives the excess of kurtosis. For a normal distribution, $\beta_2 = 3$ and the excess is zero. The curve of Fig. 7.4 (c) which is flatter than the normal curve is called Platykurtic and that of Fig. 7.4 (b) which is more peaked is called Leptokurtic. For Platykurtic curves $\beta_2 < 3$, for Leptokurtic curves $\beta_2 > 3$.

Ex. 7 : *Calculate the first four moments about the mean of the given distribution. Also find β_1 and β_2.*

(Dec. 2010, 2011; May 2015)

x	2.0	2.5	3.0	3.5	4.0	4.5	5.0
f	4	36	60	90	70	40	10

Sol. : Taking A = 3.5, h = 0.5 and $u = \dfrac{x - 3.5}{0.5}$

We prepare the table for calculating μ_1', μ_2', μ_3' and μ_4'.

x	f	$u = \dfrac{x-3.5}{0.5}$	fu	fu^2	fu^3	fu^4
2.0	4	– 3	– 12	36	– 108	342
2.5	36	– 2	– 72	144	– 288	576
3.0	60	– 1	– 60	60	– 60	60
3.5	90	0	0	0	0	0
4.0	70	1	70	70	70	70
4.5	40	2	80	160	320	640
5.0	10	3	30	90	270	810
Total	**$\Sigma f = 310$**		**$\Sigma fu = 36$**	**$\Sigma fu^2 = 560$**	**$\Sigma fu^3 = 204$**	**$\Sigma fu^4 = 2480$**

For moments about arbitrary mean A = 3.5, we use formula (D).

$$\mu_r' = h^r \frac{\Sigma fu^r}{\Sigma f}$$

$$\mu_1' = h \frac{\Sigma fu}{\Sigma f} = (0.5) \frac{36}{310} = 0.058064$$

$$\mu_2' = h^2 \frac{\Sigma fu^2}{\Sigma f} = (0.5)^2 \frac{560}{310} = 0.451612$$

$$\mu_3' = h^3 \frac{\Sigma fu^3}{\Sigma f} = (0.5)^3 \frac{204}{310} = 0.082259$$

$$\mu_4' = h^4 \frac{\Sigma fu^4}{\Sigma f} = (0.5)^4 \frac{2480}{310} = 0.5$$

Using relations A, B, C of section 7.5 central moments are

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = (0.451612) - (0.058064)^2 = 0.44824$$

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

$$= (0.082259) - 3\,(0.451612)\,(0.058064) + 2\,(0.058064)^3$$

$$= 0.082259 - 0.078668 + 0.0003916$$

$$= 3.9826 \times 10^{-3} = 0.0039826$$

$$\mu_4 = \mu_4' - 4\,\mu_3'\,\mu_1' + 6\,\mu_4'\,(\mu_1')^2 - 3\,(\mu_1')^4$$

$$= 0.5 - 0.01911 + 0.009136 - 0.0000341$$

$$= 0.48999$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0.0000159}{0.0900599} = 1.76549$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{0.48999}{0.20092} = 2.43874$$

Ex. 8 : *For the following distribution, find (i) first 4 moments about the mean, (ii) β_1 and β_2, (iii) arithmetic mean, (iv) standard deviation.* ***(Dec. 2014)***

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

Sol. : Let $u = \frac{x - 3.5}{0.5}$, $A = 3.5$, $h = 0.5$.

x	f	u	fu	fu^2	fu^3	fu^4
2	5	– 3	– 15	45	– 135	405
2.5	38	– 2	– 76	152	– 304	608
3	65	– 1	– 65	65	– 65	65
3.5	92	0	0	0	0	0
4	70	1	70	70	70	70
4.5	40	2	80	160	320	640
5	10	3	30	90	270	810
Total	**320**	**–**	**24**	**582**	**156**	**2598**

(i) Moment about the mean M.

When assumed mean is A = 3.5 and using (D), we have

$$\mu_1' = h\,\frac{\sum fu}{\sum f} = 0.5\left(\frac{24}{320}\right) = 0.0375$$

$$\mu_2' = h^2\,\frac{\sum fu^2}{\sum f} = (0.5)^2\left(\frac{582}{320}\right) = 0.4546$$

$$\mu_3' = h^3 \frac{\sum fu^3}{\sum f} = (0.5)^3 \left(\frac{156}{320}\right) = 0.0609$$

$$\mu_4' = h^4 \frac{\sum fu^4}{\sum f} = (0.5)^4 \left(\frac{2598}{320}\right) = 0.5074$$

Using results (A), (B), (C), we have four moments about the mean M

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = (0.4546) - (0.0375)^2 = 0.453$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3$$
$$= (0.0609) - 3(0.4546)(0.0375) + 2(0.0375)^3$$
$$= 0.0600$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_2' + 6(\mu_1')^2\mu_2' - 3(\mu_1')^4$$
$$= (0.5074) - 4(0.0609)(0.0375) + 6(0.0375)^2(0.4546) - 3(0.0375)^4$$
$$= 0.502$$

(ii) By definition of β_1 and β_2, we have

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.0600)^2}{(0.453)^3} = 0.0387$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{0.502}{(0.453)^2} = 2.4463$$

Since $\beta_2 < 3$, the distribution is platy kurtic i.e. it is flatter than the normal distribution.

(iii) Arithmetic Mean : Using result (B), we have

$$A = \frac{\sum fu}{\sum f} = \frac{24}{320} = 0.075$$

(iv) Standard Deviation :

$$\sigma^2 = h^2 \left\{\frac{\sum fu^2}{\sum f} - \left(\frac{\sum fu}{\sum f}\right)^2\right\}$$
$$= (0.5)^2 \left\{\frac{582}{320} - \left(\frac{24}{320}\right)^2\right\} = 0.453$$

$$\sigma = 0.673$$

Ex. 9 : *Calculate the first four moments about the mean of the given distribution. Find β_1, β_2 and comment on skewness and kurtosis.*

x	*5*	*7*	*13*	*24*	*29*	*36*	*40*	*45*	*50*
f	*4*	*6*	*17*	*25*	*18*	*12*	*9*	*3*	*2*

Sol.: Taking A = 24, d = x – 24

x	f	d = x – 24	fd	fd^2	fd^3	fd^4
5	4	–19	–76	1444	–27436	521284
7	6	–17	–102	1734	–29478	501126
13	17	–11	–187	2057	–22627	248897
24	25	0	0	0	0	0
29	18	5	90	450	2250	11250
36	12	12	144	1728	20736	248832
40	9	16	144	2304	36864	589924
45	3	21	63	1323	27783	583443
50	2	26	52	1352	35152	913952
Total	**$\sum f = 96$**	–	**$\sum fd = 128$**	**$\sum fd^2 = 12392$**	**$\sum fd^3 = 43244$**	**$\sum fd^4 = 3618608$**

$$\mu_1' = \frac{\sum fd}{\sum f} = 1.33$$

$$\mu_2' = \frac{\sum fd^2}{\sum f} = 129.08$$

$$\mu_3' = \frac{\sum fd^3}{\sum f} = 450.46$$

$$\mu_4' = \frac{\sum fd^4}{\sum f} = 37693.83$$

$$\mu_1 = 0,$$

$$\mu_2 = \mu_2' - \mu_1'^2 = 127.31$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3 = -59.86$$

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

$$= 36657.97$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^2} = 0.001736$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = 2.262$$

β_1 is very small, so the curve is symmetrical, skewness is negative.

$\beta_2 = 2.262 < 3$, curve is playticurtic type.

Ex. 10 : *Compute the first four central moments for the following frequencies :*

No. of Jobs Completed	0-10	10-20	20-30	30-40	40-50
No. of Workers	6	26	47	15	6

Sol. :

Class	Mid-Pts. (x)	Freq. (f)	$u = \frac{x-25}{10}$	fu	fu^2	fu^3	fu^4
0-10	5	6	– 2	– 12	24	– 48	96
10-20	15	26	– 1	– 26	26	– 26	26
20-30	25	47	0	0	0	0	0
30-40	35	15	1	15	15	15	15
40-50	45	6	2	12	24	48	96
Total	–	**100**	–	**– 11**	**89**	**– 11**	**233**

For moments about arbitrary mean A = 25 we use the formula (D)

$$\mu_r' = h^r \frac{\sum fu^r}{\sum f}$$

$\therefore$

$$\mu_1' = h \frac{\sum fu}{\sum f} = 10\left(\frac{-11}{100}\right) = 10\,(-0.11) = -1.1$$

$$\mu_2' = h^2 \frac{\sum fu^2}{\sum f} = (10)^2\left(\frac{89}{100}\right) = 100\,(0.89) = 89$$

$$\mu_3' = h^3 \frac{\sum fu^3}{\sum f} = (10)^3\left(\frac{-11}{100}\right) = (1000)\,(-0.11) = -110$$

$$\mu_4' = h^4 \frac{\sum fu^4}{\sum f} = (10)^4 \left(\frac{233}{100}\right) = (10)^4 (2.33) = 23300$$

Using relations A, B, C of article 7.5 central moments are

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 89 - (-1.1)^2 = 89 - (1.21) = 87.79$$

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

$$= -110 - 3(89)(-1.1) + 2(-1.1)^3$$

$$= -110 + 293.7 + 2(-1.331)$$

$$= -110 + 293.7 - 2.662$$

$$= 181.038$$

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

$$= 23300 - 4(-110)(-1.1) + 6(89)(-1.1)^2 - 3(-1.1)^4$$

$$= 23300 - 484 + 646.14 - 4.3923$$

$$= 23457.7477$$

Ex. 11 : *If $\sum f = 27$, $\sum fx = 91$, $\sum fx^2 = 359$, $\sum fx^3 = 1567$, $\sum fx^4 = 7343$. Find first four moments about origin. Find A.M., S.D., μ_3 and μ_4. Find coefficients of skewness and kurtosis. Comment on skewness and kurtosis.*

Sol.:

$$\mu_1' = \frac{\sum fx}{\sum f} = \frac{91}{27} = 3.37$$

$$\mu_2' = \frac{\sum fx^2}{\sum f} = \frac{359}{27} = 13.297$$

$$\mu_3' = \frac{\sum fx^3}{\sum f} = \frac{1567}{27} = 58.04$$

$$\mu_4' = \frac{\sum fx^4}{\sum f} = \frac{7343}{27} = 271.963$$

$$\text{A.M.} = \mu_1' = 3.37$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 13.297 - (3.37)^2 = 1.94$$

$$\text{S.D.} = \sqrt{\mu_2} = 1.3928$$

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

$$= 58.04 - 3 \times 3.37 \times 13.297 + 2 \times (3.37)^3$$

$$= 58.04 - 134.43267 + 38.2727 \times 2$$

$$= 58.04 - 134.43267 + 76.5455 = 0.15283$$

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

$$= 271.963 - 4 \times 58.04 \times 3.37 + 6(13.37)^2 \times 13.297 - 3 \times (3.37)^4$$

$$= 8.7311$$

$$\beta_1 = \text{coefficient of skewness} = \frac{\mu_3^2}{\mu_2^3} = 0.003197$$

Skewness is very small and curve is symmetrical

$$\beta_2 = \text{coefficient of kurtosis} = \frac{\mu_4}{\mu_2^2} = 2.3198$$

$$\gamma_2 = \beta_2 - 3 \text{ given excess of kurtosis, which is small}$$

Since $\beta_2 < 3$, it is playtikurtic type curve.

