CHAPTER TEN

Correlation and Regression

10·1. Bivariate Distribution, Correlation. So far we have confined ourselves to unvariate distributions, *i.e.*, the distributions involving only one variable. We may, however, come across certain series where each term of the series may assume the values of two or more variables. For example, if we measure the heights and weights of a certain group of persons, we shall get what is known as *Bivariate distribution*—one variable relating to height and other variable relating to weight.

In a bivariate distribution we may be interested to find out if there is any correlation or covariation between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables deviate in the same direction, *i.e.*, if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be *direct* or *positive*. But if they constantly deviate in the opposite directions, *i.e.*, if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be *diverse* or *negative*. For example, the correlation between (*i*) the heights and weights of a group of persons, (*ii*) the income and expenditure is positive and the correlation between (*i*) price and demand of a commodity, (*ii*) the volume and pressure of a perfect gas, is negative. Correlation is said to be *perfect* if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

10·2. Scatter Diagram. It is the simplest way of the diagrammatic representation of bivariate data. Thus for the bivariate distribution (x_i, y_i); $i = 1, 2, \ldots, n$, if the values of the variables X and Y be plotted along the x-axis and y-axis respectively in the xy plane, the diagram of dots so obtained is known as *scatter diagram*. From the scatter diagram, we can form a fairly good, though vague, idea whether the variables are correlated or not, *e.g.*, if the points are very dense, *i.e.*, very close to each other, we should expect a fairly good amount of correlation between the variables and if the points are widely scattered, a poor correlation is expected. This method, however, is not suitable if the number of observations is fairly large.

10·3. Karl Pearson Coefficient of Correlation. As a measure of itensity or degree of linear relationship between two variables, Karl Pearson (1867—1936), *a British Biometrician*, developed a formula called *Correlation Coefficient*.

Correlation coefficient between two random variables X and Y, usually denoted by $r(X, Y)$ or simply r_{XY}, is a numerical measure of *linear relationship* between them and is defined as

$$r(X, Y) = \frac{\text{Cov}\,(X, Y)}{\sigma_X \sigma_Y} \tag{10·1}$$

If (x_i, y_i) ; $i = 1, 2, ..., n$ is the bivariate distribution, then

$$\left.\begin{aligned} \text{Cov}(X, Y) &= E[\{X - E(X)\}\{(Y - E(Y)\}] \\ &= \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \mu_{11} \\ \sigma_X^2 &= E\{X - E(X)\}^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2 \\ \sigma_Y^2 &= E\{Y - E(Y)\}^2 = \frac{1}{n}\Sigma(y_i - \bar{y})^2 \end{aligned}\right\}, \qquad \text{... (10·2)}$$

the summation extending over i from 1 to n.

Another convenient form of the formula (10·2) for computational work is as follows :

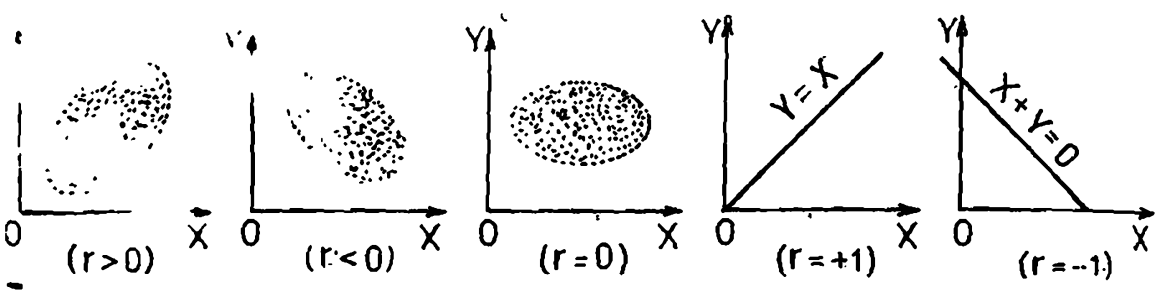
$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}(x_i y_i - x_i\bar{y} - \bar{x} y_i + \bar{x}\bar{y})$$

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

$$\therefore \qquad \text{Cov}(X, Y) = \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y}, \; \sigma_X^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2$$

and $$\sigma_Y^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2 \qquad \text{... (10.2}a\text{)}$$

Remarks 1. Following are the figures of the standard data for $r > 0$, < 0, $= 0$, and $r = \pm 1$.



2. It may be noted that $r(X, Y)$ provides a measure of *linear relationship* between X and Y. For nonlinear relationship, however, it is not very suitable.

3. Sometimes, we write : $\text{Cov}(X, Y) = \sigma_{XY}$

4. Karl Pearson's correlation coefficient is also called *product-moment correlation coefficient,* since

$\text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}] = \mu_{11}$.

10·3·1. Limits for Correlation Coefficient. We have

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\left[\frac{1}{n}\Sigma(x_i - \bar{x})^2 \cdot \frac{1}{n}\Sigma(y_i - \bar{y})^2\right]^{1/2}},$$

$$\therefore \quad r^2(X, Y) = \frac{(\Sigma a_i b_i)^2}{(\Sigma a_i^2)(\Sigma b_i^2)}, \text{ where } \begin{pmatrix} a_i = x_i - \bar{x} \\ b_i = y_i - \bar{y} \end{pmatrix} \qquad \ldots(*)$$

We have the Schwartz inequality which states that if $a_i, b_i; i = 1, 2, \ldots, n$ are real quantities then

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \le \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

the sign of equality holding if and only if

$$\frac{a_1}{b_1} = \frac{a_2}{b_2} = \ldots = \frac{a_n}{b_n}$$

Using Schwartz inequality, we get from (*)

$$r^2(X, Y) \le 1 \text{ i.e., } |r(X, Y)| \le 1 \Rightarrow -1 \le r(X, Y) \le 1 \qquad \ldots(10\cdot3)$$

Hence correlation coefficient cannot exceed unity numerically. It always lies between -1 and $+1$. If $r = +1$, the correlation is perfect and positive and if $r = -1$, correlation is perfect and negative.

Aliter. If we write $E(X) = \mu_X$ and $E(Y) = \mu_Y$, then we have

$$E\left[\left(\frac{X - \mu_X}{\sigma_X}\right) \pm \left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right]^2 \ge 0$$

$$\Rightarrow \quad E\left(\frac{X - \mu_X}{\sigma_X}\right)^2 + E\left(\frac{Y - \mu_Y}{\sigma_Y}\right)^2 \pm 2\frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \ge 0$$

$$\Rightarrow \quad 1 + 1 \pm 2r(X, Y) \ge 0$$

$$\Rightarrow \quad -1 \le r(X, Y) \le 1.$$

Theorem 10·1. *Correlation coefficient is independent of change of origin and scale.*

Proof. Let $U = \dfrac{X - a}{h}$, $V = \dfrac{Y - b}{k}$, so that $X = a + hU$ and $Y = b + kV$, where a, b, h, k are constants; $h > 0, k > 0$.

We shall prove that $r(X, Y) = r(U, V)$

Since $X = a + hU$ and $Y = b + kV$, on taking expectations, we get

$$E(X) = a + hE(U) \quad \text{and} \quad E(Y) = b + kE(V)$$

$$\therefore \quad X - E(X) = h[U - E(U)] \quad \text{and} \quad Y - E(Y) = k[V - E(V)]$$

$$\Rightarrow \text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

$$= E[h\{U - E(U)\}\{k\{V - E(V)\}]$$

$$= hk\,E[\{U - E(U)\}\{V - E(V)\}] = hk\ \text{Cov}(U, V) \qquad \ldots(10\cdot4)$$

$$\sigma_X^2 = E[\{X - E(X)\}^2] = E[h^2\{U - E(U)\}^2] = h^2\sigma_U^2$$

$$\Rightarrow \quad \sigma_X = h\sigma_U, (h > 0) \qquad \ldots(10\cdot4a)$$

$$\text{and} \quad \sigma_Y^2 = E[\{Y - E(Y)\}^2] = E[k^2\{V - E(V)\}^2] = k^2\sigma_V^2$$

$$\Rightarrow \quad \sigma_Y = k\sigma_V, (k > 0) \qquad \ldots(10\cdot4b)$$

Substituting from (10·4), (10·4*a*) and (10·4*b*) in (10·1), we get

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{hk.\ \text{Cov}(U, V)}{hk.\sigma_U \sigma_V} = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = r(U, V)$$

This theorem is of fundamental importance in the numerical computation of the correlation coefficient.

Corollary. *If X and Y are random variables and a, b, c, d are any numbers provided only that $a \neq 0$, $c \neq 0$, then*

$$r(aX + b, cY + d) = \frac{ac}{|ac|} r(X, Y)$$

Proof. With usual notations, we have

$$\text{Var}(aX + b) = a^2\sigma_X^2\ ;\ \text{Var}(cY + d) = c^2\sigma_Y^2\ ;$$

$$\text{Cov}(aX + b, cY + d) = ac\sigma_{XY}$$

$$\therefore\quad r(aX + b,\ cY + d) = \frac{\text{Cov}(aX + b, cY + d)}{[\text{Var}(aX + b)\ \text{Var}(cY + d)]^{1/2}}$$

$$= \frac{ac\ \sigma_{XY}}{|a|\,|c|\,\sigma_X \sigma_Y} = \frac{ac}{|ac|} r(X, Y)$$

If $ac > 0$, *i.e.*, if a and c are of same signs, then $ac/|ac| = +1$

If $ac < 0$, *i.e.*, if a and c are of opposite signs, then $ac/|ac| = -1$.

Theorem 10·2. *Two independent variables are uncorrelated.*

Proof. If X and Y are independent variables, then

$$\text{Cov}(X, Y) = 0 \qquad (c.f.\ \S\ 6\text{·}4)$$

$$\therefore\quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Hence two independent variables are uncorrelated.

But the converse of the theorem is not true, *i.e.*, two uncorrelated variables may not be independent as the following example illustrates :

X	−3	−2	−1	1	2	3	Total $\Sigma X = 0$
Y	9	4	1	1	4	9	$\Sigma Y = 28$
XY	−27	−8	−1	1	8	27	$\Sigma XY = 0$

$$\bar{X} = \frac{1}{n}\Sigma X = 0,\ \text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y} = 0$$

$$\therefore\quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Thus in the above example, the variables X and Y are uncorrelated. But on careful examination we find that X and Y are not independent but they are connected by the relation $Y = X^2$. Hence two uncorrelated variables need not necessarily be independent. A simple reasoning for this strange conclusion is that $r(X, Y) = 0$, merely implies the absence of any linear relationship between

the variables X and Y. There may, however, exist some other form of relationship between them, *e.g.*, quadratic, cubic or trigonometric.

Remarks. 1. Following are some more examples where two variables are uncorrelated but *not* independent.

(*i*) $X \sim N(0, 1)$ and $Y = X^2$

Since $X \sim N(0, 1)$, $E(X) = 0 = E(X^3)$

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

$$= E(X^3) - E(X)E(Y) = 0 \qquad (\because Y = X^2)$$

$$\Rightarrow \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = 0$$

Hence X and Y are uncorrelated but not independent.

(*ii*) Let X be a r.v. with p.d.f.

$$f(x) = \frac{1}{2}, -1 \le x \le 1$$

and let $Y = X^2$. Here we shall get

$$E(X) = 0 \text{ and } E(XY) = E(X^3) = 0, \Rightarrow r(X, Y) = 0$$

2. However, the converse of the theorem holds in the following cases :

(*a*) If X and Y are jointly normally distributed with $\rho = \rho(X, Y) = 0$, then they are independent. If $\rho = 0$, then [*c.f.* § 10·10, Equation (10·25)]

$$f(x, y) = \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X - \mu_X}{\sigma_X}\right)^2\right] \times \frac{1}{\sigma_Y \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y - \mu_Y}{\sigma_Y}\right)^2\right]$$

$$\therefore \quad f(x, y) = f_1(x) f_2(y)$$

$\Rightarrow$ X and Y are independent.

