12 Correlation and Regression

12.1 INTRODUCTION

Till now, we have been working on one set of observations or measurements — *e.g.* heights of students in a class, marks of students in an exam, weekly wages of workers etc. Now we shall study *joint distributions* — two or more sets of observations or measurements on the same sample — and *relationships* between them, *e.g.* relation between heights and weights of children, relation between income and expenditure, between demand and supply and so on.

If (x, y) are pairs of measurements, we say that the two measurements are *statistically related*, if knowing the value of one member x of any pair increases the precision of estimate of the second member y.

Simple Correlation is defined as the amount of similarly in direction and degree of variations in corresponding pairs of observations of two variables.

The relation between two series, or *Correlation* has following aspects:

(*i*) determining whether a relation exists, and measuring it (strength or magnitude).

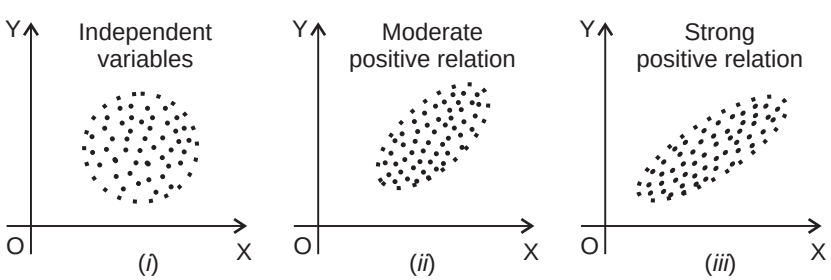
(*ii*) the *direction* of the relation (positive or negative); *e.g.* expenditure goes up as income increases; demand for a product goes down as its price increases.

(*iii*) testing whether it is significant.

(*iv*) establishing the cause and effect relation, if any.

12.2 SCATTER DIAGRAMS

If pairs (x, y) of joint observations of N samples are represented as points on the X and Y axes of a plane, we get a scatter plot, or scatter diagram. They give a good indication of relation between two variables.



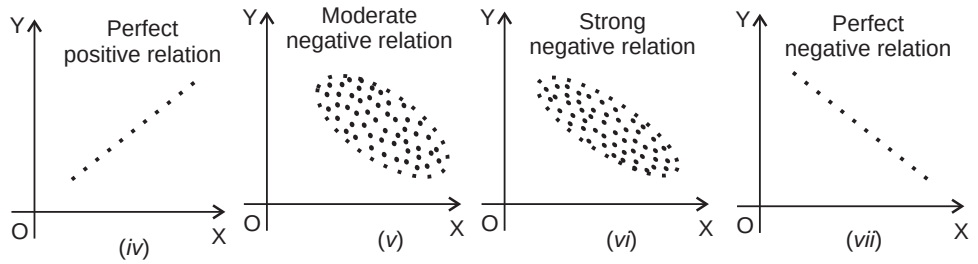


Fig. 12.1.

For example, let us draw a scatter diagram for observations
(1, 10), (2, 9), (3, 8), (4, 7), (5, 6), (6, 5), (7, 4), (8, 3), (9, 2), (10, 1).

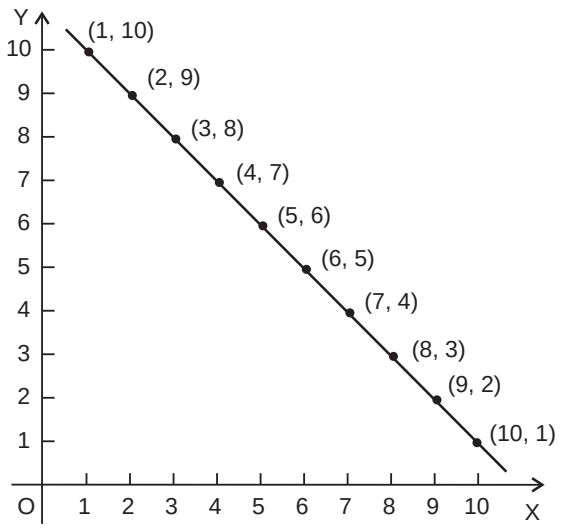


Fig. 12.2.