Ex. 12 : *The first four moments about the working mean 30.2 of a distribution are 0.255, 6.222, 30.211 and 400.25. Calculate the first four moments about the mean. Also evaluate β_1, β_2 and comment upon the skewness and kurtosis of the distribution.* ***(May 2015)***

Sol. : The first four moments about the arbitrary origin 30.2 are

$$\mu_1' = 0.255, \quad \mu_2' = 6.222, \quad \mu_3' = 30.211, \quad \mu_4' = 400.25$$

$$\therefore \quad \mu_1' = \frac{1}{N}\sum f_i (x_i - 30.2) = \frac{1}{N}\sum f_i x_i - 30.2 = \bar{x} - 30.2 = 0.255$$

or $$\bar{x} = 30.455$$

$$\mu_2 = \mu_2' - (\mu_1')^2 = 6.222 - (0.255)^2 = 6.15698$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 = 30.211 - 3\,(6.222)\,(0.255) + 2\,(0.255)^3$$

$$= 30.211 - 4.75983 + 0.03316275$$

$$\mu_3 = 25.48433$$

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

$$= 440.25 - 4\,(30.211)\,(0.255) + 6\,(6.222)\,(0.255) - 3\,(0.255)^4$$

$$\mu_4 = 378.9418$$

$$\therefore \quad \beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(25.48433)^2}{(6.15698)^3} = 2.78255$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{378.9418}{(6.15698)^2}$$

$$\beta_2 = 9.99625$$

$$\therefore \quad \gamma_1 = \sqrt{\beta_1} = \sqrt{2.78255} = 1.6681$$

which indicates considerable positive skewness of the distribution.

$$\gamma_2 = \beta_2 - 3 = 9.99625 - 3 = 6.99625$$

which shows that the distribution is leptokurtic.

Ex. 13 : *The first four moments of a distribution about the value 5 are 2, 20, 40 and 50. From the given information obtain the first four central moments, mean, standard deviation and coefficient of skewness and kurtosis.* ***(Dec. 2007; May 2018)***

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

On the basis of given information we can calculate the various central moments, mean, standard deviation and coefficient of skewness and kurtosis.

The first moment about zero gives the value of the distribution.

$$\therefore \quad \text{Mean} = \bar{x} = A + \mu_1' = 5 + 2 = 7$$

Now we calculate central moments.

$$\mu_1 = 0$$

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

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

$$= 40 - 3\,(2)\,(20) + 2\,(2)^3$$

$$= 40 - 120 + 16$$

$$= -64$$

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

$$= 50 - 4\,(2)\,(40) + 6\,(2)^2\,(20) - 3\,(2)^4$$

$$= 50 - 320 + 480 - 48$$

$$= 162$$

The second central moment gives the value of variance.

$\therefore$ Variance $= \mu_2 = 16$

$\therefore$ Standard deviation $= \sqrt{\mu_2} = \sqrt{16} = 4$

Coefficient of skewness is given by,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(-64)^2}{(16)^3} = 1$$

Since μ_3 is negative, the distribution is negatively skewed. Coefficient of kurtosis is given by,

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{162}{(16)^2} = 0.63$$

Since the value of β_2 is less than 3, hence the distribution is platykurtic.

Ex. 14 : *The first four central moments of distribution are 0, 2.5, 0.7 and 18.75. Comment on the skewness and kurtosis of the distribution.*

Sol. : Testing of Skewness : $\mu_1 = 0$, $\mu_2 = 2.5$, $\mu_3 = 0.7$ and $\mu_4 = 18.75$

Coefficient of skewness is given by,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.0314$$

Since, μ_3 is positive, the distribution is positively skewed slightly.

Testing of Kurtosis : Coefficient of kurtosis is given by

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

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3$$

Since, β_2 is exactly three, the distribution is mesokurtic.

EXERCISE 7.1

1. Find the Arithmetic Mean, Median and Standard deviation for the following frequency distribution.

x	5	9	12	15	20	24	30	35	42	49
f	3	6	8	8	9	10	8	7	6	2

Ans. $\bar{x} = 22.9851$, M = 20, $\sigma = 11.3538$

2. Age distribution of 150 life insurance policy-holders is as follows :

Age as on Nearest Birthday	Number
15 – 19.5	10
20 – 24.5	20
25 – 29.5	14
30 – 34.5	30
35 – 39.5	32
40 – 44.5	14
45 – 49.5	15
50 – 54.5	10
55 – 59.5	5

Calculate mean deviation from median age.

Ans. M.D. = 8.4284

3. The Mean and Standard deviation of 25 items is found to be 11 and 3 respectively. It was observed that one item 9 was incorrect. Calculate the Mean and Standard deviation if :

(i) The wrong item is omitted.

(ii) It is replaced by 13. **(May 2012)**

Ans. (i) $\bar{x} = 11.08, \sigma = 3.345$, (ii) $\bar{x} = 11.16, \sigma = 2.9915$

4. Following table gives the Marks obtained in a paper of statistics out of 50, by the students of two divisions :

C.I.	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50
Div. A(f)	2	6	8	8	15	18	12	11	9	4
Div. B(f)	3	5	7	9	12	16	11	5	6	2

Find out which of the two divisions show greater variability. Also find the common mean and standard deviation.

Ans. B has greater variability, $\bar{x} = 26.1458$, $\sigma = 11.1267$

5. Calculate the first four moments about the mean of the following distribution. Find the coefficient of Skewness and Kurtosis.

x	1	2	3	4	5	6	7	8	9	10
f	6	15	23	42	62	60	40	24	13	5

Ans. $\mu_1 = 0, \mu_2 = 3.703, \mu_3 = 0.04256, \mu_4 = 37.5, \beta_1 = 0.00005572, \beta_2 = 2.8411$

6. The first four moments of a distribution about the mean value 4 are −1.5, 17, −30 and 108. Find the moments about the mean and β_1 and β_2. **Ans.** $\mu_1 = 0, \mu_2 = 14.75, \mu_3 = 39.75, \mu_4 = 142.31; \beta_1 = 0.4926, \beta_2 = 0.6543.$

7.7 CURVE FITTING

7.7.1 Least Square Approximation

As a result of certain experiment suppose the values of the variables (x_i, y_i) are recorded for i = 1, 2, 3, ... n.

If these points are plotted, usually it is observed that a smooth curve passes through most of these points, while some the points are slightly away from this curve. The curve passing through these points may be a first degree curve i.e. a straight line say $y = ax + b$ or a second degree parabola such as $y = ax^2 + bx + c$ or in general an n^{th} degree curve.

$$y = a_0x^n + a_1x^{n-1} + a_2x^{n-2} \ldots a_n$$

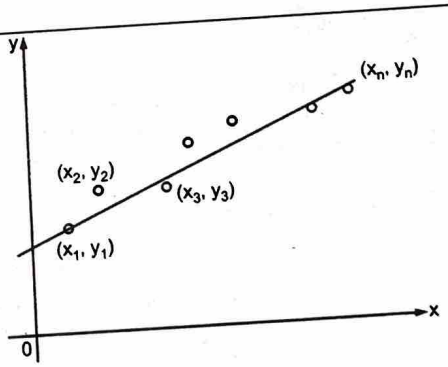


Fig. 7.5

To determine the equation of the curve which very nearly passes through the set of points, we assume some form of relation between x and y may be a straight line or a parabola of second degree, third degree and so on, which we expect to be the best fit.

In Fig. 7.5, we observe that a straight line very nearly passes through the set of points. We may assume the equation of the straight line as

$$y = ax + b \quad \ldots (1)$$

If point (x_i, y_i) is assumed to lie on (1) then y co-ordinate of the point can be calculated as,

$$y_i' = ax_i + b$$

If point actually lies on (1) then,

$$y_i = y_i'$$

otherwise $y_i - y_i'$ will represent the deviation of observed value y_i from the calculated value of y_i' using the formula (1).

In method of least squares we take the sum of the squares of these deviations and minimize this sum using the principle of maxima or minima. Values of a, b in (1) are calculated using this criteria. This is called least square criteria. Curve (1) can be of any degree using least square criteria we can find its equation.

In what follows we shall discuss fitting of straight line and second degree parabola to a given set of points.

7.7.2 Fitting Straight Line

Let (x_i, y_i); i = 1, 2, 3, ... n be the observed values of (x, y).

To fit the straight line,

$$y = ax + b$$

using least square criteria

$(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$ are the observed values.

$$y_1' = ax_1 + b,\ y_2' = ax_2 + b \ldots y_n' = ax_n + b$$

are the calculated values of y co-ordinates under the presumption that the points lie on the straight line y = ax + b.

Let $s = (y_1' - y_1)^2 + (y_2' - y_2)^2 + (y_3' - y_3)^2 \ldots (y_n' - y_n)^2$

i.e. $s = (ax_1 + b - y_1)^2 + (ax_2 + b - y_2)^2 \ldots (ax_n + b - y_n)^2$

For minimum s, $\frac{\partial s}{\partial b} = 0,\ \frac{\partial s}{\partial a} = 0.$

$\frac{\partial s}{\partial b} = 0$ gives,

$$2(ax_1 + b - y_1) + 2(ax_2 + b - y_2) \ldots 2(ax_n + b - y_n) = 0$$

or $a(x_1 + x_2 \ldots x_n) + nb - (y_1 + y_2 \ldots y_n) = 0$

$\therefore \quad a\Sigma x + nb = \Sigma y \quad \ldots (1)$

$\frac{\partial s}{\partial a} = 0$ gives

$$2(ax_1 + b - y_1)x_1 + 2(ax_2 + b - y_2)x_2 \ldots 2(ax_n + b - y_n)x_n = 0$$

or $a(x_1^2 + x_2^2 + \ldots + x_n^2) + b(x_1 + x_2 \ldots x_n) - (x_1y_1 + x_2y_2 \ldots x_ny_n) = 0$

$\therefore \quad a\Sigma x^2 + b\Sigma x = \Sigma xy \quad \ldots (2)$

Solving (1) and (2), we determine the values of a and b which gives the straight line y = ax + b, best fit for the given data.

Some times to reduce the volume of calculations we shift the origin by taking

$$x = X + h \quad \text{or} \quad X = x - h$$

where, h is generally the centrally located value in the set of observed values of x.

We find the straight line

$y = aX + b$ and then replace X by $x - h$.

7.7.3 Fitting Second Degree Parabola

We now explain the method of fitting a parabola of the form,

$$y = ax^2 + bx + c$$

to the set of observed values

(x_i, y_i) ; i = 1, 2, 3, ... n.

