$$\rho(z, x) = 1 - \frac{6\Sigma d_3^2}{n(n-1)} = 1 - \frac{6 \times 60}{10 \times 99} = 0.6$$

Since $\rho(z, x)$ is maximum, the pair of judges A and C have the nearest common approach.

Problems 25.4

1. Find the correlation coefficient between x and y from the given data :

x :	78	89	97	69	59	79	68	57
y :	125	137	156	112	107	138	123	108

(*J.N.T.U., 2005*)

2. Find the correlation co-efficient between x and y for the given value. Find also the two regression lines.

x :	1	2	3	4	5	6	7	8	9	10
y :	10	12	16	28	25	36	41	49	40	50

(*Osmania, 2003 S ; V.T.U., 2000 S*)

3. Find the co-efficient of correlation between industrial production and export using the following data and comment on the result.

Production (in crore tons) :	55	56	58	59	60	60	62
Exports (in crore tons) :	35	38	38	39	44	43	45

(*Madras, 2000 ; Cochin, 1999*)

4. Ten people of various heights as under, were requested to read the letters on a car at 25 yards distance. The number of letters correctly read is given below :

Height (in feet) :	5·1	5·3	5·6	5·7	5·8	5·9	5·10	5·11	6·0	6·1
No. of letters :	11	17	19	14	8	15	20	6	18	12

Is there any correlation between heights and visual power ?

5. Using the formula $r = \dfrac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$, find r from the following data :

x :	92	89	87	86	83	77	71	63	53	50
y :	86	88	91	77	68	85	52	82	37	57

6. Find the correlation between x (marks in Mathematics) and y (marks in Engineering Drawing) given in the following data :

y \ x	*10—40*	*40—70*	*70—100*	*Total*
0—30	5	20	—	25
30—60	—	28	2	30
60—90	—	32	13	45
Total	5	80	15	100

7. If the coefficient of correlation between two variables x and y is 0.5 and the acute angle between the lines of regression is $\tan^{-1}(3/8)$, show that $\sigma_x = \frac{1}{2}\sigma_y$.

(*V.T.U., 200*

8. Following table gives the data on rainfall and discharge in a certain river. Obtain the line of regressi of y on x.

Rainfall x (inches) :	1·53	1·78	2·60	2·95	3·42 ,
Discharge y (1000 cc) :	33·5	36·3	40·0	45·8	53·5

9. Two random variables have the regression lines with equations $3x + 2y = 26$ and $6x + y = 31$. Find mean values and the correlation coefficient between x and y.

(*Madras, 20*

Multiplying (ii) by 0·87 and subtracting from (i), we have

$$\therefore \quad [1-(0{\cdot}87)(0{\cdot}50)]\,\bar{x} = 19{\cdot}13 - (11{\cdot}64)(0{\cdot}87) \text{ or } 0{\cdot}57\,\bar{x} = 9{\cdot}00 \text{ or } \bar{x} = 15{\cdot}79$$

$$\therefore \text{ Regression co-efficient of } y \text{ on } x \text{ is } \bar{y} = 11{\cdot}64 - (0{\cdot}50)(15{\cdot}79) = 3{\cdot}74$$

Now since the co-efficient of correlation is the geometric mean between the two regression co-efficients.

$$\therefore \quad r = \sqrt{[(-0{\cdot}50)(-0{\cdot}87)]} = \sqrt{(0{\cdot}43)} = -0{\cdot}66.$$

[– ve sign is taken since both the regression co-efficients are – ve]

Example 25·16. *In the following table are recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales :*

Salesmen	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Test scores	*40*	*70*	*50*	*60*	*80*	*50*	*90*	*40*	*60*	*60*
Sales (000)	*2·5*	*6·0*	*4·5*	*5·0*	*4·5*	*2·0*	*5·5*	*3·0*	*4·5*	*3·0*

Calculate the regression line of sales on test scores and estimate the most probable weekly sales volume if a salesman makes a score of 70.

Sol. With the help of the table below, we have

$\bar{x}$ = mean of x (test scores) = 60 + 0/10 = 60.

$\bar{y}$ = mean of y (sales) = 4·5 + (– 4·5)/10 = 4·05.

