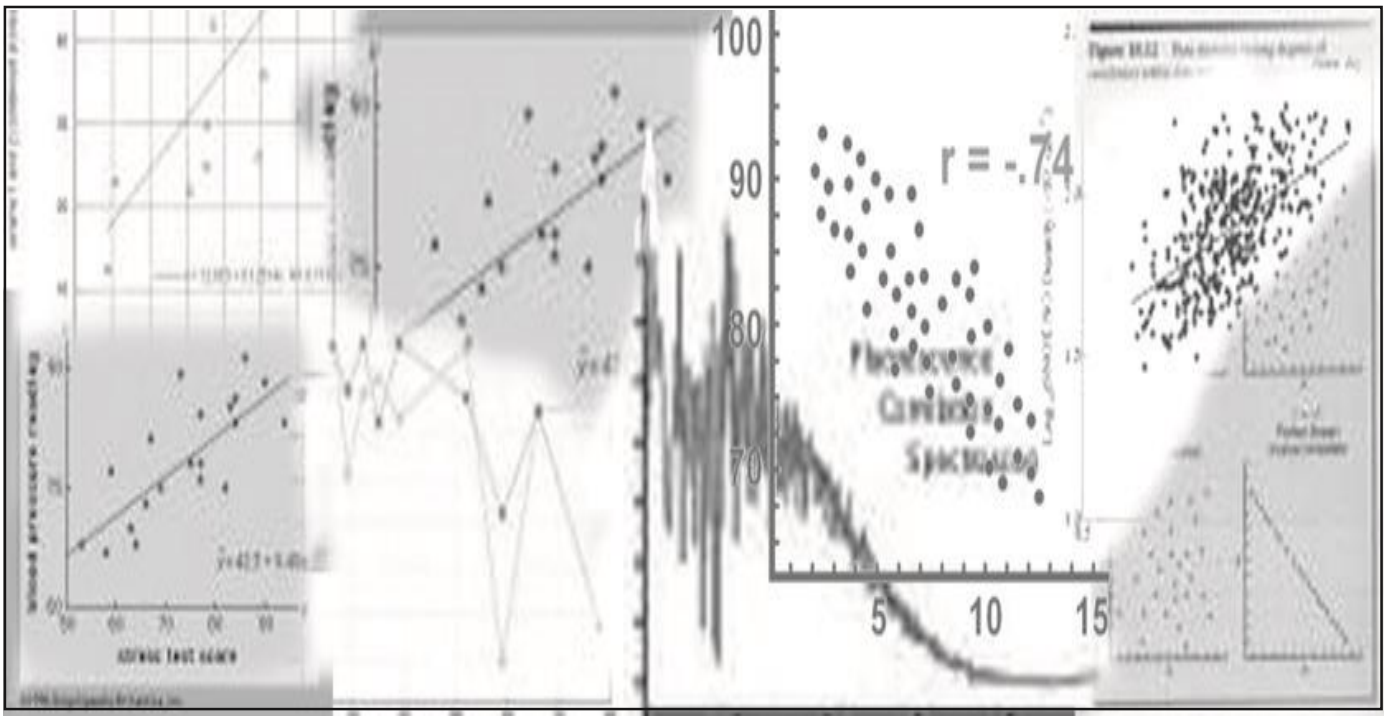


Correlation Analysis

The **Correlation** is a statistical tool used to measure the relationship between two or more variables, i.e. the degree to which the variables are associated with each other, such that the change in one is accompanied by the change in another.

In all sciences, natural, social, or biological as well as in business, we are largely concerned with the study of interrelationships among variables. For example, we may be intersected to know the relationship between the age and weight of the students, between family income and expenditure, or the relationship between the price of a commodity and the amount demand.

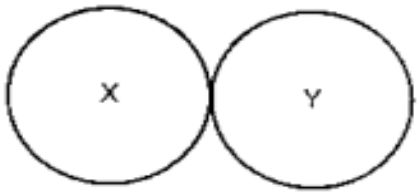
Correlation analysis enables us to measure or quantify the relationship between the pairs of variables.