As before calculated values of y co-ordinates under the assumption that points satisfy the equation of parabola are given by

$$y_1' = ax_1^2 + bx_1 + c,\ y_2' = ax_2^2 + bx_2 + c''' \ldots y_n' = ax_n^2 + bx_n + c \quad \ldots (1)$$

Let $s = (y_1' - y_1)^2 + (y_2' - y_2)^2 \dots (y_n' - y_n)^2$

or $s = (ax_1^2 + bx_1 + c - y_1)^2 + (ax_2^2 + bx_2 + c - y_2)^2 \dots (ax_n^2 + bx_n + c - y_n)^2$

For minimum s, $\frac{\partial s}{\partial c} = 0, \quad \frac{\partial s}{\partial b} = 0, \quad \frac{\partial s}{\partial a} = 0,$

$\frac{\partial s}{\partial c} = 0$ gives,

$$2 (ax_1^2 + bx_1 + c - y_1) + 2 (ax_2^2 + bx_2 + c - y_2) \dots 2 (ax_n^2 + bx_n + c - y_n) = 0$$

or $$a (x_1^2 + x_2^2 + \dots x_n^2) + b (x_1 + x_2 \dots x_n) + nc - (y_1 + y_2 \dots y_n) = 0 \quad \dots (2)$$

$\therefore$ $$a \Sigma x^2 + b \Sigma x + nc = \Sigma y$$

$\frac{\partial s}{\partial b} = 0$ gives,

$$2 (ax_1^2 + bx_1 + c - y_1) x_1 + 2 (ax_2^2 + bx_2 + c - y_2) x_2 \dots 2 (ax_n^2 + bx_n + c - y_n) x_n = 0$$

or $$a (x_1^3 + x_2^3 + \dots x_n^3) + b (x_1^2 + x_2^2 \dots x_n^2) + c (x_1 + x_2 \dots x_n) - x_1y_1 - x_2y_2 \dots - x_ny_n = 0 \quad \dots (3)$$

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

Similarly, $\frac{\partial s}{\partial a} = 0$, gives,

$$2 (ax_1^2 + bx_1 + c - y_1) x_1^2 + 2 (ax_2^2 + bx_2 + c - y_2) x_2^2 \dots 2 (ax_n^2 + bx_n + c - y_n) x_n^2 = 0$$

$\therefore$ $$a (x_1^4 + x_2^4 \dots x_n^4) + b (x_1^3 + x_2^3 \dots x_n^3) + c (x_1^2 + x_2^2 \dots x_n^2) = y_1x_1^2 + y_2x_2^2 \dots y_nx_n^2$$

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

(2), (3), (4) are three simultaneous equations in three unknowns a, b, c. Solving these equations we determine a, b, c, which gives best fit parabola for the given data :

As before, to reduce the calculations, we shift the origin by taking

$$x = X + h \quad \text{or} \quad X = x - h \quad \dots (5)$$

Equations (2), (3), (4) ataken the form

$$a \Sigma X^2 + b \Sigma X + nc = \Sigma y \quad \dots (6)$$

$$a \Sigma X^3 + b \Sigma X^2 + c \Sigma X = \Sigma Xy \quad \dots (7)$$

$$a \Sigma X^4 + b \Sigma X^3 + c \Sigma X^2 = \Sigma X^2y$$

Solving (5), (6) and (7), we find a, b, c and determine the parabola.

$$y = aX^2 + bX + c$$

then we replace X by x – h and get the parabola in terms of x and y.

Illustrations

Ex. 1 : *Fit a straight line of the form y = mx + c to the following data, by using the method of least squares.*

x	0	1	2	3	4	5	6	7
y	– 5	– 3	– 1	1	3	5	7	9

Sol. : Preparing the table as

x	y	xy	x^2
0	–5	0	0
1	–3	–3	1
2	–1	–2	4
3	1	3	9
4	3	12	16
5	7	25	25
6	9	42	36
7		63	49
$\Sigma x = 28$	$\Sigma y = 16$	$\Sigma xy = 140$	$\Sigma x^2 = 140$

n = 8 (Total number of points)

Substituting in (1) and (2) of (5.6.2) after replacing a by m and b by c we get,

$$28m + 8c = 16 \quad \text{or} \quad 140m + 28c = 140 \quad \ldots(2)$$

$$7m + 2c = 4 \quad \ldots (1) \quad \text{or} \quad 5m + c = 5$$

Solving (1) and (2), we get m = 2, c = –5.

Hence the equation of the straight line is

$$y = 2x - 5$$

Ex. 2 : *Fit a parabola of the form $y = ax^2 + bx + c$ to the following data using least square criteria.*

x	1	2	3	4	5	6	7
y	–5	–2	5	16	31	50	73

Sol. :

x	y	X = x – 4	X^2	X^3	X^4	Xy	X^2y
1	–5	–3	9	–27	81	15	–45
2	–2	–2	4	–8	16	4	–8
3	5	–1	1	–1	1	–5	5
4	16	0	0	0	0	0	0
5	31	1	1	1	1	31	31
6	50	2	4	8	16	100	200
7	73	3	9	27	81	219	657
	$\Sigma y = 168$	$\Sigma X = 0$	$\Sigma X^2 = 28$	$\Sigma X^3 = 0$	$\Sigma X^4 = 196$	$\Sigma Xy = 364$	$\Sigma X^2y = 840$

n = 7.

Substituting in equations (5), (6) and (7) of (5.6.3)

$$28a + 0b + 7c = 168 \quad \ldots (1)$$

$$a \cdot 0 + 28b + c0 = 364 \quad \ldots (2)$$

$$196a + 0 \cdot b + 28c = 840 \quad \ldots (3)$$

(1), (2), (3) can be written as

$$4a + 0 \cdot b + c = 24 \quad \ldots (4)$$

$$a \cdot 0 + b + c \cdot 0 = 13 \quad \ldots (5)$$

$$7a + 0 \cdot b + c = 30 \quad \ldots (6)$$

From (5) b = 13 and from (4) and (5) we get,

$$a = 2, \; c = 16$$

Equation of parabola in terms of variable X is

$$y = 2X^2 + 13X + 16$$

Putting X = x – 4

$$y = 2(x-4)^2 + 14(x-4) + 16$$

$$y = 2x^2 - 3x - 4$$

is the required fit for the data.

Ex. 3 : *A simply supported beam carries a concentrated load P(kg) its middle point. Corresponding to various values of P, the maximum deflection y cms is tabulated as :*

P	100	120	140	160	180	200
Y	0.90	1.10	1.20	1.40	1.60	1.70

Find a law of the form $y = aP + b$ by using least square criteria.

Sol. : Preparing the table as

P(x)	y	X = P – 140	X^2	Xy
100	0.90	– 40	1600	–36
120	1.10	– 20	400	–22
140	1.20	0	0	0
160	1.40	20	400	28
180	1.60	40	1600	64
200	1.70	60	3600	102
	$\Sigma y = 7.9$	$\Sigma X = 60$	$\Sigma X^2 = 7600$	$\Sigma xy = 136$

n = 6 (No. of points)

From (1) and (2) of (5.6.2)

$$60a + 6b = 7.9 \quad \ldots (1)$$

$$7600a + 60b = 136 \quad \ldots (2)$$

Solving (1) and (2) we get,

$$a = 0.008143 \quad b = 1.2352$$

$\therefore$ $y = 0.008143\,X + 1.2352$

but $X = P - 140$

$$y = 0.008143\,(P - 140) + 1.2352$$

$$y = 0.008143\,P + 0.9518$$

is the required result.

Ex. 4 : Values of x and y are tabulated as under :

x	1	1.5	2.0	2.5
y	25	56.2	100	156

Find the law of the form $x = ay^n$ to satisfy the given by data

Sol. : Taking logarithms, we get,

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

which can be written as,

$$X = nY + c$$

where X = log x; Y = log y.

x	y	X	Y	Y^2	XY
1.0	25	0.0	1.3979	1.9541	0
1.5	56.2	0.1761	1.7497	3.0615	0.3081
2.0	100	0.301	2.0	4.0	0.602
2.5	156	0.3979	2.1931	4.8097	0.8726
		0.875	7.3407	13.8253	1.7827

Substituting in (1) and (2) of (5.6.1) where x is replaced by Y and y by X, a by n, b by log a = c. n in (1) of (6.1) = 4 (No. of points)

$$7.3407n + 4c = 0.875 \quad \ldots (1)$$

$$13.8253n + 7.3407c = 1.7827 \quad \ldots (2)$$

Solving (1) and (2) we get,

$$n = 0.5, \quad c = \log a = -0.6988375 \quad \therefore \quad a = 0.2$$

Hence required law of the form $x = ay^n$ is $x = 0.2\,y^{0.5}$.

Ex. 5 : *Following is the data given for values of X and Y. Fit a second degree polynomial of the type $ax^2 + bx + c$ where a, b, c are constants.* **(May 2000)**

X	–3	–2	–1	0	1	2	3
Y	12	4	1	2	7	15	30

Sol. : For best fitted parabola a, b and c should satisfy equation.

$$a\sum x^4 + b\sum x^3 + c\sum x^2 = \sum x^2y$$

$$a\sum x^3 + b\sum x^2 + c\sum x = \sum xy$$

$$a\sum x^2 + b\sum x + cn = \sum y$$

where, n = number of point = 7

x	y	xy	x^2	x^2y	x^3	x^4
–3	12	–36	9	108	–27	81
–2	4	–8	4	16	–8	16
–1	1	–1	1	1	–1	1
0	2	0	0	0	0	0
1	7	7	1	7	1	1
2	15	30	4	60	8	16
3	30	90	9	270	27	81
$\sum x$	$\sum y$	$\sum xy$	$\sum x^2$	$\sum x^2y^2$	$\sum x^3$	$\sum x^4$
0	71	82	28	462	0	196

Substituting we get,

$$196a + 0b + 28c = 462 \quad \ldots (i)$$

$$0a + 28b + 0c = 82 \quad \ldots (ii)$$

$$28a + 0b + 7c = 71 \quad \ldots (iii)$$

Solving (i) and (iii) simultaneously,

$$196a + 0b + 28c = 462 \quad \ldots (i)$$

$$28a + 0b + 7c = 71 \quad \ldots (ii)$$

$$a = 2.119 \approx 2.12$$

$$c = 1.66$$

From equation (ii), we get b = 2.92.

Now solving equations (i) and (iii) simultaneously,

$$196a + 0b + 28c = 362$$

and we get, $a = 2.119$

$c = 1.00$

Hence, a = 2.12, b = 2.92, c = 1.66.

Ex. 6 : *Fit a curve $y = ax^b$ using the following data :*

x	2000	3000	4000	5000	6000
y	15	15.5	16	17	18

Find out values of a and b.

Sol. : Given curve : $y = ax^b$

Taking logarithms we get,

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

which can be written as, $Y = bX + C$

where, $Y = \log y;\ X = \log x$

By least square regression we get,

$$\Sigma Y = b\Sigma X + Cn \quad \ldots (1)$$

$$\Sigma XY = b\Sigma X^2 + C\Sigma X \quad \ldots (2)$$

$\therefore$ From given data :

x	y	X = log x	Y = log y	X^2	XY
2000	15	3.30103	1.17609	10.8967	3.88498
3000	15.5	3.47712	1.19033	12.09037	4.13892
4000	16	3.60206	1.20412	12.97483	4.33731
5000	17	3.69897	1.23044	13.682379	4.55139
6000	18	3.77815	1.2552	14.27442	4.7426
		$\Sigma X = 17.85$	$\Sigma Y = 6.056$	$\Sigma X^2 = 63.9188$	$\Sigma XY = 21.652$

Putting these values in equation (1) and (2)

$$6.056 = b \cdot 17.85 + 5 \cdot C \quad \ldots (3)$$

$$21.625 = b \cdot 63.9188 + 17.85 \cdot C \quad \ldots (4)$$

Multiplying (3) by 17.85 and (4) and 5

$$108.0996 = b \cdot 318.6225 + 89.25 \cdot C$$

$$108.125 = b \cdot 319.594 + 89.25 \cdot C$$

$\therefore$ Solving these equation we get,

$$b = 0.026, \; C = 1.117 \text{ but } C = \log a$$

$\therefore$ $$a = 13.09$$

Ex. 7 : *Given the table of points :*

x	0	2	4	6	8	12	20
y	10	12	18	22	20	30	30

Use least square method to fit a straight line to the data and find the value of y (22).