Regression line of sales (y) on scores (x) is given by

$$y - \bar{y} = r(\sigma_y/\sigma_x)(x - \bar{x})$$

where

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma XY}{\sigma_x\sigma_y} \times \frac{\sigma_y}{\sigma_x} = \frac{\Sigma XY}{(\sigma_x)^2} = \left[\Sigma d_x d_y - \frac{\Sigma d_x \Sigma d_y}{n}\right] \Big/ \left[\Sigma d_x^2 - (\Sigma d_x)^2/n\right]$$

$$= \frac{140 - \dfrac{0 \times (-4.5)}{10}}{2400 - 0^2/10} = \frac{140}{2400} = 0{\cdot}06$$

$\therefore$ The required regression line is

$$y - 4{\cdot}05 = 0{\cdot}06(x - 60) \quad \text{or} \quad y = 0{\cdot}06x + 0{\cdot}45.$$

For $x = 70$, $y = 0{\cdot}06 \times 70 + 0{\cdot}45 = 4{\cdot}65$.

Thus the most probable weekly sales volume for a score of 70 is 4·65.

Test scores x	*Sales* y	*Deviation of x from assumed mean (= 60)* d_x	*Deviation of y from assumed average (= 4.5)* d_y	$d_x \times d_y$	d_x^2	d_y^2
40	2·5	–20	–2	40	400	4
70	6·0	10	1·5	15	100	2·25
50	4·5	–10	0	0	100	0
60	5·0	0	0·5	0	0	2·25
80	4·5	20	0	0	400	0
50	2·0	–10	–2·5	25	100	6·25
90	5·5	30	1	30	900	1·00
40	3·0	–20	–1·5	30	400	2·25
60	4·5	0	0	0	0	0
60	3·0	0	–1·5	0	0	2·25
		$\Sigma d_x = 0$	$\Sigma d_y = -4.5$	$\Sigma d_x d_y = 140$	$\Sigma d_x^2 = 2400$	$\Sigma d_y^2 = 18{\cdot}25$

$$\rho(B, D) = 1 - \frac{6 \times 4}{5(5^2 - 1)} = 1 - \frac{6 \times 40}{120} = 0.8$$

Since $\rho(A, B)$ is maximum, the pair of judges A and C have the nearest common approach.

Problems 25.4

1. Find the correlation coefficient between x and y from the given data :

x	78	89	97	69	59	79	68	57
y	125	137	156	112	107	138	123	108

(J.N.T.U., 2002)

2. Find the correlation co-efficient between x and y for the given value. Find also the two regression lines.

x	1	2	3	4	5	6	7	8	9	10
y	10	12	16	28	25	36	41	49	40	50

(Osmania, 2003 S ; V.T.U., 2000 S)

3. Find the co-efficient of correlation between industrial production and export using the following data and comment on the result.

Production (in crore tons)	55	56	58	59	60	60	62
Exports (in crore tons)	35	36	38	39	44	43	45

(Madras, 2000 ; Cochin, 1999)

4. Ten people of various heights as under, were requested to read the letters on a car at 25 yards distance. The number of letters correctly read is given below :

Height (in feet)	5·1	5·3	5·6	5·7	5·8	5·9	5·10	5·11	6·0	6·1
No. of letters	11	17	19	14	8	15	20	6	18	12

Is there any correlation between heights and visual power ?

5. Using the formula $r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}$, find r from the following data :

x	92	89	87	86	83	77	71	63	53	50
y	86	88	91	77	68	85	52	82	37	57

6. Find the correlation between x (marks in Mathematics) and y (marks in Engineering Drawing) given in the following data :

y \ x	10—40	40—70	70—100	Total
0—30	5	20	—	25
30—60	—	28	2	30
60—90	—	32	13	45
Total	5	80	15	100

7. If the coefficient of correlation between two variables x and y is 0·5 and the acute angle between their lines of regression is $\tan^{-1}(3/5)$, show that $\sigma_x = \frac{1}{2}\sigma_y$.

(V.T.U., 2004)

8. Following table gives the data on rainfall and discharge in a certain river. Obtain the line of regression of y on x.

Rainfall x (inches) :	1·53	1·78	2·60	2·95	3·42 .
Discharge y (1000 cc) :	33·5	36·3	40·0	45·8	53·5

9. 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.