From the scatter diagram, we see there is a perfect negative relation between X and Y.

12.3 COVARIANCE OF X AND Y

If we plot the scatter diagrams with arithmetic means $(\bar{x}, \bar{y})$ as the origin, we get results like

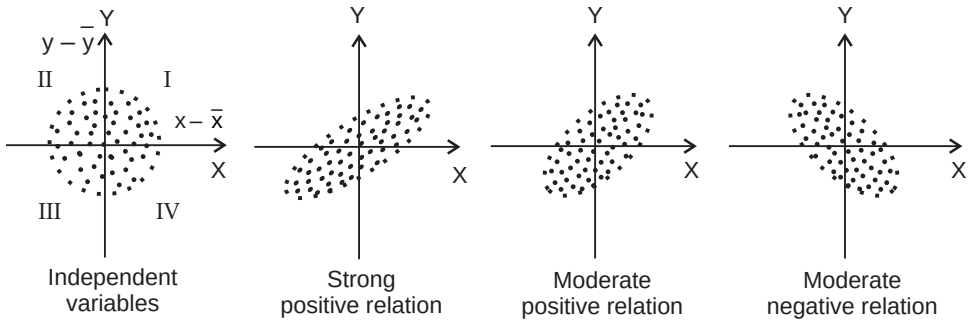


Fig. 12.3.

In first case, products $(x - \bar{x})(y - \bar{y})$ are positive in first and third quadrants, and negative in 2nd and 4th quadrants and $\Sigma(x - \bar{x})(y - \bar{y}) = 0$.

In second case, of strong positive relation, more pairs are in first and third quadrant, so $\Sigma(x - \bar{x})(y - \bar{y})$ yields a high positive result. Similarly in third case of moderate positive relation, $\Sigma(x - \bar{x})(y - \bar{y})$ yields a moderate positive result. In fourth case of moderate negative relation, more pairs lie in second and fourth quadrants, so $\Sigma(x - \bar{x})(y - \bar{y})$ yields a moderate negative result. This analysis leads us to define covariance as :

Definition. *The mean of the product of deviation scores* $(x_i - \bar{x})$ *and* $(y_i - \bar{y})$ *is called the covariance of X and Y i.e.*

$$\textbf{Cov(X, Y) or } C_{XY} = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{N} \quad \ldots(1)$$

It is easy to see that

$$\textbf{Cov(X, Y)} = \frac{1}{N}\left[\Sigma xy - \frac{1}{N}\Sigma x \, \Sigma y\right] \quad \ldots(2)$$

If x, y are small numbers, it is easier to calculate Cov (X, Y) using formula (2); if $x - \bar{x}$, $y - \bar{y}$ are small fractionless numbers, it is easy to use formula (1); in other cases, we can assume means A and B, use $u = x - A$, $v = y - B$, then

$$\textbf{Cov(X, Y)} = \frac{1}{N}\left[\Sigma uv - \frac{1}{N}\Sigma u \, \Sigma v\right] \quad \ldots(3)$$

ILLUSTRATIVE EXAMPLES

Example 1. *Find Cov(X, Y) for the following data :*

X	*3*	*4*	*5*	*6*	*7*
Y	*8*	*7*	*6*	*5*	*4*

Solution. Here N = 5. Construct the following table :

						Total
x	3	4	5	6	7	25
y	8	7	6	5	4	30
xy	24	28	30	30	28	140

Here $\Sigma x = 25$, $\Sigma y = 30$, $\Sigma xy = 140$

$$\therefore \quad \text{Cov(X, Y)} = \frac{1}{N}\left[\Sigma xy - \frac{1}{N}\Sigma x \, \Sigma y\right] = \frac{1}{5}\left[140 - \frac{1}{5}(25)(30)\right] = -2.$$

Hence, we see a negative relation between X and Y.

