When three or more variables are involved in correlation a the correlation between the dependent variable and only one pa independent variable is called partial correlation. The influence o independent variable is excluded.

For example, the yield of rice is related with the applica fertilizers and the rainfall. In this case, the relation of yield to f excluding the effect of rainfall, the relation of yield to rainfall ex the usage of fertilizer are partial correlations.

Linear and non-linear correlation :

If the amount of change in one variable tends to bear a ratio to the amount of change in the other variable, then the cor is said to be linear.

For example, if

X :	1	2	3	4	
Y :	5	10	15	20	

the variation between X and Y is a straight line.

A correlation is said to be non-linear or curvi linear if the of change in one variable does not bear a constant ratio to the of change in the other variable. For example, if rainfall is doubl production of rice would not necessarily be doubled.

Methods of studying correlation :

The following are some of the methods used for studyi correlation.

(i) Scatter diagram method
(ii) Graphic method
(iii) Karl pearson's co-efficient of correlation
(iv) Rank method
(v) Concurrent deviation method
(vi) Method of least squares.

Karl Pearson's co-efficient of correlation :

Let X and Y be given random variables. The Karl Pearson's co-efficient of correlation is denoted by r_{XY} or $r(X, Y)$ and defined as

$$r(X, Y) = r_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

where $\text{Cov}(X, Y) = E(XY) - E(X)\,E(Y)$

$$= \frac{1}{n}\Sigma xy - \bar{X}\bar{Y}, \quad \bar{X} = \frac{\Sigma x}{n}$$

and $\bar{Y} = \frac{\Sigma y}{n}$ and 'n' is the no. of items in the given data.

$$\sigma_X^2 = \text{Var}(X) = \frac{1}{n}\Sigma X^2 - (\bar{X})^2$$

and $$\sigma_Y^2 = \text{Var}(Y) = \frac{1}{n}\Sigma y^2 - (\bar{Y})^2$$

Note that correlation co-efficient always lies between –1 to +1.

Note : Two random variables with non zero correlation are said to be correlated.

3. RANK CORRELATION

Let us suppose that a group of 'n' individuals is arranged in order of merit or proficiency in possession of two characteristics A and B. These ranks in the two characteristics will, in general, be different. For example, if we consider the relation between intelligence and beauty, it is not necessary that a beautiful individual is intelligent also.

If (X_i, Y_i), $i = 1, 2, \ldots n$ are the ranks of the individuals in two characteristics A and B respectively, then the rank correlation co-efficient is given by,

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

$$d_i = x_i - y_i \; ; \qquad n = \text{no. of items}$$

where d_i is the different between the ranks. This formula is called Karl Pearson's formula for the rank correlation co-efficient.

4. REPEATED RANKS

If any two or more individuals are equal in any classification with respect to characteristic A or B, or if there is more than one item with the same value in the series then Spearman's formula for calculating the rank correlation coefficients breaks down. In this case common ranks are given to the repeated ranks. This common rank is the average of the ranks which these items would have assumed if they are slightly different from each other and the next item will get the rank next to the ranks already assumed. As a result of this, following adjustment or correction is made in the correlation formula.

In the correlation formula, we add the factor $\frac{m\,(m^2-1)}{12}$ to Σd^2 where m is the number of times an item is repeated. This correction factor is to be added for each repeated value.

Example 2.2.1

Calculate the correlation co-efficient for the following heights (in inches) of fathers X their sons Y. **[A.U. N/D 2004]**

X :	65	66	67	67	68	69	70	72
Y :	67	68	65	68	72	72	69	71

Solution :

X	Y	XY	X^2	Y^2
65	67	4355	4225	4489
66	68	4488	4356	4624
67	65	4355	4489	4225
67	68	4556	4489	4624
68	72	4896	4624	5184
69	72	4968	4761	5184
70	69	4830	4900	4761
72	71	5112	5184	5041
544	552	37560	37028	38132

$$\bar{X} = \frac{\Sigma X}{n} = \frac{544}{8} = 68 \; ; \; \bar{Y} = \frac{\Sigma Y}{n} = \frac{552}{8} = 69$$

$$\bar{X}\bar{Y} = 68 \times 69 = 4692$$

$$\sigma_X = \sqrt{\frac{1}{n}\Sigma X^2 - (\bar{X})^2} = \sqrt{\frac{1}{8}(37028) - 68^2} = \sqrt{4628.5 - 4624} = 2.121$$

$$\sigma_Y = \sqrt{\frac{1}{n}\Sigma Y^2 - (\bar{Y})^2} = \sqrt{\frac{1}{8}(38132) - 69^2} = \sqrt{4766.5 - 4761} = 2.345$$

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y} = \frac{1}{8}(37560) - 68 \times 69$$

$$= 4695 - 4692 = 3$$

The correlation co-efficient of X and Y is given by,

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{3}{(2.121)(2.345)} = \frac{3}{4.973} = 0.6032$$

Note : Correlation co-efficient is independent of change of origin and scale.

i.e., $r(X, Y) = r(U, V)$ where $U = \frac{X - a}{h}$; $V = \frac{Y - b}{K}$

where a and b are some arbitrary constants usually the mid-values of the given data X and Y respectively.

Example 2.2.2

Find the rank correlation co-efficient from the following data :

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

Solution :

X	Y	$d_i = x_i - y_i$	d_i^2
1	4	−3	9
2	3	−1	1
3	1	2	4
4	2	2	4
5	6	−1	1
6	5	1	1
7	7	0	0
		0	20

Here, $n = 7$

$\therefore$ Rank correlation co-efficient

$$r(X, Y) = 1 - \frac{6 \Sigma d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \times 20}{7(49 - 1)} = 0.6429$$

Example 2.2.3

Ten participants were ranked according to their performance in a mustical test by the 3 Judges in the following data.