(Madras, 2002)

$Y_i^2 - 2\Sigma X_i Y_i$

$\Sigma Y_i^2 - \Sigma d_i^2) = \frac{1}{12}(n^3 - n) - \frac{1}{2}\Sigma d_i^2.$

t between these variables is

$$= \frac{\frac{1}{12}(n^3 - n) - \frac{1}{2}\Sigma d_i^2}{\frac{1}{12}(n^3 - n)} = 1 - \frac{6\Sigma d_i^2}{n^3 - n}$$

ion coefficient and is denoted by ρ.

nts in a contest are ranked by two judges as follows :

10	*3*	*2*	*4*	*9*	*7*	*8*
8	*1*	*2*	*3*	*10*	*5*	*7*

coefficient ρ. *(V.T.U., 2002)*

then $d_i = -5, 2, -4, 2, 2, 0, 1, -1, 2, 1$

$+ 16 + 4 + 4 + 0 + 1 + 1 + 4 + 1 = 60$

$\frac{\Sigma d_i^2}{-n} = 1 - \frac{6 \times 60}{990} = 0{\cdot}6$ nearly.

A, B, C, give the following ranks. Find which pair of judges

)	*3*	*2*	*4*	*9*	*7*	*8,*
	7	*10*	*2*	*1*	*6*	*9,*
	1	*2*	*3*	*10*	*5*	*7*

(J.N.T.U., 2003)

d_1	d_2	d_3	d_1^2	d_2^2	d_3^2
$x - y$	$y - z$	$z - x$			
–2	–3	5	4	9	25
1	1	–2	1	1	4
–3	–1	4	9	1	16
6	–4	–2	36	16	4
–4	6	–2	16	36	4
–8	8	0	64	64	0
2	–1	–1	4	1	1
8	–9	1	64	81	1
1	1	–2	1	1	4
–1	2	–1	1	4	1
0	0	0	200	214	60

Multiplying (ii) by [illegible] and subtracting from (i), we have

[illegible]

[illegible]

Regression co-efficient of [illegible] and that of [illegible]

Now since the co-efficient of correlation is the geometric mean between the two regression co-efficients

$r = $ [illegible]

[illegible] sign is taken since both the regression co-efficients are [illegible]

Example [illegible]. *In the following table are recorded data showing the test scores made by salesmen on an intelligence test and their weekly sales:*

Salesmen	1	2	3	4	5	6	7	8	9	10
Test scores	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Sales ([illegible])	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Calculate the regression line of sales on test scores and estimate the most probable weekly sales volume if a salesman makes a score of 70.

Sol. With the help of the table below, we have

$\bar{x}$ = mean of x (test scores) = 60 + 0/10 = 60

$\bar{y}$ = mean of y (sales) = 4.5 + (− 4.5)/10 = 4.05

Regression line of sales (y) on scores (x) is given by

$$y - \bar{y} = r(\sigma_y/\sigma_x)(x - \bar{x})$$

where

$$r\frac{\sigma_y}{\sigma_x} = \frac{\Sigma XY}{n\sigma_x\sigma_y}\cdot\frac{\sigma_y}{\sigma_x} = \frac{\Sigma XY}{n(\sigma_x)^2} = \left[\Sigma d_x d_y - \frac{\Sigma d_x \Sigma d_y}{n}\right]\Big/\left[\Sigma d_x^2 - (\Sigma d_x)^2/n\right]$$

$$= \frac{140 - \dfrac{0 \times (-4.5)}{10}}{2400 - 0^2/10} = \frac{140}{2400} = 0.06$$

∴ The required regression line is

$$y - 4.05 = 0.06(x - 60) \qquad \text{or} \qquad y = 0.06x + 0.45.$$

For $x = 70$, $y = 0.06 \times 70 + 0.45 = 4.65$.

Thus the most probable weekly sales volume for a score of 70 is 4.65.

Test scores x	Sales y	Deviation of x from assumed mean (= 60) d_x	Deviation of y from assumed average (= 4.5) d_y	$d_x \times d_y$	d_x^2	d_y^2
40	2.5	−20	−2	40	400	4
70	6.0	10	1.5	15	100	2.25
50	4.5	−10	0	0	100	0
60	5.0	0	0.5	0	0	0.25
80	4.5	20	0	0	400	0
50	2.0	−10	−2.5	25	100	6.25
90	5.5	30	1	30	900	1.00
40	3.0	−20	−1.5	30	400	2.25
60	4.5	0	0	0	0	0
60	3.0	0	−1.5	0	0	2.25
		$\Sigma d_x = 0$	$\Sigma d_y = -4.5$	$\Sigma d_x d_y = 140$	$\Sigma d_x^2 = 2400$	$\Sigma d_y^2 = 18.25$

or $$\frac{\Sigma(z-\bar{z})^2}{n} = \frac{\Sigma(x-\bar{x})^2}{n} + \frac{\Sigma(y-\bar{y})^2}{n} - 2\frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

i.e. $$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 - 2r\sigma_x\sigma_y \qquad \left[\because\ r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}\right.$$

which is the required result.