Example 2. *Compute Cov(X, Y) for following pairs of observations :*
(15, 44), (20, 43), (25, 45), (30, 37), (40, 34), (50, 37).

Solution. Here $\bar{x}$ = mean of values of X-variable

$$= \frac{15 + 20 + 25 + 30 + 40 + 50}{6} = \frac{180}{6} = 30$$

$$\bar{y} = \frac{44 + 43 + 45 + 37 + 34 + 37}{6} = \frac{240}{6} = 40.$$

Construct the following table :

x	$x - \bar{x}$	y	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
15	– 15	44	4	– 60
20	– 10	43	3	– 30
25	– 5	45	5	– 25
30	0	37	– 3	0
40	10	34	– 6	– 60
50	20	37	– 3	– 60
Total				– 235

So $$\text{Cov(X, Y)} = \frac{1}{N}\Sigma(x - \bar{x})(y - \bar{y})$$

$$= \frac{1}{6}(-235) = -39{\cdot}17.$$

Example 3. *Calculate the covariance for the following bivariate data :*

X	11	12	13	14	15	17	18	19	20	21
Y	14	8	12	21	19	19	23	22	17	25

Solution. Assume mean of X-variate A = 16, and for Y-variate B = 19.

x	$u = x - 16$	y	$v = y - 19$	uv
11	– 5	14	– 5	25
12	– 4	8	– 11	44
13	–3	12	– 7	21
14	– 2	21	2	– 4
15	– 1	19	0	0
17	1	19	0	0
18	2	23	4	8
19	3	22	3	9
20	4	17	– 2	– 8
21	5	25	6	30
Total	0		– 10	125

$$\text{Cov(X, Y)} = \frac{1}{N}\left[\Sigma uv - \frac{1}{N}\Sigma u \Sigma v\right] = \frac{1}{10}\left[125 - \frac{1}{10}(0)(-10)\right]$$

$$= 12{\cdot}5.$$

EXERCISE 12.1

1. Find the covariance of the data given below :
(1, 5), (2, 7), (3, 9), (4, 11), (5, 10), (6, 9), (7, 8), (8, 7), (9, 6), (10, 5).

2. Calculate the covariance of the following bivariate data :

X	4	5	6	7	8	9	10	11	12	13	14	15
Y	78	72	66	60	54	48	42	36	30	24	18	12

3. Calculate the covariance of observations (3, 5), (6, 7), (9, 9), (12, 11), (15, 13), (18, 15), (21, 17), (24, 19) using assumed means A = 13 and B = 12.

4. Calculate covariance for the following data :

X	1	2	3	4	6	7	8	9
Y	16	9	4	1	1	4	9	16

5. Prove that though covariance is independent of the choice of origin, it depends upon the scale. If $u = ax + b$, $v = cy + d$, show that $\text{cov}(u, v) = a.c.\ \text{cov}(x, y)$.

12.4 KARL PEARSON'S COEFFICIENT OF CORRELATION

Though covariance is independent of the choice of the origin, it depends on the scale of measurement. To standardise it further, we use the following formula for (Karl Pearson's) coefficient of correlation (sometimes called **Product Moment** Correlation).

$$r \text{ or } \rho(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}\, x}\sqrt{\text{Var}\, y}} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

It is easy to see that if

$u = ax + b$ and $v = cy + d$, then

$$\rho(u, v) = \frac{\text{Cov}(u, v)}{\sigma_u \sigma_v} = \frac{\Sigma(u-\bar{u})(v-\bar{v})}{N\sqrt{\frac{(u-\bar{u})^2}{N}}\sqrt{\frac{(v-\bar{v})^2}{N}}}$$

$$= \frac{\Sigma\, a(x-\bar{x})\, b(y-\bar{y})}{N\sqrt{\frac{a^2(x-\bar{x})^2}{N}}\sqrt{\frac{b^2(y-\bar{y})^2}{N}}} = \frac{ab}{|ab|}\frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \pm\rho(x, y)$$

Therefore, coefficient of correlation is independent of choice of origin and scale. Note that $-1 \le r \le 1$ (Proof is beyond the scope of this book).

