5. CORRELATION

Let's Study

- Concept of Correlation
- Methods of computing correlation
- Properties of Covariance
- Karl Pearson's coefficient of correlation
- Scatter diagram
- Interpretation

Let's Observe…

So far, we have studied the statistical methods used for analysis of data involving only one variable. There are situations where two variables are involved

For example consider following data,

(i) Demand and Price of a certain commodity over a specified period of time.

(ii) Weight of a person and Blood Pressure of the person.

(iii) Quantity of water and crop yield.

(iv) Sales of cosmetics and advertisements

(v) Monthly income and expenditure of a family.

A set of observations made on two variables is called Bivariate Data. The two variables are denoted by X and Y respectively. Then observations on two variables X and Y can be represented by *n ordered pairs* (x_1,y_1), (x_2,y_2),...... (x_n,y_n),.. *The pair* (x_i,y_i),values of the variables for ith observation. For example X denotes demand of the commodity and Y denotes price of the commodity then x_i *denotes demand of* i^{th} commodity and y_i *denotes price of* ith commodity.

Let's Learn

5.1 Concept of Correlation

In a bivariate data, we may be interested in finding if there is any relationship or association between the two variables.“A correlation is a measure of association or relation”. If we observe in the bivariate data the changes in one variable are accompanied by changes in the other variable then the two variables are said to be correlated. In this case we say that there is a correlation between two underlying variables.

For example,

i) Intelligence Quotient (IQ) and marks of a student.

ii) Demand and price of a commodity.

Let's discuss:

5.2 Covariance:-

Covariance is a measure of joint variation between the two variables. If (x_1, y_1), (x_2, y_2), (x_n, y_n) are n ordered pairs of values of x and y, then covariance between X and Y is defined by

$$\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right),$$

$$\text{where } \bar{x} = \frac{\sum x_i}{n} \text{ and } \bar{y} = \frac{\sum y_i}{n}.$$

The above formula can be simplified as follows :

$$\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}\right)$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{x}\bar{y}\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i y_i\right] - \left[\bar{y}\bar{x} - \bar{x}\bar{y} + \bar{x}\bar{y}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

This formula is used in practice.

5.3 Properties of covariance:

(i) Cov (X,Y) = Cov (Y,X)

(ii) Cov (X,C) = 0 where C is a constant

(iii) Covariance may be positive, negative or zero.

(iv) Cov (X,X) = Var (X)

(v) Covariance is invariant under change of origin but is affected by change of scale.

That is if $U = \frac{x-a}{h}$ and $V = \frac{y-b}{k}$, where *a*, *b*, *h*, *k* are constants and $h \neq 0$, $k \neq 0$ then,

$$\text{Cov}\,(U,V) = \frac{1}{hk}\,\text{Cov}\,(X,Y)$$

$$\text{Cov}(X,Y) = hk\text{Cov}(U,V)$$

Note that these are standard deviations of X and Y respectively.

$$\sigma x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}$$

$$\sigma y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2}$$

Compare and understand the difference between variance and co-variance.

5.4 Correlation coefficient:-

Karl Pearson (1867-1936) developed a measure for the degree of relation between two variables. This measure is called correlation coefficient.

Correlation coefficient between two random variables X and Y denoted by r_{xy} or $r(x,y)$ is defined by

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

Properties of correlation coefficient r(*x*, *y*):-

(i) $r_{xy} = r_{yx}$ (order of variable is not important).

(ii) Change of origin and scale :

Correlation coefficient (r) does not change its magnitude under the change of origin and scale.

(iii) But if one of the change of scale has Negative sign then correlation coefficient becomes negative.

Correlation $\left(\frac{x-a}{h}, \frac{y-b}{k}\right)$ = Correlation (*x*,*y*)

if h, k has same algebraic sign $r_{uv} = r_{xy}$ also h, k $\neq$ 0, $r_{uv} = -r_{xy}$ if h, k have opposite algebraic sign

(iv) Correlation (x, x) = 1.

(v) r lies between −1 and 1 that is $-1 \leq r \leq 1$

5.5 Scatter Diagram:-

A scatter diagram is a graphical method of presenting bivariate data (grouped and ungrouped).

