Correlation

If the change in one variable affects a change in the value of the other variable, the variables are said to be correlated.

- **Coefficient of correlation**

  It is a numerical measure of linear relationship of two variables x and y, it is denoted by $r(x, y)$.

That is, $r(x,y) = \dfrac{\dfrac{\sum xy}{n} - \bar{x}.\bar{y}}{\sqrt{\dfrac{\sum x^2}{n} - (\bar{x})^2}.\sqrt{\dfrac{\sum y^2}{n} - (\bar{y})^2}}$ where $\bar{x} = \dfrac{\sum x}{n}$, $\bar{y} = \dfrac{\sum y}{n}$.

Problem

1. Calculate the correlation coefficient for the following data

x	1	3	5	8	9
y	3	4	8	10	12

x	y	xy	x^2	y^2
1	3	3	1	9
3	4	12	9	16
5	8	40	25	64
8	10	80	64	100
9	12	108	81	144
10	11	110	100	121
36	**48**	**353**	**280**	**454**

Now, $\bar{x} = \dfrac{\sum x}{n} = \dfrac{36}{6} = 6$, $\bar{y} = \dfrac{\sum y}{n} = \dfrac{48}{6} = 8$

$$r(x,y) = \frac{\dfrac{\sum xy}{n} - \bar{x}.\bar{y}}{\sqrt{\dfrac{\sum x^2}{n} - (\bar{x})^2}.\sqrt{\dfrac{\sum y^2}{n} - (\bar{y})^2}} = \frac{\dfrac{353}{6} - 48}{\sqrt{\dfrac{280}{6} - 36}.\sqrt{\dfrac{454}{6} - 64}} = 0.97$$

Practice Problem

1. Find the correlation coefficient for the following data:

x	65	66	67	67	68	69	70	72
y	67	68	65	68	72	72	69	71

Solution: $r(x,y) = 0.603$

- **Rank Correlation Coefficient:**

Rank correlation coefficient, is given by

$$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where $d_i = (x_i - y_i)$, that is difference between the ranks.

Problem

1. The rankings of ten students in two subjects A and B are as follows:

A	3	5	8	4	7	10	2	1	6	9
B	6	4	9	8	1	2	3	10	5	7

Find the rank correlation coefficient.

Solution:

$A(x_i)$	$B(y_i)$	$d_i = (x_i - y_i)$	d_i^2
3	6	-3	9
5	4	1	1
8	9	-1	1
4	8	-4	16
7	1	6	36
10	2	8	64
2	3	-1	1
1	10	-9	81
6	5	1	1
9	7	2	4
		0	$\Sigma d_i^2 = 214$

The rank correlation coefficient is $r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6(214)}{10(10^2-1)} = -0.297$ =

Note: When we have repeated ranks, we have to use correction factor $= \frac{m(m^2-1)}{12}$

Where *m* represents the number of times the data repeated.

2. The following table gives the number of units rejected by two operators *X* and *Y* in 8 inspections:

X	15	20	28	12	40	60	20	80
Y	40	30	50	30	20	10	30	60

Obtain rank correlation coefficient between *X* and *Y* with respect to the quality of the product.

Solution:

X	Y	Ranks in X (x_i)	Ranks in Y (y_i)	$d_i = (x_i - y_i)$	d_i^2
15	40	2	6	-4	16
20	30	3.5	4	-0.5	0.25
28	50	5	7	-2	4
12	30	1	4	-3	9
40	20	6	2	4	16
60	10	7	1	6	36
20	30	3.5	4	-0.5	0.25
80	60	8	8	0	0
					$\sum d_i^2 = 81.5$

In X series 20 repeated twice, correction factor $= \frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = \frac{1}{2}$

In Y series 30 repeated thrice, correction factor $= \frac{m(m^2-1)}{12} = \frac{3(3^2-1)}{12} = 2$

Therefore, $r(X,Y) = 1 - \frac{6\left[81.5 + \frac{1}{2} + 2\right]}{8(8^2-1)} = 0$

Practice Problem:

1. The marks secured by recruits in the selection test X and in the proficiency test Y are given below:

Serial No.	1	2	3	4	5	6	7	8	9
X	10	15	12	17	13	16	24	14	22
Y	30	42	45	46	33	34	40	35	39

Calculate the rank correlation coefficient.
Solution:0.4

2. Find the rank correlation coefficient for the following.

X	92	89	87	86	86	77	71	63	53	50
Y	86	83	91	77	68	85	52	82	37	57

Solution: 0.727

3. Following are the marks obtained by 10 students in a class in two tests.

Students	A	B	C	D	E	F	G	H	I	J
Test 1	70	68	67	55	60	60	75	63	60	72
Test 2	65	65	80	60	68	58	75	63	60	70

Find the rank correlation for tied observations.
Solution: r = 0.68.

- **Multiple and Partial Correlation**

Partial correlation coefficient provides a measure of the relationship between the dependent variable and other variable, with the effect of the rest of the variables eliminated.