Correlation can be measured by two ways-

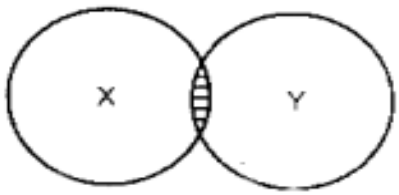
1. Graphically (Van Diagram, Scatter Diagram, & Pair panel correlation Plot)
2. Algebraically (Actual mean method, Assumed mean method, Direct method, & Variance covariance method)

Graphical Method-

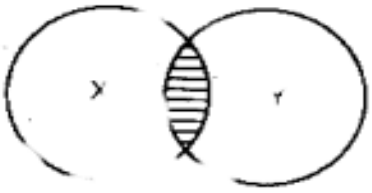
1) **Van Diagram**: **Venn diagrams** are used to provide an intuitive understanding of simple & multiple correlations & regression **analysis** and these diagrams work well with two variables.



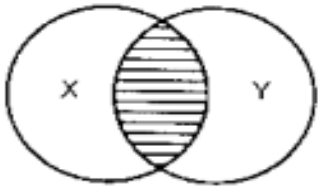
$r = .00$ $r^2 = .00$
SHARED VARIANCE = 0%



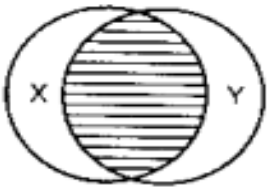
$r = \pm.20$ $r^2 = +.04$
SHARED VARIANCE = 4%



$r = \pm.40$ $r^2 = +.16$
SHARED VARIANCE = 16%



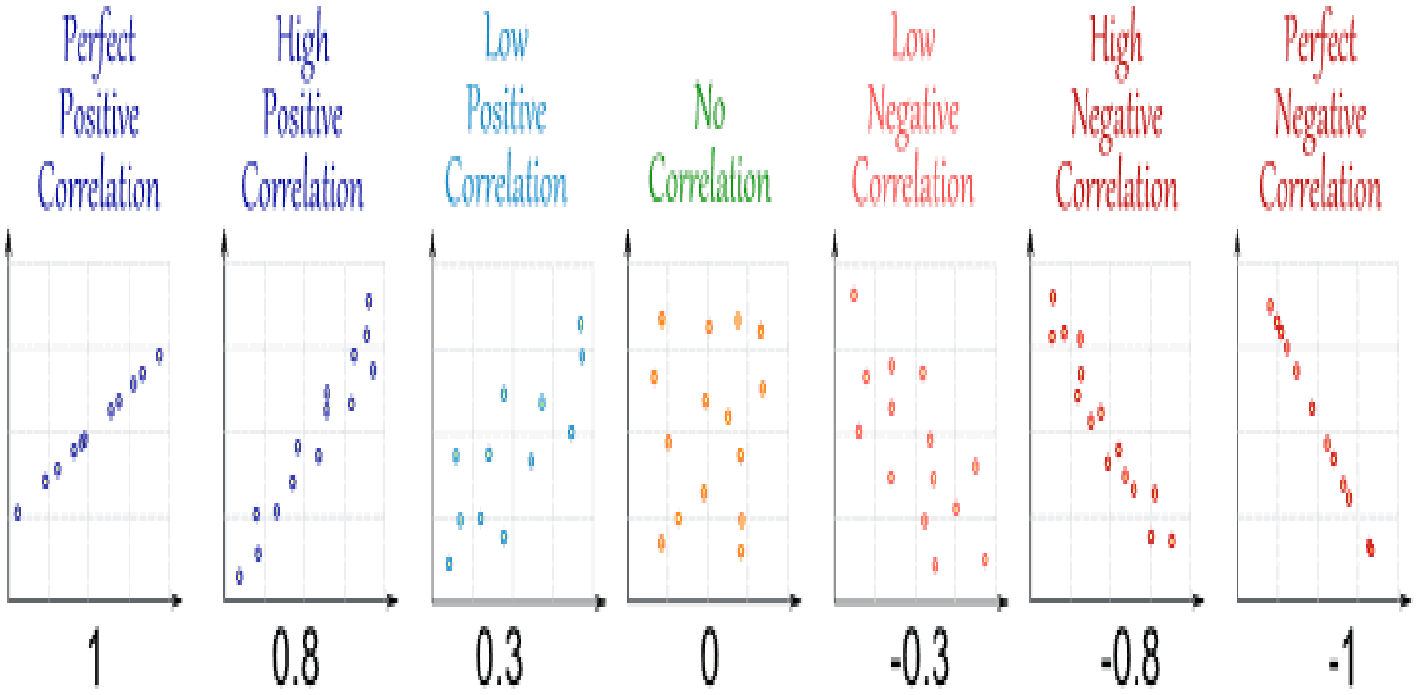
$r = \pm.60$ $r^2 = +.36$
SHARED VARIANCE = 36%