(*b*) *If each of the two variables X and Y takes two values, 0, 1 with positive probabilities, then r (X, Y) = 0 ⇒ X and Y are independent.*

Proof. Let X take the values 1 and 0 with positive probabilities p_1 and q_1 respectively and let Y take the values 1 and 0 with positive probabilities p_2 and q_2 respectively. Then

$$r(X, Y) = 0 \quad \Rightarrow \quad \text{Cov}(X, Y) = 0$$

$$\Rightarrow \quad 0 = E(XY) - E(X)E(Y)$$

$$= 1 \,.\, P(X = 1 \cap Y = 1) - [1 \,.\, P(X) = 1) \times 1 \,.\, P(Y = 1)]$$

$$= P(X = 1 \cap Y = 1) - p_1 p_2$$

$$\Rightarrow \quad P(X = 1 \cap Y = 1) = p_1 p_2 = P(X = 1) \,.\, P(Y = 1)$$

$\Rightarrow$ X and Y are independent.

10·3·2. Assumptions Underlying Karl Pearson's Correlation Coefficient. Pearsonian correlation coefficient r is based on the following assumptions :

(*i*) *The variables X and Y under study are linearly related.* In other words, the scatter diagram of the data will give a straight line curve.

(*ii*) *Each of the variables (series) is being affected by a large number of independent contributory causes of such a nature as to produce normal distribution.* For example, the variables (series) relating to ages, heights, weight, supply, price, etc., conform to this assumption. In the words of Karl Pearson :

"The sizes of the complex of organs (something measurable) are determined by a great variety of independent contributory causes, for example, climate, nourishment, physical training and innumerable other causes which cannot be individually observed or their effects measured." Karl Pearson further observes, "The variations in intensity of the contributory causes are small as compared with their absolute intensity and these variations follow the normal law of distribution."

(*iii*) *The forces so operating on each of the variable series are not independent of each other but are related in a causal fashion.* In other word, cause and effect relationship exists between different forces operating on the items of the two variable series. These forces must be common to both the series. If the operating forces are entirely independent of each other and not related in any fashion, then there cannot be any correlation between the variables under study.

For example, the correlation coefficient between,

(*a*) the series of heights and incomes of individuals over a period of time,

(*b*) the series of marriage rate and the rate of agricultural production in a country over a period of time,

(*c*) the series relating to the size of the shoe and intelligence of a group of individuals,

should be zero, since the forces affecting the two variable series in each of the above cases are entirely independent of each other.

However, if in any of the above cases the value of r for a given set of data is not zero, then such correlation is termed as *chance correlation* or *spurious* or *non-sense correlation.*

Example 10·1. *Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons (Y) :*

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

Solution.

CALCULATIONS FOR CORRELATION COEFFICIENT

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

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{544}{8} = 68, \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{8} \times 552 = 69$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma XY - \bar{X}\,\bar{Y}}{\sqrt{\left(\frac{1}{n}\Sigma X^2 - \bar{X}^2\right)\left(\frac{1}{n}\Sigma Y^2 - \bar{Y}^2\right)}}$$

$$= \frac{\frac{1}{8} \times 37560 - 68 \times 69}{\sqrt{\left[\frac{37028}{8} - (68)^2\right]\left[\frac{38132}{8} - (69)^2\right]}}$$

$$= \frac{4695 - 4692}{\sqrt{(4628{\cdot}5 - 4624)\,(4766{\cdot}5 - 4761)}} = \frac{3}{\sqrt{4{\cdot}5 \times 5{\cdot}5}} = 0{\cdot}603$$

Aliter.

(SHORT-CUT METHOD)

X	Y	$U = X - 68$	$V = Y - 69$	U^2	V^2	UV
65	67	−3	− 2	9	4	6
66	68	−2	− 1	4	1	2
67	65	−1	− 4	1	16	4
67	68	−1	− 1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
Total		0	0	36	44	24

$$\bar{U} = \frac{1}{n}\Sigma U = 0, \quad \bar{V} = \frac{1}{n}\Sigma V = 0$$

$$\text{Cov}(U, V) = \frac{1}{n}\Sigma UV - \bar{U}\,\bar{V} = \frac{1}{8} \times 24 = 3$$

$$\sigma_U^2 = \frac{1}{n}\Sigma U^2 - (\bar{U})^2 = \frac{1}{8} \times 36 = 4{\cdot}5$$

$$\sigma_V^2 = \frac{1}{n}\Sigma V^2 - (\bar{V})^2 = \frac{1}{8} \times 44 = 5{\cdot}5$$

$$\therefore \quad r(U, V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{3}{\sqrt{4{\cdot}5 \times 5{\cdot}5}} = 0{\cdot}603 = r(X, Y)$$

Remark. The reader is advised to calculate the correlation coefficient by arbitrary origin method rather than by the direct method; since the latter leads to much simpler arithmetical calculations.

Example 10·2. *A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results :*

$$n = 25, \Sigma X = 125, \Sigma X^2 = 650, \Sigma Y = 100, \Sigma Y^2 = 460, \Sigma XY = 508$$

It was, however, later discovered at the time of checking that he had copied down two pairs as

X	Y
6	14
8	6

while the correct values were

X	Y
8	12
6	8

Obtain the correct value of correlation coefficient.

[***Calcutta Univ. B.Sc. (Maths. Hons.), 1988, 1991***]

Solution.

$$\text{Corrected } \Sigma X = 125 - 6 - 8 + 8 + 6 = 125$$
$$\text{Corrected } \Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$$
$$\text{Corrected } \Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$$
$$\text{Corrected } \Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$$
$$\text{Corrected } \Sigma XY = 508 - 6\times 14 - 8\times 6 + 8\times 12 + 6\times 8 = 520$$

$$\bar{X} = \frac{1}{n}\Sigma X = \frac{1}{25}\times 125 = 5, \quad \bar{Y} = \frac{1}{n}\Sigma Y = \frac{1}{25}\times 100 = 4$$

$$\text{Cov}(X, Y) = \frac{1}{n}\Sigma XY - \bar{X}\bar{Y} = \frac{1}{25}\times 520 - 5\times 4 = \frac{4}{5}$$

$$\sigma_X^2 = \frac{1}{n}\Sigma X^2 - \bar{X}^2 = \frac{1}{25}\times 650 - (5)^2 = 1$$

$$\sigma_Y^2 = \frac{1}{n}\Sigma Y^2 - \bar{Y}^2 = \frac{1}{25}\times 436 - 16 = \frac{36}{25}$$

$$\therefore \quad \text{Corrected } r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\frac{4}{5}}{1 \times \frac{6}{5}} = \frac{2}{3} = 0{\cdot}67$$

Example 10·3. *Show that if X', Y' are the deviations of the random variables X and Y from their respective means then*

(*i*) $$r = 1 - \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2$$

(*ii*) $$r = -1 + \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y}\right)^2$$

Deduce that $-1 \le r \le +1$.

[***Delhi Univ. B.Sc. Oct. 1992; Madras Univ. B.Sc., Nov. 1991***]

Solution. (*i*) Here $X_i' = (x_i - \bar{X})$ and $Y_i' = (Y_i - \bar{Y})$

$$\text{R.H.S.} = 1 - \frac{1}{2N}\sum_i \left(\frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y}\right)^2$$

$$= 1 - \frac{1}{2N}\sum_i \left[\frac{X_i'^2}{\sigma_X^2} + \frac{Y_i'^2}{\sigma_Y^2} - \frac{2X_i'Y_i'}{\sigma_X\sigma_Y}\right]$$

$$= 1 - \frac{1}{2N}\left[\frac{1}{\sigma_X^2}\sum_i X_i'^2 + \frac{1}{\sigma_Y^2}\sum_i Y_i'^2 - \frac{2}{\sigma_X\sigma_Y}\sum_i X_i'Y_i'\right]$$

$$= 1 - \frac{1}{2N}\left[\frac{1}{\sigma_X^2}\sum_i (X_i - \bar{X})^2 + \frac{1}{\sigma_Y^2}\sum_i (Y_i - \bar{Y})^2 - \frac{2}{\sigma_X\sigma_Y}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})\right]$$

$$= 1 - \frac{1}{2}\left[\frac{1}{\sigma_X^2}\cdot\sigma_X^2 + \frac{1}{\sigma_Y^2}\cdot\sigma_Y^2 - \frac{2}{\sigma_X\sigma_Y}\cdot r\sigma_X\sigma_Y\right]$$

$$= 1 - \frac{1}{2}[1 + 1 - 2r] = r$$

(*ii*) Proceeding similarly, we will get

$$\text{R.H.S.} = -1 + \frac{1}{2}(1 + 1 + 2r) = r$$

Deduction. Since $\left(\frac{X_i'}{\sigma_X} \pm \frac{Y_i'}{\sigma_Y}\right)^2$, being the square of a real quantity is always non-negative, $\sum_i \left(\frac{X_i'}{\sigma_X} \mp \frac{Y_i'}{\sigma_Y}\right)^2$ is also non-negative. From part (*i*), we get

$$r = 1 - (\text{some non-negative quantity}) \quad \Rightarrow \quad r \le 1 \qquad \ldots(*)$$

Also from part (*ii*), we get

$$r = -1 + (\text{some non-negative quantity}) \quad \Rightarrow \quad -1 \le r \qquad \ldots(**)$$

The sign of equality in (*) and (**) holds if and only if

$$\left.\begin{aligned} \frac{X_i'}{\sigma_X} - \frac{Y_i'}{\sigma_Y} &= 0 \\ \text{and} \quad \frac{X_i'}{\sigma_X} + \frac{Y_i'}{\sigma_Y} &= 0 \end{aligned}\right\} \forall\, i = 1, 2, \ldots, n$$

respectively.

From (*) and (**), we get

$$-1 \le r \le 1$$

Example 10·4. *The variables X and Y are connected by the equation $aX + bY + c = 0$. Show that the correlation between them is –1 if the signs of a and b are alike and +1 if they are different.*

[***Nagpur Univ. B.Sc. 1992; Delhi Univ. B.Sc. (Stat. Hons.) 1992***]

Solution. $aX + bY + c = 0 \Rightarrow aE(X) + bE(Y) + c = 0$

$$\therefore \quad a\{X - E(X)\} + b\{Y - E(Y)\} = 0$$

$$\Rightarrow \quad \{X - E(X)\} = -\frac{b}{a}\{Y - E(Y)\}$$

$$\therefore \quad \text{Cov}(X, Y) = E[\{X - E(X)\}\{Y - E(Y)\}]$$

$$= -\frac{b}{a} E[\{Y - E(Y)\}^2] = -\frac{b}{a} \cdot \sigma_Y^2$$

$$E\{X - E(X)\}^2 = \frac{b^2}{a^2} E[\{Y - E(Y)\}^2] = \frac{b^2}{a^2} \cdot \sigma_Y^2$$

$$\therefore \quad r = \frac{-\frac{b}{a} \cdot \sigma_Y^2}{\sqrt{\sigma_Y^2}\sqrt{\frac{b^2}{a^2} \cdot \sigma_Y^2}} = \frac{-\frac{b}{a}\sigma_Y^2}{\left|\frac{b}{a}\right| \sigma_Y^2}$$

$$= \begin{cases} +1, \text{ if } b \text{ and } a \text{ are of opposite signs.} \\ -1, \text{ if } b \text{ and } a \text{ are of same sign.} \end{cases}$$

Example 10·5. (a) *If $Z = aX + bY$ and r is the correlation coefficient between X and Y, show that*

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X \sigma_Y$$

(b) *Show that the correlation coefficient r between two random variables X and Y is given by*

$$r = (\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2) / 2\sigma_X\sigma_Y$$

where σ_X, σ_Y and σ_{X-Y} are the standard deviations of X, Y and $X - Y$ respectively.

[***Calcutta Univ. B.Sc., 1992; M.S. Baroda Univ. B.Sc. 1992***]

Solution. Taking expectation of both sides of $Z = aX + bY$, we get

$$E(Z) = aE(X) + bE(Y)$$

$$\therefore \quad Z - E(Z) = a\{X - E(X)\} + b\{Y - E(Y)\}$$

Squaring and taking expectation of both sides, we get

$$\sigma_Z^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab \text{ Cov } (X, Y)$$

$$= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X \sigma_Y$$

(*b*) Taking $a = 1$, $b = -1$ in the above case, we have

$$Z = X - Y \text{ and } \sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2r\sigma_X \sigma_Y$$

$$\therefore \quad r = \frac{\sigma_X^2 + \sigma_Y^2 - \sigma_{X-Y}^2}{2\sigma_X \sigma_Y}$$

Remark. In the above example, we have obtained

$$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2ab \text{ Cov } (X, Y)$$

Similarly, we could obtain the result

$$V(aX - bY) = a^2 V(X) + b^2 V(Y) - 2ab \text{ Cov } (X, Y).$$

The above results are useful in solving theoretical problems.