	1	2	3	4	5	6	7	8	9	10
Rank by X	1	6	5	10	3	2	4	9	7	8
Rank by Y	3	5	8	4	7	10	2	1	6	9
Rank by Z	6	4	9	8	1	2	3	10	5	7

Using rank correlation method, discuss which pair of judges has the nearest approach to common likings of music.

Solution :

x_i	y_i	z_i	$d_1 = x_i - y_i$	$d_2 = y_i - z_i$	$d_3 = x_i - z_i$	d_1^2	d_2^2	d_3^2
1	3	6	−2	−3	−5	4	9	25
6	5	4	1	1	2	1	1	4
5	8	9	3	−1	−4	9	1	16
10	4	8	6	−4	2	36	16	4
3	7	1	−4	6	2	16	36	4
2	10	2	−8	8	0	64	69	0
4	2	3	2	−1	1	4	1	1
9	1	10	8	−9	−1	64	81	1
7	6	5	1	1	2	1	1	4
8	9	7	−1	2	1	1	4	1

The rank correlation co-efficient between X and Y is given by

$$r(X, Y) = 1 - \frac{6 \Sigma d_1^2}{n(n^2 - 1)}$$

Here, $n = 10$

$$= 1 - \frac{6 \times 200}{10\,(100-1)} = -0.212$$

The rank correlation co-efficient between Y and Z is given by

$$r(Y, Z) = 1 - \frac{6 \Sigma d_2^2}{n\,(n^2-1)} = 1 - \frac{6 \times 214}{10\,(100-1)} = -0.296$$

The rank correlation co-efficient between X and Z is given by

$$r(X, Z) = 1 - \frac{6 \Sigma d_3^2}{n\,(n^2-1)} = 1 - \frac{6 \times 60}{10\,(100-1)} = 0.636$$

Since the rank correlation coefficient between X and Z is positive and maximum, we conclude that the pair of judges X and Z has the nearest approach to common likings in music.

Example 2.2.4

Obtain the rank correlation coefficient for the following data :

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

Solution :

X	Y	Rank X (x_i)	Rank Y (y_i)	$d_i = x_i - y_i$	d^2
68	62	4	5	−1	1
64	58	6	7	−1	1
75	68	2.5	3.5	−1	1
50	45	9	10	−1	1
64	81	6	1	5	25
80	60	1	6	−5	25
75	68	2.5	3.5	−1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16
					72

In X series 75 is repeated twice which are in the positions 2nd and 3rd ranks. Therefore common ranks 2.5 (which is the average of 2 and 3) is to be given for each 75. Also in X series 64 is repeated thrice which are in the position 5th, 6th and 7th ranks.

Therefore common ranks 6 (which is the average of 5, 6 and 7) is to be given for each 64.

Similarly in Y series 68 is repeated twice which are in the positions 3rd and 4th ranks. Therefore common ranks 3.5 (which is the average of 3 and 4) is to be given for each 68.

Correction factors

In X series 75 is repeated twice

$$\therefore \text{C.F} = \frac{m\,(m^2 - 1)}{12}$$

Here, $m = 2$

$$\therefore \text{C.F}_1 = \frac{2\,(2^2 - 1)}{12} = \frac{2\,(4 - 1)}{12} = \frac{6}{12} = \frac{1}{2}$$

In X series 64 is repeated thrice

$$\therefore \text{C.F} = \frac{m\,(m^2 - 1)}{12}$$

Here, $m = 3$

$$\text{C.F}_2 = \frac{3\,(3^2 - 1)}{12} = \frac{3\,(9 - 1)}{12} = \frac{24}{12} = 2$$

In Y series 68 is repeated twice

$$\therefore \text{C.F} = \frac{m\,(m^2 - 1)}{12}$$

Here, $m = 2$

$$\therefore \text{C.F}_3 = \frac{2\,(2^2 - 1)}{12} = \frac{6}{12} = \frac{1}{2}$$

$$\therefore r\,[X, Y] = 1 - \frac{6\,(\Sigma d^2 + C.F_1 + C.F_2 + C.F_3)}{n\,(n^2 - 1)}$$

Here, $n = 10$ (10 data's given)

$$\therefore\ r[X, Y] = 1 - \frac{6\left(72 + \frac{1}{2} + 2 + \frac{1}{2}\right)}{10\,(10^2 - 1)}$$

$$= 1 - \frac{6\,(75)}{(10)\,(99)} = 1 - \frac{450}{990} = 0.5454$$

Example 2.2.5

The joint probability mass function of X and Y is given below.

y \ x	-1	$+1$
0	$\frac{1}{8}$	$\frac{3}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$

Find the correlation coefficient of (X, Y). **[A.U. N/D 2004]**

Solution :

y \ x	-1	1	$p(y) = p_{*j}$
0	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{4}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{4}{8}$
$p_{i\bullet} = p(x)$	$\frac{3}{8}$	$\frac{5}{8}$	1

(i) $E[X] = \Sigma x_i p(x_i) = (-1)\left(\frac{3}{8}\right) + (1)\left(\frac{5}{8}\right) = \frac{2}{8}$

(ii) $E[X^2] = \Sigma x_i^2 p(x_i) = (1)\frac{3}{8} + (1)\left(\frac{5}{8}\right) = 1$

(iii) $E[Y] = \Sigma y_j p(y_j) = (0)\left(\frac{4}{8}\right) + (1)\left(\frac{4}{8}\right) = \frac{4}{8} = \frac{1}{2}$

