# *CORRELATION ANALYSIS*

# Correlation Analysis

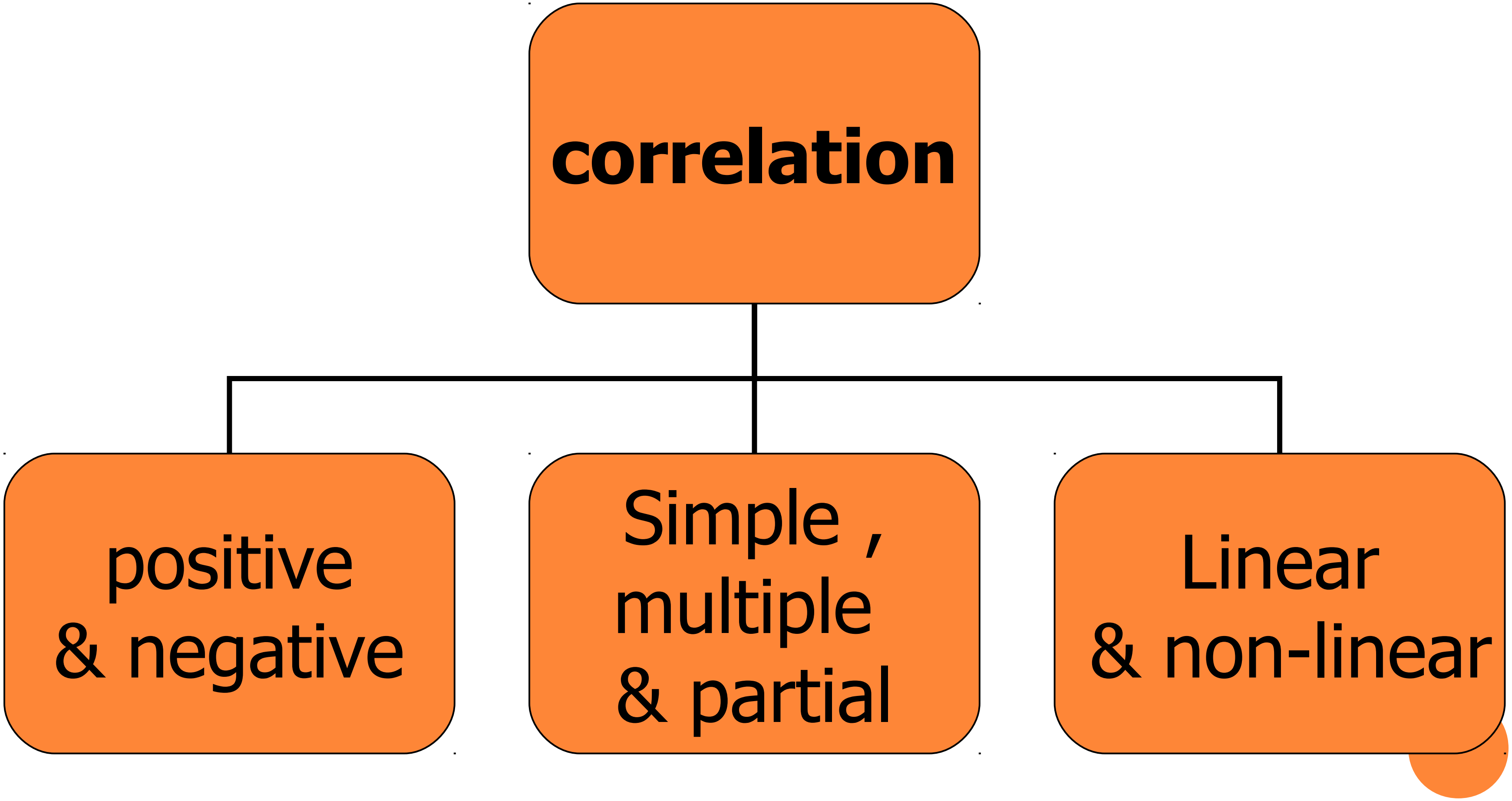
- The *sample correlation coefficient* ($r$) measures the degree of linearity in the relationship between $X$ and $Y$.

$$-1 \leq r \leq +1$$

Strong negative relationship — Strong positive relationship

- $r = 0$ indicates no linear relationship
- In Excel, use =CORREL(array1,array2), where array1 is the range for $X$ and array2 is the range for $Y$.