If $x - \bar{x}$, $y - \bar{y}$ are small fractionless numbers, we use

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}} \quad \ldots(1)$$

If x, y are small numbers, we use

$$r = \frac{\Sigma xy - \frac{1}{N}\Sigma x \Sigma y}{\sqrt{\Sigma x^2 - \frac{1}{N}(\Sigma x)^2}\sqrt{\Sigma y^2 - \frac{1}{N}(\Sigma y)^2}} \quad \ldots(2)$$

Otherwise, we use assumed means A and B, and $u = x - A$, $v = y - B$,

$$r = \frac{\Sigma uv - \frac{1}{N}\Sigma u \Sigma v}{\sqrt{\Sigma u^2 - \frac{1}{N}(\Sigma u)^2}\sqrt{\Sigma v^2 - \frac{1}{N}(\Sigma v)^2}} \quad \ldots(3)$$

Some remarks regarding coefficient of correlation

1. The square of r *i.e.* r^2 is called **coefficient of determination.** Obviously $0 \le r^2 \le 1$. Variation between X and Y is indicated by r^2 and not r. For example, if $r = 0{\cdot}9$, there is strong positive relation between X and Y, but as $r^2 = (0{\cdot}9)^2 = 0{\cdot}81$, only 81 percent variation in Y is explained due to variation in X.

2. Correlation is said to be of high degree if $\frac{3}{4} \le |r| \le 1$, of moderate degree if $\frac{1}{4} \le |r| < \frac{3}{4}$ and of low degree if $0 \le |r| < \frac{1}{4}$.

3. If X and Y are independent variables then cov(X, Y) = 0 and coefficient of correlation $r = 0$. Inversely, if $r = 0$, then X and Y have no *linear* relation. However, Y may still have a curved relation with X. For example, for observations (– 4, 16), (– 3, 9), (– 2, 4), (– 1, 1), (1, 1), (2, 4), (3, 9), (4, 16), we find that $r = 0$ (Do it !). However, we also see that $Y = X^2$. Hence, though $r = 0$, we can still accurately predict the value of Y, given the value of X.

4. Correlation coefficient is highly abused by researchers and advertisers. It may or may not indicate cause and effect relationship. For example, in any school, you will find a high positive correlation between children's shoe size and spelling ability. Does it mean that bigger feet lead to better brains or that if you learn to spell better, your feet will get bigger? May be a third factor, that is, age of children, affects both these factors.

ILLUSTRATIVE EXAMPLES

Example 1. *Find Karl Pearson's coefficient of correlation between X and Y for the following data :*

X	*5*	*4*	*3*	*2*	*1*
Y	*4*	*2*	*10*	*8*	*6*

Solution. Here N = 5, and X, Y are small numbers. So we use formula (2).

We construct the following table :

	x	x^2	y	y^2	xy
	5	25	4	16	20
	4	16	2	4	8
	3	9	10	100	30
	2	4	8	64	16
	1	1	6	36	6
Total	15	55	30	220	80

$$\therefore \quad r = \frac{\Sigma xy - \frac{1}{N}\Sigma x \Sigma y}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}\sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{N}}}$$

$$= \frac{80 - \frac{1}{5}(15)(30)}{\sqrt{55 - \frac{1}{5}(15)^2}\sqrt{220 - \frac{1}{5}(30)^2}} = \frac{-10}{\sqrt{10}\sqrt{40}} = -\frac{10}{20}$$

$$= -0{\cdot}5.$$

Example 2. *Calculate coefficient of correlation from the following data :*

X	12	13	14	15	16	17	18
Y	14	17	18	19	20	24	28

Solution. Here $\overline{X} = \frac{\Sigma X}{N} = \frac{105}{7} = 15,$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{140}{7} = 20.$$