(iv) $E[Y^2] = \Sigma y_j^2 p(y_j) = (0)\left(\frac{4}{8}\right) + (1)\left(\frac{4}{8}\right) = \frac{4}{8} = \frac{1}{2}$

(v) $E[XY] = \sum_i \sum_j x_i y_j \, p(x_i y_j)$

$= (0)(-1)\left(\frac{1}{8}\right) + (0)(1)\left(\frac{3}{8}\right) + (1)(-1)\left(\frac{2}{8}\right) + (1)(1)\left(\frac{2}{8}\right)$

$= 0 + 0 - \frac{2}{8} + \frac{2}{8} = 0$

(vi) $\sigma_x^2 = E[X^2] - [E(X)]^2 = 1 - \left(\frac{1}{4}\right)^2 = 1 - \frac{1}{16} = \frac{15}{16}$

$\sigma_x = \frac{\sqrt{15}}{4}$

(vii) $\sigma_y^2 = E[Y^2] - [E[Y]]^2 = \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

$\sigma_y = \frac{1}{2}$

(viii) $r_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y}$

$= \frac{0 - \left(\frac{1}{4}\right)\left(\frac{1}{2}\right)}{\left(\frac{\sqrt{15}}{4}\right)\left(\frac{1}{2}\right)} = \frac{-\left(\frac{1}{8}\right)}{\left(\frac{\sqrt{15}}{8}\right)} = \frac{-1}{\sqrt{15}} = -0.258$

Example 2.2.6

Let X and Y be discrete RVs with probability function

$f(x, y) = \frac{x + y}{21}$, $x = 1, 2, 3$; $y = 1, 2$.

Find (i) Mean and Variance of X and Y. (ii) Cov (X, Y) (iii) Correlation of X and Y. **[A.U CBT A/M 2011]**

Y	X	1	2	3	$f(y)$
	1	$\frac{2}{21}$	$\frac{3}{21}$	$\frac{4}{21}$	$\frac{9}{21}$
	2	$\frac{3}{21}$	$\frac{4}{21}$	$\frac{5}{21}$	$\frac{12}{21}$
	$f(x)$	$\frac{5}{21}$	$\frac{7}{21}$	$\frac{9}{21}$	1

$$E(X) = \Sigma x f(x) = 1 \times \frac{5}{21} + 2 \times \frac{7}{21} + 3 \times \frac{9}{21}$$

$$= \frac{5}{21} + \frac{14}{21} + \frac{27}{21} = \frac{46}{21}$$

$$E(Y) = \Sigma y f(y) = 1 \times \frac{9}{21} + 2 \times \frac{12}{21} = \frac{9}{21} + \frac{24}{21} = \frac{33}{21}$$

$$E(X^2) = \Sigma x^2 f(x) = 1^2 \times \frac{5}{21} + 2^2 \times \frac{7}{21} + 3^2 \times \frac{9}{21}$$

$$= \frac{5}{21} + \frac{28}{21} + \frac{81}{21} = \frac{114}{21}$$

$$E(Y^2) = \Sigma y^2 f(y) = 1^2 \times \frac{9}{21} + 2^2 \times \frac{12}{21} = \frac{9}{21} + \frac{48}{21} = \frac{57}{21}$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{114}{21} - \left(\frac{46}{21}\right)^2 = \frac{278}{441}$$

$$V(Y) = E(Y^2) - [E(Y)]^2 = \frac{57}{21} - \left(\frac{33}{21}\right)^2 = \frac{108}{441}$$

$$E(XY) = 1.1.\frac{2}{21} + 1.2.\frac{3}{21} + 1.3.\frac{4}{21} + 2.1.\frac{3}{21} + 2.2.\frac{4}{21} + 2.3.\frac{5}{21}$$

$$= \frac{2}{21} + \frac{6}{21} + \frac{12}{21} + \frac{6}{21} + \frac{16}{21} + \frac{30}{21} = \frac{72}{21}$$

$$\text{Cov}(X, Y) = \frac{72}{21} - \frac{46}{21} \cdot \frac{33}{21} = \frac{1512 - 1518}{441} = \frac{-6}{441} = -0.0136$$

$$\therefore r_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{\frac{-6}{441}}{\frac{\sqrt{278}}{21} \cdot \frac{\sqrt{108}}{21}} = \frac{-6}{(16.673)(10.392)}$$

$$= \frac{-6}{173.265} = -0.0346$$

Example 2.2.7

Two random variables X and Y have the joint density

$f(x, y) = \begin{cases} 2-x-y, 0 < x < 1, 0 < y < 1 \\ 0 \quad \text{otherwise} \end{cases}$ **Show that** $\text{cov}(X, Y) = \frac{-1}{144}$.

[AU, N/D. 2004, M/J 2006, N/D 2010, Tvli A/M 2009, M/J 2010] [A.U. CBT M/J 2010] [A.U N/D 2011]

Solution : The marginal density function of X is,

$$f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^1 (2 - x - y)\, dy$$

$$= \left[(2-x)y - \frac{y^2}{2}\right]_0^1 = 2 - x - \frac{1}{2} = \frac{3}{2} - x, \; 0 < x < 1.$$