# TYPES OF CORRELATION

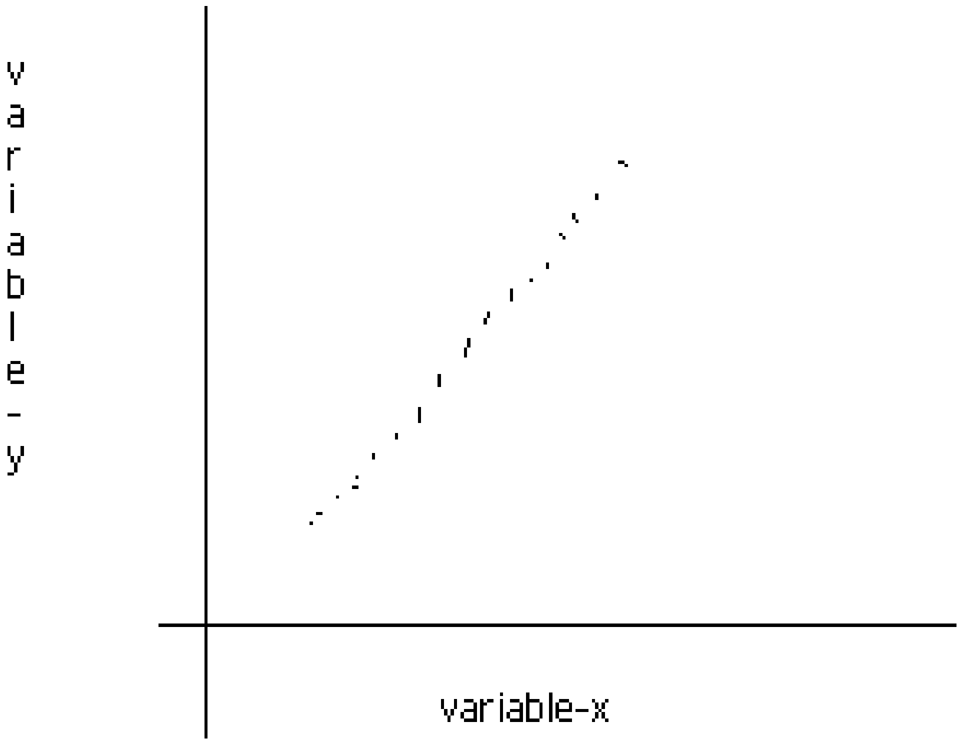


# METHODS OF CORRELATION

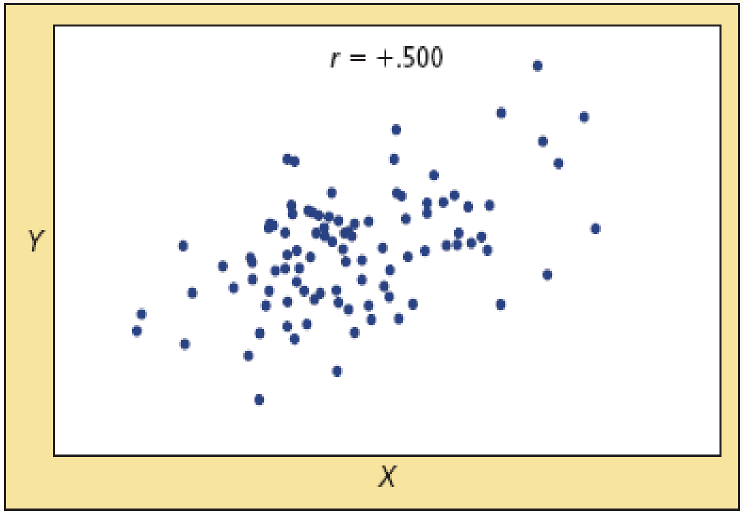
- Scatter diagram
- Product moment or covariance
- Rank correlation
- Concurrent deviation