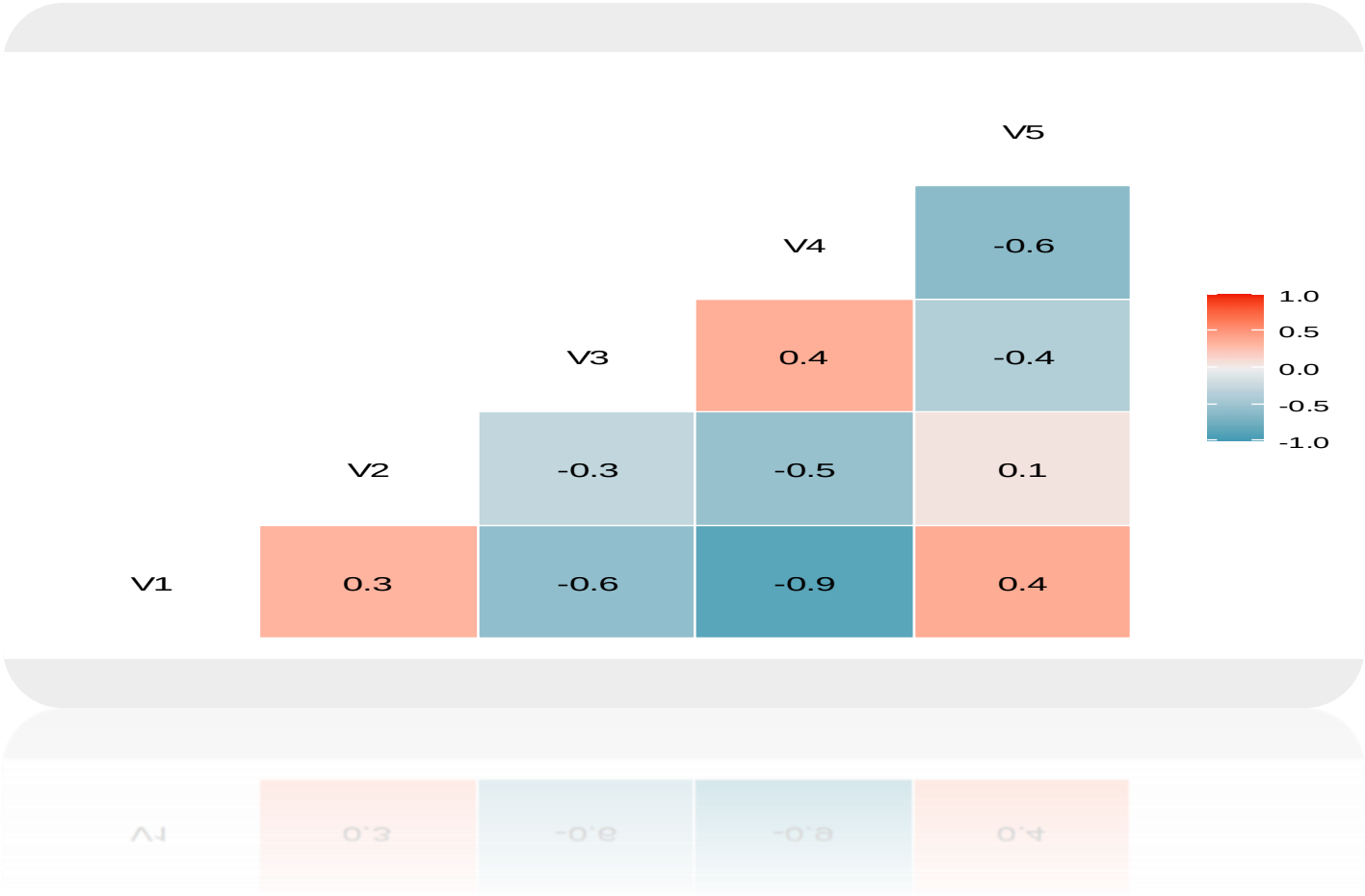
$r = \pm.80$ $r^2 = +.64$
SHARED VARIANCE = 64%