Example 10·6. *X and Y are two random variables with variances σ_X^2 and σ_Y^2 respectively and r is the coefficient of correlation between them. If $U = X + kY$ and $V = X + \frac{\sigma_X}{\sigma_Y} Y$, find the value of k so that U and V are uncorrelated.* [***Delhi Univ. B.Sc. 1992; Andhra Univ. B.Sc. 1993***]

Solution. Taking expectations of $U = X + kY$ and $V = X + \frac{\sigma_X}{\sigma_Y} Y$, we get

$$E(U) = E(X) + kE(Y) \text{ and } E(V) = E(X) + \frac{\sigma_X}{\sigma_Y} E(Y)$$

$$U - E(U) = \{X - E(X)\} + k\{Y - E(Y)\} \text{ and}$$

$$V - E(V) = \{X - E(X)\} + \frac{\sigma_X}{\sigma_Y} \{Y - E(Y)\}$$

$$\text{Cov}(U, V) = E[\{U - E(U)\}\{V - E(V)\}]$$

$$= E\left[\{X - E(X)\} + k\{Y - E(Y)\}\right] \times \left[\{X - E(X)\} + \frac{\sigma_X}{\sigma_Y}\{Y - E(Y)\}\right]$$

$$= \sigma_X^2 + \frac{\sigma_X}{\sigma_Y} \text{Cov}(X, Y) + k \text{ Cov}(X, Y) + k \frac{\sigma_X}{\sigma_Y} . \sigma_Y^2$$

$$= \left[\sigma_X^2 + k\sigma_X\sigma_Y\right] + \left[\frac{\sigma_X}{\sigma_Y} + k\right] \text{Cov}(X, Y)$$

$$= \sigma_X(\sigma_X + k\sigma_Y) + \left[\frac{\sigma_X + k\sigma_Y}{\sigma_Y}\right] \text{Cov}(X, Y)$$

$$= (\sigma_X + k\sigma_Y)\left[\sigma_X + \frac{\text{Cov}(X, Y)}{\sigma_Y}\right] = (\sigma_X + k\sigma_Y)(1 + r)\sigma_X$$

U and V will be uncorrelated if

$$r(U, V) = 0 \Rightarrow \text{Cov}(U, V) = 0$$

i.e., if $\quad (\sigma_X + k\sigma_Y)(1 + r)\sigma_X = 0$

$\Rightarrow \quad \sigma_X + k\sigma_Y = 0 \qquad (\because \sigma_X \neq 0, r \neq -1)$

$\Rightarrow \quad k = -\frac{\sigma_X}{\sigma_Y}$

Example 10·7. *The random variables X and Y are jointly normally distributed and U and V are defined by*

$$U = X \cos\alpha + Y \sin\alpha,$$
$$V = Y \cos\alpha - X \sin\alpha$$

Show that U and V will be uncorrelated if

$$\tan 2\alpha = \frac{2r\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2},$$

where r = corr. (X, Y), σ_X^2 = Var (X) and σ_Y^2 = Var (Y). Are U and V then independent?

[***Delhi Univ. B.Sc. (Stat. Hons.) 1989; (Maths. Hons.), 1990***]

Solution. We have

$$\text{Cov}(U, V) = E[\{U - E(U)\}\{V - E(V)\}]$$

$$= E\Big[[\{X - E(X)\}\cos\alpha + \{Y - E(Y)\}\sin\alpha] \times [\{Y - E(Y)\}\cos\alpha - \{X - E(X)\}\sin\alpha]\Big]$$

$$= \cos^2\alpha \text{ Cov } (X, Y) - \sin\alpha\cos\alpha.\sigma_X^2 + \sin\alpha\cos\alpha.\sigma_Y^2 - \sin^2\alpha \text{ (Cov } (X, Y)$$

$$= (\cos^2\alpha - \sin^2\alpha) \text{ Cov } (X, Y) - \sin\alpha\cos\alpha\ (\sigma_X^2 - \sigma_Y^2)$$

$$= \cos 2\alpha \text{ Cov } (X, Y) - \sin\alpha\cos\alpha\ (\sigma_X^2 - \sigma_Y^2)$$

U and V will be uncorrelated if and only if

$$r(U, V) = 0, \quad i.e., \quad \text{iff Cov } (U, V) = 0$$

i.e., if $\cos 2\alpha \text{ Cov } (X, Y) - \sin\alpha\cos\alpha\ (\sigma_X^2 - \sigma_Y^2) = 0$

or if $$\cos 2\alpha\, r\,\sigma_X\sigma_Y = \frac{\sin 2\alpha}{2}.(\sigma_X^2 - \sigma_Y^2)$$

or if $$\tan 2\alpha = \frac{2r\,\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}$$

However, $r(U, V) = 0$ does not imply that the variables U and V are independent. [For detailed discussion, see Theorem 10·2, page 10·4.].

Example 10·8. *If X, Y are standardized random variables, and*

$$r(aX + bY, bX + aY) = \frac{1 + 2ab}{a^2 + b^2} \qquad \ldots(*)$$

find $r(X, Y)$, the coefficient of correlation between X and Y.

[***Sardar Patel Univ. B.Sc., 1993; Delhi Univ. B.Sc. (Stat. Hons.), 1989***]

Solution. Since X and Y are standardised random variables, we have

$$\left.\begin{array}{lll} & E(X) = E(Y) = 0 & \\ \text{and} & \text{Var } (X) = \text{Var } (Y) = 1 \Rightarrow & E(X^2) = E(Y^2) = 1 \\ \text{and} & \text{Cov}(X, Y) = E(XY) \quad \Rightarrow & E(XY) = r(X,Y).\sigma_X\sigma_Y \\ & & \quad = r(X,Y) \end{array}\right] \qquad \ldots(**)$$

Also we have

$$r(aX + bY, bX + aY)$$

$$= \frac{E[(aX + bY)(bX + aY)] - E(aX + bY)\,E(bX + aY)}{[\text{Var}(aX + bY).\ \text{Var}(bX + aY)]^{1/2}}$$

$$= \frac{E[abX^2 + a^2XY + b^2YX + abY^2] - 0}{\{[a^2 \text{ Var } (X) + b^2 \text{ Var } (Y) + 2ab \text{ Cov } (X,Y)] \times [b^2 \text{ Var } (X) + a^2 \text{ Var } Y + 2ba \text{ Cov } (X,Y)]\}^{1/2}}$$

$$= \frac{ab.1 + a^2 r(X, Y) + b^2 r(X, Y) + ab.1}{\{[a^2 + b^2 + 2ab\, r(X, Y)][b^2 + a^2 + 2ba\, r(X, Y)]\}^{1/2}}$$

[Using (**)]

$$= \frac{2ab + (a^2 + b^2).\, r(X, Y)}{a^2 + b^2 + 2ab.\, r(X, Y)}$$

From (*) and (**), we get

$$\frac{1 + 2ab}{a^2 + b^2} = \frac{(a^2 + b^2).\, r(X, Y) + 2ab}{a^2 + b^2 + 2ab.\, r(X, Y)}$$

Cross multiplying, we get

$$(a^2+b^2)(1+2ab)+2ab.\,r(X,Y)(1+2ab)=(a^2+b^2)^2.\,r(X,Y)+2ab\,(a^2+b^2)$$

$$\Rightarrow \quad (a^4+b^4+2a^2b^2-2ab-4a^2b^2).\,r(X,Y)=(a^2+b^2)$$

$$\Rightarrow \quad [(a^2-b^2)^2-2ab]\,r(X,Y)=a^2+b^2$$

$$\Rightarrow \quad r(X,Y)=\frac{a^2+b^2}{(a^2-b^2)^2-2ab}$$

Example 10·9. *If X and Y are uncorrelated random variables with means zero and variances σ_1^2 and σ_2^2 respectively, show that*

$$U=X\cos\alpha+Y\sin\alpha,\ V=X\sin\alpha-Y\cos\alpha$$

have a correlation coefficient ρ given by

$$\rho=\frac{\sigma_1^2-\sigma_2^2}{[(\sigma_1^2-\sigma_2^2)^2+4\sigma_1^2\sigma_2^2\,\mathrm{cosec}^2 2\alpha]^{1/2}}$$

Solution. We are given that

$$r(X,Y)=0 \Rightarrow \mathrm{Cov}(X,Y)=0,\ \sigma_X^2=\sigma_1^2 \text{ and } \sigma_Y^2=\sigma_2^2 \qquad \ldots(1)$$

We have

$$\begin{aligned}\sigma_U^2 &= V(X\cos\alpha+Y\sin\alpha)\\ &=\cos^2\alpha\,V(X)+\sin^2\alpha\,V(Y)+2\sin\alpha\cos\alpha\,\mathrm{Cov}(X,Y)\\ &=\cos^2\alpha\,\sigma_1^2+\sin^2\alpha\,\sigma_2^2 \qquad \text{[Using (1)]}\end{aligned}$$

Similarly,

$$\sigma_V^2 = V(X\sin\alpha-Y\cos\alpha)=\sin^2\alpha.\sigma_1^2+\cos^2\alpha.\sigma_2^2$$

$$\begin{aligned}\mathrm{Cov}(U,V) &= E[\{U-E(U)\}\{V-E(V)\}]\\ &=E\Big[\{(X-E(X))\cos\alpha+\{(Y-E(Y))\sin\alpha\}\\ &\qquad\qquad\times\{(X-E(X))\sin\alpha-(Y-E(Y))\cos\alpha\}\Big]\\ &=\sin\alpha\cos\alpha\,V(X)-\cos^2\alpha\,\mathrm{Cov}(X,Y)\\ &\qquad\qquad+\sin^2\alpha\,\mathrm{Cov}(X,Y)-\sin\alpha\cos\alpha\,V(Y)\\ &=(\sigma_1^2-\sigma_2^2)\sin\alpha\cos\alpha \qquad \text{[Using (1)]}\end{aligned}$$

Now
$$\rho^2=\frac{[\mathrm{Cov}(U,V)]^2}{\sigma_U^2\sigma_V^2}$$

where
$$\begin{aligned}\sigma_U^2\sigma_V^2 &= (\cos^2\alpha\,\sigma_1^2+\sin^2\alpha\,\sigma_2^2)(\sin^2\alpha\,\sigma_1^2+\cos^2\alpha\,\sigma_2^2)\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^4+\sigma_2^4)+\sigma_1^2\sigma_2^2(\cos^4\alpha+\sin^4\alpha)\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^4+\sigma_2^4)+\sigma_1^2\sigma_2^2[(\sin^2\alpha+\cos^2\alpha)^2-2\sin^2\alpha\cos^2\alpha]\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^4+\sigma_2^4-2\sigma_1^2\sigma_2^2)+\sigma_1^2\sigma_2^2\\ &=\sin^2\alpha\cos^2\alpha(\sigma_1^2-\sigma_2^2)^2+\sigma_1^2\sigma_2^2\end{aligned}$$

$$\therefore \quad \rho^2=\frac{(\sigma_1^2-\sigma_2^2)^2.\sin^2\alpha\cos^2\alpha}{\sigma_1^2\sigma_2^2+\sin^2\alpha\cos^2\alpha(\sigma_1^2-\sigma_2^2)^2}$$

$$=\frac{\frac{1}{4}(\sigma_1^2-\sigma_2^2)^2\sin^2 2\alpha}{\sigma_1^2\sigma_2^2+\sin^2 2\alpha.\,\frac{1}{4}(\sigma_1^2-\sigma_2^2)^2}$$

$$= \frac{(\sigma_1^2 - \sigma_2^2)^2}{4\sigma_1^2\sigma_2^2 \operatorname{cosec}^2 2\alpha + (\sigma_1^2 - \sigma_2^2)^2}$$

$$\Rightarrow \qquad \rho = \frac{\sigma_1^2 - \sigma_2^2}{[(\sigma_1^2 - \sigma_2^2)^2 + 4\sigma_1^2\sigma_2^2 \operatorname{cosec}^2 2\alpha]^{1/2}}$$

Example 10·10. *If $U = aX + bY$ and $V = cX + dY$, where X and Y are measured from their respective means and if r is the correlation coefficient between X and Y, and if U and V are uncorrelated, show that*

$$\sigma_U \sigma_V = (ad - bc)\, \sigma_X \sigma_Y (1 - r^2)^{1/2}$$

[***Poona Univ. B.Sc., 1990; Delhi Univ. B.Sc. (Stat. Hons.), 1986***]