Sol. : We fit straight line y = ax + b for n = 7 points.

The various summations are as follow :

x	y	x^2	xy
0	10	0	0
2	12	4	24
4	18	16	72
6	22	36	132
8	20	64	160
12	30	144	360
20	30	400	600
$\Sigma x = 52$	$\Sigma y = 142$	$\Sigma x^2 = 664$	$\Sigma xy = 1348$

From (1) and (2) of 5.6.2 we obtain

$$52a + 7b = 142 \quad \ldots(1)$$

$$664a + 52\,b = 1348 \quad \ldots(2)$$

Solving (1) and (2),

$$a = 1.0555 \text{ and } b = 12.4485$$

Therefore the straight line equation y = ax + b is

$$y = 1.0555x + 12.4485$$

Alternatively, we can also find the values of a and b in y = ax + b on solving (1) and (2 of 5.6.2).

$$a = \frac{\Sigma x \Sigma y - n \Sigma xy}{(\Sigma x)^2 - n\Sigma x^2} \text{ and } b = \frac{\Sigma y}{n} - a\frac{\Sigma x}{n}$$

Thus, $$a = \frac{(52) \times (142) - 7\,(1348)}{(52)^2 - 7\,(664)} = 1.0555$$

and $$b = \frac{142}{7} - 1.0555\,\frac{52}{7} = 12.4485$$

Ex. 8 : *For the tabulated values of x and y given below fit a linear curve of the type y = mx + c.*

x	1.0	3.0	5.0	7.0	9.0
y	1.5	2.8	4.0	4.7	6.0

Sol. : Here n = 5 . The various summations are as follows :

x	y	x^2	xy
1.0	1.5	1	1.5
3.0	2.8	9.	8.4
5.0	4.0	25	20.0
7.0	4.7	49	32.9
9.0	6.0	81	54.0
$\Sigma x = 25$	$\Sigma y = 19$	$\Sigma x^2 = 165$	$\Sigma xy = 116.8$

From (1) and (2) of 5.6.2 ,we obtain

$$25a + 5b = 19 \quad \ldots(1)$$

$$165a + 25b = 116.8 \quad \ldots(2)$$

Solving (1) and (2),

$$a = 0.545 \text{ and } b = 1.075$$

Therefore the equation of straight line y = ax + b is

$$y = 0.545\,x + 1.075$$

Alternatively, $$a = \frac{\Sigma x \Sigma y - n\Sigma xy}{(\Sigma x)^2 - n\Sigma x^2} = 0.545$$

$$b = \frac{\Sigma y}{n} - a\frac{\Sigma x}{n}$$

$$b = 1.075$$

Ex. 9 : *Following data refers to the load lifted and corresponding force applied in a pulley system. If the load lifted and effort required are related by equation effort = a . (load lifted) b, where a and b are constants. Evaluate a and b by linear curve fitting.*

Load lifted in kN	10	15.0	20.0	25.0	30.0
Effort applied in kN	0.750	0.935	1.100	1.200	1.300

Sol. : Here n = 5 preparing table as,

Sr. No.	Load x_i	Effort y_i	x_iy_i	x_i^2
1	10	0.75	7.5	100
2	15	0.935	14.025	225
3	20	1.1	22	400
4	25	1.2	30	625
5	30	1.3	39	900
	$\Sigma x_i = 90$	$\Sigma y_i = 5285$	$\Sigma x_iy_i = 112.525$	$\Sigma x_i^2 = 2250$

From (1) and (2) of 5.6.2, we obtain

$$90a + 5b = 5.285 \quad \ldots(1)$$

$$2250a + 90b = 112.525 \quad \ldots(2)$$

Solving (1) and (2), $a = 0.02761$ and $b = 0.56$

Effort = 0.02761 (load lifted) + 0.56

Ex. 10 : *If the relation between x and y is of the type $y = a\,b^x$. Using following values of x and y, find he values of constants a and b for the best fitting curve.*

x	2.1	2.5	3.1	3.5	4.1
y	5.14	6.788	10.29	13.58	20.578

Sol. : $y = a \cdot b^x$

Taking logarithms, $\log y = \log a + x \log b$

$\therefore$ $Y = C + Bx$

where, $Y = \log y$, $C = \log a$, $B = \log b$

By least square regression we get

$$\sum Y = nC + B \sum x \quad \ldots (1)$$

and $$\sum xY = (\sum x)C + B \sum x^2 \quad \ldots (2)$$

x	y	Y = log y	x^2	xY
2.1	5.14	0.71096	4.41	1.493016
2.5	6.788	0.83174	6.25	2.07935
3.1	10.29	1.01241	9.61	3.138471
3.5	13.58	1.13289	12.25	3.965115
4.1	20.578	1.3134	16.81	5.38494
$\sum x = 15.3$		$\therefore \sum Y = 5.0014$	$\sum x^2 = 49.33$	$\sum xY = 16.060892$

Substituting in equations (1) and (2)

$$5.0014 = 5C + B(15.3) \quad \ldots (3)$$

$$16.060892 = 15.3C + B(49.33) \quad \ldots (4)$$

Solving (3) and (4) $B = 0.301197$ and $C = 0.00786157$

$B = \log b$

$\therefore$ $b = 10^B = 10^{0.3011974} = 2$

$C = \log a$

$\therefore$ $a = 10^C = 10^{0.0786157} = 1.1984386$

Therefore the required relation $y = ab^x$ is $y = (1.1984386)\,2^x$.

Ex. 11 : *If X and Y are connected by the relation $ax^2 + by^2 = X$, find the value of a and b by rearranging the relation into linear form by using least square criteria for following data :*

x	1	2	3	4	5
y	3.35	5.92	8.43	10.93	13.45

Sol. : Function to be fitted is,

$$ax^2 + by^2 = x$$

Deviation is, $D = ax^2 + by^2 - x$

Sum of square of deviations is,

$$D^2 = \sum (ax^2 + by^2 - x)^2$$

Differentiating with respect to a and explain to zero,

$$\frac{\partial D^2}{\partial a} = \sum 2\,(ax^2 + by^2 - x)\,x^2 = 0$$

$\frac{\partial D^2}{\partial a} = 0$ gives,

$$\Sigma ax^4 + b\Sigma x^2y^2 - \Sigma x^3 = 0 \quad \ldots(1)$$

Similarly, differentiating with respect to b and equation to zero gives,

$$\frac{\partial D^2}{\partial b} = \Sigma 2 (ax^2 + by^2 - x) y^2 = 0$$

$$\Sigma (ax^2y^2 + by^4 - xy^2) = 0$$

$\frac{\partial D^2}{\partial b} = 0$ gives,

$$a \Sigma x^2y^2 + b \Sigma y^4 - \Sigma xy^2 = 0 \quad \ldots (2)$$

Here n = 5. Various summations are as follows :

x	y	x^4	x^2y^2	x^3	y^4	xy^2
1	3.35	1	11.22	1	125.94	11.22
2	5.92	16	140.19	8	1228.25	70.095
3	8.43	81	639.58	27	5050.22	213.193
4	10.93	256	1911.44	64	14271.86	477.86
5	13.45	625	4522.56	125	32725.72	904.512
	Σ	$\Sigma x^4 = 979$	$\Sigma x^2y^2 = 7224.99$	$\Sigma x^3 = 224$	$\Sigma y^4 =$ 53401.99	$\Sigma xy^2 =$ 1676.88

Using in equations (1) and (2), we obtain

$$979 (a) + 7224.99 (b) - 224 = 9$$

$$979a + 7224.99b = 224$$

$$a + 7.37997b = 0.2288 \quad \ldots (3)$$

$$\text{and}7224.99a + 53401.99b = 1676.88$$

$$a + 7.391289b = 0.232094 \quad \ldots (4)$$

Solving (3) and (4)

$$0.011319b = 0.0032944$$

$$b = 0.29105$$

Using value of b in equation (3)

$$a + 7.37997 \times 0.29105 = 0.2288$$

$$a = -1.91914$$

$\therefore$ Equation of the required curve is,

$$-1.91914x^2 + 0.29105y^2 = x$$

7.8 CORRELATION

We have already considered distributions involving one variable or what we call as univariate distributions. In many problems of practical nature, we are required to deal with two or more variables. If we consider the marks obtained by a group of students in two or more subjects, the distribution will involve two or more variables. Distributions using two variables are called *Bivariate distributions*. In such distributions, we are often interested in knowing whether there exists some kind of relationship between the two variables involved. In language of statistics, this means whether there is correlation or co-variance between the two variables. If the change in one variable affects the change in the other variable, the variables are said to be ***correlated***. For example, change in rainfall will affect the crop output and thus the variables 'Rainfall recorded' and 'crop output' are correlated. Similarly, for a group of workers, the variables 'income' and 'expenditure' would be correlated. If the increase (or decrease) in one variable causes corresponding increase (or decrease) in the other, the correlation is said to be ***positive*** or ***direct***. On the other hand, if increase in the value of one variable shows a corresponding decrease in the value of the other or vice versa, the correlation is called ***negative*** or ***inverse***. As the income of a worker increases, as a natural course his expenditure also increases, hence the correlation between income and expenditure is positive or direct. Correlation between heights and weights of a group of students will also be positive. If we consider the price and demand of a certain commodity then our experience tells us that as

the price of a commodity rises, its demand falls and thus the correlation between these variables is negative or inverse. Several such examples can be given. Correlation can also be classified as linear and non-linear. It is based upon the constancy of the ratio of change between the two variables. As an example, consider the values assumed by variables x and y.

x	5	8	11	15	17	19	20
y	10	16	22	30	34	38	40

Here the ratio $\frac{y}{x}$ is equal to 2 for all the values of x and y.

Correlation in such case is called *linear*.

When the amount of change in one variable is not in a constant ratio to the amount of change in other variable, the correlation is called *non-linear*. In such a case, the relationship between the variables x and y is not of the form y = mx (or of the form y = mx + c). In practical situations, the correlation is generally non-linear, but its analysis is quite complicated. Usually, it is assumed that the relation between x and y is linear and further analysis is made. There are different methods to determine whether the two variables are correlated. Some of these methods such as 'Scatter Diagram' are graphical methods and give rough idea about the correlation. These methods are not suitable if the number of observations is large. There are mathematical methods such as '*Karl Pearson's* Coefficient of Correlation', 'Concurrent Deviation Method' etc. which are more suitable. We shall discuss '*Karl Pearson's* Coefficient of Correlation' which is widely used in practice.

7.9 KARL PEARSON'S COEFFICIENT OF CORRELATION

To measure the intensity or degree of linear relationship between two variables, Karl Pearson developed a formula called *correlation coefficient.*

Correlation coefficient between two variables x and y denoted by r (x, y) is defined as

$$r(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

In bivariate distribution if (x_i, y_i) take the values $(x_1, y_1)\ (x_2, y_2) \ldots (x_n, y_n)$

$$\text{cov}(x, y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

where, $\bar{x}, \bar{y}$ are arithmetic means for x and y series respectively.