$r = \pm.100$ $r^2 = +1.00$
SHARED VARIANCE = 100%
(COMPLETE OVERLAP
OF X AND Y)

2) **Scatter Diagram**: When two numeric variables are plotted in form of dots on a graph paper is called a scatter diagram.



3) Pairs panel correlation plot



The relationship between the variables is measured by the Pearson's correlation coefficient is denoted by 'ρ' for population data and 'r' for sample data.

It was introduced by Galton in 1877 and developed later by Karl Pearson in 1890, is most widely used in practice. Thus r is also known as Karl Pearson or simply Pearson's correlation coefficient the relationship between two variables x and y, r is defined as

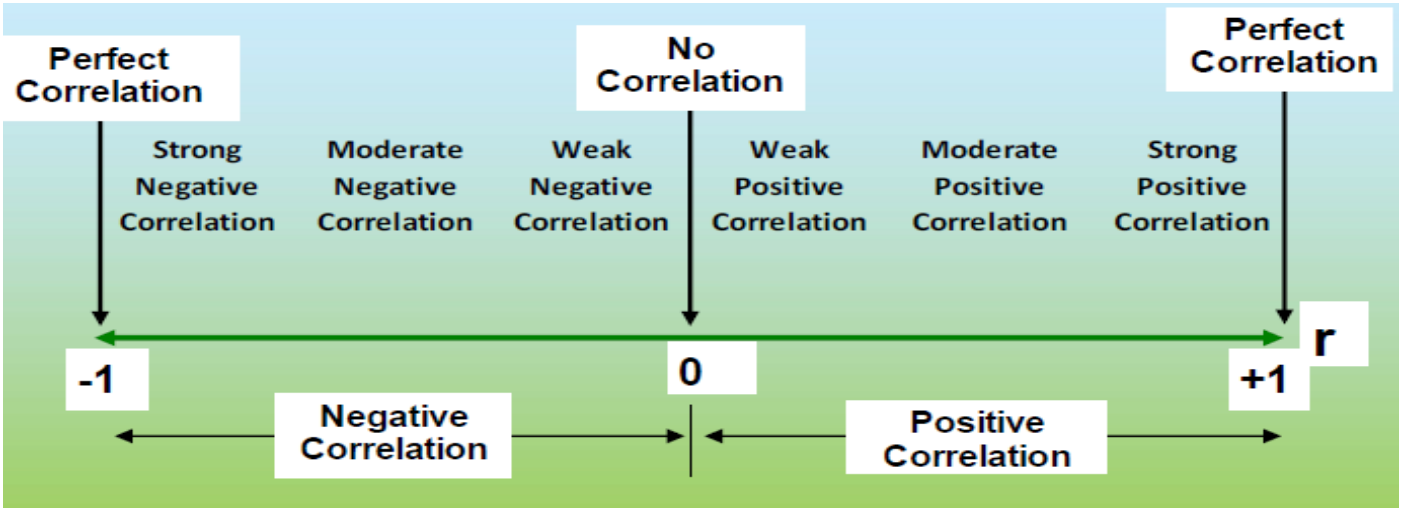
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