Solution. We have

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad \Rightarrow \quad 1 - r^2 = 1 - \frac{[\text{Cov}(X, Y)]^2}{\sigma_X^2 \sigma_Y^2}$$

$$\Rightarrow \qquad (1 - r^2)\, \sigma_X^2 \sigma_Y^2 = \sigma_X^2 \sigma_Y^2 - [\text{Cov}(X, Y)]^2 \qquad \ldots(*)$$

[This step is suggested by the answer]

$$U = aX + bY\,,\ V = cX + dY$$

Since X, Y are measured from their means,

$$\left.\begin{aligned} & E(X) = 0 = E(Y) \Rightarrow E(U) = 0 = E(V) \\ \text{and} \quad & \sigma_U^2 = E(U^2)\,;\ \sigma_V^2 = E(V^2) \end{aligned}\right\} \qquad \ldots(**)$$

Also $\quad aX + bY - U = 0$ and $cX + dY - V = 0$

$$\Rightarrow \qquad \frac{X}{-bV + dU} = \frac{Y}{-cU + aV} = \frac{1}{ad - bc}$$

$$\Rightarrow \qquad \left.\begin{aligned} X &= \frac{1}{ad - bc}(dU - bV) \\ Y &= \frac{1}{ad - bc}(-cU + aV) \end{aligned}\right\} \qquad \ldots(***)$$

$$\therefore \qquad \text{Var}(X) = \frac{1}{(ad - bc)^2}[d^2 \sigma_U^2 + b^2 \sigma_V^2 - 2\, bd\, \text{Cov}(U, V)$$

$$= \frac{1}{(ad - bc)^2}[d^2 \sigma_U^2 + b^2 \sigma_V^2]$$

[Since U, V are uncorrelated $\Rightarrow$ Cov $(U, V) = 0$]

Similarly, we have

$$\text{Var}(Y) = \frac{1}{(ad - bc)^2}(c^2 \sigma_U^2 + a^2 \sigma_V^2)$$

$$\text{Cov}(X, Y) = E(XY) - E(X)\, E(Y) = E(XY) \qquad [\because E(X) = 0 = E(Y)]$$

$$= \frac{1}{(ad - bc)^2} E[(dU - bV)(-cU + aV)] \qquad [\text{From } (***)]$$

$$= \frac{1}{(ad-bc)^2} [-cd\,\sigma_U^2 - ab\,\sigma_V^2]$$

[Using (**) and Cov $(U, V) = 0$, given]

$$= \frac{-1}{(ad-bc)^2} [cd\,\sigma_U^2 + ab\,\sigma_V^2]$$

Substituting in (*), we get

$$(1-r^2)\,\sigma_X^2\,\sigma_Y^2 = \frac{1}{(ad-bc)^4} \times \left[(d^2\,\sigma_U^2 + b^2\,\sigma_V^2)(c^2\,\sigma_U^2 + a^2\,\sigma_V^2) - (cd\sigma_U^2 + ab\,\sigma_V^2)^2\right]$$

$$= \frac{1}{(ad-bc)^4} \times \left[c^2d^2\,\sigma_U^4 + a^2b^2\sigma_V^4 + (a^2d^2 + b^2c^2)\,\sigma_U^2\,\sigma_V^2 - c^2d^2\sigma_U^4 - a^2b^2\,\sigma_V^4 - 2abcd\,\sigma_U^2\,\sigma_V^2\right]$$

$$= \frac{1}{(ad-bc)^4} \left[a^2d^2 + b^2c^2 - 2abcd\right] \sigma_U^2\,\sigma_V^2$$

$$= \frac{1}{(ad-bc)^4} (ad-bc)^2\,\sigma_U^2\,\sigma_V^2$$

$$= \frac{\sigma_U^2\,\sigma_V^2}{(ad-bc)^2}$$

Cross multiplying and taking square root, we get the required result.

Example 10·11. (a) *Establish the formula* :

$$nr\sigma_X\sigma_Y = n_1r_1\sigma_{X_1}\sigma_{Y_1} + n_2r_2\sigma_{X_2}\,\sigma_{Y_2} + n_1dx_1dy_1 + n_2dx_2dy_2 \qquad \ldots(10\cdot5)$$

where n_1, n_2 and n are respectively the sizes of the first, second and combined sample : $(\bar{x}_1, \bar{y}_1)$, $(\bar{x}_2, \bar{y}_2)$, $(\bar{x}, \bar{y})$, their means r_1, r_2 and r their coefficients of correlation; $(\sigma_{X_1}, \sigma_{Y_1})$, $(\sigma_{X_2}, \sigma_{Y_2})$, (σ_X, σ_Y) their standard deviations, and

$$dx_1 = \bar{x}_1 - \bar{x} \quad , \quad dy_1 = \bar{y}_1 - \bar{y}$$

$$dx_2 = \bar{x}_2 - \bar{x} \quad , \quad dy_2 = \bar{y}_2 - \bar{y}$$

(b) *Find the correlation co-efficient of combined sample given that*

	Sample I	*Sample II*
Sample size	*100*	*150*
Sample mean $(\bar{x})$	*80*	*72*
Sample mean $(\bar{y})$	*100*	*118*
Sample variance (σ_X^2)	*10*	*12*
Sample variance (σ_Y^2)	*15*	*18*
Correlation coefficient	*0·6*	*0·4*

Solution. (*a*) Let (x_{1i}, y_{1i}) ; $i = 1, 2, \ldots, n_1$ and (x_{2j}, y_{2j}); $j = 1, 2, \ldots, n_2$, be the two samples of sizes n_1 and n_2 respectively from the bivariate population. Then with the given notations, we have

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}, \quad \bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2}$$

$$\left.\begin{aligned} n\sigma_X^2 &= n_1(\sigma_{X_1}^2 + dx_1^2) + n_2(\sigma_{X_2}^2 + dx_2^2) \\ n\sigma_Y^2 &= n_1(\sigma_{Y_1}^2 + dy_1^2) + n_2(\sigma_{Y_2}^2 + dy_2^2) \end{aligned}\right\} \quad \ldots(1)$$

$$r_1 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1)}{n_1\sigma_{X_1}\sigma_{Y_1}}, \quad r_2 = \frac{\sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)(y_{2j} - \bar{y}_2)}{n_2\sigma_{X_2}\sigma_{Y_2}} \quad \ldots(2)$$

For the pooled sample, we have

$$r = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x})(y_{1i} - \bar{y}) + \sum_{j=1}^{n_2}(x_{2j} - \bar{x})(y_{2j} - \bar{y})}{n\sigma_X\sigma_Y} \quad \ldots(3)$$

Now

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x})(y_{1i} - \bar{y}) = \sum_{i=1}^{n_1}\left\{\{(x_{1i} - \bar{x}_1) + (\bar{x}_1 - \bar{x})\}\{(y_{1i} - \bar{y}_1) + (\bar{y}_1 - \bar{y})\}\right\}$$

$$= \sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1) + (\bar{y}_1 - \bar{y})\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)$$

$$+ (\bar{x}_1 - \bar{x})\sum_{i=1}^{n_1}(y_{1i} - \bar{y}_1) + n_1(\bar{x}_1 - \bar{x})(\bar{y}_1 - \bar{y})$$

But $\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1) = 0$ and $\sum_{i=1}^{n_1}(y_{1i} - y_1) = 0$,

being the algebraic sum of the deviations from the mean.

$$\therefore \quad \sum_{i=1}^{n_1}(x_{1i} - \bar{x})(y_{1i} - \bar{y}) = n_1 r_1 \sigma_{X_1}\sigma_{Y_1} + n_1 dx_1 dy_1 \qquad \text{[Using (2)]}$$

Similarly, we will get

$$\sum_{j=1}^{n_2}(x_{2j} - \bar{x})(y_{2j} - \bar{y}) = n_2 r_2 \sigma_{X_2}\sigma_{Y_2} + n_2 dx_2 dy_2$$

Substituting in (3), we get the required formula.

(*b*) Here we are given :

$$n_1 = 100,\ \bar{x}_1 = 80,\ \bar{y}_1 = 100,\ \sigma_{X_1}^2 = 10,\ \sigma_{Y_1}^2 = 15,\ r_1 = 0{\cdot}6$$

$$n_2 = 150,\ \bar{x}_2 = 72,\ \bar{y}_2 = 118,\ \sigma_{X_2}^2 = 12,\ \sigma_{Y_2}^2 = 18,\ r_2 = 0{\cdot}4$$

$$\therefore \quad \bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{100 \times 80 + 150 \times 72}{100 + 150} = 75{\cdot}2$$

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2}{n_1 + n_2} = \frac{100 \times 100 + 150 \times 118}{100 + 150} = 110{\cdot}8$$

$$dx_1 = \bar{x}_1 - \bar{x} = 4{\cdot}8, \; dy_1 = \bar{y}_1 - \bar{y} = 10{\cdot}8$$

$$dx_2 = \bar{x}_2 - \bar{x} = 3{\cdot}2, \; dy_2 = \bar{y}_2 - \bar{y} = 7{\cdot}2$$

$$n\sigma_X^2 = n_1(\sigma_{X_1}^2 + dx_1^2) + n_2(\sigma_{X_2}^2 + dx_2^2) = 6640$$

$$n\sigma_Y^2 = n_1(\sigma_{Y_1}^2 + dy_1^2) + n_2(\sigma_{Y_2}^2 + dy_2^2) = 23640$$

Substituting these values in the formula and simplifying, we get

$$r = \frac{n_1 r_1 \sigma_{X_1}\sigma_{Y_1} + n_2 r_2 \sigma_{X_2}\sigma_{Y_2} + n_1 dx_1 dy_1 + n_2 dx_2 dy_2}{n\sigma_X\sigma_Y} = 0{\cdot}8186$$

Example 10·12. *The independent variables X and Y are defined by :*

$$f(x) = 4ax, \; 0 \le x \le r \qquad\qquad f(y) = 4by, \; 0 \le y \le s$$
$$= 0, \; \text{otherwise} \qquad\qquad = 0, \; \text{otherwise}$$

Show that :

$$Cov\,(U, V) = \frac{b-a}{b+a},$$

where $U = X + Y$ and $V = X - Y$

[I.I.T. (B. Tech.), Nov. 1992]

Solution. Since the total area under probability curve is unity (one), we have :

$$\int_0^r f(x)dx = 4a\int_0^r x dx = 1 \quad \Rightarrow \quad 2ar^2 = 1 \;\Rightarrow\; a = \frac{1}{2r^2} \qquad \ldots(i)$$

$$\int_0^r f(y)dy = 4b\int_0^s y dy = 1 \quad \Rightarrow \quad 2bs^2 = 1 \;\Rightarrow\; b = \frac{1}{2s^2} \qquad \ldots(ii)$$

$$\therefore\; f(x) = 4ax = \frac{2x}{r^2}, \; 0 \le x \le r; \text{ and } f(y) = 4by = \frac{2y}{s^2}, \; 0 \le y \le s \qquad \ldots(iii)$$

Since X and Y are independent variates,

$$r(X, Y) = 0 \;\Rightarrow\; \text{Cov}\,(X, Y) = 0 \qquad \ldots(iv)$$

$$\begin{aligned}
\text{Cov}\,(U, V) &= \text{Cov}\,(X + Y, X - Y) \\
&= \text{Cov}\,(X, X) - \text{Cov}\,(X, Y) + \text{Cov}\,(Y, X) - \text{Cov}\,(Y, Y) \\
&= \sigma_X^2 - \sigma_Y^2 && \text{[Using (iv)]} \\
\text{Var}\,(U) &= \text{Var}\,(X + Y) = \text{Var}\,(X) + \text{Var}\,(Y) + 2\,\text{Cov}\,(X, Y) \\
&= \sigma_X^2 + \sigma_Y^2 && \text{[Using (iv)]} \\
\text{Var}\,(Y) &= \text{Var}\,(X - Y) = \text{Var}\,(X) + \text{Var}\,(Y) - 2\,\text{Cov}\,(X, Y) \\
&= \sigma_X^2 + \sigma_Y^2 && \text{[Using (iv)]}
\end{aligned}$$

$$\therefore \quad r(U,V) = \frac{\text{Cov}(U, V)}{\sigma_U \sigma_V} = \frac{\sigma_X^2 - \sigma_Y^2}{\sigma_X^2 + \sigma_Y^2} \qquad \ldots (v)$$

We have :

$$E(X) = \int_0^r x f(x)\, dx = \frac{2}{r^2}\int_0^r x^2\, dx = \frac{2r}{3} \qquad \text{[From } (iii)\text{]}$$

$$E(X^2) = \int_0^r x^2 f(x)\, dx = \frac{2}{r^2}\int_0^r x^3 dx = \frac{r^2}{2}$$

$$\therefore \quad \text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{r^2}{2} - \frac{4r^2}{9} = \frac{r^2}{18} = \frac{1}{36a} \qquad \text{[From } (i)\text{]}$$

Similarly, we shall get

$$E(Y) = \frac{2s}{3},\ E(Y^2) = \frac{s^2}{2} \text{ and Var } (Y) = \frac{s^2}{18} = \frac{1}{36b}$$

Substituting in (v), we get

$$r(U, V) = \frac{1/(36a) - 1/(36b)}{1/(36a) + 1/(36b)} = \frac{b-a}{b+a}.$$

Example 10·13. *Let the random variable X have the marginal density*

$$f_1(x) = 1, -\frac{1}{2} < x < \frac{1}{2}$$

and let the conditional density of Y be

$$\left.\begin{aligned} f(y \mid x) &= 1,\ x < y < x+1,\ -\frac{1}{2} < x < 0 \\ &= 1,\ -x < y < 1-x,\ 0 < x < \frac{1}{2} \end{aligned}\right\} \qquad (*)$$

Show that the variables X and Y are uncorrelated.