Also $X - \overline{X}$, $Y - \overline{Y}$ are small numbers, so we use formula (1).
We construct the following table :

X	$X - \overline{X}$ *i.e.* $X - 15$	$(X - \overline{X})^2$	Y	$Y - \overline{Y}$ *i.e.* $Y - 20$	$(Y - \overline{Y})^2$	$(X - \overline{X})(Y - \overline{Y})$
12	−3	9	14	−6	36	18
13	−2	4	17	−3	9	6
14	−1	1	18	−2	4	2
15	0	0	19	−1	1	0
16	1	1	20	0	0	0
17	2	4	24	4	16	8
18	3	9	28	8	64	24
		$\Sigma(X-X)^2 = 28$			$\Sigma(Y-Y)^2 = 130$	$\Sigma(X - \overline{X})(Y - \overline{Y}) = 58$

$$\therefore \quad r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma(X - \overline{X})^2}\sqrt{\Sigma(Y - \overline{Y})^2}} = \frac{58}{\sqrt{28}\sqrt{130}} = \frac{58}{\sqrt{3640}} = 0{\cdot}961.$$

Example 3. *Find the Karl Pearson's coefficient of correlation between x and y for the following data :*

x	16	18	21	20	22	26	27	15
y	22	25	24	26	25	30	33	18

(*I.S.C. 2013*)

Solution. Assume mean A = 20 for the x-variate and B = 25 for y-variate and we shall use the formula (3).

x	$u = x - 20$	u^2	y	$v = y - 25$	v^2	uv
16	– 4	16	22	– 3	9	12
18	– 2	4	25	0	0	0
21	1	1	24	–1	1	–1
20	0	0	26	1	1	0
22	2	4	25	0	0	0
26	6	36	30	5	25	30
27	7	49	33	8	64	56
15	– 5	25	14	–11	121	55
	5	135		– 1	221	152

$$\text{Hence, } \rho(X, Y) = \frac{\Sigma u v - \frac{1}{N}\Sigma u \Sigma v}{\sqrt{\Sigma u^2 - \frac{1}{N}(\Sigma u)^2}\sqrt{\Sigma v^2 - \frac{1}{N}(\Sigma v)^2}}$$

$$= \frac{152 - \frac{1}{8}(5)(-1)}{\sqrt{135 - \frac{1}{8}(5)^2}\sqrt{221 - \frac{1}{8}(-1)^2}} = \frac{1221}{\sqrt{1055}\sqrt{1767}}$$

$$= 0{\cdot}894.$$

Example 4. *Find the correlation coefficient between the heights of husbands and wives based on the following data (given in inches) and interpret the result.*

Couple	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Height of husband	*76*	*75*	*75*	*72*	*72*	*71*	*71*	*70*	*68*	*68*	*68*	*68*	*67*	*67*	*62*
Height of wife	*71*	*70*	*70*	*67*	*71*	*65*	*65*	*67*	*64*	*65*	*65*	*66*	*63*	*65*	*61*

Solution. We use assumed means A = 70, B = 66, and shall use the formula (3)

Couple	x	$u = x - A = x - 70$	u^2	y	$v = y - B = y - 66$	v^2	uv
1	76	6	36	71	5	25	30
2	75	5	25	70	4	16	20
3	75	5	25	70	4	16	20
4	72	2	4	67	1	1	2
5	72	2	4	71	5	25	10
6	71	1	1	65	– 1	1	– 1
7	71	1	1	65	– 1	1	– 1
8	70	0	0	67	1	1	0
9	68	– 2	4	64	– 2	4	4
10	68	– 2	4	65	– 1	1	2
11	68	– 2	4	65	– 1	1	2
12	68	– 2	4	66	0	0	0
13	67	– 3	9	63	– 3	9	9
14	67	– 3	9	65	– 1	1	3
15	62	– 8	64	61	– 5	25	40
Total		0	194		5	127	140