# SCATTER DIAGRAM

- Perfectly +ve

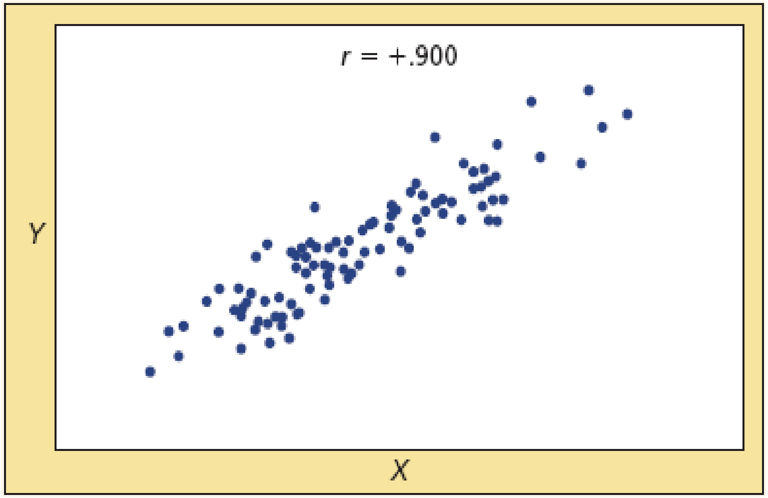


# LESS-DEGREE +VE



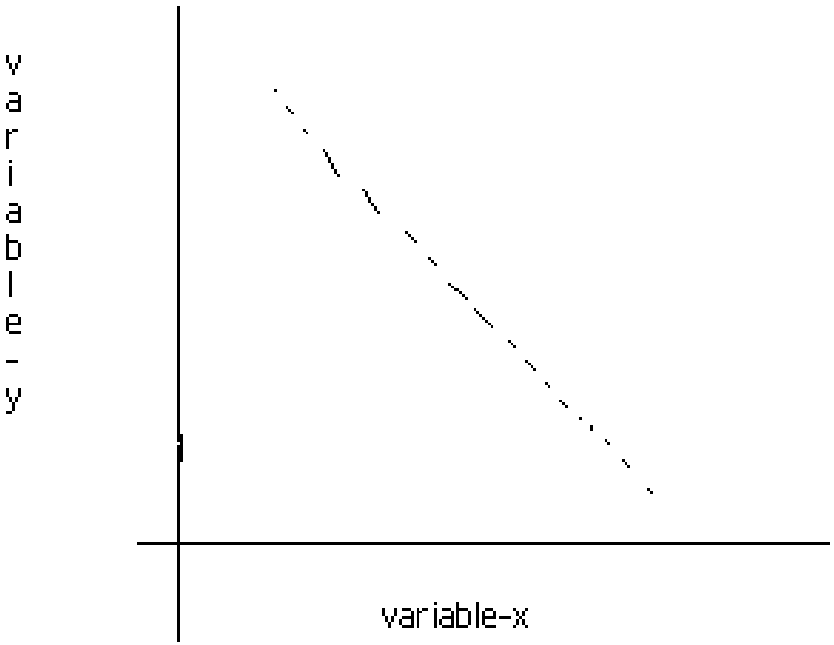

Weak Positive Correlation

# HIGH DEGREE +VE



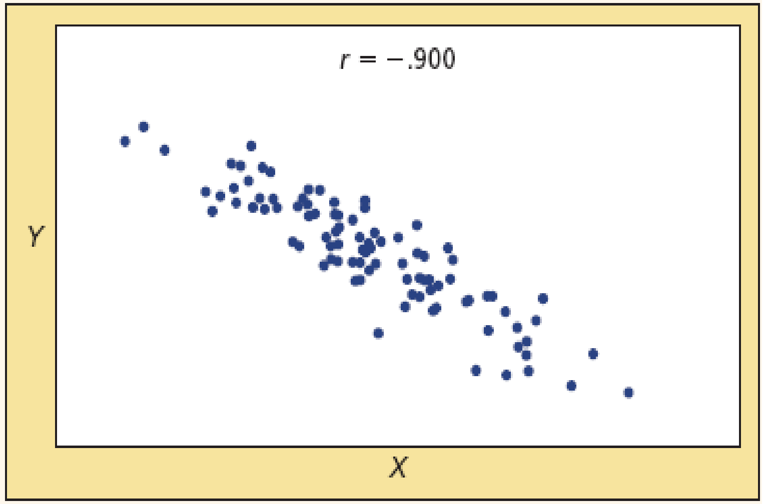

Strong Positive Correlation

# PERFECTLY -VE

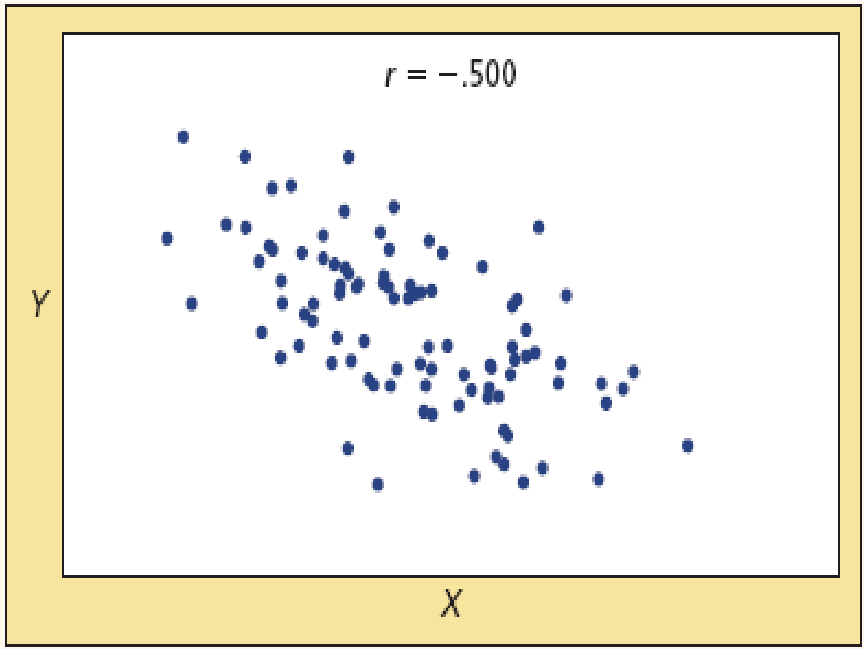


# HIGH DEGREE -VE

Strong Negative Correlation

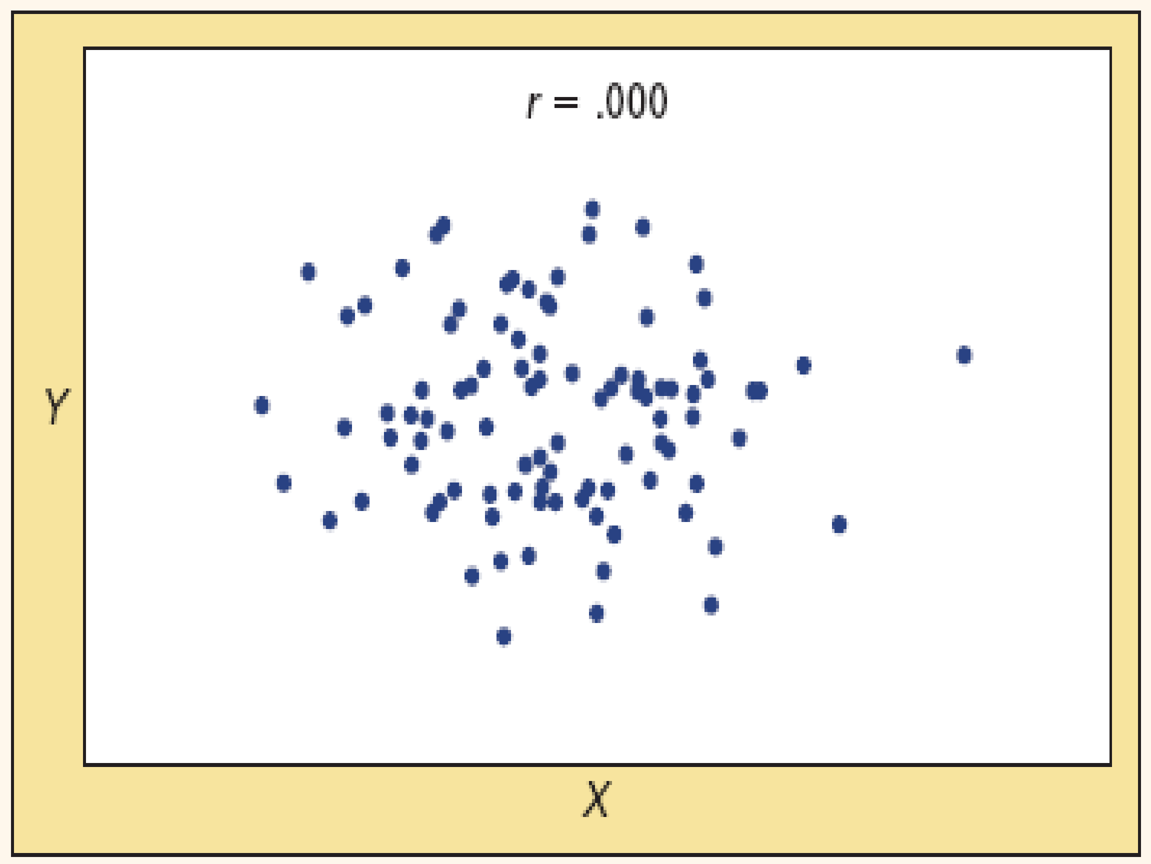


# LESS DEGREE -VE




Weak Negative Correlation

# ZERO DEGREE



# KARL PEARSON CORRELATION COEFFICIENT

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

$$r = \frac{\sum x.y}{\sqrt{\sum x^2 . \sum y^2}}$$

# WHERE

$$x = X - \overline{X}$$

$$and$$

$$y = Y - \overline{Y}$$