Correlation can be observed in a scatter diagram.

Following are examples of different situations given different types of scatter diagrams.

(I) a) Perfect positive correlation:-

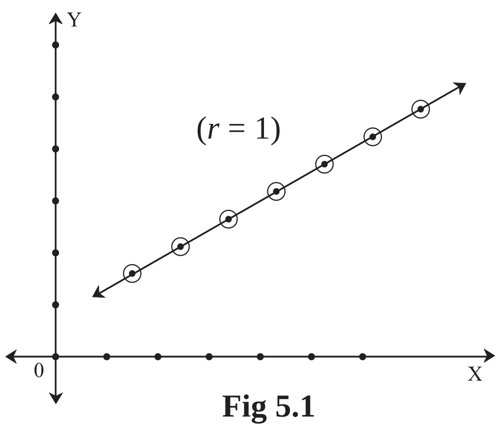

Fig 5.1

If the points are rising from left to right in a straight line then scatter diagram indicates a perfect positive correlation.

b) Positive correlation with high degree:-

If the band is rising from left to right then it indicates positive correlation. If the width of the band is smaller, then the correlation is of high degree.

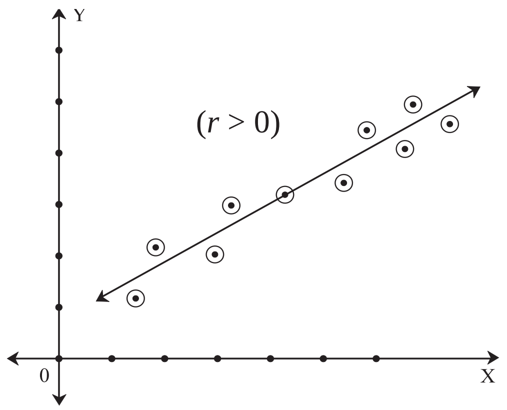

Fig 5.2

c) Positive correlation with low degree:-

If the band is rising from left to right then it indicates positive correlation. If the width of the band is bigger, then the correlation is of low degree.

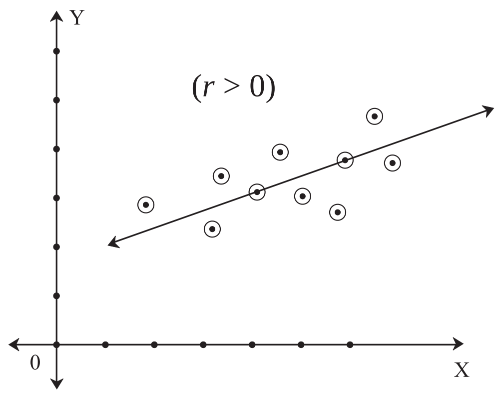

Fig 5.3

(II) a) Perfect negative correlation:-

If the points are falling from left to right in a straight line then scatter diagram indicates a perfect negative correlation.

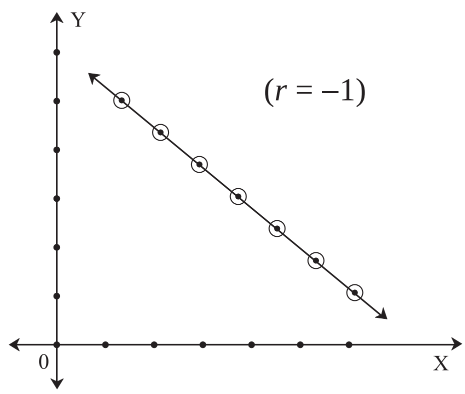

Fig 5.4

b) Negative correlation with high degree:-

If the band is falling down from left to right it indicates negative correlation.If the width of the band is smaller, then the correlation is of high degree.

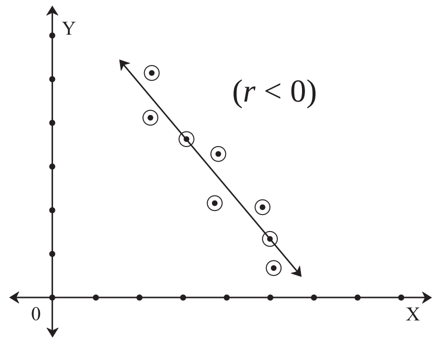

Fig 5.5

c) Negative correlation with low degree:-

If the band is falling down from left to right it indicates negative correlation.If the width of the band is bigger then the correlation is of low degree.