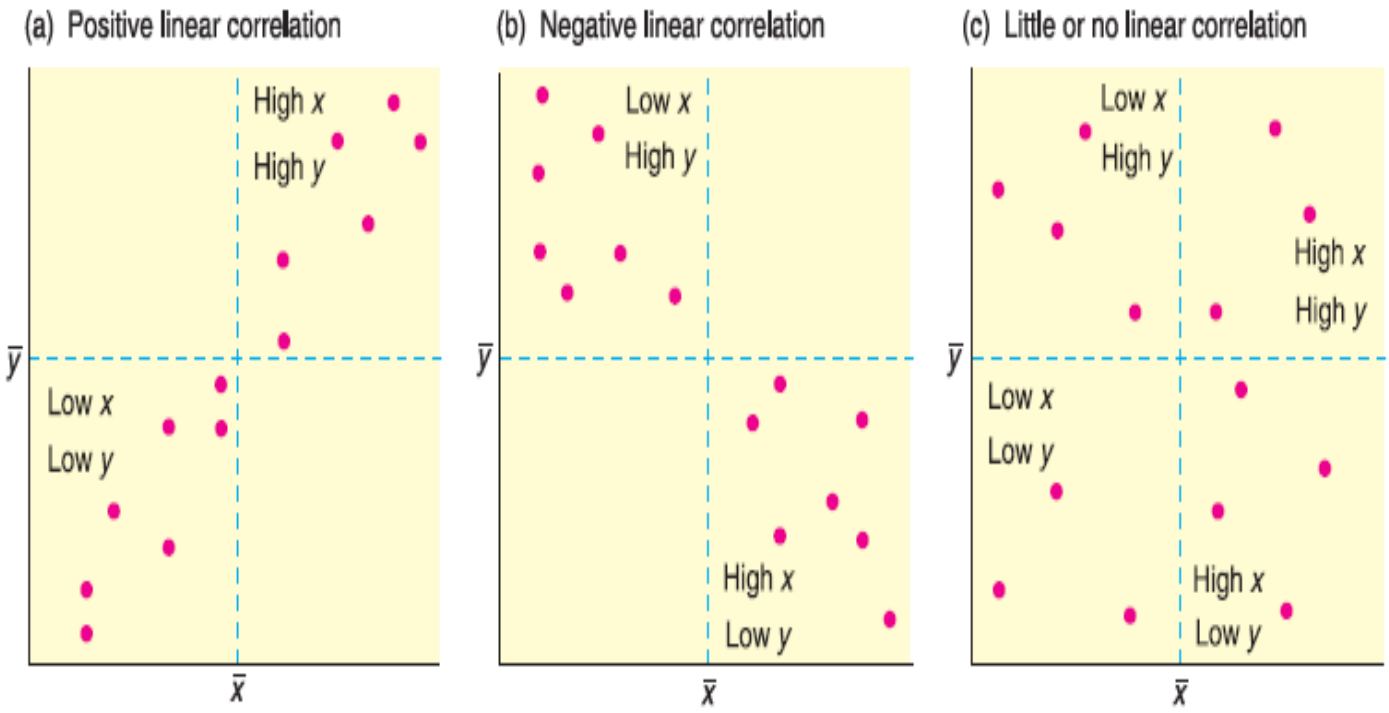
Equation (1) is also called Karl pearson's coefficient of correlation formula given by 1890.