Similarly, the marginal density function of Y is,

$$f(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^1 (2 - x - y)\, dx$$

$$= \left[(2-y)x - \frac{x^2}{2}\right]_0^1 = 2 - y - \frac{1}{2} = \frac{3}{2} - y, \; 0 < y < 1$$

$$\text{Now, } E(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_0^1 x\left(\frac{3}{2} - x\right) dx = \int_0^1 \left(\frac{3}{2}x - x^2\right) dx$$

$$= \left[\frac{3x^2}{4} - \frac{x^3}{3}\right]_0^1 = \frac{3}{4} - \frac{1}{3} = \frac{9-4}{12} = \frac{5}{12}$$

$$E(Y) = \int_{-\infty}^{\infty} y f(y)\, dy = \int_0^1 y\left(\frac{3}{2} - y\right) dy = \int_0^1 \left(\frac{3}{2}y - y^2\right) dx$$

$$= \left[\frac{3y^2}{4} - \frac{y^3}{3}\right]_0^1 = \frac{3}{4} - \frac{1}{3} = \frac{9-4}{12} = \frac{5}{12}$$

$$E(XY) = \int_0^1 \int_0^1 xy f(x, y)\, dx\, dy = \int_0^1 \int_0^1 xy\,(2 - x - y)\, dx\, dy$$

$$= \int_0^1 \int_0^1 (2xy - x^2y - xy^2)\, dx\, dy$$

$$= \int_0^1 \left[x^2y - \frac{x^3y}{3} - \frac{x^2y^2}{2}\right]_0^1 dy$$

$$= \int_0^1 \left(y - \frac{y}{3} - \frac{y^2}{2}\right) dy = \int_0^1 \left(\frac{2y}{3} - \frac{y^2}{2}\right) dy$$

$$= \left[\frac{y^2}{3} - \frac{y^3}{6}\right]_0^1 = \frac{1}{3} - \frac{1}{6} = \frac{2-1}{6} = \frac{1}{6}$$

$\therefore$ Cov (X, Y) = E(XY) − E(X) . E(Y)

$$= \frac{1}{6} - \frac{5}{12} \cdot \frac{5}{12} = \frac{1}{6} - \frac{25}{144} = \frac{24-25}{144} = \frac{-1}{144}$$

Example 2.2.8

Suppose that the 2D RVs (X, Y) has the joint p.d.f.

$$f(x, y) = \begin{cases} x+y, & 0 < x < 1, 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Obtain the correlation co-efficient between X and Y.

Check whether X and Y are independent.

[AU, N/D, 2003, 2004] [A.U Tvli M/J 2010] [A.U A/M 2010] [A.U CBT N/D 2011]

Solution : The marginal density function of X is given by,

$$f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_0^1 (x + y)\, dy = \left[xy + \frac{y^2}{2}\right]_0^1$$

$$= x + \frac{1}{2},\ 0 < x < 1$$

The marginal density function of Y is given by,

$$f(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^1 (x + y)\, dx = \left[\frac{x^2}{2} + xy\right]_0^1 = y + \frac{1}{2},\ 0 < y < 1$$

$$E(X) = \int_0^1 xf(x)\,dx = \int_0^1 x\left(x+\frac{1}{2}\right)dx = \int_0^1 \left(x^2+\frac{x}{2}\right)dx$$

$$= \left[\frac{x^3}{3}+\frac{x^2}{4}\right]_0^1 = \frac{1}{3}+\frac{1}{4} = \frac{7}{12}$$

$$E(Y) = \int_0^1 yf(y)\,dy = \int_0^1 y\left(y+\frac{1}{2}\right)dy = \left[\frac{y^3}{3}+\frac{y^2}{4}\right]_0^1 = \frac{7}{12}$$

$$E(X^2) = \int_0^1 x^2 f(x)\,dx = \int_0^1 x^2\left(x+\frac{1}{2}\right)dx = \int_0^1 \left(x^3+\frac{x^2}{2}\right)dx$$

$$= \left[\frac{x^4}{4}+\frac{x^3}{6}\right]_0^1 = \frac{1}{4}+\frac{1}{6} = \frac{3+2}{12} = \frac{5}{12}.$$

Similarly $E(Y^2) = \frac{5}{12}$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{5}{12} - \frac{49}{144} = \frac{11}{144}$$

$$\therefore \sigma_X = \frac{\sqrt{11}}{12} \text{ and } |||\text{ly } \sigma_Y = \frac{\sqrt{11}}{12}$$

$$\therefore r_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{E(XY) - E(X)\,E(Y)}{\sigma_X \cdot \sigma_Y}$$

$$E(XY) = \int_0^1 \int_0^1 (xy)(x+y)\,dxdy = \int_0^1 \int_0^1 (x^2y + xy^2)\,dx\,dy$$

$$= \int_0^1 \left[\frac{x^3y}{3}+\frac{x^2y^2}{2}\right]_0^1 dy = \int_0^1 \left(\frac{y}{3}+\frac{y^2}{2}\right)dy$$

$$= \left[\frac{y^2}{6}+\frac{y^3}{6}\right]_0^1 = \frac{1}{6}+\frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

$\therefore$ Cov (X, Y) = E(XY) − E(X) E(Y)

$$= \frac{1}{3} - \frac{7}{12} \cdot \frac{7}{12} = \frac{1}{3} - \frac{49}{144} = \frac{48 - 49}{144} = \frac{-1}{144}$$

$Cov(X, Y) \neq 0$

$\therefore$ X and Y are not independent.

$$\therefore r_{XY} = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{\frac{-1}{144}}{\frac{\sqrt{11}}{12} \cdot \frac{\sqrt{11}}{12}} = \frac{-1}{11} = 0.0909 \ (+ve)$$

Example 2.2.9

Let X be a random variable with p.d.f $f(x) = \frac{1}{2}$, $-1 \leq x \leq 1$ and let $Y = X^2$. Prove that, the correlation co-efficient between X and Y is zero.