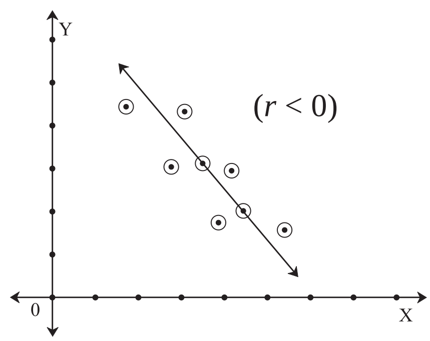

Fig 5.6

(III) No correlation (Zero correlation):-

In this case no trend line is observed.

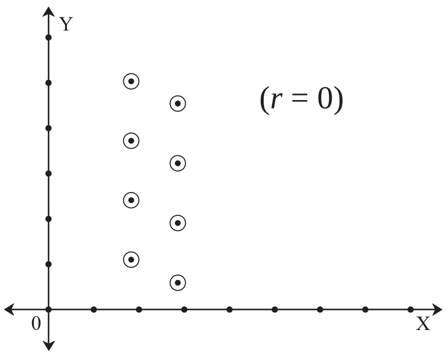

Fig 5.7

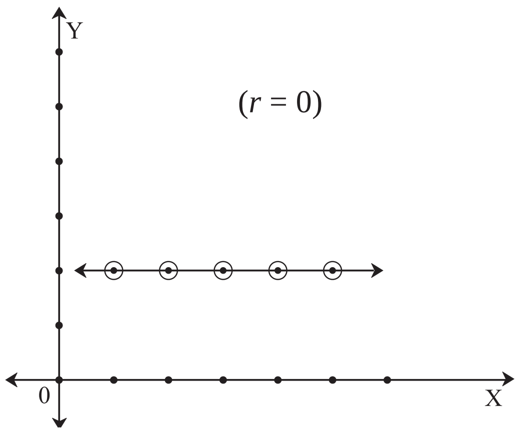

Fig 5.8

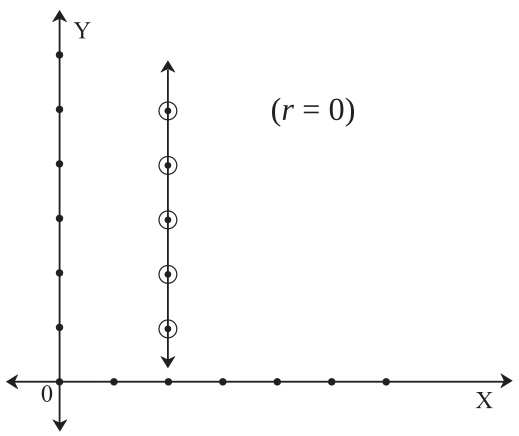

Fig 5.9

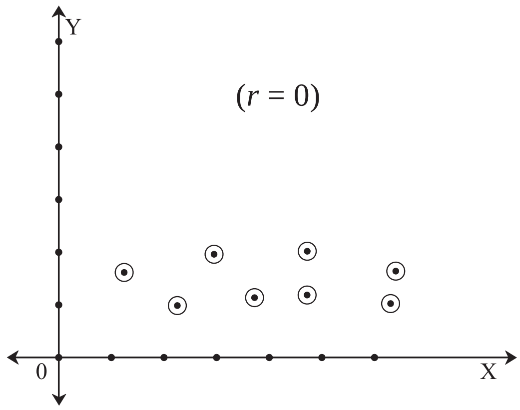

Fig 5.10

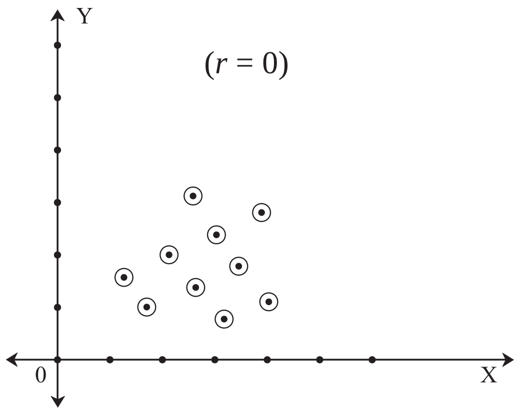

Fig 5.11

5.6 Interpretation of value of correlation coefficient:-

If $r > 0$, the correlation is positive.