Algebraically (1) reduces to

$$r = \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$$

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





Strength of correlation

Perfect	+1	−1
Strong	+0.9	−0.9
	+0.8	−0.8
	+0.7	−0.7
Moderate	+0.6	−0.6
	+0.5	−0.5
	+0.4	−0.4
Weak	+0.3	−0.3
	+0.2	−0.2
	+0.1	−0.1
Zero	0	

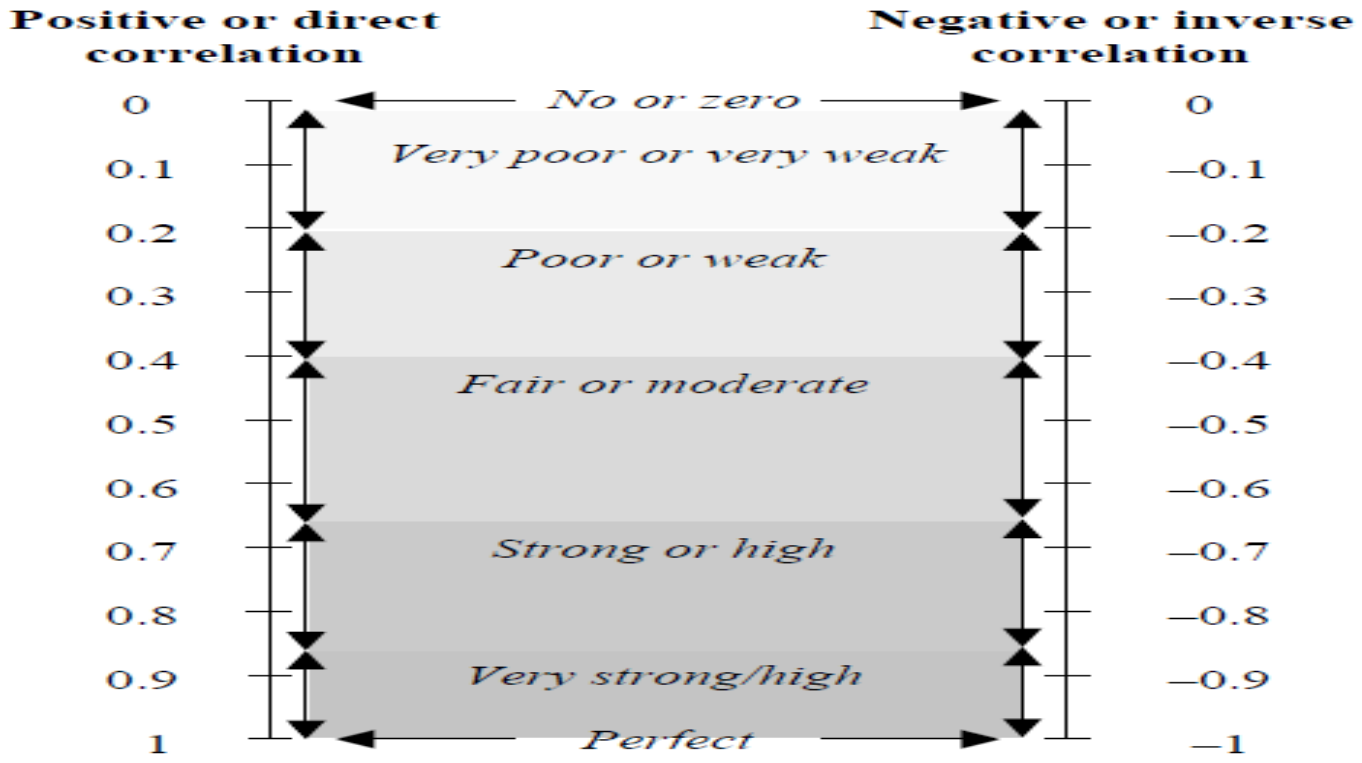


Fig. 2.1 Interpretation of correlation coefficient

Positive or negative

If the two variables deviate in the same direction, that is if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be director positive. But if they constantly deviate in the opposite directions, that is if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be inverse or negative. If the variables are independent, there cannot be any correlation and the variables are said to be zero correlation.

For example, the correlation between (1) the heights and weights of a group of persons, (2) the income and expenditure is positive and the correlation between (1) price and demand of a commodity, (2) the volume and pressure of a perfect gas is negative. And there is no correlation between income and height.

Simple correlation and Multiple Correlation

Correlation only between two variables is called simple correlation. For example, correlation between income and expenditure.

In multiple correlation there involve more than two variables.