Similarly, $\sigma_x = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}$ and $\sigma_y = \sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2}$

which are the standard deviations for x and y series.

$$\text{cov}(x, y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{1}{n} \sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x}\bar{y})$$

$$= \frac{1}{n} \sum x_i y_i - \bar{y} \frac{1}{n} \sum x_i - \bar{x} \frac{1}{n} \sum y_i + \frac{1}{n} \sum \bar{x}\bar{y}$$

$$= \frac{1}{n} \sum x_i y_i - \bar{y}\bar{x} - \bar{x}\bar{y} + \frac{1}{n}(n\bar{x}\bar{y})$$

$$= \frac{1}{n} \sum x_i y_i - \bar{x}\bar{y}$$

$$\sigma_x^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \frac{1}{n} \sum x_i^2 - 2x_i \bar{x} + \bar{x}^2$$

$$= \frac{1}{n} \sum x_i^2 - 2\frac{\bar{x}}{n} \sum x_i + \frac{1}{n} \sum \bar{x}^2$$

$$= \frac{1}{n} \sum x_i^2 - 2\bar{x}^2 + \frac{1}{n}(n\bar{x}^2)$$

$$= \frac{1}{n} \sum x_i^2 - \bar{x}^2$$

Similarly, $\sigma_y^2 = \frac{1}{n}\sum y_i^2 - \bar{y}^2$

$r(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$ can then be calculated.

If we put, $u_i = x_i - A$ or $\frac{x_i - A}{h}$ and $v_i = y_i - B$ or $\frac{y_i - B}{k}$

then $\text{cov}(u, v) = \frac{1}{n}\sum u_i v_i - \bar{u}\bar{v}$, $\sigma_u^2 = \frac{1}{n}\sum u_i^2 - \bar{u}^2$, $\sigma_v^2 = \frac{1}{n}\sum v_i^2 - \bar{v}^2$

r (u, v) is given by, $r(u, v) = \frac{\text{cov}(u, v)}{\sigma_u \sigma_v}$

It can be established that r (x, y) = r (u, v). Calculation of r (u, v) is simpler as compared to r (x, y).

Correlation coefficient 'r' always lies between –1 and 1 i.e. $-1 \le r \le 1$.

If r > 0, the correlation is positive and if r < 0, the correlation is negative. If r = 0, we say the variables are uncorrelated. In general, if |r| > 0.8, we consider high correlation. If |r| is between 0.3 to 0.8, we say that, correlation is considerable. If |r| < 0.3, we say that correlation is negligible. If r = 1, we say that there is perfect positive correlation; where as if r = –1, we say that there is perfect negative correlation.

We also, note that covariance can also be considered as a joint central moment of order (1, 1) of (X, Y). Hence, we denote μ_{11} = cov (X, Y).

ILLUSTRATIONS

Ex. 1 : *Following are the values of import of raw material and export of finished product in suitable units.*

Export	10	11	14	14	20	22	16	12	15	13
Import	12	14	15	16	21	26	21	15	16	14

Calculate the coefficient of correlation between the import values and export values. ***(Dec. 2010, May 2014)***

Sol. : Let X : Quantity exported, Y : Quantity imported, Preparing table as follows calculations can be made simple.

x	y	x^2	y^2	xy
10	12	100	144	120
11	14	121	196	154
14	15	196	225	210
14	16	196	256	224
20	21	400	441	420
22	26	484	676	572
16	21	256	441	336
12	15	144	225	180
15	16	225	256	240
13	14	169	196	182
Total = 147	**170**	**2291**	**3056**	**2638**

Here, n = 10, hence $\bar{x} = \frac{\sum x}{N} = \frac{147}{10} = 14.7$

and $\bar{y} = \frac{\sum y}{N} = \frac{170}{10} = 17$

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2) \times (\sum y^2 - n\bar{y}^2)}}$$

$$= \frac{2638 - 10 \times 14.7 \times 17}{\sqrt{(2291 - 10 \times 14.7^2)\,(3056 - 10 \times 17^2)}}$$

$$= \frac{139}{\sqrt{130.1 \times 166}} = 0.9458$$

Ex. 2 : *Calculate the correlation coefficient for the following weights (in kg) of husband (x) and wife (y).* **(May 2010)**

x	65	66	67	67	68	69	70	72
y	55	58	72	55	66	71	70	50

Sol. :

x	y	x^2	y^2	xy
65	55	4225	3025	3575
66	58	4356	3364	3828
67	72	4489	5184	4824
67	55	4489	3025	3685
68	66	4624	4354	4488
69	71	4761	5041	4899
70	70	4900	4900	4900
72	50	5184	2500	3600
544	**497**	**37028**	**31393**	**33799**

$$\bar{x} = \frac{\sum x}{n} = \frac{544}{8} = 68$$

$$\bar{y} = \frac{\sum y}{n} = \frac{497}{8} = 62.125$$

Correlation coefficient between x and y is given by

$$r(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\sum x^2 - (\bar{x})^2\right)\left(\frac{1}{n}\sum y^2 - (\bar{y})^2\right)}}$$

$$= \frac{\frac{1}{8}(33799) - 68(62.125)}{\sqrt{\left(\frac{37028}{8} - (68)^2\right)\left(\frac{31398}{8} - (62.125)^2\right)}}$$

$$= \frac{4224.875 - 4224.5}{\sqrt{(4628.5 - 4624)(3924.125 - 3859.52)}}$$

$$= \frac{0.375}{\sqrt{4.5 \times 64.605}} = \frac{0.375}{\sqrt{290.7225}} = \frac{0.375}{17.051}$$

$$r(x, y) = 0.022$$

Ex. 3 : *From a group of 10 students, marks obtained by each in papers of Mathematics and Applied Mechanics are given as :*

x Marks in Maths	23	28	42	17	26	35	29	37	16	46
y Marks in App. Mech.	25	22	38	21	27	39	24	32	18	44

Calculate Karl Pearson's Coefficient of correlation.

Sol. : The data is tabulated as :

x	y	u = x – 35	v = y – 39	u^2	v^2	uv
16	18	– 19	– 21	361	441	399
17	21	– 18	– 18	324	324	324
23	25	– 12	– 14	144	196	168
26	27	– 09	– 12	81	144	108
28	22	– 07	– 17	49	289	119
29	24	– 06	– 15	36	225	90
35	39	– 00	00	00	00	00
37	32	02	– 07	04	49	– 14
42	38	07	– 01	49	01	– 07
46	44	11	05	121	25	55
Total		**∑ u = – 51**	**∑ v = – 100**	**∑ u^2 = 1169**	**∑ v^2 = 1694**	**∑ uv = 1242**

$$\bar{u} = \frac{-51}{10} = -5.1, \quad \bar{u}^2 = 26.01$$

$$\bar{v} = \frac{-100}{10} = -10, \quad \bar{v}^2 = 100$$

$$\text{cov}(u, v) = \frac{1}{n}\sum u_i v_i - \bar{u}\,\bar{v} = \frac{1}{10}(1242) - 51 = 73.2$$

$$\sigma_u^2 = \frac{1}{n}\sum u_i^2 - \bar{u}^2 = \frac{1169}{10} - 26.01 = 90.89$$

$$\sigma_u = \sqrt{90.89} = 9.534$$

$$\sigma_v^2 = \frac{1}{n}\sum v_i^2 - \bar{v}^2 = \frac{1694}{10} - 100 = 69.4$$

$$\sigma_v = \sqrt{69.4} = 8.33$$

$$r(x, y) = r(u, v) = \frac{\text{cov}(u, v)}{\sigma_u\,\sigma_v} = \frac{73.2}{9.534 \times 8.33} = 0.9217$$

Ex. 4 : *Compute correlation coefficient between supply and price of commodity using following data.*

Supply	152	158	169	182	160	166	182
Price	198	178	167	152	180	170	162

Sol. : Let x = Supply, u = x – 150, y = price, v = y – 160

x	y	u	v	u^2	v^2	uv
152	198	2	38	4	1444	76
158	178	8	18	64	324	144
169	167	19	7	361	49	133
182	152	32	– 8	1024	64	– 256
160	180	10	20	100	400	200
166	170	16	10	256	100	160
182	162	32	2	1024	4	64
Total	–	**119**	**87**	**2833**	**2385**	**521**

Here, n = 7, $\sum u = 119$, $\sum v = 87$, $\sum u^2 = 2833$, $\sum v^2 = 2385$, $\sum uv = 521$

$\therefore$ $\bar{u} = 17$, $\bar{v} = 12.4286$

$$r = \frac{\sum uv - n\,\bar{u}\,\bar{v}}{\sqrt{\left(\sum u^2 - n\bar{u}^2\right) \times \left(\sum v^2 - n\bar{v}^2\right)}}$$

$$r = \frac{521 - 7 \times 17 \times 12.4286}{\sqrt{(2833 - 7 \times 17^2)\,(2.385 - 7 \times 12.4286)^2}}$$

$$r = \frac{-958}{\sqrt{810 \times 1303.7142}} = \frac{-958}{1027.6227}$$

$$= -0.9322$$

Interpretation : There is high negative correlation between supply and price.

Ex. 5 : *Obtain correlation coefficient between population density (per square miles) and death rate (per thousand persons) from data related to 5 cities.*

Population Density	200	500	400	700	800
Death Rate	12	18	16	21	10

Sol. : Let x = Population density and y = Death rate.

Let, $u = x - a$ and $v = y - b$

$= x - 500$ $= y - 15$

x	y	$u = x - 500$	v	u^2	v^2	uv
200	12	– 300	– 3	90000	9	900
500	18	0	3	0	9	0
400	16	– 100	1	10000	1	– 100
700	21	200	6	40000	36	1200
800	10	300	– 5	90000	25	– 1500
Total	**–**	**100**	**2**	**230000**	**80**	**500**

Here, $n = 5$, $\Sigma u = 100$, $\Sigma v = 2$, $\Sigma u^2 = 230000$, $\Sigma v = 80$, $\Sigma uv = 500$

$$\bar{u} = \frac{\Sigma u}{n} = \frac{100}{5} = 20$$

$$\bar{v} = \frac{\Sigma v}{n} = \frac{2}{5} = 0.4$$

$$r(u, v) = \frac{\Sigma uv - n\,\bar{u}\,\bar{v}}{\sqrt{\left[\Sigma u^2 - n(\bar{u})^2\right]\left[\Sigma v^2 - n(\bar{v})^2\right]}}$$

$$= \frac{500 - 5(20)(0.4)}{\sqrt{230000 - 5(20)^2}\sqrt{80 - 5(0.4)^2}}$$

$$= \frac{460}{\sqrt{228000}\sqrt{79.2}}$$

$$= \frac{460}{4249.42} = 0.1082$$

Ex. 6 : *Calculate the coefficient of correlation for the following distribution.*

x	*5*	*9*	*15*	*19*	*24*	*28*	*32*
y	*7*	*9*	*14*	*21*	*23*	*29*	*30*
f	*6*	*9*	*13*	*20*	*16*	*11*	*7*