If $r < 0$, the correlation is negative.

If $r = 0$, there is no correlation.

If $r > 0.8$, there is high positive correlation.

If $0.3 < r < 0.8$, there is moderate positive correlation.

If $|r| < 0.3$, the correlation is insignificant or poor.

If $r = 1$, the correlation is perfect positive.

If $r = -1$, the correlation is perfect negative.

Alternative formula of correlation coefficient:-

(i) $r = \dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{(\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$

when $\bar{x}$, $\bar{y}$ are integers & small nos.

(ii) $r = \dfrac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2} \times \sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

when $\bar{x}$, $\bar{y}$ are decimals $\sum x$, $\sum y$ are comparitively small nos.

(iii) For change of origin & scale

$$r_{uv} = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2} \times \sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

SOLVED EXAMPLES

1) A train travelled between two stations and distance and time were recorded as below,

Distance(km)	80	120	160	200	240
Time(Hr)	2	3	4	5	6

Draw scatter diagram and identify the type of correlation.

Solution: Here we take distance on X- axis and Time on Y- axis and plot the points as below,

Graph:-

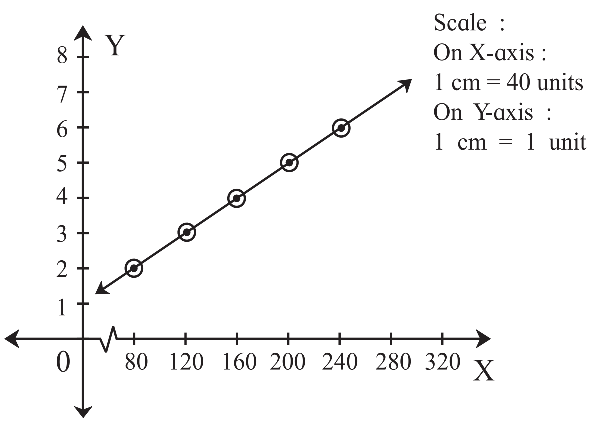

Fig 5.12

Since all the points lie on the straight line rising from left to right, there is perfect positive correlation between distance and time for the train.

2) Draw scatter diagram for the following data and identify the type of correlation.

Capi-tal (in crores Rs.)	2	3	4	5	6	8	9
Profit (in lakh Rs.)	6	5	7	7	8	9	10

Solution: Here we take capital on X- axis and profit on Y- axis and plot the points as below,

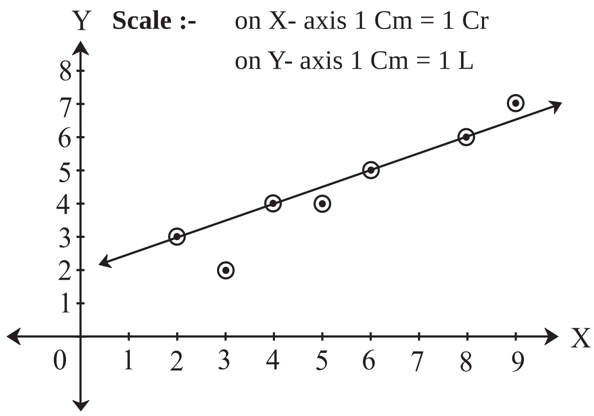

Fig 5.13

We get a band of points rising left to right. This indicates the positive correlation between capital and profit.

3) Compute correlation coefficient for the following data,

$n = 100$, $\bar{x} = 62$, $\bar{y} = 53$, $\sigma_x = 10$,

$\sigma_y = 12$, $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 8000$.

Solution: Given that $n = 100$, $\bar{x} = 62$, $\bar{y} = 53$, $\sigma_x = 10$, $\sigma_y = 12$,

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 8000$.