Solution :

$$E(X) = \int_{-1}^{1} x f(x)\, dx = \frac{1}{2} \int_{-1}^{1} x\, dx = \frac{1}{2} \left[\frac{x^2}{2}\right]_{-1}^{1} = \frac{1}{2} \left[\frac{1}{2} + \frac{1}{2}\right] = 0$$

$$E(XY) = E(X^3) = \int_{-1}^{1} x^3 f(x)\, dx = \int_{-1}^{1} x^3 \frac{1}{2} dx$$

$$= \frac{1}{2} \int_{-1}^{1} x^3\, dx = 0 \ [\because \text{for odd fn.}]$$

$$E(Y) = E(X^2) = \int_{-1}^{1} x^2 f(x)\, dx = \frac{1}{2} \left[\frac{x^3}{3}\right]_{-1}^{1} = \frac{1}{2} \left[\frac{1}{3} + \frac{1}{3}\right] = \frac{1}{3}$$

$$Cov(X, Y) = E(XY) - E(X)\, E(Y) = 0 \Rightarrow r_{XY} = 0$$

Example 2.2.10

Two independent random variables X and Y are defined by,

$f(x) = 4ax$, $0 \leq x \leq 1$
$= 0$, otherwise

$f(y) = 4by$, $0 \leq y \leq 1$
$= 0$, otherwise

Show that $U = X + Y$ and $V = X - Y$ are uncorrelated.

[AU A/M 2003, N/D 2012]

Solution :

(i) $f(x) = 4ax,\ 0 \le x \le 1$
$= 0$, otherwise

$f(x)$ is the density function of X

$$\int_0^1 f(x)\,dx = 1$$

$$\int_0^1 4ax\,dx = 1$$

$$4a\left[\frac{x^2}{2}\right]_0^1 = 1$$

$$2a = 1\ ;\ a = \frac{1}{2}$$

$f(y) = 4by,\ 0 \le y \le 1$
$= 0$, otherwise

$f(y)$ is the density function of Y

$$\int_0^1 f(y)\,dy = 1$$

$$\int_0^1 4by\,dy = 1$$

$$4b\left[\frac{y^2}{2}\right]_0^1 = 1$$

$$2b = 1\ ;\ b = \frac{1}{2}$$

To prove $U = X + Y$ and $V = X - Y$ are uncorrelated

i.e., to prove Cov $(U, V) = 0$

Cov (U, V) $= E(UV) - E(U)\,E(V)$

$E(U) = E[X + Y] = E(X) + E(Y)$

$E(V) = E[X - Y] = E(X) - E(Y)$

$E(UV) = E[X^2 - Y^2]$

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx = \int_0^1 x(4ax)\,dx = 4a\int_0^1 x^2\,dx = 4a\left[\frac{x^3}{3}\right]_0^1 = \frac{4a}{3} = \frac{2}{3}$$

$$E(Y) = \int_{-\infty}^{\infty} y f(y)\,dy = \int_0^1 y(4by)\,dy = 4b\int_0^1 y^2\,dy = 4b\left[\frac{y^3}{3}\right]_0^1 = \frac{4b}{3} = \frac{2}{3}$$

$E[XY] = E[X]\,E[Y]$ [$\because$ X and Y are independent]

$$= \frac{2}{3}\cdot\frac{2}{3} = \frac{4}{9}$$

$$E[U] = E[X + Y] = E[X] + E[Y] = \frac{2}{3} + \frac{2}{3} = \frac{4}{3}$$

$$E[V] = E[X - Y] = E[X] - E[Y] = \frac{2}{3} - \frac{2}{3} = 0$$

$$E[UV] = E[X^2 - Y^2] = E[X^2] - E[Y^2] = \frac{1}{2} - \frac{1}{2} = 0$$

$$E[X^2] = \int_0^1 x^2 f(x)\,dx = \int_0^1 x^2 (2x)\,dx = \int_0^1 2x^3\,dx = 2\left[\frac{x^4}{4}\right]_0^1$$
$$= \frac{1}{2}(1-0) = \frac{1}{2}$$

$$E[Y^2] = \int_0^1 y^2 f(y)\,dy = \int_0^1 y^2 (2y)\,dy = \int_0^1 2y^3\,dy = 2\left[\frac{y^4}{4}\right]_0^1$$
$$= \frac{1}{2}(1-0) = \frac{1}{2}$$

$$Cov(U, V) = E[UV] - E[U]\,E[V] = 0 - \frac{4}{3}(0) = 0$$

$\therefore$ U and V are uncorrelated.

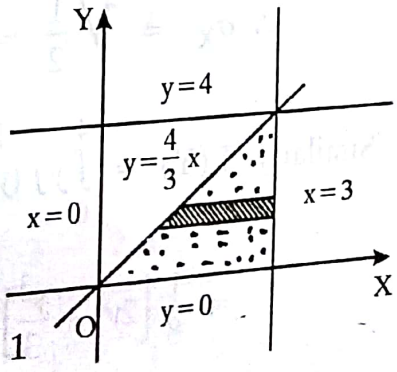
Example 2.2.11

If (X, Y) is a two-dimensional random variable uniformly distributed over the triangular region R bounded by y=0, x = 3, and $y = \frac{4}{3}x$. Find the correlation coefficient r_{xy}. **[A.U.]**

Sol. (X, Y) is uniformly distributed, $f(x, y) = K$, constant (say)