Solution. We have

$$E(X) = \int_{-\frac{1}{2}}^{\frac{1}{2}} x f_1(x)\, dx = \int_{-\frac{1}{2}}^{\frac{1}{2}} x.1.dx = \left[\frac{x^2}{2}\right]_{-1/2}^{1/2} = 0$$

If $f(x, y)$ is the joint p.d.f. of X and Y, then

$$f(x, y) = f(y \mid x) f_1(x) = f(y \mid x). \qquad (**) \qquad [\because f_1(x) = 1]$$

$$E(XY) = \int_{-\frac{1}{2}}^{0}\int_{x}^{x+1} xy f(x, y)\, dx\, dy + \int_{0}^{\frac{1}{2}}\int_{-x}^{1-x} xy f(x, y)\, dx\, dy$$

$$= \int_{-\frac{1}{2}}^{0}\left[x\int_{x}^{x+1} y\, dy\right] dx + \int_{0}^{\frac{1}{2}}\left[x\int_{-x}^{1-x} y\, dy\right] dx \text{ [From } (*) \text{ and}$$

$$= \frac{1}{2}\int_{-\frac{1}{2}}^{0} x(2x+1)dx + \frac{1}{2}\int_{0}^{\frac{1}{2}} x\,(1-2x)\,dx$$

$$= \frac{1}{2}\left[\frac{2}{3}x^3 + \frac{x^2}{2}\right]_{-\frac{1}{2}}^{0} + \frac{1}{2}\left[\frac{x^2}{2} - \frac{2}{3}x^3\right]_{0}^{\frac{1}{2}}$$

$$= \frac{1}{2}\left[\frac{1}{12} - \frac{1}{8} - \frac{1}{12} + \frac{1}{8}\right] = 0$$

$\therefore \quad \text{Cov}\,(XY) = E(XY) - E(X)\,E(Y) = 0 \quad \Rightarrow \quad r(X, Y) = 0$

Hence the variables X and Y are uncorrelated.

EXERCISE 10(a)

1. (*a*) Show that the co-efficient of correlation r is independent of a change of scale and origin of the variables. Also prove that for two independent variables $r = 0$. Show by an example that the converse is not true. State the limits between which r lies and give its proof.

[***Delhi Univ. M.Sc. (O.R.), 1986***]

(*b*) Let ρ be the correlation coefficient between two jointly distributed random variables X and Y. Show that $|\rho| \leq 1$ and that $|\rho| = 1$ if and only if X and Y are linearly related. [***Indan Forest Service, 1991***]

2. (*a*) Calculate the coefficient of correlation between X and Y for the following :

X...	1	3	4	5	7	8	10
Y...	2	6	8	10	14	16	20

Ans. $r(X, Y) = +1$

(*b*) Discuss the statistical validity of the following statements :

(*i*) "High positive coefficient of correlation between increase in the sale of newspapers and increase in the number of crimes leads to the conclusion that newspaper reading may be responsible for the increase in the number of crimes."

(*ii*) "A high positive value of r between the increase in cigarette smoking and increase in lung cancer establishes that cigarette smoking is responsible for lung cancer."

(*c*) (*i*) Do you agree with the statement that "$r = 0{\cdot}8$ implies that 80% of the data are explained."

(*ii*) Comment on the following :

"The closeness of relationship between two variables is proportional to r".

Hint. (*a*) No (*b*) Wrong.

(*d*) By effecting suitable change of origin and scale, compute the product moment correlation coefficient for the following set of 5 observations on (X, Y) :

X :	−10	−5	0	5	10
Y :	5	9	7	11	13

Ans. $r(X, Y) = 0{\cdot}34$

3. The marks obtained by 10 students in Mathematics and Statistics are given below. Find the coefficient of correlation between the two subjects.

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks in Mathematics :	75	30	60	80	53	35	15	40	38	48
Marks in Statistics :	85	45	54	91	58	63	35	43	45	44

4. (*a*) The following table gives the number of blind per lakh of population in different age-groups. Find out the correlation between age and blindness.

Age in years :	0—10	10—20	20—30	30—40	40—50
Number of blind per lakh :	55	67	100	111	150
Age in year :	50—60	60—70	70—80		
Number of blind per lakh :	200	300	500		

Ans. 0·89

(*b*) The following table gives the distribution of items of production and also the relatively defective items among them, according to size-groups. Is there ay correlation between size and defect in quality ?

Size-Group :	15—16	16—17	17—18	18—19	19—20	20—21
No. of Items :	200	270	340	360	400	300
No. of defective items	150	162	170	180	180	120

Hint. Here we have to find the correlation coefficient between the size-group (X) and the percentage of defectives (Y) given below.

X	15·5	16·5	17·5	18·5	19·5	20·5
Y	75	60	50	50	45	40

Ans. $r = 0{\cdot}94$.

5. Using the formula

$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\, r(X, Y)\, \sigma_X\, \sigma_Y$$

obtain the correlation coefficient between the heights of fathers (X) and of the sons (Y) from the following data :

X :	65	66	67	68	69	70	71	67
Y :	67	68	64	72	70	67	70	68

6. (*a*) From the following data, compute the co-efficient of correlation between X and Y.

	X series	*Y series*
No. of items	15	15
Arithmetic mean	25	18
Sum of squares of deviations from mean	136	138

Summation of product of deviations of X and Y series from the respective arithmetic means = 122.

Ans. $r(X, Y) = 0{\cdot}891$

(*b*) Coefficient of correlation between two variables X and Y is 0·32. Their covariance is 7·86. The variance of X is 10. Find the standard deviation of Y series.

(*c*) In two sets of variables X and Y with 50 observations each, the following data were observed :

$$\bar{X} = 10, \sigma_X = 3, \bar{Y} = 6, \sigma_Y = 2 \text{ and } r(X, Y) = 0{\cdot}3$$

But on subsequent verification it was found that one value of X (= 10) and one value of Y (= 6) were inaccurate and hence weeded out. With the remaining 49 pairs of values, how is the original value of r affected ?

(Nagpur Univ. B.Sc., 1990)

Hint. $\Sigma X = n\bar{X} = 500, \Sigma Y = n\bar{Y} = 300$

$$\Sigma X^2 = n(\sigma_X^2 + \bar{X}^2) = 5450, \Sigma Y^2 = 50(4 + 36) = 2000$$

$$r\,\sigma_X \sigma_Y = \text{Cov}(X, Y) = \frac{\Sigma XY}{n} - \bar{X}\,\bar{Y}$$

$$\Rightarrow \quad 0{\cdot}3 \times 3 \times 2 = \frac{\Sigma XY}{50} - 10 \times 6$$

$$\Rightarrow \quad \Sigma XY = 50(1{\cdot}8 + 60) = 3090$$

After weeding out the incorrect pair of observation, *viz.*, $(X = 10, Y = 6)$, the corrected values of $\Sigma X, \Sigma Y, \Sigma X^2, \Sigma Y^2$ and ΣXY for the remaining 50 – 1 = 49 pairs of observations are given below :

Corrected Values :

$\Sigma X = 500 - 10 = 490$; $\Sigma Y = 300 - 6 = 294$

$\Sigma XY = 3090 - 10 \times 6 = 3090 - 60 = 3030$

$\Sigma X^2 = 5450 - 10^2 = 5350, \Sigma Y^2 = 2000 - 6^2 = 1964$

$$\therefore r = \frac{\text{Corrected Cov}(X, Y)}{(\text{Corrected } \sigma_X) \times (\text{Corrected}) \sigma_Y} = \frac{90/49}{\sqrt{\frac{450}{49} \times \frac{200}{49}}} = 0{\cdot}3$$

Hence the correlation coefficient is invariant in this case.

(*d*) A prognostic test in Mathematics was given to 10 students who were about to begin a course in Statistics. The scrores (X) in their test were examined in relations to scores (Y) in the final examination in Statistics. The following results were obtained :—

$$\Sigma X = 71, \Sigma Y = 70, \Sigma X^2 = 555, \Sigma Y^2 = 526 \text{ and } \Sigma XY = 527$$

Find the coefficient of correlation between X and Y.

(Kerala Univ. B.Sc., 1990)

7. (*a*) X_1 and X_2 are independent variables with means 5 and 10 and standard deviations 2 and 3 respectively. Obtain $r(U, V)$ where

$$U = 3X_1 + 4X_2 \text{ and } V = 3X_1 - X_2$$

Ans. 0 ***(Delhi Univ. B.Sc. ,1988)***

(*b*) If X and Y are normal and independent with zero means and standard deviations 9 and 12 respectively, and if $X + 2Y$ and $kX - Y$ are non-correlated, find k.

(*c*) X, Y, Z are random variables each with expectation 10 and variances 1, 4 and 9 respectively. The correlation coefficients are

$$r(X, Y) = 0, r(Y, Z) = r(X, Y) = 1/4$$

Obtain the numerical values of :

(*i*) $E(X + Y - 2Z)$, (*ii*) Cov $(X + 3, Y + 3)$, (*iii*) $V(X - 2Z)$ and (*iv*) Cov $(3X, 5Z)$

Ans. (*i*) = 0, (*ii*) 0, (*iii*) 34, and (*iv*) 45/4.

(*d*) X and Y are discrete random variables. If Var (X) = Var$(Y) = \sigma^2$, Cov $(X, Y) = \frac{\sigma^2}{2}$, find (*i*) Var $(2X - 3Y)$, (*ii*) Corr $(2X + 3, 2Y - 3)$.

8. (*a*) Prove that :

$$V(aX \pm bY) = a^2V(X) + b^2V(Y) \pm 2ab \text{ Cov } (X, Y)$$

Hence deduce that if X and Y are independent

$$V(X \pm Y) = V(X) + V(Y)$$

(*b*) Prove that correlation coefficient between X and Y is positive or negative according as

$$\sigma_{X+Y} > \text{ or } < \sigma_{X-Y}$$

9. Show that if X and Y are two random variables each assuming only two values and the correlation co-efficient between them is zero, then they are independent. Indicate with justification whether the result is true in general.

Find the correlation coefficient between X and $a - X$, where X is any random variable and a is constant.

10. (*a*) X_i (i =1, 2, 3) are uncorrelated variables each having the same standard deviation. Obtain the correlation between $X_1 + X_2$ and $X_2 + X_3$.

Ans. 1/2

(*b*) If X_i (i =1, 2, 3) are three uncorrelated variables having standard deviations σ_1, σ_2 and σ_3 respectively, obtain the coefficient of correlation between $(X_1 + X_2)$ and $(X_2 + X_3)$.

Ans. $\sigma_2^2/\sqrt{(\sigma_1^2 + \sigma_2^2)(\sigma_2^2 + \sigma_3^2)}$

(*c*) Two random variables X and Y have zero means, the same variance σ^2 and zero correlation. Show that

$U = X \cos \alpha + Y \sin \alpha$ and $V = X \sin \alpha - Y \cos \alpha$

have the same variance σ^2 and zero correlation.