For finding correlation coefficient, we require cov(x,y) and σ_x and σ_y

require cov(x,y) and σ_x and σ_y

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

$$= \frac{8000}{100}$$

$$= 80.$$

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{80}{10\times12} = 0.67$$

4) Find correlation coefficient between x and y for the following data and interpret it.

x	1	2	3	4	5	6	7	8	9
y	12	11	13	15	14	17	16	19	18

$(\sqrt{666} = 25.80)$

Solution: For finding correlation coefficient, we require cov (x,y) and x and y We construct the following table,

x_i	y_i	x_i^2	y_i^2	x_iy_i
1	12	1	144	12
2	11	4	121	22
3	13	9	169	39
4	15	16	225	60
5	14	25	196	70
6	17	36	289	102
7	16	49	256	112
8	19	64	361	152
9	18	81	324	162
Total 45	135	285	2085	731

Table 5.1

From table we have,

$\Sigma x_i = 45$, $\Sigma y_i = 135$, $\Sigma x_i^2 = 285$, $\Sigma y_i^2 = 2085$, $\Sigma x_i y_i = 731$.

$$\therefore \bar{x} = \frac{\Sigma x_i}{n} = \frac{45}{9} = 5. \quad \bar{y} = \frac{\Sigma y_i}{n} = \frac{135}{9} = 15.$$

$$cov(x,y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\,\bar{y} = \frac{731}{9} - (5)(15)$$

$$= 81.22 - 75 = 6.22.$$

$$\sigma_x = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n} x_i^2 - (\bar{x})^2} = \sqrt{\frac{285}{9} - 5^2}$$

$$= \sqrt{31.66 - 25} = \sqrt{6.66}$$

$$\sigma_y = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n} y_i^2 - (\bar{y})^2} = \sqrt{\frac{2085}{9} - 15^2}$$

$$= \sqrt{231.66 - 225} = \sqrt{6.66}$$

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{6.22}{\sqrt{6.66}\sqrt{6.66}} = \frac{6.22}{6.66} = 0.93$$

There is high degree positive correlation between x and y.

5) Calculate correlation coefficient from the following data,

$n = 10$, $\Sigma x_i = 140$, $\Sigma y_i = 150$, $\Sigma(x_i - 10)^2 = 180$, $\Sigma(y_i - 15)^2 = 500$, and

$\Sigma(x_i - 10)(y_i - 15) = 60.$

Solution: We are given that $\Sigma x_i = 140$, $\Sigma y_i = 150$, $\Sigma(x_i - 10)^2 = 180$, $\Sigma(y_i - 15)^2 = 500$, and $\Sigma(x_i - 10)(y_i - 15) = 60$.

Let us define $u_i = x_i - 10$ and $v_i = y_i - 15$, then we have,

$\Sigma u_i = \Sigma(x_i - 10) = \Sigma x_i - \Sigma 10 = \Sigma x_i - 10n$

$= 140 - 10\times10 = 40.$

$\Sigma v_i = \Sigma(y_i - 15) = \Sigma y_i - \Sigma 15 = \Sigma y_i - 15n$

$= 150 - 150 = 0.$

$\Sigma u_i^2 = \Sigma(x_i - 10)^2 = 180.$

$\Sigma v_i^2 = \Sigma(y_i - 15)^2 = 500.$

$\Sigma u_i v_i = \Sigma(x_i - 10)(y_i - 15) = 60.$

$$\bar{u} = \frac{\Sigma u_i}{n} = \frac{40}{10} = 4. \quad \bar{v} = \frac{\Sigma v_i}{n} = \frac{0}{10} = 0.$$

$$\sigma_u = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n} u_i^2 - \bar{u}^2} = \sqrt{\frac{180}{10} - 4^2}$$

$$= \sqrt{18 - 16} = \sqrt{2}$$

$$6_v = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n} v_i^2 - \bar{v}^2} = \sqrt{\frac{500}{10} - 0^2}$$

$= \sqrt{50-0} = \sqrt{50}$

$\therefore \text{cov(u,v)} \ \frac{1}{n}\Sigma u_i v_i - \bar{u}\ \bar{v} = \frac{60}{10} - (4)(0) = 6$

$r_{uv} = \frac{cov(u,v)}{\sigma_u \sigma_v} = \frac{6}{\sqrt{2}\sqrt{50}} = 0.6$

But $r_{xy} = r_{uv} = 0.6$.