Linear correlation and Non Linear correlation

Correlation is said to be linear when the amount of change in one variable tends to bear a constant ratio to the amount of change in the other. The graph of the variables having a linear relationship will form a straight line.

Example: X = 1, 2, 3, 4, 5, 6, 7, 8,

Y = 5, 7, 9, 11, 13, 15, 17, 19,

Y = 3 +2x

The correlation would be non linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

Karl Pearson's Coefficient of Correlation:

Karl Pearson's method of calculating coefficient of correlation is based on the covariance of the two variables in a series. This method is widely used in practice and the coefficient of correlation is denoted by the symbol "***r***". If the two variables under study are X and Y, the following formula suggested by Karl Pearson can be used for measuring the degree of relationship of correlation.

$$r = \frac{Covariance\ (x,y)}{S.D.(x)S.D.(y)}$$

$$r = \frac{Cov(X, Y)}{\sigma_x\ \sigma_y}$$

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \qquad \text{where } X = x - \bar{x},\ Y = y - \bar{y}$$

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2]\,[n(\sum y^2) - (\sum y)^2]}}$$

$$r = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum (X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

$$r = \frac{\sum f(dx)(dy) - \frac{\sum fdx)(\sum fdy)}{N}}{\sqrt{\sum (fdx) \wedge 2 - \frac{(\sum fdx) \wedge 2}{N}}\sqrt{\sum (dy) \wedge 2 - \frac{(\sum fdy) \wedge 2}{N}}}$$

$$d_x = X - A$$

$$d_y = Y - A$$

Application Problem-1: A research physician recorded the pulse rates and the temperatures of water submerging the faces of ten small children in cold water to control the abnormally rapid heartbeats. The results are presented in the following table. Calculate the correlation coefficient between temperature of water and reduction in pulse rate.

Temperature of water	68	65	70	62	60	55	58	65	69	63
Reduction in pulse rate.	2	5	1	10	9	13	10	3	4	6

Solve the above problem:

By using actual mean

$$r = \frac{\Sigma\left(X-\bar{X}\right)\times\left(Y-\bar{Y}\right)}{\sqrt{\Sigma\left(X-\bar{X}\right)^2}\times\sqrt{\Sigma\left(Y-\bar{Y}\right)^2}} \qquad ...(2)$$

By assumed mean method

$$r = \frac{\Sigma dxdy - \frac{\Sigma dx\cdot\Sigma dy}{N}}{\sqrt{\Sigma dx^2 - \frac{\left(\Sigma dx\right)^2}{N}}\times\sqrt{\Sigma dy^2 - \frac{\left(\Sigma dy\right)^2}{N}}} \qquad ...(3)$$

By direct method

$$r = \frac{N\Sigma XY - [\Sigma X][\Sigma Y]}{\sqrt{N\Sigma X^2 - \left(\Sigma X\right)^2}\times\sqrt{N\Sigma Y^2 - \left(\Sigma Y\right)^2}} \qquad ...(4)$$

Now covariance of X and Y is defined as

$$\mathrm{cov}\left(X,Y\right) = \frac{\Sigma\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{N}$$

$$\therefore \qquad r = \frac{\mathrm{cov}\left(X,Y\right)}{\sigma_X\sigma_Y}$$

Where N is the number of pairs of data.

$$d_x = X - A_X$$
$$d_y = Y - A_Y$$

Solution: Calculating table of correlation coefficient.

x_i	y_i	x_i^2	y_i^2	x_iy_i
68	2	4624	4	136
65	5	4225	25	325
70	1	4900	1	70
62	10	3844	100	620
60	9	3600	81	540
55	13	3025	169	715
58	10	3364	100	580
65	3	4225	9	195
69	4	4761	16	276
63	6	3969	36	378
$\sum x_i = 635$	$\sum y_i = 63$	$\sum x_i^2 = 40537$	$\sum y_i^2 = 541$	$\sum x_i y_i = 3835$