(*b*) To find r, we have to calculate σ_x, σ_y and σ_{x-y}. We make the following table :

x	$X = x - 54$	X^2	y	$Y = y - 100$	Y^2	$y - x$	$(x-y)^2$
21	−33	1089	60	−40	1600	39	1521
23	−31	961	71	−29	841	48	2304
30	−24	576	72	−28	784	42	1764
54	0	0	83	−17	289	29	841
57	3	9	110	10	100	53	2809
58	4	16	84	−16	256	26	676
72	18	324	100	0	0	28	784
78	24	576	92	−8	64	14	196
87	33	1089	113	13	169	26	676
90	36	1296	135	35	1225	45	2025
Total	30	5936		−80	5328	350	13596

$$\therefore \quad \sigma_x^2 = \frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2 = \frac{5636}{10} - \left(\frac{30}{10}\right)^2 = 593.6 - 9 = 584.6$$

$$\sigma_y^2 = \frac{\Sigma Y^2}{N} - \left(\frac{\Sigma Y}{N}\right)^2 = \frac{5328}{10} - \left(\frac{-80}{10}\right)^2 = 532.8 - 64 = 468.8$$

$$\sigma_{x-y}^2 = \frac{\Sigma(x-y)^2}{N} - \left\{\frac{\Sigma(x-y)}{N}\right\}^2 = 1359.6 - 1225 = 134.6$$

From the above formula,

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y} = \frac{584.6 + 468.8 - 134.6}{2 \times 24.18 \times 23.85} = 0.876$$

Example 25·20. *While calculating correlation coefficient between two variables x and y from 25 pairs of observations, the following results were obtained : $n = 25$, $\Sigma x = 125$, $\Sigma x^2 = 650$, $\Sigma y = 100$, $\Sigma y^2 = 460$, $\Sigma xy = 508$. Later it was discovered at the time of checking that the pairs of values*

x	y
8	12
6	8

were copied down as

x	y
6	14
8	6

Obtain the correct value of correlation coefficient.

Sol. To get the correct results, we subtract the incorrect values and add the corresponding correct values.

∴ The correct results would be

$$\Sigma n = 25,\ \Sigma x = 125 - 6 - 8 + 8 + 6 = 125,\ \Sigma x^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$$

$$\Sigma y = 100 - 14 - 6 + 12 + 8 = 100,\ \Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$$

$$\Sigma xy = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$$

Example 25·17. *If θ is the angle between the two regression lines, show that*

$$\tan\theta = \frac{1-r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}.$$ *(V.T.U., 2007 ; Andhra, 2000 ; Ranchi, 1999)*

Explain the significance when $r = 0$ and $r = \pm 1$. *(V.T.U., 2001)*

Sol. The equations to the line of regression of y on x and x on y are

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

$\therefore$ Their slopes are $m_1 = r\,\sigma_y/\sigma_x$ and $m_2 = \sigma_y/r\,\sigma_x$

Thus $\tan\theta = \dfrac{m_2-m_1}{1+m_1m_2} = \dfrac{\sigma_y/r\,\sigma_x - r\sigma_y/\sigma_x}{1+\sigma_y^2/\sigma_x^2} = \dfrac{1-r^2}{r}\cdot\dfrac{\sigma_x\sigma_y}{\sigma_x^2+\sigma_y^2}$

When $r = 0$, $\tan\theta \to \infty$ or $\theta = \pi/2$ i.e. when the variables are independent, the two lines of regression are perpendicular to each other.

When $r = \pm 1$, $\tan\theta = 0$ i.e. $\theta = 0$ or π. Thus the lines of regression coincide i.e. there is perfect correlation between the two variables.