Sol. : Tabulating the data as

x	y	f	$u = x - 19$	$v = y - 21$	fu	fv	fu^2	fv^2	fuv
5	7	6	– 14	– 14	– 84	– 84	1176	1176	1176
9	9	9	– 10	– 12	– 90	– 108	900	1296	1080
15	14	13	– 4	– 7	– 52	– 91	208	637	364
19	21	20	0	0	0	0	0	0	0
24	23	16	5	2	80	32	400	64	160
28	29	11	9	8	99	88	891	704	792
32	30	7	13	9	91	63	1183	567	819
Total		**$\Sigma f = 82$**			**$\Sigma fu = 44$**	**$\Sigma fv = -100$**	**$\Sigma fu^2 = 4758$**	**$\Sigma fv^2 = 4444$**	**$\Sigma fuv = 4391$**

$$\bar{u} = \frac{\Sigma fu}{\Sigma f} = \frac{44}{82} = 0.5366; \quad \bar{u}^{-2} = 0.288$$

$$\bar{v} = \frac{\Sigma fv}{\Sigma f} = \frac{-100}{82} = 1.2195; \quad \bar{v}^{2} = 1.4872$$

$$\text{cov}(u, v) = \frac{1}{\sum f}\sum f u_i v_i - \bar{u}\,\bar{v} = \frac{4391}{82} - 0.6544 = 52.89$$

$$\sigma_u^2 = \frac{1}{\sum f}\sum f u_i^2 - \bar{u}^2 = \frac{758}{82} - 0.288 = 57.7364$$

$$\sigma_v^2 = \frac{1}{\sum f}\sum f v_i^2 - \bar{v}^2 = \frac{4444}{82} - 1.4872 = 52.708$$

$$\sigma_u = 7.598$$

$$\sigma_v = 7.26$$

$$r(u, v) = \frac{\text{cov}(u, v)}{\sigma_u \sigma_v} = \frac{52.89}{55.16} = 0.9588$$

$\therefore$ Coefficient of correlation = $r(x, y) = 0.9588$

Ex. 7: *Find correlation coefficient between X and Y, given that, $n = 25$, $\sum x = 75$, $\sum y = 100$, $\sum x^2 = 250$, $\sum y^2 = 500$, $\sum xy = 325$.*

Sol. : Here $\bar{x} = \frac{75}{25} = 3$, $\bar{y} = \frac{100}{25} = 4$.

$\therefore$

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2) \times \sum(y^2 - n\bar{y}^2)}}$$

$$r = \frac{325 - 25 \times 3 \times 4}{\sqrt{(250 - 25 \times 9)\,(500 - 25 \times 16)}} = \frac{25}{\sqrt{25 \times 100}} = \frac{25}{50} = 0.5$$

Ex. 8 : *Calculate the coefficient of correlation from the following information.*

$n = 10$, $\sum x = 40$, $\sum x^2 = 190$, $\sum y^2 = 200$, $\sum xy = 150$, $\sum y = 40$.

Sol. : Here,

$$\bar{x} = \frac{\sum x}{n} = \frac{40}{10} = 4$$

$$\bar{y} = \frac{\sum y}{n} = \frac{40}{10} = 4$$

Coefficient of correlation is given by,

$$r = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n(\bar{x})^2) \times (\sum y^2 - n(\bar{y})^2)}}$$

$$r = \frac{150 - 10\,(4)\,(4)}{\sqrt{190 - 10\,(4)^2}\sqrt{200 - 10\,(4)^2}}$$

$$= \frac{150 - 160}{\sqrt{30}\,\sqrt{40}} = \frac{-10}{34.6410} = -0.2850$$

Ex. 9 : *Given : $n = 6$, $\sum (x - 18.5) = -3$, $\sum (y - 50) = 20$, $\sum (x - 18.5)^2 = 19$, $\sum (y - 50)^2 = 850$, $\sum (x - 18.5)(y - 50) = -120$. Calculate coefficient of correlation.*

Sol. : Let $u = x - 18.5$ and $v = y - 50$

$$\bar{u} = \frac{-3}{6} = -0.5$$

and

$$\bar{v} = \frac{20}{6} = 3.33$$

From the given data $\sum u = -3$, $\sum v = 20$, $\sum u^2 = 19$, $\sum v^2 = 850$ and $\sum uv = -120$

Coefficient of correlation is given by

$$r = \frac{\sum uv - n\bar{u}\bar{v}}{\sqrt{[\sum u^2 - n(\bar{u})^2] \times [\sum v^2 - n(\bar{v})^2]}}$$

$$= \frac{-120 - 6\,(-0.5) \times (3.33)}{\sqrt{[19 - 6\,(-0.5)^2] \times [850 - 6\,(3.33)^2]}}$$

$$= \frac{-120 + 9.99}{\sqrt{(17.5)\,(783.47)}} = \frac{-110.01}{117.0928}$$

$$r = -0.9395$$

Ex. 10 : *Given : $r = 0.9$, $\sum XY = 70$, $\sigma_y = 3.5$, $\sum X^2 = 100$.*

Find the number of items, if X and Y are deviations from arithmetic mean.

Sol. : $\sum X^2 = 100$ $\qquad \sum XY = 70$

$r = 0.9$ $\qquad \sigma_y = 3.5$

We have to find the value of n

$$\sigma_x^2 = \frac{1}{n}\sum (x - \bar{x})^2 = \frac{1}{n}\sum X^2 = \frac{100}{n}$$

$$r = \frac{\sum (x - \bar{x})\,(y - \bar{y})}{n\,\sigma_x\,\sigma_y} = \frac{\sum XY}{n\,\sigma_x\,\sigma_y}$$

Squaring we get, $\quad r^2 = \dfrac{(\sum XY)^2}{n^2\,\sigma_x^2\,\sigma_y^2}$

$$(0.9)^2 = \frac{(70)^2}{n^2 \times \left(\frac{100}{n}\right) \times (3.5)^2}$$

$$0.81 = \frac{4900}{1225 \cdot n}$$

$$0.81 \times 1225\, n = 4900$$

$$992.25\, n = 4900$$

$$n = 4.9383$$

$$n \approx 5$$

Ex. 11 : *Find the coefficient of correlation for distribution in which S.D. of $x = 4$, and S.D. of $y = 1.8$. Coefficient of regression of y on x is 0.32.*

Sol. : $\sigma_x = 4$, $\sigma_y = 1.8$ and $b_{yx} = 0.32$

We have, $\quad b_{yx} = r\dfrac{\sigma_y}{\sigma_x}$

$$0.32 = r \times \frac{1.8}{4}$$

$\therefore$ $\quad r = \dfrac{0.32 \times 4}{1.8} = 0.711$

7.10 REGRESSION

After having established that the two variables are correlated, we are generally interested in estimating the value of one variable for a given value of the other variable.

For example, if we know that rainfall affects the crop output then it is possible to predict the crop output at the end of a rainy season. If the variables in a bivariate distribution are related, the points in scatter diagram cluster round some curve called the curve of regression or the regression curve. If the curve is a straight line, it is called the ***line of regression*** and in such case the regression between two variables is linear. The line of regression gives best estimate for the value of one variable for some specified value of the other variable. If correlation is not perfect (i.e. $r \neq \pm 1$), then several lines can be drawn through given points. Being the line of best fit, the regression line is obtained by using the method of least squares.

Consider the set of values of (x_i, y_i), $i = 1, 2, \ldots n$. Let the line of regression of y on x be $y = mx + c$

From the method of least squares, the normal equations for estimating unknown m and c are given by

$$\sum y_i = nc + m \sum x_i \qquad \ldots (1)$$

$$\sum x_i y_i = c \sum x_i + m \sum x_i^2 \qquad \ldots (2)$$

Dividing (1) by n, we get

$$\frac{1}{n}\sum y_i = c + m\left(\frac{1}{n}\sum x_i\right)$$

i.e. $$\bar{y} = c + m\bar{x} \quad \text{...(3)}$$

which shows that the point $(\bar{x}, \bar{y})$ lies on the line of regression.

We know that, $$\mu_{11} = \text{cov}(x, y) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$$

$\therefore$ $$\frac{1}{n}\sum x_i y_i = \mu_{11} + \bar{x}\bar{y} \quad \text{... (4)}$$

Also $$\sigma_x^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2$$

$\therefore$ $$\frac{1}{n}\sum x_i^2 = \sigma_x^2 + \bar{x}^2 \quad \text{...(5)}$$

Dividing (2) by n, we get

$$\frac{1}{n}\sum x_i y_i = c\frac{\sum x_i}{n} + m\frac{\sum x_i^2}{n} \quad \text{... (6)}$$

Substituting from (4) and (5) in (6),

$$\mu_{11} + \bar{x}\bar{y} = c\bar{x} + m\left(\sigma_x^2 + \bar{x}^2\right) \quad \text{... (7)}$$

Multiplying (3) by $\bar{x}$ and subtracting from (7), we get

$$\mu_{11} = m\sigma_x^2$$

$$m = \frac{\mu_{11}}{\sigma_x^2}$$

Equation of regression line which passes through $(\bar{x}, \bar{y})$ and which has slope $\frac{\mu_{11}}{\sigma_x^2}$ is thus given by the equation

$$y - \bar{y} = \frac{\mu_{11}}{\sigma_x^2}(x - \bar{x}) \quad \text{... (8)}$$

This equation gives regression line of y on x.

Similarly, if we start with regression line of x on y as x = my + c same procedure will give

$$x - \bar{x} = \frac{\mu_{11}}{\sigma_y^2}(y - \bar{y}) \quad \text{... (9)}$$

Also, we know that $$r(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\mu_{11}}{\sigma_x \sigma_y}$$

Putting for μ_{11} in (8) and (9), we have

$$y - \bar{y} = \frac{r\sigma_x \sigma_y}{\sigma_x^2}(x - \bar{x}) \quad \text{and} \quad x - \bar{x} = \frac{r\sigma_x \sigma_y}{\sigma_x^2}(y - \bar{y})$$

Thus, the regression line of y on x is

$$\boxed{y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) = b_{yx}(x - \bar{x})} \quad \text{... (10)}$$

Similarly, the regression line of x on y is

$$\boxed{x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) = b_{xy}(y - \bar{y})} \quad \text{... (11)}$$

The coefficient b_{yx} involved in the equation (10) is known as ***regression coefficient of y on x*** and the coefficient b_{xy} involved in the equation (11) is known as ***regression coefficient of x on y***.

Remark 1 : For obtaining (10) and (11) we have to calculate r = r (x, y) the correlation coefficient, which can be also determined using change of origin and scale property.

Thus, $$r = r(x, y) = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\text{cov}(u, v)}{\sigma_u \sigma_v} = r(u, v)$$

If $$u = \frac{x - a}{h}, \quad v = \frac{y - b}{k}, \text{ then } \sigma_x = h\,\sigma_u, \quad \sigma_v = k\,\sigma_v$$

and $$\sigma_u^2 = \frac{1}{n}\sum u_i^2 - \bar{u}^2 \text{ and } \sigma_v^2 = \frac{1}{n}\sum v_i^2 - \bar{v}^2$$

and $$\bar{x} = a + h\,\bar{u}, \text{ and } \bar{y} = b + k\,\bar{v}$$

In particular, if $$u = x - a,\ v = y - b \text{ then, } h = k = 1 \text{ and } \sigma_x = \sigma_u \text{ and } \sigma_y = \sigma_v$$

and $$\bar{x} = a + \bar{u}, \quad \bar{y} = b + \bar{v}$$

These results help us to determine (10) and (11).

Remark 2 : Correlation coefficient and regression coefficients have same algebraic signs. If r > 0, then $b_{yx} > 0$ and $b_{xy} > 0$. If r < 0, then $b_{yx} < 0$ and $b_{xx} < 0$.