# PROBLEM

From the following data find the coefficient of correlation by Karl Pearson method

X:6 2 10 4 8

Y:9 11 5 8 7

# SOL.

| X | Y | X-6 | Y-8 | $x^2$ | $y^2$ | $x.y$ |
|---|---|---|---|---|---|---|
| 6 | 9 | 0 | 1 | 0 | 1 | 0 |
| 2 | 11 | -4 | 3 | 16 | 9 | -12 |
| 10 | 5 | 4 | -3 | 16 | 9 | -12 |
| 8 | 8 | -2 | 0 | 4 | 0 | 0 |
| 4 | 7 | 2 | -1 | 4 | 1 | -2 |
| 30 | 40 | 0 | 0 | 40 | 20 | -26 |

# SOL.CONT.

$$\bar{X} = \frac{\sum X}{N} = \frac{30}{5} = 6$$

$$\bar{Y} = \frac{\sum Y}{N} = \frac{40}{5} = 8$$

$$r = \frac{\sum x.y}{\sqrt{\sum x^2 . \sum y^2}} = \frac{-26}{\sqrt{40.20}} = \frac{-26}{\sqrt{800}} \approx -0.92$$

# DIRECT METHOD

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

# SHORT-CUT METHOD

$$r = \frac{N\sum d_x . d_y - \sum d_x . \sum d_y}{\sqrt{N\sum d^2{}_x - \left(\sum d_x\right)^2} . \sqrt{N\sum d^2{}_y - \left(\sum d_y\right)^2}}$$

# WHERE

$$d_x = X - A$$

&

$$d_y = Y - A$$

$$A = assume\ mean$$

# PRODUCT MOMENT METHOD

$$r = \sqrt{b_{xy} . b_{yx}}$$

*where*

$$b_{xy} = \frac{\sum xy}{\sum y^2}$$

$$b_{yx} = \frac{\sum xy}{\sum x^2}$$

# SPEARMAN'S RANK CORRELATION(WHEN RANKS ARE NOT EQUAL)

$$R = 1 - \frac{6\sum D^2}{N(N^2 - 1)}$$

$$where$$

$$D = R_x - R_y$$

$$R_x = rank.of.X$$

$$R_y = rank.of.y$$

# PROBLEM

Calculate spearman's rank correlation coefficient between advt.cost & sales from the following data