Example 25·18. *In a partially destroyed laboratory record, only the lines of regression of y on x and x on y are available as $4x - 5y + 33 = 0$ and $20x - 9y = 107$ respectively. Calculate $\bar{x}$, $\bar{y}$ and the coefficient of correlation between x and y.* *(V.T.U., 2005 ; Anna, 2003 S ; U.P.T.U., 2003)*

Sol. Since the regression lines pass through $(\bar{x}, \bar{y})$, therefore,

$$4\bar{x} - 5\bar{y} + 33 = 0,\quad 20\bar{x} - 9\bar{y} = 107.$$

Solving these equations, we get $\bar{x} = 13$, $\bar{y} = 17$.

Rewriting the line of regression of y on x as $y = \frac{4}{5}x + \frac{33}{5}$, we get

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{4}{5} \qquad \ldots(i)$$

Rewriting the line of regression of x on y as $x = \frac{9}{20}y + \frac{107}{9}$, we get

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{9}{20} \qquad \ldots(ii)$$

Multiplying (i) and (ii), we get

$$r^2 = \frac{4}{5}\times\frac{9}{20} = 0.36 \qquad \therefore\ r = 0.6$$

Hence $r = 0.6$, the positive sign being taken as b_{yx} and b_{xy} both are positive.

Example 25·19. *Establish the formula* $r = \dfrac{\sigma_x^2+\sigma_y^2-\sigma_{x-y}^2}{2\,\sigma_x\,\sigma_y}$

Hence calculate r from the following data :

x :	21	23	30	54	57	58	72	78	87	90
y :	60	71	72	83	110	84	100	92	113	135

(U.P.T.U. 2002)

(a) Let $z = x - y$ so that $\bar{z} = \bar{x} - \bar{y}$.

$$z - \bar{z} = (x-\bar{x}) - (y-\bar{y})$$

$$(z-\bar{z})^2 = (x-\bar{x})^2 + (y-\bar{y})^2 - 2(x-\bar{x})(y-\bar{y})$$

umming up for n terms, we have

$$\Sigma(z-\bar{z})^2 = \Sigma(x-\bar{x})^2 + \Sigma(y-\bar{y})^2 - 2\Sigma(x-\bar{x})(y-\bar{y})$$

$$\therefore \quad r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{[n\Sigma u^2 - (\Sigma u)^2]\,[n\Sigma v^2 - (\Sigma v)^2]}} = \frac{25 \times 520 - 125 \times 100}{\sqrt{[(25 \times 650 - (125)^2]\,[25 \times 436 - (100)^2]}}$$

$$= \frac{20}{\sqrt{(25 \times 36)}} = \frac{2}{3}.$$

25.15. STANDARD ERROR OF ESTIMATE

The sum of the squares of the deviations of the points from the line of regression of y on x is $\Sigma(y - a - bx)^2 = \Sigma(Y - bX)^2$, where $X = x - \bar{x}$, $Y = y - \bar{y}$

$$= \sum\left(Y - r\frac{\sigma_y}{\sigma_x}X\right)^2 = \Sigma Y^2 - 2r(\sigma_y/\sigma_x)\Sigma XY + r^2(\sigma_y^2/\sigma_x^2)\Sigma X^2$$

$$= n\sigma_y^2 - 2r(\sigma_y/\sigma_x)\, r \cdot n\sigma_x\sigma_y + r^2(\sigma_y^2/\sigma_x^2) \cdot n\sigma_x^2 = n\sigma_y^2(1 - r^2).$$

Denoting this sum of squares by nS_y^2, we have $S_y = \sigma_y\sqrt{(1 - r^2)}$ ...(1)

Since S_y is the root mean square deviation of the points from the regression line of y on x, it is called the *standard error of estimate of y*. Similarly the standard error of estimate of x is given by

$$S_x = \sigma_x\sqrt{(1 - r^2)} \qquad ...(2)$$

Since the sum of the squares of deviations cannot be negative, it follows that

$$r^2 \le 1 \quad \text{or} \quad -1 \le r \le 1.$$

i.e. correlation coefficient lies between – 1 and 1. (*J.N.T.U., 2006*)

If $r = 1$ or -1, the sum of the squares of deviations from either line of regression is zero. Consequently each deviation is zero and all the points lie on both the lines of regression. These two lines coincide and we say that the correlation between the variables is *perfect*. The nearer r^2 is to unity the closer are the points to the lines of regression. Thus *the departure of r^2 from unity is a measure of departure from linearity of the relationship between the variables.*

25.16. RANK CORRELATION

A group of n individuals may be arranged in order of merit with respect to some characteristic. The same group would give different orders for different characteristics. Considering the orders corresponding to two characteristics A and B, the correlation between these n pairs of ranks is called the *rank correlation* in the characteristics A and B for that group of individuals.