Remark 3 : Since $b_{yx} \times b_{xy} = r^2$ therefore correlation coefficient = $r = \sqrt{b_{xy} \times b_{yx}}$. i.e. geometric mean of regression coefficients. Choose positive square root, if regression coefficients are positive, otherwise negative.

Remark 4 : The acute angle θ between the regression lines is given by,

$$\theta = \tan^{-1}\left\{\frac{1 - r^2}{|r|} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\right\}$$

Remark 5 : The point of intersection of two regression line is $(\bar{x}, \bar{y})$.

ILLUSTRATIONS

Ex. 1 : *Obtain regression lines for the following data :* ***(Dec. 2006, 2007, 2008, 2011; Nov. 2015, 2016 ; May 2009, 2017)***

x	6	2	10	4	8
y	9	11	5	8	7

Sol. : To find regression lines we require to calculate regression coefficient b_{xy} and b_{yx}. These coefficients depend upon $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ and $\sum xy$. So we prepare the following table and simplify the calculations.

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
$\sum x_i = 30$	$\sum y_i = 40$	$\sum x_i^2 = 220$	$\sum y_i^2 = 340$	$\sum x_i y_i = 214$

No. of observations = n = 5

$$\bar{x} = \frac{\sum x_i}{n} = \frac{30}{5} = 6 \text{ and } \bar{y} = \frac{\sum y_i}{n} = \frac{40}{5} = 8$$

$$\sigma_x^2 = \frac{\sum x_i^2}{n} - (\bar{x})^2 = \frac{220}{5} - (6)^2 = 44 - 36 = 8$$

$$\sigma_y^2 = \frac{\sum y_i^2}{n} - (\bar{y})^2 = \frac{340}{5} - (8)^2 = 68 - 64 = 4$$

$$\text{Cov}(x, y) = \frac{\sum (x_i y_i)}{n} - \bar{x}\,\bar{y} = \frac{214}{5} - 6 \times 8$$

$$\text{Cov}(x, y) = 42.8 - 48 = -5.2$$

$$b_{yx} = \frac{\text{Cov}(x, y)}{\sigma_x^2} = \frac{-5.2}{8} = -0.65$$

$$b_{xy} = \frac{\text{Cov}(x, y)}{\sigma_y^2} = \frac{-5.2}{6} = -1.3$$

Regression line of Y on X is

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

$$y - 8 = -0.65(x - 6)$$

$$y = -0.65x + 3.9 + 8$$

$$y = -0.65x + 11.9$$

Regression line of X on Y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$x - 6 = -1.3(y - 8)$$

$$x - 6 = -1.3y + 10.4$$

$$x = -1.3y + 10.4 + 6$$

$$x = -1.3y + 16.4$$

Ex. 2 : *Obtain regression lines for the following data :*

X	2	3	5	7	9	10	12	15
Y	2	5	8	10	12	14	15	16

Find estimate of (i) Y when X = 6 and (ii) X when Y = 20. ***(Nov. 2015; May 2016)***

Sol. : To find regression lines we require to calculate regression coefficients b_{xy} and b_{yx}. These coefficients depend upon $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$, $\sum xy$. So we prepare the following table and simplify the calculations :

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
2	2	4	4	4
3	5	9	25	15
5	8	25	64	40
7	10	49	100	70
9	12	81	144	108
10	14	100	196	140
12	15	144	225	180
15	16	225	256	240
Total = 63	**82**	**637**	**1014**	**797**

n = number of pairs of observations = 8

$$\bar{x} = \frac{\sum x_i}{n} = \frac{63}{8} = 7.875$$

$$\sigma_x^2 = \frac{\sum x_i^2}{n} - (\bar{x})^2$$

$$= \frac{637}{8} - (7.875)^2 = 17.6094$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{82}{8} = 10.25$$

$$\sigma_y^2 = \frac{\sum y_i^2}{n} - (\bar{y})^2$$

$$= \frac{1014}{8} - (10.25)^2 = 21.6875$$

$$\text{Cov}(x, y) = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y} = \frac{797}{8} - 7.875 \times 10.25$$

$$= 18.9063$$

$$b_{yx} = \frac{\text{Cov}(x, y)}{\sigma_x^2} = \frac{18.9063}{17.6094} = 1.0736$$

$$b_{xy} = \frac{\text{Cov}(x, y)}{\sigma_y^2} = \frac{18.9063}{21.6875} = 0.8718$$

Regression line of Y on X : $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$y - 10.25 = 1.0736 (X - 7.875)$$

$$Y = 1.0736 X + 1.7954$$

(i) Estimate of y for $x = 6$ can be obtained by substituting $x = 6$ in the above regression equation.

$\therefore$ $$Y = 1.0736 \times 6 + 1.7954 = 8.237$$

Regression line of X on Y :

$$X - \bar{X} = b_{xy}(Y - \bar{Y})$$

$$X - 7.875 = 0.8718 (Y - 10.25)$$

$$X = 0.8718 Y - 1.06095$$

(ii) Estimate of x can be obtained by substituting $y = 20$ in the above equation.

$$X = 16.37505$$

Note : For estimation of x and estimation of y, separate equations are to be used.

Ex. 3 : *Find the lines of regression for the following data :*

x	*10*	*14*	*19*	*26*	*30*	*34*	*39*
y	*12*	*16*	*18*	*26*	*29*	*35*	*38*

and estimate y for x = 14.5 and x for y = 29.5. ***(Dec. 2005)***

Sol. : Tabulating the data as :

x	y	u = x – 26	v = y – 26	u^2	v^2	uv
10	12	– 16	– 14	256	196	224
14	16	– 12	– 10	144	100	120
19	18	– 7	– 8	49	64	56
26	26	0	0	0	0	0
30	29	4	3	16	9	12
34	35	8	9	64	81	72
39	38	13	12	169	144	156
Total	–	$\sum u = -10$	$\sum v = -8$	$\sum u^2 = 698$	$\sum v^2 = 594$	$\sum uv = 640$

Here $n = 7$, $\bar{u} = \frac{-10}{7} = -1.429$, $\bar{v} = \frac{-8}{7} = -1.143$

$\bar{u}^2 = 2.042$, $\bar{v}^2 = 1.306$

$$\text{cov}(u, v) = \frac{1}{n}\sum uv - \bar{u}\,\bar{v}$$

$$= \frac{1}{7}(640) - (1.429)(1.143) = 89.795$$

$$\sigma_u^2 = \frac{1}{n}\sum u_i^2 - \bar{u}^2 = \frac{1}{7}(698) - 2.042 = 97.672$$

$\therefore$ $\sigma_u = 9.883$

$$\sigma_v^2 = \frac{1}{n}\sum v_i^2 - \bar{v}^2 = \frac{1}{7}(594) - 1.306 = 83.551$$

$\therefore$ $\sigma_v = 9.14$

$$r = r(x, y) = r(u, v) = \frac{\text{cov}(u, v)}{\sigma_u \sigma_v} = \frac{89.795}{9.883 \times 9.14}$$

$$= \frac{89.795}{90.33062} = 0.9941$$

$$r \times \frac{\sigma_y}{\sigma_x} = r \times \frac{\sigma_v}{\sigma_u} = 0.9941 \times \frac{9.14}{9.883} = 0.9194$$

$$r \times \frac{\sigma_x}{\sigma_y} = r \times \frac{\sigma_u}{\sigma_v} = 0.9941 \times \frac{9.883}{9.14} = 1.0749$$

$$\bar{x} = a + \bar{u} = 26 - 1.429 = 24.571$$

$$\bar{y} = b + \bar{v} = 26 - 1.143 = 24.857$$

Regression line of y on x is given by equation (10)

$$y - 24.857 = 0.9194\,(x - 24.571) \quad \text{... (i)}$$

Regression line of x on y is given by equation (11)

$$x - 24.571 = 1.0749\,(y - 24.857) \quad \text{... (ii)}$$

To estimate y for x = 14.5

put x = 14.5 in (i), $\therefore$ $y = 24.857 + 0.9194\,(14.5 - 24.571) = 15.5977$

Estimate of x for y = 29.5 is obtained from (ii).

$$x = 24.571 + 1.0749\,(29.5 - 24.857)$$

$$= 29.56176$$

Ex. 4 : *The table below gives the respective heights x and y of a sample of 10 fathers and their sons :*

(i) Find regression line of y on x.

(ii) Find regression line of x on y.

(iii) Estimate son's height if father's height is 65 inches.

(iv) Estimate father's height if son's height is 60 inches.

(v) Compute correlation coefficient between x and y.

(vi) Find the angle between the regression lines.

Height of Father x (Inches)	65	63	67	64	68	62	70	66	68	67
Height of Son y (Inches)	68	66	68	65	69	66	68	65	71	67

Sol. : Let $u = x - 62$, $v = y - 65$. We prepare the table to simplify the computations.

x	y	u	v	u^2	v^2	uv
65	68	3	3	9	9	9
63	66	1	1	1	1	1
67	68	5	3	25	9	15
64	65	2	0	4	0	0
68	69	6	4	36	16	24
62	66	0	1	0	1	0
70	68	8	3	64	9	24
66	65	4	0	16	0	0
68	71	6	6	36	36	36
67	67	5	2	25	4	10
Total		**40**	**23**	**216**	**85**	**119**

n = Number of pairs = 10

$$\bar{u} = \frac{40}{10} = 4, \qquad \sigma_u^2 = \frac{216}{10} - 4^2 = 5.6$$

$$\bar{v} = \frac{23}{10} = 2.3, \qquad \sigma_v^2 = \frac{85}{10} - (2.3)^2 = 3.21$$

$$\text{Cov}(u, v) = \frac{119}{10} - 4 \times 2.3 = 2.7$$

$$\therefore \quad b_{xy} = b_{uv} = \frac{2.7}{3.21} = 0.8411, \text{ and } b_{yx} = b_{vu} = \frac{2.7}{5.6} = 0.4821$$

$$\bar{x} = \bar{u} + 62 = 66, \quad \bar{y} = \bar{v} + 65 = 67.3$$

(i) Regression line of y on x is $Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$\therefore \quad Y - 67.3 = 0.4821 (X - 66)$$

$$\therefore \quad Y = 0.4821X + 35.4814$$

(ii) Regression line of x on y is $X - \bar{X} = b_{xy}(Y - \bar{Y})$

$$\therefore \quad X - 66 = 0.8411 (Y - 67.3)$$

$$\therefore \quad X = 0.8411\, Y + 9.3940$$

(iii) Estimate of son's height Y for x = 65

$$Y = 0.4821 \times 65 + 35.4814 = 66.8179 \text{ inches}$$

(iv) Estimate of father's height x for y = 60

$$X = 0.8411 \times 60 + 9.394 = 59.86 \text{ inches}$$

(v) Correlation coefficient,

$$r = \sqrt{b_{xy} \cdot b_{yx}} = \sqrt{0.8411 \times 0.4821} = 0.63678$$

We choose positive square root because regression coefficients are positive.