6) Find correlation coefficient between x and y for the following data

$n = 25$, $\Sigma x_i = 75$, $\Sigma y_i = 100$, $\Sigma x_i^2 = 250$, $\Sigma y_i^2 = 500$, $\Sigma x_i y_i = 325$.

Solution : We are given that, $n = 25$, $\Sigma x_i = 75$, $\Sigma y_i = 100$, $\Sigma x_i^2 = 250$, $\Sigma y_i^2 = 500$, $\Sigma x_i y_i = 325$.

$\therefore \ \bar{x} = \frac{\Sigma x_i}{n} = \frac{75}{25} = 3. \ \bar{y} = \frac{\Sigma y_i}{n} = \frac{100}{25} = 4.$

$\sigma_x = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n} x_i^2 - (\bar{x})^2} = \sqrt{\frac{250}{25} - 3^2}$

$= \sqrt{10-9} = 1$

$\sigma_y = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n} y_i^2 - (\bar{y})^2} = \sqrt{\frac{500}{25} - 4^2}$

$= \sqrt{20-16} = \sqrt{4} = 2$

$\therefore \text{cov (x,y)} = \frac{1}{n}\Sigma x_i y_i - \bar{x}\ \bar{y} = \frac{325}{25} - (3)(4)$

$= 13 - 12 = 1$

$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{1}{2} = 0.5$

7) Calculate correlation coefficient between age of husbands and age of wives.

Age of husbands	23	27	28	29	30	31	33	35	36	39
Age of wives	18	22	23	24	25	26	28	30	31	34

Solution: Here the change of origin and scale property can be used to find the correlation coefficient. We construct the table as below,

x_i	y_i	$u_i=x_i-31$	$v_i=y_i-25$	u_i^2	v_i^2	$u_i v_i$
23	18	-8	-7	64	49	56
27	22	-4	-3	16	9	12
28	23	-3	-2	9	4	6
29	24	-2	-1	4	1	2
30	25	-1	0	1	0	0
31	26	0	1	0	1	0
33	28	2	3	4	9	6
35	30	4	5	16	25	20
36	31	5	6	25	36	30
39	34	8	9	64	81	72
Total -	-	1	11	203	215	204

Table 5.2

From table we have,

$\Sigma u_i = 1$, $\Sigma v_i = 11$, $\Sigma u_i^2 = 203$, $\Sigma v_i^2 = 215$,

$\Sigma u_i v_i = 204$.

$\bar{u} = \frac{\Sigma u_i}{n} = \frac{1}{10} = 0.1. \ \bar{v} = \frac{\Sigma v_i}{n} = \frac{11}{10} = 1.1.$

$\therefore \text{cov (u,v)} = \frac{1}{n}\Sigma u_i v_i - \bar{u}\ \bar{v}$

$= \frac{204}{10} - (0.1)(1.1) = 20.4 - 0.11 = 20.29$

$\sigma_u = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n} u_i^2 - (\bar{u})^2} = \sqrt{\frac{203}{10} - 0.1^2}$

$= \sqrt{20.3 - 0.01} = \sqrt{20.29}$

$\sigma_v = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n} v_i^2 - (\bar{v})^2} = \sqrt{\frac{215}{10} - 1.1^2}$

$= \sqrt{21.5 - 1.21} = \sqrt{20.29}$

$r_{uv} = \frac{cov(u,v)}{\sigma_u \sigma_v} = \frac{20.29}{\sqrt{20.29}\sqrt{20.29}} = 1$

But $r_{xy} = r_{uv} = 1$

EXERCISE 5.1

1) Draw scatter diagram for the data given below and interpret it.

x	10	20	30	40	50	60	70
y	32	20	24	36	40	28	38

2) For the following data of marks of 7 students in Physics (x) and Mathematics (y), draw scatter diagram and state the type of correlation.

x	8	6	2	4	7	8	9
y	6	5	1	4	4	7	8

3) Draw scatter diagram for the data given below. Is there any correlation between Aptitude score and Grade points?