Let x_i, y_i be the ranks of the ith individuals in A and B respectively. Assuming that no two individuals are bracketed equal in either case, each of the variables taking the values 1, 2, 3, ..., n, we have

$$\bar{x} = \bar{y} = \frac{1 + 2 + 3 + \ldots\ldots + n}{n} = \frac{n(n+1)}{2n} = \frac{n+1}{2}$$

If X, Y be the deviations of x, y from their means, then

$$\Sigma X_i^2 = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 + n(\bar{x})^2 - 2\bar{x}\Sigma x_i = \Sigma n^2 + \frac{n(n+1)^2}{4} - 2\frac{n+1}{2} \cdot \Sigma n$$

$$= \frac{n(n+1)(2n+1)}{6} + \frac{n(n+1)^2}{4} - \frac{n(n+1)^2}{2} = \frac{1}{12}(n^3 - n)$$

Similarly $\Sigma Y_i^2 = \frac{1}{12}(n^3 - n)$

Now let $d_i = x_i - y_i$ so that $d_i = (x_i - \bar{x}) - (y_i - \bar{y}) = X_i - Y_i$

With the help of the above correlation table, we have

$$r = \frac{n(\Sigma f d_x d_y) - (\Sigma f d_x)(\Sigma f d_y)}{\sqrt{[\{n\Sigma f d_x^2 - (\Sigma f d_x)^2\} \times \{n\Sigma f d_y^2 - (\Sigma f d_y)^2\}]}}$$

$$= \frac{53 \times 86 - 10 \times 16}{\sqrt{[(53 \times 98 - 100) \times (53 \times 92 - 256)]}} = \frac{4398}{\sqrt{(5094 \times 4620)}} = \frac{4398}{4850} = 0{\cdot}91 \text{ (approx.)}$$

25·14. LINES OF REGRESSION

It frequently happens that the dots of the scatter diagram generally, tend to cluster along a well defined direction which suggests a linear relationship between the variables x and y. Such a line of best-fit for the given distribution of dots is called the *line of regression* (Fig. 25·6). In fact there are two such lines, one giving the best possible mean values of y for each specified value of x and the other giving the best possible mean values of x for given values of y. The former is known as the *line of regression of y on x* and the latter as the *line of regression of x on y*.

Consider first the line of regression of y on x. Let the straight line satisfying the general trend of n dots in a scatter diagram be

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

We have to determine the constants a and b so that (1) gives for each value of x, the best estimate for the average value of y in accordance with the *principle of least squares* (page 20), therefore, the normal equations for a and b are

$$\Sigma y = na + b\Sigma x \qquad \ldots(2)$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad \ldots(3)$$

(2) gives $\frac{1}{n}\Sigma y = a + b \cdot \frac{1}{n}\Sigma x$ *i.e.* $\bar{y} = a + b\bar{x}$.

This shows that $(\bar{x}, \bar{y})$, *i.e.* the means of x and y, lie on (1).

Shifting the origin to $(\bar{x}, \bar{y})$, (3) takes the form

$$\Sigma(x - \bar{x})(y - \bar{y}) = a\Sigma(x - \bar{x}) + b\Sigma(x - \bar{x})^2, \quad \text{But} \quad a\Sigma(x - \bar{x}) = 0,$$

$$\therefore \quad b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\Sigma XY}{\Sigma X^2} = \frac{\Sigma XY}{n\sigma_x^2} = r\frac{\sigma_y}{\sigma_x} \qquad \left[\because r = \frac{\Sigma XY}{n\sigma_x\sigma_y}\right]$$

Thus the line of best fit becomes $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$ ...(4)

which is the equation of *the line of regression of y on x*. Its slope is called the *regression co-efficient of y on x*.

Interchanging x and y, we find that the line of regression of x on y is

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \ldots(5)$$

Thus the *regression co-efficient of y on x* $= r\sigma_y/\sigma_x$ ...(6)

and the *regression co-efficient of x on y* $= r\sigma_x/\sigma_y$ ...(7)

Cor. *The correlation co-efficient r is the geometric mean between the two regression co-efficients.*

For $r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y} = r^2$.