(vi) The acute angle between the regression lines is given by

$$\tan\theta = \frac{1 - r^2}{|r|} \cdot \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} = \frac{1 - (0.63678)^2}{0.63678} \times \frac{\sqrt{5.6 \times 3.21}}{(5.6 + 3.21)}$$

$$= 0.933621 \times \frac{4.2398}{8.81} = 0.4493$$

$$\therefore \quad \theta = \tan^{-1}(0.4493) = 24.19^{\circ}$$

Ex. 5 : *The following are marks obtained by 10 students in Statistics and Economics.*

No.	1	2	3	4	5	6	7	8	9	10
Marks in Economics	25	28	35	32	31	36	29	38	34	32
Marks in Statistics	43	46	49	41	36	32	31	30	33	39

Marks are out of 50. Obtain regression equation to estimate marks in Statistics if marks in Economics are 30. **(Dec. 2005)**

Sol. : n = 10. Let us denote marks in Economics by x and marks in Statistics by y.

Let u = x – 30 and v = y – 35.

x	y	u = x – 30	v = y – 35	u^2	v^2	uv
25	43	– 5	8	25	64	– 40
28	46	– 2	11	4	121	– 22
35	49	5	14	25	196	70
32	41	2	6	4	36	12
31	36	1	1	1	1	1
36	32	6	– 3	36	9	– 18
29	31	– 1	– 4	1	16	4
38	30	8	– 5	64	25	– 40
34	33	4	– 2	16	4	– 8
32	39	2	4	4	16	8
–	–	$\sum u = 2$	$\sum v = 30$	$\sum u^2 = 180$	$\sum v^2 = 488$	$\sum uv = -33$

$$\bar{u} = \frac{\sum u}{n} = \frac{20}{10} = 2 \text{ and } \bar{v} = \frac{\sum v}{n} = \frac{30}{10} = 3$$

$u = x - 30$ $\therefore$ $u = \bar{x} - 30$

$\therefore$ $\bar{x} = \bar{u} - 30 = 2 + 30 = 32$

$v = y - 35$ $\therefore$ $\bar{v} = \bar{y} - 35$

$\therefore$ $\bar{y} = \bar{v} + 35 = 3 + 35 = 38$

$$\sigma_u^2 = \frac{\sum u^2}{n} - (\bar{u})^2 = \frac{180}{10} - (2)^2 = 18 - 4 = 14$$

$$\sigma_v^2 = \frac{\sum v^2}{n} - (\bar{v})^2 = \frac{488}{10} - (3)^2 = 48.8 - 9 = 39.8$$

$\therefore$ $\sigma_u = 3.742$ and $\sigma_v = 6.309$

Standard deviation is invariant to the change of origin.

$\therefore$ $\sigma_x = 3.742$ and $\sigma_y = 6.309$

$\therefore$ $\sigma_x^2 = 14$ and $\sigma_y^2 = 39.8$

$$\text{Cov}(u, v) = \frac{\sum uv}{n} - \bar{u}\,\bar{v} = \frac{-33}{10} - 2(3) = -3.3 - 6$$

$\therefore$ $\text{Cov}(u, v) = -9.3$

Covariance is invariant to the change of origin.

$\therefore$ $\text{Cov}(x, y) = \text{Cov}(u, v) = -9.3$

We have to find regression equation of y on x. It is given by

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

$$b_{yx} = \frac{\text{Cov}(x, y)}{\sigma_x^2} = \frac{-9.3}{14} = -0.664$$

$\therefore$ Regression equation becomes,

$$y - 38 = -0.664\,(x - 32)$$

$$y = -0.664\,x + 21.248 + 38$$

$$y = -0.664\,x + 59.248$$

Now, we have to estimate marks in Statistics if marks in Economics are 30, i.e. we have to find value of y when x = 30.

Substituting x = 30 in above equation, we get

$$y = -0.664 \times 30 + 59.248$$

$$y = 39.328$$

$\therefore$ Marks in Economics are 39.328 i.e. approximately 39.

Ex. 6 : *Determine regression line for price, given the supply, hence estimate price when supply is 180 units, from the following information : x = supply, y = Price, n = 7, $\Sigma(x - 150) = 119$, $\Sigma(y - 160) = 84$, $\Sigma(x - 150)^2 = 2835$, $\Sigma(y - 160)^2 = 2387$, $\Sigma(x - 150)\,\Sigma(y - 160) = 525$. Also, find correlation coefficient between price and supply.*

Sol. : Let $u = x - 150$, $v = y - 160$

$\therefore$ $$\bar{u} = \frac{119}{7} = 17, \quad \bar{v} = \frac{84}{7} = 12$$

$$\sigma_u^2 = \frac{1}{7}(2835) - (17)^2 = 405 - 289 = 116$$

$$\sigma_v^2 = \frac{1}{7}(2835) - (12)^2 = 341 - 144 = 197$$

$$\text{cov}(x, y) = \text{cov}(u, v) = \frac{1}{7}(525) - 17 \times 12 = -129$$

$$\bar{x} = 150 + \bar{u} = 167, \bar{y} = 160 + \bar{v} = 172$$

$$b_{xy} = b_{vu} = \frac{\text{cov}\,(u, v)}{\sigma_u^2} = \frac{-129}{116} = -1.1121$$

Equation of regression line y on x is

$$y - \bar{y} = b_{yx}\,(x - \bar{x})$$

$$y - 172 = (-1.1121)\,(x - 167)$$

Correlation coefficient r is obtain as

$$r = \frac{\text{cov}\,(u, v)}{\sigma_u\,\sigma_v} = \frac{-129}{\sqrt{116 \times 197}} = -0.8534$$

Since both the regression coefficients are negative, we take r = – 0.663.

Ex. 7 : *Given $x - 4y = 5$ and $x - 16y = -64$ are the regression lines, fine regression coefficient of x on y, regression coefficient of y on x and $\bar{x}$, $\bar{y}$.*

Sol.: Here by looking at the equations we cannot decide which of the equation is regression equation of x on y and which is of y on x. We arbitrarily decide one line as regression line of y on x and find regression coefficients. Then we verify whether these values are admissible.

Suppose the equation $x - 16y = -64$ represent regression line of x on y. The equation can be written as

$$x = 16y - 64 \qquad \therefore\ b_{xy} = 16 \qquad \ldots (1)$$

Next, let the equation $x - 4y = 5$ will be regression line of y on x. The equation can be written as

$$y = \frac{1}{4}x - \frac{5}{4} \qquad \therefore\ b_{yx} = \frac{1}{4} \qquad \ldots (2)$$

From (1) and (2), we have

$$r^2 = b_{xy} \times b_{yx} = 16 \times \frac{1}{4} = 4 > 1$$

Hence, our choice of regression lines is incorrect.

Next, exchanging the choice,

Suppose $x - 16y = -64$ as regression line x on y. The equation can be written as

$$y = \frac{1}{16}x + 4 \qquad \therefore\ b_{yx} = \frac{1}{16} \qquad \ldots (3)$$

and, let $x - 4y = 5$ as regression line x on y. The equation can be written as

$$x = 4y + 5 \qquad \therefore\ b_{xy} = 4 \qquad \ldots (4)$$

From (3) and (4), we have

$$r^2 = b_{yx} \times b_{xy} = \frac{1}{16} \times 4 = \frac{1}{4} < 1$$

Thus, from (3) and (4) $b_{yx} = \frac{1}{16}$ and $b_{xy} = 4$ are correct regression coefficients.

$$\text{correlation coefficient} = r^2 = b_{yx} \times b_{xy} = \frac{1}{16} \times 4 = \frac{1}{4} < 1 \quad \therefore r = \frac{1}{2} \qquad \ldots (5)$$

We choose positive square root because regression coefficients are positive.

We also, note that since $(\bar{x}, \bar{y})$ is the point of intersection of regression lines. Thus, $(\bar{x}, \bar{y})$ will satisfy both the equations

$$\therefore \qquad \bar{x} - 4\bar{y} = 5 \qquad \ldots (6)$$

and $$\bar{x} - 16\bar{y} = -64 \qquad \ldots (7)$$

Solving equations (6) and (7), we get

$$\bar{x} = 28,\ \bar{y} = \frac{23}{4}$$

Ex. 8 : *If the two lines of regression are $9x + y - \lambda = 0$ and $4x + y = \mu$ and the means of x and y are 2 and –3 respectively, find the values of λ, μ and the coefficient of correlation between x and y.* **(May 2009)**

Sol. : $\bar{x} = 2$ and $\bar{y} = -3$.

The lines of regression are $9x + y = \lambda$ and $4x + y = \mu$.

The point of intersection of two regression lines is (x, y) i.e. $(\bar{x}, \bar{y})$ lies on both the regression lines.

$$9\bar{x} + \bar{y} = \lambda \qquad \ldots (1)$$

$$4\bar{x} + \bar{y} = \mu \qquad \ldots (2)$$

Substituting values of $\bar{x}$ and $\bar{y}$, we get

$$9(2) + (-3) = \lambda$$

$$\lambda = 18 - 3 = 15$$

and $$4(2) + (-3) = \mu$$

$$\therefore \qquad \mu = 8 - 3 = 5$$

Thus, the regression lines are,

$$9x + y = 15 \quad \text{and} \quad 4x + y = 5$$

Let $9x + y = 15$ be the regression line of x on y, so it can be written as

$$x = \frac{15}{9} - \frac{y}{9}$$

$$\therefore \quad b_{xy} = -\frac{1}{9} = -0.11$$

Let $4x + y = 5$ be the regression line of y on x. So it can be written as $y = 5 - 4x$.

$$\therefore \quad b_{yx} = -4$$

Correlation coefficient between x and y is given as,

$$r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{(-4) \times (-0.11)} = \sqrt{0.44} = 0.663$$

Since both the regression coefficients are negative, we take $r = -0.663$.

Ex. 9 : *The regression equations are $8x - 10y + 66 = 0$ and $40x - 18y = 214$. The value of variance of x is 9. Find :*

(1) The mean values of x and y.

(2) The correlation x and y and

(3) The standard deviation of y. **(May 2006, Dec. 2009)**

Sol. : (1) Since both the regression lines pass through the point $(\bar{x}, \bar{y})$, we have

$$8\bar{x} - 10\bar{y} + 66 = 0 \text{ and } 40\bar{x} - 18\bar{y} = 214$$

Solving these two equations, we get

$$\bar{x} = 13 \text{ and } \bar{y} = 17$$

(2) Let $8x - 10y + 66 = 0$ be the line of regression of y on x and $40x - 18y = 214$ be the line of regression of x on y. These equations can be written in the form

$$y = \frac{8}{10}x + \frac{66}{10} \quad \text{and} \quad x = \frac{18}{40}y + \frac{214}{40}$$

i.e. $y = 0.8x + 6.6$ and $x = 0.45y + 5.35$

$\therefore$ b_{yx} = Regression coefficient of y on x

$= 0.8$

and b_{xy} = Regression coefficient of x on y

$= 0.45$

Correlation coefficient between x and y is given by

$$r = \sqrt{b_{xy} \times b_{yx}} = \sqrt{0.45 \times 0.8} = \pm 0.6$$

But since both the regression coefficients are positive, we take

$$r = +0.6$$

(3) Variance of $x = 9$, i.e. $\sigma_x^2 = 9$

$\therefore$ $\sigma_x = 3$

We have, $b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$

$$0.8 = 0.6 \times \frac{\sigma_y}{3}$$

$$\sigma_y = 4$$

Ex. 10 : *Given the following information*

	Variable x	*Variable y*
Arithmetic Mean	*8.2*	*12.4*
Standard Deviation	*6.2*	*20*

Coefficient of correlation between x and y is 0.9. Find the linear regression estimate of x, given y = 10. **(May 2019)**