(Bangalore Univ. B.Sc., 1991)

(*d*) Let X and Y be uncorrelated random variables. If $U = X + Y$ and $V = X - Y$, prove that the coefficient of correlation between U and V is $(\sigma_X^2 - \sigma_Y^2)/(\sigma_X^2 + \sigma_Y^2)$, where σ_X^2 and σ_Y^2 are variances of X and Y respectively.

(*e*) Two independent random variables X and Y have the following variances : $\sigma_X^2 = 36$, $\sigma_Y^2 = 16$. Calculate the coefficient of correlation between

$$U = X + Y \text{ and } V = X - Y$$

(*f*) Random variables X and Y have zero means and non-zero variances σ_X^2 and σ_Y^2. If $Z = Y - X$, then find σ_Z and the correlation coefficient $\rho(X, Z)$ of X and Z in terms of σ_X, σ_Y and the correlation coefficient $\rho(X, Y)$ of X and Y.

(*g*) If the independant random variables X_1, X_2 and X_3 have the means 4, 9 and 3 and variances 3, 7, 5, respectively, obtain the mean and variance of

(*i*) $Y = 2X_1 - 3X_2 + 4X_3$, (*ii*) $Z = X_1 + 2X_2 - X_3$, and

(*iii*) Calculate the correlation between Y and Z.

[***Delhi Univ. M.A.(Eco.), 1989***]

11. (*a*) $X_1, X_2, \ldots, X_n$ are uncorrelated random variables, all with the same distribution and zero means. Let $\bar{X} = \Sigma X_i/n$

Find the correlation co-efficient between (*i*) X_i and $\bar{X}$ and (*ii*) $X_i - \bar{X}$ and $\bar{X}$.

[***Delhi Univ. B.Sc. (Stat. Hons.), 1993***]

Hint. $$r(X_i, \bar{X}) = \frac{\sigma^2/n}{\sqrt{\sigma^2 . \sigma^2/n}} = \frac{1}{\sqrt{n}}$$

$$\text{Cov}(X_i - \bar{X}, \bar{X}) = \text{Cov}(X_i, \bar{X}) - \text{Var}(\bar{X})$$
$$= (\sigma^2/n) - (\sigma^2/n) = 0$$

$$\therefore \quad r(X_i - \bar{X}, \bar{X}) = 0$$

(*b*) $X_1, X_2, \ldots \ldots, X_n$ are random variables each with the same expected value μ and s.d. σ. The correlation coefficient between any two X's is ρ. Show that (*i*) $\text{Var}(\bar{X}) = \frac{\sigma^2}{n} + \left(1 - \frac{1}{n}\right)\rho\sigma^2$,

(*ii*) $E \sum_1^n (X_i - \bar{X})^2 = (n-1)(1-\rho)\sigma^2$, and (*iii*) $\rho > -\frac{1}{n-1}$

12. (*a*) If X and Y are independent random variables, show that

$$r(X + Y, X - Y) = r^2(X, X + Y) - r^2(Y, X + Y),$$

where $r(X + Y, X - Y)$ denotes the co-efficient of correlation between $(X + Y)$ and $(X - Y)$. (***Meerut Univ. B.Sc., 1991***)

(*b*) Let X and Y be random variables having mean 0, variance 1 and correlation r. Show that $X - rY$ and Y are uncorrelated and that $X - rY$ has mean zero and variance $1 - r^2$.

13. X_1 and X_2 are two variables with zero means, variances σ_1^2 and σ_2^2 respectively and r is the correlation coefficient between them. Determine the values of the constants a and b which are independent of r such that $X_1 + aX_2$ and $X_1 + bX_2$ are uncorrelated.

14. (*a*) If X_1 and X_2 are two random variables with means μ_1 and μ_2, variances σ_1^2, σ_2^2 and correlation coefficient r, find the correlation co-efficient between

$$U = a_1X_1 + a_2X_2 \text{ and } V = b_1X_1 + b_2X_2,$$

where a_1, a_2 and b_1, b_2 are constants.

(*b*) Let X_1, X_2 be independent random variables with means μ_1, μ_2 and non-zero variances σ_1^2, σ_2^2 respectively. Let $U = X_1 - X_2$ and $V = X_1 X_2$. Find the correlation coefficient between (*i*) X_1 and U, (*ii*) X_1 and V, in terms of μ_1, μ_2, σ_1^2, σ_2^2.

15. (*a*) If $U = aX + bY$ and $V = bX - aY$, where X and Y are measured from their respective means and if U and V are uncorrelated, r the co-efficient of correlation between X and Y is given by the equation.

$$\sigma_U \sigma_V = (a^2 + b^2)\, \sigma_X \sigma_Y (1 - r^2)^{1/2}$$ **(Utkal Univ. B. Sc., 1993)**

(*b*) Let $U = aX + bY$ and $V = aX - bY$ where X, Y represent deviations from the means of two measurements on the same individual. The coefficient of correlation between X and Y is ρ. If U, V are uncorrelated, show that

$$\sigma_U \sigma_V = 2ab\sigma_X \sigma_Y (1 - r^2)^{1/2}$$

16. Show that, if a and b are constants and r is the correlation coefficient between X and Y, then the correlation coefficient between aX and bY is equal to r if the signs of a and b are alike, and to $-r$ if they are different.

Also show that, if constants a, b and c are positive, the correlation coefficient between $(aX + bY)$ and cY is equal to

$$(ar\sigma_X + b\sigma_Y) / \sqrt{(a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abr\sigma_X\sigma_Y)}$$

17. If X_1, X_2 and X_3 are three random variables measured from their respective means as origin and of equal variances, find the coefficient of correlation between $X_1 + X_2$ and $X_2 + X_3$ in terms of r_{12}, r_{13} and r_{23} and show that it is equal to

$$(i)\ \frac{r_{12} + 1}{2}, \text{ if } r_{13} = r_{23} = 0, \text{ and } (ii) = \frac{r_{12} + 3}{4}, \text{ if } r_{13} = r_{23} = 1$$

18. (*a*) For a weighted distribution (x_i, w_i), $(i = 1, 2, \ldots, n)$ show that the weighted arithmetic mean $\bar{x}_w = \Sigma\, w_i x_i / \Sigma\, w_i$ > or < the unweighted mean $\bar{x} = \Sigma\, x_i / n$ according as $r_{xw} >$ or < 0.

(*b*) Given N values $x_1, x_2, \ldots, x_N$ of variable X and weights $w_1, w_2, \ldots, w_N$, express the coefficient of correlation between X and W in terms involving the difference between the arithmetic mean and the weighted mean of X.

19. (*a*) A coin is tossed n times. If X and Y denote the (random) number of heads and number of tails turned up respectively, show that $r(X, Y) = -1$.

Hint. Note that $X + Y = n \Rightarrow Y = n - X$

$$\therefore \quad r(X, Y) = r(X, n - X) = r(X, -X) = -r(X, X) = -1.$$

(*b*) Two dice are thrown, their scores being a and b. The first die is left on the table while the second is picked up and thrown again giving the score c. Suppose the process is repeated a large number of times. What is the correlation coefficient between $X = a + b$ and $Y = a + c$?

Ans. $r(X, Y) = \dfrac{1}{2}$

20. (*a*) If X and Y are independent random variables with means μ_1 and μ_2 and variances σ_1^2, σ_2^2 respectively, show that the correlation coefficient between $U = X$ and $V = X - Y$ in terms of μ_1, μ_2, σ_1^2 and σ_2^2 is $\sigma_1 / \sqrt{\sigma_1^2 + \sigma_2^2}$.

(*b*) If X and Y are independent random variables with non-zero variances, show that the correlation coefficient between $U = XY$ and $V = X$ in terms of mean and variance of X and Y is given by

$$\mu_2\sigma_1 / \sqrt{\sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2}$$

[*Delhi Univ. B.Sc. (Stat Hons.), 1987*]

21. If X_i, Y_j and Z_k are all independent random variables with mean zero and unit variance, find the correlation coefficient between

$$U = \sum_{i=1}^{m} X_i + \sum_{j=1}^{n} Y_j \text{ and } V = \sum_{i=1}^{m} X_i + \sum_{k=1}^{n} Z_k$$

Ans. $r(U, V) = m/(m + n)$ (*Bombay Univ., B.Sc, 1990*)

22. (*a*) Find the value of l so that the correlation coefficient between $(X - lY)$ and $(X + Y)$ is maximum, where X, Y are independent random variables each with mean zero and variance 1. [**Ans.** $l = -1$]

Hint. $U = X - lY$; $V = X + Y$. Now find l so that $r(U, V) = 1$.

(*b*) If $U = X + kY$ and $V = X + mY$ and r is the correlation coefficient between X and Y, find the correlation coefficient between U and V. Show that U and V are uncorrelated if $k = \dfrac{-\sigma_X(\sigma_X + rm\,\sigma_Y)}{\sigma_Y(r\,\sigma_X + m\sigma_Y)}$

and further if $m = \dfrac{\sigma_X}{\sigma_Y}$, then $k = -\dfrac{\sigma_X}{\sigma_Y}$. (*Gujarat Univ. M.A., 1993*)

23. X_1, X_2, X_3 are three variables, each with variance σ^2 and the correlation coefficient between any two of them is r. If $\bar{X} = (X_1 + X_2 + X_3)/3$, show that

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{3}(1 + 2r)$$

Deduce that $r \geq -1/2$.

24. (*a*) If $U = aX + bY$ and $V = bX - aY$, show that U and V are uncorrelated if $\dfrac{ab}{a^2 - b^2} = \dfrac{\rho\sigma_X\sigma_Y}{\sigma_X^2 - \sigma_Y^2}$

where ρ is the coefficient of correlation between X and Y. Show further that, in this case

$$\sigma_U^2 + \sigma_V^2 = (a^2 + b^2)(\sigma_X^2 + \sigma_Y^2) \text{ and } \sigma_U\sigma_V = (a^2 + b^2)\,\sigma_X\sigma_Y\sqrt{1 - \rho^2}$$

(*b*) If $u = aX + bY$, $v = cX + dY$, show that

$$\begin{vmatrix} \text{var}(u) & \text{cov}(u,v) \\ \text{cov}(u,v) & \text{var}(v) \end{vmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix}^2 \begin{vmatrix} \text{var}(X) & \text{cov}(X, Y) \\ \text{cov}(X, Y) & \text{var}(Y) \end{vmatrix}$$

25. If X is a standard normal variate and $Y = a + bX + cX^2$, where a, b, c are constants, find the correlation coefficient between X and Y. Hence or otherwise obtain the conditions when (*i*) X and Y are uncorrelated and (*ii*) X and Y are perfectly correlated.

26. (*a*) If $X \sim N(0, 1)$, find corr (X, Y) where $Y = a + bX + cX^2$.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1985*]

Ans. $r(X, Y) = \dfrac{b}{\sqrt{b^2 + 2c^2}}$

(*b*) If X has Laplace distribution with parameters $(\lambda, 0)$ and $Y = a + bX + cX^2$, find $\rho(X, Y)$

[*Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1989*]

Hint. $p(x) = \frac{1}{2}\lambda \exp[-\lambda |x|], \ -\infty < x < \infty.$

$E(X^{2k+1}) = 0 = \mu_{2k+1}$; $E(X^{2k}) = \mu_{2k} = (2k)! / \lambda^{2k}$

$$\rho_{XY} = \frac{\lambda b}{\sqrt{b^2\lambda^2 + 10c^2}}$$

27. In a sample of n random observations from exponential distribution with parameter λ, the number of observations in $(0, 1/\lambda)$ and $(1/\lambda, 2/\lambda)$, denoted by X and Y are noted. Find $\rho(X, Y)$.

Hint. $p_1 = p(0 < X < 1/\lambda) = \displaystyle\int_0^{1/\lambda} \lambda e^{-\lambda x} dx = \frac{e-1}{e}$

$$p_2 = p(1/\lambda < Y < 2/\lambda) = \int_{1/\lambda}^{2/\lambda} \lambda e^{-\lambda y} dy = \frac{e-1}{e^2}$$

Then (X, Y) has a trinomial distribution with parameters $(n = 3, p_1, p_2, p_3 = 1 - p_1 - p_2)$.