$$\therefore \quad r = \frac{\Sigma uv - \frac{1}{N}\Sigma u \Sigma v}{\sqrt{\Sigma u^2 - \frac{(\Sigma u)^2}{N}}\sqrt{\Sigma v^2 - \frac{(\Sigma v)^2}{N}}} = \frac{140 - \frac{(0)(5)}{15}}{\sqrt{194 - \frac{(0)^2}{15}}\sqrt{(127)^2 - \frac{(5)^2}{15}}}$$

$= 0{\cdot}89$, which is a strong positive correlation.

This shows that tall men usually marry tall women and short men marry short women (called *assortive mating*).

Example 5. *The following table shows the percentage of cats killed while falling from various storeys from skyscrapers in Newyork. Calculate correlation coefficient and draw scatter diagram. Comment on the results.*

Fallen from number of storeys	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Percentage killed	*3*	*9*	*12*	*15*	*18*	*15*	*12*	*9*	*3*

Solution. We use assumed means A = 5 and B = 12.

x	$u = x - 5$	u^2	y	$v = y - 12$	v^2	uv
1	−4	16	3	−9	81	36
2	−3	9	9	−3	9	9
3	−2	4	12	0	0	0
4	−1	1	15	3	9	−3
5	0	0	18	6	36	0
6	1	1	15	3	9	3
7	2	4	12	0	0	0
8	3	9	9	−3	9	−9
9	4	16	3	−9	81	−36
Total	0	60		−12	234	0

$$\therefore \quad r = \frac{\Sigma uv - \frac{1}{N}\Sigma u \Sigma v}{\sqrt{\Sigma u^2 - \frac{1}{N}(\Sigma u)^2}\sqrt{\Sigma v^2 - \frac{1}{N}(\Sigma v)^2}} = 0, \text{ as both } \Sigma uv = 0 \text{ and } \Sigma u = 0.$$

Though $r = 0$ means there is no *linear* relationship between X and Y, the scatter diagram clearly shows a *non-linear* relationship between X and Y. This shows that upto 5th storey, percentage of cats getting killed increases, but thereafter it decreases, so much so that only 3% cats are killed when they fall from 9th storey. Cats have highly flexible bodies — when they fall from higher storeys, they get sufficient time to stretch their bodies like parachutes, which saves them from getting killed — even from getting their legs broken.

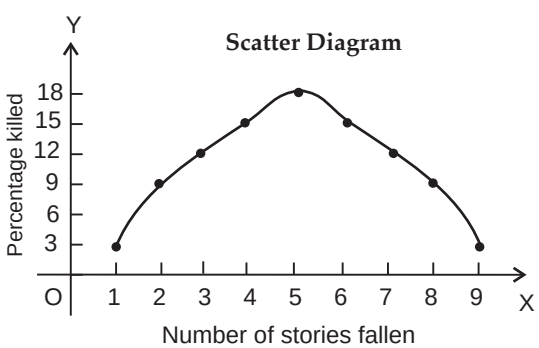


Fig. 12.4.

Example 6. *A student while calculating correlation coefficient between two variables x and y for 25 pairs of observations obtained the following results :*

$\Sigma x = 125$, $\Sigma x^2 = 650$, $\Sigma y = 100$, $\Sigma y^2 = 460$ *and* $\Sigma xy = 508$.

On rechecking, it was found that he had wrongly copied two pairs as (6, 14) and (8, 6) whereas values were (8, 12) and (6, 8). Calculate the correct correlation coefficient between x and y.