Advt.cost :39 65 62 90 82 75 25 98 36 78

Sales(lakhs): 47 53 58 86 62 68 60 91 51 84

## SOL.

| X | Y | R-x | R-y | D | $D^2$ |
|---|---|---|---|---|---|
| 39 | 47 | 8 | 10 | -2 | 4 |
| 65 | 53 | 6 | 8 | -2 | 4 |
| 62 | 58 | 7 | 7 | 0 | 0 |
| 90 | 86 | 2 | 2 | 0 | 0 |
| 82 | 62 | 3 | 5 | -2 | 4 |
| 75 | 68 | 5 | 4 | 1 | 1 |
| 25 | 60 | 10 | 6 | 4 | 16 |
| 98 | 91 | 1 | 1 | 0 | 0 |
| 36 | 51 | 9 | 9 | 0 | 0 |
| 78 | 84 | 4 | 3 | 1 | 1 |
| | | | | | 30 |

# SOL.CONT.

$$R = 1 - \frac{6\sum D^2}{N^3 - N}$$

$$\Rightarrow R = 1 - \frac{6.30}{10^3 - 10}$$

$$\Rightarrow R = 1 - \frac{2}{11}$$

$$\Rightarrow R = \frac{9}{11} = 0.82$$

# IN CASE OF EQUAL RANK

$$R = 1 - \frac{6\left\{\sum D^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right) + \ldots\ldots\ldots\right\}}{N\left(N^2 - 1\right)}$$

*where*

$m = no.of\ repeated\ items$

# PROBLEM

*A psychologist wanted to compare two methods A & B of teaching. He selected a random sample of 22 students. He grouped them into 11 pairs so that the students in a pair have approximately 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 as follows*

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

SOL.

| A | B | R-A | R-B | D | $D^2$ |
|---|---|---|---|---|---|
| 24 | 37 | 6 | 1 | 5 | 25 |
| 29 | 35 | 3 | 2 | 1 | 1 |
| 19 | 16 | 8.5 | 9.5 | -1 | 1 |
| 14 | 26 | 10 | 4 | 6 | 36 |
| 30 | 23 | 1.5 | 5 | -3.5 | 12.25 |
| 19 | 27 | 8.5 | 3 | 5.5 | 30.25 |
| 27 | 19 | 5 | 8 | -3 | 9 |
| 30 | 20 | 1.5 | 7 | -5.5 | 30.25 |
| 20 | 16 | 7 | 9.5 | -2.5 | 6.25 |
| 28 | 11 | 4 | 11 | -7 | 49 |
| 11 | 21 | 11 | 6 | 5 | 25 |
| | | | | | 225 |

# SOL.CONT.

in A series the items 19 & 30 are repeated twice and in B series 16 is repeated twice $\therefore$

$$R = 1 - \frac{6\left[\sum D^2 + \frac{2(4-1)}{12} + \frac{2(4-1)}{12} + \frac{2(4-1)}{12}\right]}{11(121-1)}$$

$$\Rightarrow R = -0.0225$$

# PROPERTIES OF CORRELATION COEFFICIENT

- r always  lies between +1 & -1

i.e.   $-1<r<+1$

- Two independent variables are uncorrelated but converse is not true
- r is independent of change in origin and scale
- r is the G.M. of two regression coefficients
- r is symmetrical

# PROBABLE ERROR(PE)

$$\text{Standard Error } SE(r) = \frac{1 - r^2}{\sqrt{n}}$$

$$PE(r) = 0.6745 \times SE(r)$$

$$or$$

$$PE(r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}}$$

# PARTIAL CORRELATION COEFFICIENT

$$r_{12.3} = \frac{r_{12} - r_{13} \times r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$r_{13.2} = \frac{r_{13} - r_{12} \times r_{23}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{23}^2}}$$

$$r_{23.1} = \frac{r_{23} - r_{12} \times r_{13}}{\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13}^2}}$$

# MULTIPLE CORRELATION COEFFICIENT

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

$$r_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}}$$

$$r_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}$$