Hence we have

$$\rho(X, Y) = -\left[\frac{p_1 p_2}{(1-p_1)(1-p_2)}\right]^{1/2} = -\frac{e-1}{\sqrt{e^2 - e + 1}}.$$

28. Prove that :

$$r(X, Y+Z) = \frac{\sigma_Y}{\sigma_{Y+Z}} \cdot r(X, Y) + \frac{\sigma_Z}{\sigma_{Y+Z}} \cdot r(X, Z)$$

29. If X and Y are independent random variables, find $\text{Corr}(X, XY)$. Deduce the value of $\text{Corr}(X, X/Y)$.

Ans. $r(X, XY) = \sigma_X \mu_Y / [\sigma_X^2 \sigma_Y^2 + \mu_X^2 \sigma_Y^2 + \mu_Y^2 \sigma_X^2]^{1/2}$

30. Prove or Disprove :

(*a*) $r(X, Y) = 0 \quad \Rightarrow \quad r(|X|, Y) = 0$

(*b*) $r(X, Y) = 0, r(Y, Z) = 0 \Rightarrow r(X, Z) = 0.$

Ans. (*a*) False, unless X and Y are independent.

(*b*) **Hint.** Let $Z \equiv X$, and X and Y be independent. Then $r(X, Y) = 0 = r(Y, Z)$. But $r(X, Z) = r(X, X) = 1$.

31. Let random variable X have a p.d.f. $f(.)$ with distribution function $F(.)$, mean μ and variance σ^2. Define $Y = \alpha + \beta X$, where α and β are constants satisfying $-\infty < \alpha < \infty$, and $\beta > 0$.

(*a*) Select α and β so that Y has mean 0 and variance 1.

(*b*) What is the correlation coefficient between X and Y ?

32. Let (X, Y) be jointly discrete random variables such that each X and Y have at most two mass points. Prove or disprove : X and Y are independent if and only if they are uncorrelated.

Ans. True.

33. If the variables $X_1, X_2, \ldots, X_{2n}$ all have the same variance σ^2 and the correlation coefficient between X_i and X_j $(i \neq j)$ has the same value, show that the correlation between $\sum_{i=1}^{n} X_i$ and $\sum_{j=n+1}^{2n} X_j$ is given by $[n\rho/\{1 + (n-1)\rho\}]$.

34. The means of independent r.v's $X_1, X_2, \ldots, X_n$ are zero and variances are equal, say unity. The correlation coefficients between the sum of selected t $(< n)$ variables out of these variables and the sum of all n variables are found out. Prove that the sum of squares of all these correlation coefficients is ${}^{n-1}C_{t-1}$.

[*Burdwan Univ. B.Sc. (Hons.), 1989*]

35. Two variables U and V are made up of the sum of a number of terms as follows :

$$U = X_1 + X_2 + \ldots + X_n + Y_1 + Y_2 + \ldots + Y_a,$$
$$V = X_1 + X_2 + \ldots + X_n + Z_1 + Z_2 + \ldots + Z_b,$$

where a and b are all suffixes and where X's, Y's and Z's are all uncorrelated standardised random variables. Show that the correlation coefficient between U and V is $\dfrac{n}{\sqrt{(n+a)(n+b)}}$. Show further that

$$\left.\begin{aligned} \xi &= \sqrt{(n+b)}\,U + \sqrt{(n+a)}\,V \\ \text{and} \quad \eta &= \sqrt{(n+b)}\,U - \sqrt{(n+a)}\,V \end{aligned}\right\} \quad \ldots(*)$$

are uncorrelated [*South Gujarat Univ. B.Sc., 1989*]

36. (*a*) Let the random variables X and Y have the joint p.d.f.

$$f(x, y) = 1/3\ ;\ (x, y) = (0, 0), (1, 1)\ (2, 0)$$

Compute $E(X)$, $V(X)$, $E(Y)$, $V(Y)$ and $r(X, Y)$. Are X and Y stochastically independent ? Give reasons.

(*b*) Let (X, Y) have the probability distribution :

$$f(0, 0) = 0{\cdot}45, f(0, 1) = 0{\cdot}05, f(1, 0) = 0{\cdot}35, f(1, 1) = 0{\cdot}15.$$

Evaluate $V(X)$, $V(Y)$ and $\rho(X, Y)$.

Show that while X and Y are correlated, X and $X-5Y$ are uncorrelated. Are X and $X - 5Y$ independent ?

(*c*) Given the bivariate probability distribution :

$$\begin{array}{lll} f(-1, 0) = 1/15, & f(-1, 1) = 3/15, & f(-1, 2) = 2/15 \\ f(0, 0) = 2/15, & f(0, 1) = 2/15, & f(0, 2) = 1/15 \\ f(1, 0) = 1/15, & f(1, 1) = 1/15, & f(1, 2) = 2/15 \end{array}$$
$$f(x, y) = 0, \text{ elsewhere.}$$

Obtain :

(*i*) The marginal distributions of X and Y.

(*ii*) The conditional distributions of Y given $X = 0$.

(*iii*) $E(Y|X = 0)$.

(*iv*) The product moment correlation coefficient between X and Y.

Are X and Y independently distributed ?

37. If X and Y are standardised variates with correlation coefficient ρ, prove that $$E\,[\max(X^2, Y^2)] \le 1 + \sqrt{1 - \rho^2}$$

Hint. $$\max(X^2, Y^2) = \frac{1}{2}|X^2 - Y^2| + \frac{1}{2}(X^2 + Y^2) \qquad \ldots(*)$$

$$E(X) = E(Y) = 0;\ E(X^2) = E(Y^2) = 1;\ E(XY) = \rho$$

and $$[E\,|X - Y|\,.\,|X + Y|]^2 \le E\,(X - Y)^2\,.\,E\,(X + Y)^2$$

(By Cauchy-Schwartz Inequality)

38. The joint p.d.f. of two variates X and Y is given by

$$f(x, y) = k[(x + y) - (x^2 + y^2)]\ ;\ 0 < (x, y) < 1$$
$$= 0, \text{ otherwise.}$$

Show that X and Y are uncorrelated but not independent.

39(*a*). If the random variables X and Y have the joint p.d.f.,

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

then show that the correlation coefficient between X and Y is $-\frac{1}{11}$.

[***Madras Univ. B.Sc., Oct., 1990***]

(*b*) The density function f of a random variable X is given by

$$f(x) = \begin{cases} kx^2, \text{ if } -1 \le x \le 1 \\ 0, \quad \text{otherwise} \end{cases}$$

(*i*) What is the value of k ? What is the distribution function of X ?

(*ii*) Obtain the density function of the random variable $Y = X^2$.

(*iii*) Obtain the correlation coefficient between X and Y.

(*iv*) Are X and Y independently distributed ?

40(*a*). If $f(x, y) = \dfrac{6 - x - y}{8}$; $0 \le x \le 2,\ 2 \le y \le 4$,

find (*i*) Var (X), (*ii*) Var (Y) (*iii*) $r\,(X, Y)$.

Ans. (*i*) $\frac{11}{36}$, (*ii*) $\frac{11}{36}$, (*iii*) $-\frac{1}{11}$.

(*b*) Given the joint density of random variables X, Y, Z as :

$$f(x, y, z) = k\,x \exp[-(y + z)],\ 0 < x < 2,\ y \ge 0,\ z \ge 0$$
$$= 0, \text{ elsewhere}$$

Find

(*i*) k,

(*ii*) the marginal density function,

(*iii*) conditional expectation of Y, given X and Z, and

(*iv*) the product moment correlation between X and Y.

[***Madras Univ. B.Sc. (Main Stat.), 1988***]

(*c*) Suppose that the two dimensional random variable (X, Y) has p.d.f. given by $f(x, y) = ke^{-y}, 0 < x < y < 1$
$= 0$, elsewhere

Find the correlation coefficient r_{XY}. [*Delhi Univ. M.C.A., 1991*]

41. The joint density of (X, Y) is :

$$f(x, y) = \frac{1}{8}(x + y), \quad 0 \le x \le 2, 0 \le y \le 2.$$

Find $\mu'_{rs} = E(X^r Y^s)$ and hence find Corr (X, Y).

Ans. $\mu'_{rs} = 2^{r+s}\left[\frac{1}{(r+2)(s+1)} + \frac{1}{(r+1)(s+2)}\right]; r = -\frac{1}{11}$.

(*b*) Find the m.g.f. of the bivariate distribution :

$$f(x, y) = 1, 0 < (x, y) < 1$$
$$= 0, \text{otherwise}$$

and hence find $r(X, Y)$.

Ans. $M(t_1, t_2) = (e^{t_1} - 1)(e^{t_2} - 1)/(t_1 t_2)$; $t_1 \ne 0, t_2 \ne 0, r(X, Y) = 0$.

42. Let (X, Y) have joint density :

$$f(x, y) = e^{-(x+y)} I_{(0,\infty)}(x) \cdot I_{(0,\infty)}(y)$$

Find Corr (X, Y). Are X and Y independent ?

Ans. Corr $(X, Y) = 0$: X and Y are independent.

43. A bivariate distribution in two discrete random variables X and Y is defined by the probability generating function :

$$\exp[a(u - 1) + b(v - 1) + c(u - 1)(v - 1)],$$

simultaneous probability of $X = r \cap Y = s$, where r and s are integers being the coefficient of $u^r v^s$. Find the correlation coefficient between X and Y.

Hint. Put $u = e^{t_1}$ and $v = e^{t_2}$ in $\exp[a(u - 1) + b(v - 1) + c(u - 1)(v - 1)]$, the result will be the m.g.f. of a bivariate distribution and is given by

$$M(t_1, t_2) = \exp[a(e^{t_1} - 1) + b(e^{t_2} - 1) + c(e^{t_1} - 1)(e^{t_2} - 1)]$$

We have $\left[\frac{\partial M}{\partial t_1}\right]_{t_1 = t_2 = 0} = a, \quad \left[\frac{\partial^2 M}{\partial t_1^2}\right]_{t_1 = t_2 = 0} = a(a + 1)$

$$\left[\frac{\partial^2 M}{\partial t_1 \partial t_2}\right]_{\substack{t_1 = 0 \\ t_2 = 0}} = ab + c, \left[\frac{\partial M}{\partial t_2}\right]_{\substack{t_1 = 0 \\ t_2 = 0}} = b, \left[\frac{\partial^2 M}{\partial t_2^2}\right]_{\substack{t_1 = 0 \\ t_2 = 0}} = b(b + 1)$$

So we have

$E(X) = a, E(X^2) = a(a + 1), E(Y) = b, E(Y^2) = b(b + 1)$ and $E(XY) = ab + c$

$$\therefore \quad r(X, Y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{[E(X^2) - \{E(X)\}^2][E(Y^2) - \{E(Y)\}^2]}} = \frac{c}{\sqrt{ab}}$$

44. Let the number X be chosen at random from among the integers 1, 2, 3, 4 and the number Y be chosen from among those at least as large as X. Prove that Cov $(X, Y) = 5/8$. Find also the regression line of Y on X.

[*Delhi Univ. B.Sc. (Maths. Hons.), 1990*]

Hint. $P(X = k) = \frac{1}{4}$; $k = 1, 2, 3, 4$ and $Y \ge X$.

$P(Y = y \mid X = 1) = \frac{1}{4}$; $y = 1, 2, 3, 4$ ($\because y \geq x$) ;

$P(Y = y \mid X = 2) = \frac{1}{3}$, $y = 2, 3, 4$

$P(Y = y \mid X = 3) = \frac{1}{2}$, $y = 3, 4$; $P(Y = y \mid X = 4) = 1$, $y = 4$.

The joint probability distribution can be obtained on using :

$$P(X = x, Y = y) = P(X = x) \,.\, P(Y = y \mid X = x).$$

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{5/8}{\sqrt{(5/4) \times (41/48)}} = \sqrt{\frac{15}{41}}$$

Regression line of Y on X : $Y - E(Y) = \frac{r\,\sigma_Y}{\sigma_X}[X - E(X)]$

45. Two ideal dice are thrown. Let X_1 be the score on the first dice ad X_2, the score on the second dice. Let $Y = \max\{X_1, X_2\}$. Obtain the joint distribution of Y and X_1 and show that

$$\text{Corr}((Y, X_1) = \frac{3}{2\sqrt{73}}$$

46. Consider an experiment of tossing two tetrahedra. Let X be the number of the down turned face of first tetrahedron and Y, the larger of the two numbers. Obtain the joint distribution of X and Y and hence $\rho(X, Y)$.

Ans. $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{5/8}{\sqrt{5/4}.\sqrt{55/64}} = \frac{2}{\sqrt{11}}$

47. Three fair coins are tossed. Let X denote the number of heads on the first two coins and let Y denote the number of tails on the last two coins.