Aptitude score	40	50	55	60	70	80
Grade points	1.8	3.8	2.8	1.7	2.8	3.2

4) Find correlation coefficient between x and y series for the following data

$n = 15$, $\bar{x} = 25$, $\bar{y} = 18$, $\sigma_x = 3.01$, $\sigma_y = 3.03$,

$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 122$.

5) The correlation coefficient between two variables x and y is 0.48. The covariance is 36 and the variance of x is 16. Find the standard deviation of y.

6) In the following data one of the value of y is missing. Arithmetic means of x and y series are 6 and 8 respectively. $\left(\sqrt{2} = 1.4142\right)$

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

(i) Estimate missing observation

(ii) Calculate correlation coefficient.

7) Find correlation coefficient from the following data, [Given : $\sqrt{3} = 1.732$]

X	3	6	2	9	5
Y	4	5	8	6	7

8) Correlation coefficient between x and y is 0.3 and their covariance is 12. The variance of x is 9, find the standard deviation of y.

Let's Remember

1) Bivariate data is the observation recorded on two variables.

2) Correlation is the study of mutual or joint relationship between two variables.

3) There are 3 types of correlation,

i) Positive correlation ii) Negative correlation iii) No correlation.

4) Correlation coefficient between the variables x and y is given by $r_{xy} = \dfrac{cov(x, y)}{\sigma_x \sigma_y}$

5) $-1 \leq r \leq 1$.

6) Numerical value of correlation coefficient is invariant to the change of origin and scale.

7) $cov(x,y) = \dfrac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})$

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

MISCELLANEOUS EXERCISE

1) Two series of x and y with 50 items each have standard deviations 4.8 and 3.5 respectively. If the sum of products of deviations of x and y series from respective arithmetic means is 420, then find the correlation coefficient between x and y.

2) Find the number of pairs of observations from the following data,

$r = 0.15$, $\sigma_y = 4$, $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 12$, $\Sigma(x_i - \bar{x})^2 = 40$.

3) Given that r = 0.4 , $\sigma_y = 3$,
$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 108$, $\Sigma(x_i - \bar{x})^2 = 900$.
Find the number of pairs of observations.

4) Given the following information, $\Sigma x_i^2 = 90$, $\Sigma x_i y_i = 60$, r = 0.8, $\sigma_y = 2.5$, where x_i and y_i are the deviations from their respective means. Find the number of items.

5) A sample of 5 items is taken from the production of a firm. Length and weight of 5 items are given below,
[Given : $\sqrt{0.8823}$ = 0.9393]

Length(cm)	3	4	6	7	10
Weight(gm.)	9	11	14	15	16

Calculate correlation coefficient between length and weight and interpret the result.

6) Calculate correlation coefficient from the following data, and interpret it.

X	1	3	5	7	9	11	13
Y	12	10	8	6	4	2	0

7) Calculate correlation coefficient from the following data and interpret it.

x	9	7	6	8	9	6	7
y	19	17	16	18	19	16	17

8) If the correlation coefficient between x and y is 0.8,what is the correlation coefficient between i) 2x and y ii) $\frac{x}{2}$ and y iii) x and 3y iv) x-5 and y-3 v) x+7 and y+9 vi) $\frac{x-5}{7}$ and $\frac{y-3}{8}$?

9) In the calculation of the correlation coefficient between height and weight of a group of students of a college, one investigator took the measurements in inches and pounds while the other investigator took the measurements in cm. and kg. Will they get the same value of the correlation coefficient or different values? Justify your answer.

Activity 5.1

Calculate the correlation coefficient between weight and height in Activity 4.2

Activity 5.2

Calculate the correlation coefficient between age (in years) and blood pressure from Example 4 of Miscellaneous Exercise.

Activity 5.3

Using the given data plot the points & draw the scatter diagram. And identify the type of correlation.

X	8	12	16	20	24	28	32
Y	2	3	4	5	6	7	8

Activity 5.4

Select any 2 stocks and record the share prices for 10 days. Draw the scatter digram of them.

❖ ❖ ❖