We know, $r_{xy} = \dfrac{\sum x_i y_i - \dfrac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^2 - \dfrac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$

$$= \frac{3835 - \dfrac{635 \times 63}{10}}{\sqrt{\left\{40537 - \dfrac{(635)^2}{10}\right\}\left\{541 - \dfrac{(63)^2}{10}\right\}}}$$

$= -0.94$

The result -0.94, indicates that the correlation coefficient between temperature of water and reduction in pulse rate is highly negatively correlated.

i.(x, y): (1,2) , (2, 3), (3, 5), (4, 4), (5, 7)

ii. (x, y): (1,1) , (2, 3), (3, 5), (4, 7), (5, 9)

iii.(x, y): (1,10) , (2, 8), (3, 6), (4, 4), (5, 2)

iv.(x, y): (2,9) , (3, 5), (4, 6), (5, 2), (6, 1)

v.(x, y): (-2,4) , (-1, 1), (0, 0), (1, 1), (2, 4)

Solution 1: (x, y): (1,2) , (2, 3), (3, 5), (4, 4), (5, 7)

The formula for finding correlation coefficient is

$$r_{xy} = \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^{\,2} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$$

Let us make a table to calculate correlation coefficient.

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	1	4	2
2	3	4	9	6
3	5	9	25	15
4	4	16	16	16
5	7	25	49	35

$\sum x_i = 15$	$\sum y_i = 21$	$\sum x_i^2 = 55$	$\sum y_i^2 = 103$	$\sum x_i y_i = 74$

$$r_{xy} = \frac{\sum x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left\{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right\}\left\{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right\}}}$$

$$= \frac{74 - \frac{15 \times 21}{5}}{\sqrt{\left\{55 - \frac{(15)^2}{5}\right\}\left\{103 - \frac{(21)^2}{5}\right\}}}$$

$$= 0.90$$

Comment: There exists a strong positive relationship between x and y.

Problem: above ii-v (Assignment)

Assignment Problem-2: The following table gives the ages and blood pressure of 10 women:

Age in years x	56	42	36	47	49	42	72	63	55	60
Blood pressure y	147	125	118	128	125	140	155	160	149	150

Draw a scatter diagram

Find correlation coefficient between x and y and comment.

Ans: Try your-self

Assignment Problem-3: The scores of 12 students in their mathematics and physics classes are:

Mathematics	2	3	4	4	5	6	6	7	7	8	10	10
Physics	1	3	2	4	4	4	6	4	6	7	9	10

Find the correlation coefficient distribution and interpret it.

Comment on the followings:

(i) r=0 (ii) r=-1 (iii) r=1 (iv) r ≥ 1 (v) r<1

(i) r=0, indicates that the correlation coefficient between x and y is zero.

(ii) r=-1, indicates that the correlation coefficient between x and y is perfect negative.

(iii) r= 1, indicates that the correlation coefficient between x and y is perfect positive.

(iv) r ≥ 1 i.e, r=1 and r>1 i.e, r>1, is not possible, because the Correlation coefficient lies between -1 to +1.

(v) r<1, not possible because, the Correlation coefficient lies between -1 to +1.

Uses of correlation coefficient.

1. To find the relationship between two variables.
2. To find the relationship between dependent variable and combined influence of a group of independent variables.
3. To solve many problem in biology.
4. In social studies like relationships between crime and educations, correlation analysis has got definite role to play.
5. In economies this is used specially.

Assignment problem-4:

The following figures relate to advertisement expenditure and profit:

Profit (Tk.Crore):x	25	28	27	33	31	10	16	16	18	23
Adv. Exp.(Tk. Lakh):y	87	91	92	95	93	52	68	72	78	86

(i)Draw a scatter diagram and comment

(ii) Find Karl Pearson's correlation coefficient

Assignment Problem-5:

The following figures relate to advertisement expenditure and sales of a company:

Adv. Exp. (Tk. Lac)	62	67	73	78	85	78	91	92	96	98
Sales (Tk.Crore)	11	13	17	18	21	24	21	27	26	21

Find Karl Pearson's correlation coefficient