To find the point of X^n of $x = 3$ and $y = \frac{4}{3}x$

$y = \frac{4}{3}x$, where $x = 3 \Rightarrow y = 4$



$\therefore f(x, y)$ is a pdf, we have

$$\iint_R f(x, y)\,dx\,dy = 1$$

$$\therefore \int_0^4 \int_{\frac{3y}{4}}^3 K\,dx\,dy = 1 \Rightarrow K\int_0^4 \left[x\right]_{\frac{3y}{4}}^3 dy = 1$$

$$K\int_0^4 \left[3 - \frac{3y}{4}\right] dy = 1 \Rightarrow 3K\left[y - \frac{y^2}{8}\right]_0^4 = 1$$

$$3K[2] = 1 \Rightarrow K = \frac{1}{6} \quad \therefore f(x, y) = \frac{1}{6}$$

2.3 REGRESSION

■ (1) Regression ■

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

■ (2) Lines of regression ■

(1) The line of regression of y on x is given by

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}) \qquad \ldots (1)$$

(2) The line of regression of x on y is given by

$$x - \bar{x} = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \qquad \ldots (2)$$

Note : Both the lines of regression passes through $(\overline{X}, \overline{Y})$

(3) Regression coefficients

(1) Regression coefficient of y on x is $r\frac{\sigma_y}{\sigma_x} = b_{yx}$

(2) Regression coefficient of x on y is $r\frac{\sigma_x}{\sigma_y} = b_{xy}$

Correlation coefficient $r = \pm\sqrt{b_{yx}\, b_{xy}}$

where $b_{yx} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2}$; $b_{xy} = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (y-\bar{y})^2}$

(4) Properties of Regression Lines

(1) The regression lines pass through $(\bar{x}, \bar{y})$. So $(\bar{x}, \bar{y})$ is the point of intersection of the regression lines.

(2) When $r = 1$, that is when there is a perfect +ve correlation or when $r = -1$, that is when there is a perfect –ve correlation the equation (1) and (2) becomes one are the same and so the regression lines coincide

(3) When $r = 0$ the equations of the lines are $y = \bar{y}$ and $x = \bar{x}$ which represent perpendicular lines which are parallel to the axis.

(4) The slopes of the lines are $r\frac{\sigma_y}{\sigma_x}, \frac{1}{r}\frac{\sigma_y}{\sigma_x}$

Since the S.D's σ_x and σ_y are +ve, both the slopes are +ve if r is +ve and –ve if r is –ve. That is all the three, namely the two slopes and r are of same sign.

(5) Angle between the regression lines

The slopes of the regression lines are

$$m_1 = r\frac{\sigma_y}{\sigma_x}, \qquad m_2 = \frac{1}{r}\frac{\sigma_y}{\sigma_x}$$

If θ is the angle between the lines, then

$$\tan\theta = \frac{m_2 - m_1}{1 + m_1 m_2} = \frac{\sigma_y}{\sigma_x} \frac{\frac{1}{r} - r}{1 + \left(\frac{\sigma_y}{\sigma_x}\right)^2} = \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\left[\frac{1}{r} - r\right]$$

$$\tan\theta = \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}\left[\frac{1 - r^2}{r}\right]$$

Note : 1. When $r = 0$, that is, when there is no correlation between x and y.

$\tan\theta = \infty$ (or) $\theta = \frac{\pi}{2}$ and so the regression lines are perpendicular

2. When $r = 1$ or -1, that is, when their is a perfect correlation, +ve or –ve, $\theta = 0$ and so the lines coincide.

■ 6. Correlation coefficient is the geometric mean between the two regression coefficients ■

Proof : We know that, $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ and $b_{yx} = r\frac{\sigma_y}{\sigma_x}$

$$\Rightarrow (b_{xy})(b_{yx}) = r^2 \frac{\sigma_x}{\sigma_y}\frac{\sigma_y}{\sigma_x} = r^2$$

$$\Rightarrow r = \pm\sqrt{(b_{xy})(b_{yx})}$$

■ 7. If one of the regression coefficient is greater than unity the other must be less than unity. ■

Proof : We know that, $r^2 = b_{xy}\, b_{yx} \le 1$ ... (1)

Assume that $b_{xy} > 1$

we have, to prove that $b_{yx} < 1$

Since, $b_{xy} > 1$; $\frac{1}{b_{xy}} < 1$

$\therefore$ (1) $\Rightarrow b_{xy}\, b_{yx} \le 1$; $b_{yx} \le \frac{1}{b_{xy}} < 1$; $\therefore b_{yx} < 1$

8. Distinguish between correlation and regression Analysis

	Correlation		Regression
1.	Correlation means relationship between two variables	1.	Regression is a mathematical measure of expressing the average relationship between the two variables.
2.	Correlation need not imply cause and effective relationship between the variables.	2.	Regression Analysis clearly indicates the cause and effect relationship between variables.
3.	Correlation coefficient is symmetric is $r_{xy} = r_{yx}$.	3.	Regression coefficient is not symmetric i.e. $b_{xy} \pm b_{yx}$
4.	Correlation coefficient is a measure of the direction and degree of linear relationship between two variables.	4.	Using the relationship between two variables we can predict the dependent variable value for any given independent variable value.

9. Standard errors of estimate

The standard error of estimate of x is

(1) $S_x = \sigma_x \sqrt{1 - r^2}$

(2) The standard error of estimate of y is

$S_y = \sigma_y \sqrt{1 - r^2}$

10. Correlation of Grouped data

When the number of observations is large and the variables are grouped, the data can be classified into two way frequency distribution called a correlation table. If there are 'n' classes for X and 'm' classes for Y, there will be $(m \times n)$ cells in the two-way table.

The formula for calculating the co-efficient of correlation is

$$r = \frac{\rho}{\sigma_X \sigma_Y}$$

where $$\rho = \frac{\Sigma XY f_{xy}}{N} - \left(\frac{\Sigma X f_x}{N}\right)\left(\frac{\Sigma Y f_y}{N}\right)$$

$$\sigma_x^2 = \frac{\Sigma X^2 f_x}{N} - \left(\frac{\Sigma X f_x}{N}\right)^2$$

and $$\sigma^2 Y = \frac{\Sigma Y^2 f_y}{N} - \left(\frac{\Sigma Y f_y}{N}\right)^2$$

■ 11. Probable Error of correlation co-efficient ■

The probable error of correlation co-efficient is given by,

$$\text{P.E. } (r) = 0.6745 \times \text{S.E.}$$

where S.E. is the standard error and is S.E. $(r) = \frac{1 - r^2}{\sqrt{n}}$, where '$r$' is the correlation co-efficient and 'n' is the number of observation.