Example 25·15. *The two regression equations of the variables x and y are* $x = 19{\cdot}13 - 0{\cdot}87y$ *and* $y = 11{\cdot}64 - 0{\cdot}50x$. *Find (i) mean of x's, (ii) mean of y's and (iii) the correlation co-efficient between x and y.* (V.T.U., 2004 ; Anna, 2003 ; Burdwan, 2003

Sol. Since the mean of x's and the mean of y's lie on the two regression lines, we have

$$\bar{x} = 19{\cdot}13 - 0{\cdot}87\bar{y} \qquad \ldots(i$$

$$\bar{y} = 11{\cdot}64 - 0{\cdot}50\bar{x} \qquad \ldots(ii$$

With the help of the above correlation table, we have

$$r = \frac{n\Sigma f d_x d_y - (\Sigma f d_x)(\Sigma f d_y)}{\sqrt{[n\Sigma f d_x^2 - (\Sigma f d_x)^2][n\Sigma f d_y^2 - (\Sigma f d_y)^2]}}$$

$$= \frac{53 \times 86 - 10 \times 16}{\sqrt{[(53 \times 98 - 100) \times (53 \times 92 - 256)]}} = \frac{4398}{\sqrt{(5094 \times 4620)}} = \frac{4398}{4850} = 0.91 \text{ (approx.)}$$

25.14. LINES OF REGRESSION

It frequently happens that the dots of the scatter diagram generally, tend to cluster along a well defined direction which suggests a linear relationship between the variables x and y. Such a line of best-fit for the given distribution of dots is called the *line of regression* (Fig. 25.6). In fact there are two such lines, one giving the best possible mean values of y for each specified value of x and the other giving the best possible mean values of x for given values of y. The former is known as the *line of regression of y on x* and the latter as the *line of regression of x on y*.

Consider first the line of regression of y on x. Let the straight line satisfying the general trend of n dots in a scatter diagram be

$$y = a + bx \quad \ldots(1)$$

We have to determine the constants a and b so that (1) gives for each value of x, the best estimate for the average value of y in accordance with the *principle of least squares* (page 20), therefore, the normal equations for a and b are

$$\Sigma y = na + b\Sigma x \quad \ldots(2)$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \quad \ldots(3)$$

(2) gives $\quad \frac{1}{n}\Sigma y = a + b \cdot \frac{1}{n}\Sigma x \quad i.e. \quad \bar{y} = a + b\bar{x}.$

This shows that $(\bar{x}, \bar{y})$, *i.e.* the means of x and y, lie on (1).

Shifting the origin to $(\bar{x}, \bar{y})$, (3) takes the form

$$\Sigma(x-\bar{x})(y-\bar{y}) = a\Sigma(x-\bar{x}) + b\Sigma(x-\bar{x})^2, \quad \text{But} \quad a\Sigma(x-\bar{x}) = 0,$$

$$\therefore \quad b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{\Sigma XY}{\Sigma X^2} = \frac{\Sigma XY}{n\sigma_x^2} = r\frac{\sigma_y}{\sigma_x} \qquad \left[\because r = \frac{\Sigma XY}{n\sigma_x\sigma_y}\right]$$

Thus the line of best fit becomes $y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})$ $\quad \ldots(4)$

which is the equation of *the line of regression of y on x*. Its slope is called the *regression co-efficient of y on x*.

Interchanging x and y, we find that the line of regression of x on y is

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad \ldots(5)$$

Thus the regression co-efficient of y on x $= r\sigma_y/\sigma_x$ $\quad \ldots(6)$

and *the regression co-efficient of x on y* $= r\sigma_x/\sigma_y$ $\quad \ldots(7)$

Cor. The correlation co-efficient r is the geometric mean between the two regression co-efficients.