If there are three variables X_1, X_2 *and* X_3, there will be three coefficients of partial correlation, each studying the relationship between two variables when the third is held constant. If we denote by $r_{12.3,}$ that is, the coefficient of partial correlation X_1 *and* X_2 keeping X_3 constant, it is calculated as

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}, \quad r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}}, \; r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}}$$

Problem:

1. Given $r_{12} = 0.7$, $r_{13} = 0.61$ and $r_{23} = 0.4$. Find the partial correlation coefficients.

Solution: $r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}} = \frac{0.7 - (0.61 \times 0.4)}{\sqrt{1-(0.61)^2}\sqrt{1-(0.4)^2}} = 0.628$

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{23}^2}} = \frac{0.61 - (0.7 \times 0.4)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.4)^2}} = 0.504$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{13}^2}} = \frac{0.4 - (0.7 \times 0.61)}{\sqrt{1-(0.7)^2}\sqrt{1-(0.61)^2}} = -0.048$$

- **Multiple correlation**

In multiple correlation, we are trying to make estimates of the value of one of the variable based on the values of all the others. The variable whose value we are trying to estimate is called the dependent variable and the other variables on which our estimates are based are known as independent variables.

The coefficient of multiple correlation with three variables X_1, X_2 *and* X_3 are $R_{1.23,}$ $R_{2.13}$ and $R_{3.21}$. $R_{1.23,}$ is the coefficient of multiple correlation related to X_1 as a dependent variable and X_2, X_3 as two independent variables and it can be expressed interms of $r_{12,}$ r_{23} and r_{13} as $R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1-r_{23}^2}}$,

$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1-r_{13}^2}}$, $R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1-r_{12}^2}}$

Problem:

1. The following zero-order correlation coefficients are given: $r_{12} = 0.98$, $r_{13} = 0.44$ and $r_{23} = 0.54$. Find $R_{1.23}$.

Solution: $R_{1.23} = \sqrt{\dfrac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{23}^2}}$

$$= \sqrt{\frac{(0.98)^2 + (0.44)^2 - 2(0.98)(0.54)(0.44)}{1 - (0.54)^2}} = 0.986$$

Regression:

Regression is a mathematical measure of the average relationship between two or more variables in terms of the original limits of the data.

- **Lines of regression:**

1. The line of regression of Y on X is given by $y - \bar{y} = r.\dfrac{\sigma_Y}{\sigma_X}(x - \bar{x})$
2. The line of regression of X on Y is given by $x - \bar{x} = r.\dfrac{\sigma_X}{\sigma_Y}(y - \bar{y})$

- **Regression Coefficients:**

1. Regression coefficient of Y on X : $r.\dfrac{\sigma_Y}{\sigma_X} = b_{YX}$

   Where $b_{YX} = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$

2. Regression coefficient of X on Y : $r.\dfrac{\sigma_X}{\sigma_Y} = b_{XY}$

   Where $b_{XY} = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{\sum(y - \bar{y})^2}$

3. Correlation coefficient $r = \pm\sqrt{b_{XY} \times b_{YX}}$

Problem:

1. From the following data find (i) two regression equations (ii) the coefficient of correlation (iii) Find Y when $X = 30$

X	25	28	35	32	31	36	29	38	34	32
Y	43	46	49	41	36	32	31	30	33	39

Solution:

X	Y	$X-\bar{X}$ $= X-32$	$Y-\bar{Y}$ $= Y-38$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})$ $(Y-\bar{Y})$
25	43	-7	5	49	25	-35
28	46	-4	8	16	64	-32
35	49	3	11	9	121	33
32	41	0	3	0	9	0
31	36	-1	-2	1	4	2
36	32	4	-6	16	36	-24
29	31	-3	-7	9	49	21
38	30	6	-8	36	64	-48
34	33	2	-5	4	25	-10
32	39	0	1	0	1	0
320	**380**	**0**	**0**	**140**	**398**	**-93**

Here, $\bar{X} = \frac{\sum X}{n} = \frac{320}{10} = 32$, $\bar{Y} = \frac{\sum Y}{n} = \frac{380}{10} = 38$

The line of regression of X on Y is given by $x - \bar{x} = b_{XY}(y - \bar{y})$

$$b_{XY} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2} = \frac{-93}{398} = -0.2337$$

$$\Rightarrow (x-32) = -0.2337 \quad (y-38)$$

$$= -0.2337y + 0.2337 \times 38$$

$$\Rightarrow x = -0.2337y + 40.8806$$

The line of regression of Y on X is given by $y - \bar{y} = b_{YX}(x - \bar{x})$

$$b_{YX} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{-93}{140} = -0.6643$$

$$\Rightarrow (y-38) = -0.6643 \quad (x-32)$$

$$= -0.6643x + 0.6643 \times 32$$

$$\Rightarrow y = -0.6643x + 59.2576$$

Coefficient of correlation $r^2 = b_{YX} \times b_{XY}$

$$= (-0.6643)\,(-0.2337) = 0.1552$$

$$r = \pm\sqrt{0.1552} = \pm 0.394$$

2. The two lines of regression are $8x - 10y + 66 = 0$, $40x - 18y - 214 = 0$. The variance of X is 9. Find the mean values of X and Y.

Solution:

Since both the lines of regression passes through the mean values $\bar{x}$ and $\bar{y}$, the point $(\bar{x}, \bar{y})$ must satisfy the two given regression lines .

$8\bar{x} - 10\bar{y} = -66$ ……..(1)

$40\bar{x} - 18\bar{y} = 214$ ……..(2)

Solving these (1) and (2) we get, $\bar{x} = 13$, $\bar{y} = 17$

Practice Problem:

1. Find the regression equations for the following data:

X	1	3	5	7	9
Y	15	18	21	23	22

Solution: $x = 0.887y - 12.562$, $y = 0.95x + 15.05$