Solution. Correct $\Sigma x = 125 - 6 - 8 + 8 + 6 = 125$

Correct $\Sigma y = 100 - 14 - 6 + 12 + 8 = 100$

Correct $\Sigma x^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Correct $\Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2$

$= 460 - 196 - 36 + 144 + 64 = 436$

Correct $\Sigma xy = 508 - (6)(14) - (8)(6) + (8)(12) + (6)(8)$

$= 508 - 84 - 48 + 96 + 48 = 520.$

$\therefore$ Correct coefficient of correlation

$$r = \frac{\Sigma xy - \frac{1}{n}\Sigma x \Sigma y}{\sqrt{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}\sqrt{\Sigma y^2 - \frac{(\Sigma y)^2}{n}}}$$

$$= \frac{520 - \frac{1}{25}(125)(100)}{\sqrt{650 - \frac{1}{25}(125)^2}\sqrt{436 - \frac{1}{25}(100)^2}} = \frac{20}{\sqrt{25}\sqrt{36}} = \frac{20}{30}$$

$= 0{\cdot}667.$

EXERCISE 12.2

1. Find $\rho(x, y)$ if $\text{cov}(x, y) = -16{\cdot}5$, $\text{var}(x) = 2{\cdot}25$ and $\text{var}(y) = 144$.

2. If $n = 10$, $\Sigma x = 26$, $\Sigma y = -27$, $\Sigma x^2 = 226$, $\Sigma y^2 = 267$, $\Sigma xy = 7$, find correlation coefficient.

3. For the observations (1, 2), (2, 4), (3, 6), (4, 8), (5, 10), (6, 12), (7, 14), (8, 16), (9, 18), (10, 20), calculate cov(X, Y) and ρ(X, Y). Also make scatter diagram. Interpret the result.

4. Calculate the coefficient of correlation between X and Y from the following data using Karl Pearson's method.

X	1	2	3	4	5
Y	2	5	3	8	7

(*I.S.C. 2007*)

5. Compute Karl Pearson's Coefficient of Correlation between sales and expenditures of a firm for six months.

Sales (in lakh of ₹)	18	20	27	20	21	29
Expenditure (in lakh of ₹)	23	27	28	28	29	30

(*I.S.C. 2011*)

6. The coefficient of correlation between two variables X and Y is 0·64. Their covariance is 16. The variance of X is 9. Find the standard deviation of Y-series.

7. Find the covariance and the coefficient of correlation between x and y when $n = 10$, $\Sigma x = 60$, $\Sigma y = 60$, $\Sigma x^2 = 400$, $\Sigma y^2 = 580$ and $\Sigma xy = 305$.

8. Calculate correlation coefficient from the following results :
$N = 10$, $\Sigma X = 100$, $\Sigma Y = 150$, $\Sigma(X - 10)^2 = 180$, $\Sigma(Y - 15)^2 = 215$, $\Sigma(X - 10)(Y - 15) = 60$.

9. Coefficient of correlation between X and Y for 50 observations is 0·3, mean of X is 10 and that of Y is 6, S.D. of X is 3 and that of Y is 2. Later it was discovered that one pair of values (10, 6) was inaccurate and hence weeded out. Calculate the correlation between remaining 49 pairs of values.

10. Calculate Karl Pearson's coefficient of correlation between the marks in English and Mathematics obtained by 10 students :

English	20	13	18	21	11	12	17	14	19	15
Mathematics	17	12	23	25	14	8	19	21	22	19

11. Find Karl Pearson's coefficient of correlation between X and Y for the following data :

X	16	18	21	20	22	26	27	15
Y	22	25	24	26	25	30	33	14

(*I.S.C. 2003*)

12. From the following table, calculate the Karl Pearson's coefficient of correlation.

X	6	2	10	4	8
Y	9	11	?	8	7

Arithmetic means of X and Y series are 6 and 8 respectively.

13. The weights of sons and fathers (in kilograms) are given below :

Weight of father	65	66	67	67	68	69	70	72
Weight of son	67	68	65	68	72	72	69	71

Find the coefficient of correlation.