Thus $$\text{P.E. } (r) = 0.6745 \frac{(1 - r^2)}{\sqrt{n}}$$

The reason for taking the factor 0.6745 is that in a normal distribution, the range $\mu = \pm 0.6745$ covers 50% of the total area. This error enables us to find the limits within which correlation co-efficient can be expected to vary.

Example 2.3.1

From the following data, find (i) the two regression equations, (ii) the coefficient of correlation between the marks in Economics and statistics, (iii) the most likely marks in statistics when marks in Economics are 30. [A.U M/J 2007]

Marks in Economics x	25	28	35	32	31	36	29	38	34	32
Statistics y	43	46	49	41	36	32	31	30	33	39

Solution :

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
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

$$\bar{x} = \frac{\Sigma x}{n} = \frac{320}{10}, \qquad \bar{y} = \frac{\Sigma y}{n} = \frac{380}{10}$$

$$= 32, \qquad \bar{y} = 38$$

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

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

(i) (a) Equation of the line of regression of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

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

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

(i) (b) Equation of the line of regression of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

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

$$y = -0.6643x + 59.2576$$

(ii) Coefficient of correlation

$$r^2 = b_{yx}\, b_{xy} = (-0.6643)\,(-0.2337) = 0.1552$$

$$r = \pm 0.394$$

(iv) The most likely marks in statistics (y) when marks in Eco(x) are

i.e., $y = -0.6643\, x + 59.2575$

$x = 30 \Rightarrow y = 39$

Example 2.3.2

The two lines of regression are

$$8x - 10y + 66 = 0 \quad \text{... (A)}$$

$$40x - 18y - 214 = 0 \quad \text{... (B)}$$

The variance of x is 9. Find (i) The mean values of x and y (ii) Correlation coefficient between x and y **[AU N/D 200**

[A.U CBT M/J 2010, CBT N/D 2011, CBT A/M 201

Solution : (i) 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

$$8\bar{x} - 10\bar{y} = -66 \quad \text{... (1)}$$

$$40\bar{x} - 18\bar{y} = 214 \quad \text{... (2)}$$

(1) $\times$ 5 $\Rightarrow$ $40\bar{x} - 50\bar{y} = -330$ ... (3)

(2) $\times$ 1 $\Rightarrow$ $40\bar{x} - 18\bar{y} = 214$ ... (4)

(3) – (4) $\Rightarrow$ $-32\,\bar{y} = -544$

$$\bar{y} = 17$$

Sub in (1) we get

$$8\bar{x} - 10\,(17) = -66$$

$$\bar{x} = 13$$

Hence the mean values are given by $\bar{x} = 13$, $\bar{y} = 17$

(ii)

(A) $\Rightarrow 8x = 10y - 66$ $\Rightarrow x = \frac{10}{8}y - \frac{66}{8}$ i.e., $bxy = \frac{10}{8}$	(B) $\Rightarrow 18y = 40x - 214$ $\Rightarrow y = \frac{40}{18}x - \frac{214}{18}$ i.e., $byx = \frac{40}{18}$	$r^2 = bxy\ byx$ $= \left(\frac{10}{8}\right)\left(\frac{40}{18}\right)$ $= 2.77$ $r = 1.66 \not< 1$
(A) $\Rightarrow 10y = 8x + 66$ $\Rightarrow y = \frac{8}{10}x + \frac{66}{10}$ i.e., $byx = \frac{8}{10}$	(B) $\Rightarrow 40x = 18y + 214$ $\Rightarrow x = \frac{18}{40}y + \frac{214}{40}$ i.e., $bxy = \frac{18}{40}$	$r^2 = byx\ bxy$ $= \left(\frac{8}{10}\right)\left(\frac{18}{40}\right)$ $= 0.36$ $r = \pm 0.6$

Since both the regression coefficients are positive r must be positive $r = 0.6$.

Example 3.3.3

The following table gives according to age x, the frequency of marks obtained 'y' by 100 students in an intelligence test. Measure the degree of relationship between age and intelligence test.

Age/marks	16-17	17-18	18-19	19-20
30-40	20	10	3	2
40-50	4	28	6	4
50-60	0	5	11	0
60-70	0	0	2	0
70-80	0	0	0	5

The origin is taken as $\bar{x} = 18.5$ and $\bar{y} = 55$

$X = \frac{x - \bar{x}}{h} = \frac{x - 18.5}{1}$ [$\because h$ = difference between x values]

$Y = \frac{y - \bar{y}}{k} = \frac{y - 55}{10}$ [$\because k$ = difference between y values]

f_y → sum of the each row

f_x → sum of the each column

f_{xy} → Given frequency

$N = \Sigma f_x = \Sigma f_y = 100$

X / Y	16.5	17.5	18.5	19.5	f_y	Y	$Y f_y$	$Y^2 f_y$	XYf_{xy}
35	$f_{xy}=20$ $XY=4$ $XYf_{xy}=80$	10 2 =20	3 0 =0	2 −2 =−4	35	−2	−70	140	96
45	4 2 =8	28 1 =28	6 0 =0	4 −1 =−4	42	−1	−42	42	32
55	0 0 =0	5 0 =0	11 0 =0	0 0 =0	16	0	0	0	0
65	0 −2 =0	0 −1 =0	2 0 =0	0 1 =0	2	1	2	2	0
75	0 −4 =0	0 −2 =0	0 0 =0	5 2 =10	5	2	10	20	10
f_x	24	43	22	11	100	0	−100	204	138
X	−2	−1	0	1	−2				
Xf_x	−48	−43	0	11	−80				
$X^2 f_x$	96	43	0	11	150				
XYf_{xy}	88	48	0	2	138				