For $\quad r\frac{\sigma_y}{\sigma_x} \times r\frac{\sigma_x}{\sigma_y} = r^2.$

Example 25.15. *The two regression equations of the variables x and y ar $= 19.13 - 0.87y$ and $y = 11.64 - 0.50x$. Find (i) mean of x's, (ii) mean of y's and (iii) th rrelation co-efficient between x and y.* (V.T.U., 2004 ; Anna, 2003 ; Burdwan, 200

Sol. Since the mean of x's and the mean of y's lie on the two regression lines, we have

$$\bar{x} = 19.13 - 0.87\bar{y} \quad \ldots($$

$$\bar{y} = 11.64 - 0.50\bar{x} \quad \ldots($$

With the help of the above correlation table, we have

$$r = \frac{n\Sigma f_i d_{x_i} d_{y_i} - (\Sigma f_i d_{x_i})(\Sigma f_i d_{y_i})}{\sqrt{[n\Sigma f_i d_{x_i}^2 - (\Sigma f_i d_{x_i})^2]\,[n\Sigma f_i d_{y_i}^2 - (\Sigma f_i d_{y_i})^2]}}$$

$$= \frac{88 + 80 - 10 - 16}{\sqrt{(88 + 88 - 100)(168 + 92 - 256)}} = \frac{4398}{\sqrt{(4264)(4630)}} = \frac{4398}{4850} = 0.91 \text{ (approx.)}$$

25.14. LINES OF REGRESSION

It frequently happens that the dots of the scatter diagram generally, tend to cluster along a well defined direction which suggests a linear relationship between the variables x and y. Such a line of best fit for the given distribution of dots is called the *line of regression* (Fig. 25.6). In fact there are two such lines, one giving the least possible mean values of y for each specified value of x and the other giving the best possible mean values of x for given values of y. The former is known as the *line of regression of y on x* and the latter as the *line of regression of x on y*.

Consider first the line of regression of y on x. Let the straight line satisfying the general trend of n dots in a scatter diagram be

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

We have to determine the constants a and b so that (1) gives for each value of x, the best estimate for the average value of y in accordance with the *principle of least squares* (page 20), therefore, the normal equations for a and b are

$$\Sigma y = na + b\Sigma x \qquad \ldots(2)$$

and

$$\Sigma xy = a\Sigma x + b\Sigma x^2 \qquad \ldots(3)$$

(2) gives $\quad \frac{1}{n}\Sigma y = a + b\cdot\frac{1}{n}\Sigma x \quad i.e., \quad \bar{y} = a + b\bar{x}.$

This shows that $(\bar{x}, \bar{y})$, *i.e.* the means of x and y, lie on (1).

Shifting the origin to $(\bar{x}, \bar{y})$, (3) takes the form

$$\Sigma(x - \bar{x})(y - \bar{y}) = a\Sigma(x - \bar{x}) + b\Sigma(x - \bar{x})^2, \quad \text{But} \quad a\Sigma(x - \bar{x}) = 0,$$

$$\therefore \quad b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\Sigma XY}{\Sigma X^2} = \frac{\Sigma XY}{n\sigma_x^2} = r\frac{\sigma_y}{\sigma_x} \qquad \left[\because r = \frac{\Sigma XY}{n\sigma_x\sigma_y}\right]$$

Thus the line of best fit becomes $y - \bar{y} = r\dfrac{\sigma_y}{\sigma_x}(x - \bar{x})$ $\qquad \ldots(4)$

which is the equation of *the line of regression of y on x*. Its slope is called the *regression co-efficient of y on x*.

Interchanging x and y, we find that the line of regression of x on y is

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \ldots(5)$$

Thus the *regression co-efficient of y on x* $\quad = r\sigma_y/\sigma_x$ $\qquad \ldots(6)$

and the *regression co-efficient of x on y* $= r\sigma_x/\sigma_y$ $\qquad \ldots(7)$

Cor. *The correlation co-efficient r is the geometric mean between the two regression co-efficients.*

For $\quad r\dfrac{\sigma_y}{\sigma_x} \times r\dfrac{\sigma_x}{\sigma_y} = r^2.$

Example 25.15. *The two regression equations of the variables x and y are $x = 19.13 - 0.87y$ and $y = 11.64 - 0.50x$. Find* (i) *mean of x's,* (ii) *mean of y's and* (iii) *the correlation co-efficient between x and y.* (V.T.U., 2004 ; Anna, 2003 ; Burdwan, 2003)

Sol. Since the mean of x's and the mean of y's lie on the two regression lines, we have

$$\bar{x} = 19.13 - 0.87\,\bar{y}$$

$$\bar{y} = 11.64 - 0.50\,\bar{x}$$

...(i)

...(ii)