14. Calculate Karl Pearson's coefficient of correlation from the following data and interpret the result :

Serial number of student	1	2	3	4	5	6	7	8	9	10
Marks in mathematics	15	18	21	24	27	30	36	39	42	48
Marks in statistics	25	25	27	27	31	33	35	41	41	45

15. Calculate Karl Pearson's coefficient of correlation between x and y for the following data :

X	6	2	4	9	1	3	5	8
Y	13	8	12	15	9	10	11	16

[Take assumed mean for x as 5 and for y as 12.] (*I.S.C. 2005*)

16. Calculate ρ(X, Y) for the following data and comment on the result.

X	– 4	– 3	– 2	– 1	1	2	3	4
Y	16	9	4	1	1	4	9	16

17. Calculate Karl Pearson's coefficient of correlation between X and Y from the following data and comment on the result.

X	1	2	3	4	5
Y	7	6	5	4	3

18. A psychologist selected a random sample of 22 students. He grouped them in 11 pairs so that students in a pair have nearly equal scores in an intelligence test. In each pair one student was taught by method A and the other by method B and examined after the course. The marks obtained by them are tabulated below :

Pair	1	2	3	4	5	6	7	8	9	10	11
Method A (marks)	24	29	19	14	30	19	27	30	20	28	11
Method B(marks)	37	35	16	26	23	27	19	20	16	11	21

Find the rank correlation coefficient.

12.5 LINES OF REGRESSION

Very often, we are required to "estimate" things. People lay bets on cricket matches based on past performance; companies estimate (project) sales based on past data, and so on.

Given the data (x_1, y_1), (x_2, y_2), …, (x_n, y_n), we can plot a scatter diagram and then visualise a smooth curve approximating the data. This is called *curve fitting*.

7. The following table gives the result of heights of 8 athletes and the measurements of their long jumps and high jumps (all in centimeters). Calculate the Spearman's rank correlation between height and long jump and between height and high jump. Interpret the result.

Athlete	*Height*	*Long jump*	*High jump*
A	158	324	175
B	165	365	185
C	162	380	180
D	170	400	184
E	175	350	200
F	163	350	172
G	178	425	188
H	164	375	180

8. The marks of 8 students in an examination in Mathematics and Statistics are given below :

Student No.	1	2	3	4	5	6	7	8
Marks in Mathematics	70	48	58	55	54	50	60	52
Marks in Statistics	62	47	53	60	55	68	51	48

Calculate Spearman's rank correlation coefficient.

9. A national consumer protection society investigated seven brands of paint to determine their quality relative to price. The society's conclusions were ranked according to the following table :

Brand	T	U	V	W	X	Y	Z
Price/litre	192	158	135	160	205	139	177
Quality ranking	2	6	7	4	3	5	1

Using Spearman's rank correlation coefficient, determine whether the consumer generally gets value for money.

ANSWERS

EXERCISE 12.1

1. – 1·35. **2.** – 71·5. **3.** 31·5. **4.** 0.

EXERCISE 12.2

1. $r = -0{\cdot}92$. **2.** $r = 0{\cdot}44$.

3. Cov(X, Y) = 16·5 and $r = 1$, which shows perfect positive linear relation.

4. 0·806. **5.** 0·68 (approx.). **6.** 8·33.

7. – 5·5, – 0·5863. **8.** 0·305. **9.** 0·3 (no effect).

10. 0·748. **11.** 0·894. **12.** – 0·919.

13. $r = 0{\cdot}603$, which shows a moderate but significant relationship between weights of fathers and sons.

14. $r = 0{\cdot}98$, which shows a strong positive relationship, which points to application of common skills / intelligence factor. But it does *not* mean that if you study only mathematics and score good marks in it, you will automatically score well in statistics. So take care!

15. $r = 0{\cdot}946$.

16. $r = 0$, which shows lack of *linear* relationship, though we see that $Y = X^2$.

17. $r = -1$, which shows perfect negative correlation.

18. $r = -0{\cdot}227$.