In each cell upper values are f_{xy} (given), middle are XY, lower are XYf_{xy}

$$\sigma^2 X = \frac{\Sigma (X^2 f_x)}{N} - \left(\frac{\Sigma (X f_x)}{N}\right)^2 = \frac{150}{100} - \left(\frac{-80}{100}\right)^2 = 0.86 \quad \therefore \sigma_x = 0.927$$

$$\sigma^2 Y = \frac{\Sigma (Y^2 f_y)}{N} - \left(\frac{\Sigma (Y f_y)}{N}\right)^2 = \frac{204}{100} - \left(\frac{-100}{100}\right)^2 = 1.04 \quad \sigma_Y = 1.019$$

$$\rho = \frac{\Sigma (XY f_{xy})}{N} - \left(\frac{\Sigma (x f_x)}{N}\right)\left(\frac{\Sigma (y f_y)}{N}\right) = \frac{138}{100} - \left(\frac{-80}{100}\right)\left(\frac{-100}{100}\right)$$

$$= 1.38 - 0.8 = 0.58$$

$$\therefore r = \frac{\rho}{\sigma_X \sigma_Y} = \frac{0.58}{(0.927)\ (1.019)} = 0.6137$$

Example 2.3.4

Calculate the co-efficient of correlation between x and y from the following table and write down the regression equation of y on x :

y / x	0 -	40 -	80 -	120 -
10 -	9	4	1	
30 -	47	19	6	
50 -	26	18	11	
70 -	2	3	2	2

[AU. A/M. 2004]

Solution : The origin is taken as $\bar{x} = 60$

The origin is taken as $\bar{y} = 40$

$$X = \frac{x - \bar{x}}{h} = \frac{x - 60}{40} \quad [\because h = \text{difference between } x \text{ values}]$$

$$Y = \frac{y - \bar{y}}{k} = \frac{y - 40}{20} \quad [\because k = \text{difference between } y \text{ values}]$$

f_x → sum of the each column

f_y → sum of the each row

f_{xy} → given frequency [in each cell upper values]

XY → In each cell middle values

$XY f_{xy}$ → In each cell sum of lower values

Y \ X	20	60	100	140	f_y	Y	$Y f_y$	$Y^2 f_y$	$XY f_{xy}$
20	9 1 9	4 0 0	1 −1 −1	0 −2 0	14	−1	−14	14	8
40	47 0 0	19 0 0	6 0 0	0 0 0	72	0	0	0	0
60	26 −1 −26	18 0 0	11 1 11	0 2 0	55	1	55	55	−15
80	2 −2 −4	3 0 0	2 2 4	2 4 8	9	2	18	36	8
f_x	84	44	20	2	$\Sigma f_x = \Sigma f_y = 150$		$\Sigma Y f_y = 59$	$\Sigma Y^2 f_y = 105$	$\Sigma XY f_{xy} = 1$
X	−1	0	1	2					
$X f_x$	−84	0	20	4	$\Sigma X f_x = -60$				
$X^2 f_x$	84	0	20	8	$\Sigma X^2 f_x = 112$				
$XY f_{xy}$	−21	0	14	8	$\Sigma XY f_{xy} = 1$				

$$\sigma_x^2 = \frac{\Sigma (X^2 f_x)}{N} - \left(\frac{\Sigma (X f x)}{N}\right)^2 = \frac{112}{150} - \left(\frac{-60}{150}\right)^2$$

$$= \frac{112}{150} - \frac{3600}{22500} = \frac{44}{75}$$

$$\sigma_x = 0.766$$

$$\sigma_y^2 = \frac{\Sigma (Y^2 f_y)}{N} - \left[\frac{\Sigma (Y f_y)}{N}\right]^2 = \frac{105}{150} - \left(\frac{59}{150}\right)^2$$

$$= \frac{105}{150} - \frac{3481}{22500} = 0.545$$

$$\sigma_y = 0.738$$

$$\rho = \frac{\Sigma (XY f_{xy})}{N} - \left(\frac{\Sigma (X f_x)}{N}\right) \left(\frac{\Sigma (Y f_y)}{N}\right)$$

$$= \frac{1}{150} - \left(\frac{-60}{150}\right) \left(\frac{59}{150}\right) = \frac{1}{150} + \frac{3540}{22500} = \frac{41}{250} = 0.164$$

$$r = \frac{\rho}{\sigma_x \sigma_y} = \frac{0.164}{(0.766)\,(0.738)} = \frac{0.164}{0.565308} = 0.29$$

The regression equation of y on x is

$$(y - \bar{y}) = r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$$

$$(y - 40) = (0.29) \frac{0.738}{0.766} (x - 60)$$

$$(y - 40) = (0.2794)\,(x - 60)$$

$$y - 40 = (0.28)\,(x - 60)$$

$$y - 40 = 0.28x - 16.8$$

$$y = 0.28x + 23.2$$

Note : The regression equation x on y is

$$(x - \bar{x}) = r \cdot \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

$$(x - 60) = (0.29) \frac{0.766}{0.738} (y - 40)$$

$$(x - 60) = 0.3\,(y - 40)$$

$$x - 60 = 0.3y - 12$$

$$x = 0.3y + 48$$