(*a*) Find the joint distribution of X and Y.

(*b*) Find the conditional distribution of Y given that $X = 1$.

(*c*) Find Cov. (X, Y)

Ans. Cov. $(X, Y) = -1/4$.

48. For the trinomial distribution of two random variables X and Y :

$$f(x, y) = \frac{n!}{x!\,y!\,(n - x - y)!} p^x q^y (1 - p - q)^{n - x - y}$$

for $x, y = 0, 1, 2, \ldots, n$ and $x + y \leq n$, $p \geq 0$, $q \geq 0$ and $p + q \leq 1$.

(*a*) Obtain the marginal distribution of Y

(*b*) Obtain $E(X \mid Y = y)$.

(*c*) Find $\rho(X, Y)$.

Ans. (*a*) $X \sim B(n, p)$, $Y \sim B(n, q)$

(*b*) $(X \mid Y = y) \sim B\left(n - y, \frac{p}{1 - q}\right)$

(Note : $p + q \neq 1$)

$$\therefore E(X \mid Y = y) = (n - y)\left(\frac{p}{1 - q}\right)$$

(c) $\text{Cov}(X, Y) = -npq$; $\rho(X, Y) = -\left[\dfrac{pq}{(1-p)(1-q)}\right]^{1/2}$

OBJECTIVE TYPE QUESTIONS

I. Comment on the following :

(i) $r_{XY} = 0 \Rightarrow X$ and Y are independent.

(ii) If $r_{XY} > 0$ then $r_{X,-Y} > 0$, $r_{-X,Y} > 0$ and $r_{-X,-Y} > 0$

(iii) $r_{XY} > 0 \Rightarrow E(XY) > E(X)\,E(Y)$

(iv) Pearson's coefficient of correlation is independent of origin but not of scale.

(v) The numerical value of product moment correlation coefficient 'r' between two variables X and Y cannot exceed unity.

(vi) If the correlation coefficient between the variables X and Y is zero then the correlation coefficient between X^2 and Y^2 is also zero.

(vii) If $r > 0$, then as X increases, Y also increases.

(viii) "The closeness of relationship between two variables is proportional to r."

(ix) r measures every type of relationship between the two variables.

II. Comment on the following values of 'r' (correlation coefficient) :

1, – 0·95, 0, –1·64, 0·87, 0·32, –1, 2·4.

III. (i) If $\rho_{XY} = -0{\cdot}9$, then for large values of X, what sort of values do we expect for Y ?

(ii) If $\rho_{XY} = 0$, what is the value of cov (X, Y) and how are X and Y related ?

IV. Indicate the correct answer :

(i) The coefficient of correlation will have positive sign when
(a) X is increasing, Y is decreasing, (b) both X and Y are increasing, (c) X is decreasing, Y is increasing, (d) there is no change in X and Y.

(ii) The coefficient of correlation (a) can take any value between –1 and +1 (b) is always less than –1, (c) is always more than +1, (d) cannot be zero.

(iii) The coefficient of correlation (a) cannot be positive, (b) cannot be negative, (c) is always positive, (d) can be both positive as well as negative.

(iv) Probable error of r is

(a) $0{\cdot}6475\dfrac{1-r^2}{\sqrt{n}}$, (b) $0{\cdot}6754\dfrac{1+r^2}{\sqrt{n}}$, (c) $0{\cdot}6547\dfrac{1-r^2}{n}$,

(d) $0{\cdot}6754\dfrac{1-r^2}{n}$.

(v) The coefficient of correlation between X and Y is 0·6. Their covariance is 4·8. The variance of X is 9. Then the S.D. of Y is

(a) $\dfrac{4{\cdot}8}{3 \times 0{\cdot}6}$, (b) $\dfrac{0{\cdot}6}{4{\cdot}8 \times 3}$, (c) $\dfrac{3}{4{\cdot}8 \times 0{\cdot}6}$, (d) $\dfrac{4{\cdot}8}{9 \times 0{\cdot}6}$.

(*vi*) The coefficient of correlation is independent of (*a*) change of scale only, (*b*) change of origin only, (*c*) both change of scale and origin, (*d*) neither change of scale nor change of origin.

V. Fill in the blanks :

(*i*) The Karl Pearson coefficient of correlation between variables X and Y is ... ...

(*ii*) Two independent variables are ... ...

(*iii*) Limits for correlation coefficient are ... ...

(*iv*) If r be the correlation coefficient between the random variables X and Y then the variance of $X + Y$ is ... ...

(*v*) The absolute value of the product moment correlation coefficient is less than ... ...

(*vi*) Correlation coefficient is i3nvariant under changes of .. and ... ...

VI. How can you use scatter diagram to obtain an idea of extent and nature (direction) of the correlation coefficient ?

10·4. Calculation of the Correlation Coefficient for a Bivariate Frequency Distribution. When the data are considerably large, they may be summarised by using a two-way table. Here, for each variable a suitable number of classes are taken, keeping in view the same considerations as in the univariate case. If there are n classes for X and m classes for Y, there will be in all $m \times n$ cells in the two-way table. By going through the pairs of values of X and Y, we can find the frequency for each cell. The whole set of cell frequencies will then define a *bivariate frequency distribution*. The column totals and row totals will give us the marginal distributions of X and Y. A particular column or row will be called the conditional distribution of Y for given X or of X for given Y respectively.

Suppose that the bivariate data on X and Y are presented in a two-way correlation table (shown on page 10·33) where there are m classes of Y placed along the horizontal line and n classes of X along a vertical line and f_{ij} is the frequency of individuals lying in the (i, j)th cell.

Here
$$\sum_x f(x, y) = g(y)$$
is the sum of the frequencies along any row and
$$\sum_y f(x, y) = f(x)$$
is the sum of the frequencies along any column. We observe that

$$\sum_x \sum_y f(x, y) = \sum_y \sum_x f(x, y) = \sum_x f(x) = \sum_y g(y) = N$$

Thus
$$\bar{x} = \frac{1}{N} \sum_x \sum_y x f(x, y) = \frac{1}{N}\left[\sum_x \{x \sum_y f(x, y)\}\right] = \frac{1}{N} \sum_x x f(x)$$

Similarly
$$\bar{y} = \frac{1}{N} \sum_x \sum_y y f(x, y) = \frac{1}{N} \sum_y y \, g(y)$$
$$\sigma_X^2 = \frac{1}{N} \sum_x \sum_y x^2 f(x, y) - \bar{x}^2 = \frac{1}{N} \sum_x x^2 f(x) - \bar{x}^2$$

BIVARIATE FREQUENCY TABLE (CORRELATION TABLE)

Y Series ↓ \ X Series →	Classes: Mid Points x_1	x_2	... x_i ...	...	x_m	Total of frequencies of Y $g(y)$
y_1						$g(y) = \sum_x f(x, y)$
y_2						
⋮						
y_j			$f(x, y)$			
⋮						
⋮						
y_n						
Total of frequencies of X $f(x)$	$f(x) = \sum_y f(x, y)$					$N \rightarrow \sum_x \sum_y f(x, y)$ ↓ $\sum_y \sum_x f(x, y)$

Example 10·14. *The following table gives, according to age, the frequency of marks obtained by 100 students in an intelligence test.*

Marks ↓ \ Ages in years →	18	19	20	21	Total
10—20	4	2	2	—	8
20—30	5	4	6	4	19
30—40	6	8	10	11	35
40—50	4	4	6	8	22
50—60	—	2	4	4	10
60—70	—	2	3	1	6
Total	19	22	31	28	100

Calculate the correlation coefficient.

Solution.

CORRELATION TABLE

v	y	Marks \ u	-1	0	1	2	Total $f(v)$	$vf(v)$	$v^2f(v)$	$\sum_u uv\,f(u,v)$
		x	18	19	20	21				
-2	15	10—20	(8) 4	(0) 2	(−4) 2		8	−16	32	4
-1	25	20—30	(5) 5	(0) 4	(−6) 6	(−8) 4	10	−19	19	−9
0	35	30—40	(0) 6	(0) 8	(0) 10	(0) 11	35	0	0	0
1	45	40—50	(−4) 4	(0) 4	(6) 6	(16) 8	22	22	22	18
2	55	50—60		(0) 2	(8) 4	(16) 4	10	20	40	24
3	65	60—70		(0) 2	(9) 3	(6) 1	6	18	54	15
		Total $f(u)$	19	22	31	28	100	25	167	52
		$uf(u)$	−19	0	31	56	68			
		$u^2f(u)$	19	0	31	112	162			
		$u\sum_v vf(u,v)$	9	0	13	30	52			

Let $\quad U = X - 19,\ V = \{(Y - 35)/10\}$

$$\bar{u} = \frac{1}{N}\sum_u u\,f(u) = \frac{68}{100} = 0{\cdot}68,\ \bar{v} = \frac{1}{N}\sum_v v\,g(v) = \frac{25}{100} = 0{\cdot}25$$

$$\text{Cov}\,(u,v) = \frac{1}{N}\sum_u\sum_v uv\,f(u,v) - \bar{u}\,\bar{v} = \frac{1}{100}\times 52 - 0{\cdot}68\times 0{\cdot}25 = 0{\cdot}35$$

$$\sigma_U^2 = \frac{1}{N}\sum_u u^2 f(u) - \bar{u}^2 = \frac{162}{100} - (0{\cdot}68)^2 = 1{\cdot}1576$$

$$\sigma_V^2 = \frac{1}{N}\sum_v v^2 g(v) - \bar{v}^2 = \frac{167}{100} - (0{\cdot}25)^2 = 1{\cdot}6075$$

$$\therefore \quad r\,(U,V) = \frac{\text{Cov}\,(U,V)}{\sigma_U\,\sigma_V} = \frac{0{\cdot}35}{\sqrt{1{\cdot}1576\times 1{\cdot}6075}} = 0{\cdot}25$$

Since correlation coefficient is independent of change of origin and scale,

$$r(X, Y) = r(U, V) = 0{\cdot}25$$

Remark. Figures in circles in the table on page 10·34 are the product terms $uvf(u, v)$:

Example 10·15. *The joint probability distribution 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 between X and Y.

Solution.

COMPUTATION OF MARGINAL PROBABILITIES

Y \ X	−1	+1	$g(y)$
0	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{4}{8}$
1	$\frac{2}{8}$	$\frac{2}{8}$	$\frac{4}{8}$
$p(x)$	$\frac{3}{8}$	$\frac{5}{8}$	1

We have :

$$E(X) = \Sigma\, xp(x) = (-1) \times \frac{3}{8} + 1 \times \frac{5}{8} = \frac{1}{4}$$

$$E(X^2) = \Sigma\, x^2 p(x) = (-1)^2 \times \frac{3}{8} + 1^2 \times \frac{5}{8} = 1$$

$$\therefore \quad \text{Var}(X) = E(X^2) - [E(X)]^2 = 1 - \frac{1}{16} = \frac{15}{16}$$

$$E(Y) = \Sigma\, y\, g(y) = 0 \times \frac{4}{8} + 1 \times \frac{4}{8} = \frac{1}{2}$$

$$E(Y^2) = \Sigma\, y^2\, g(y) = 0^2 \times \frac{4}{8} + 1^2 \times \frac{4}{8} = \frac{1}{2}$$

$$\therefore \quad \text{Var}(Y) = E(Y^2) - [E(Y)]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

$$E(XY) = 0 \times (-1) \times \frac{1}{8} + 0 \times 1 \times \frac{3}{8} + 1 \times (-1) \times \frac{2}{8} + 1 \times 1 \times \frac{2}{8}$$

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

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\,E(Y) = 0 - \frac{1}{4} \times \frac{1}{2} = -\frac{1}{8}$$

$$\therefore \quad r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{-\frac{1}{8}}{\sqrt{\frac{15}{16} \times \frac{1}{4}}} = \frac{-1}{\sqrt{15}} = \frac{-1}{3{\cdot}873}$$

$$= -0{\cdot}2582$$

EXERCISE 10(b)

1. Write a brief note on the correlation table :

The following are the marks obtained by 24 students in a class test of Statistics and Mathematics :

Role No. of Students :	1	2	3	4	5	6	7	8	9	10	11	12
Marks in Statistics :	15	0	1	3	16	2	18	5	4	17	6	19
Marks in Mathematics :	13	1	2	7	8	9	12	9	17	16	6	18
Roll No. of Students :	13	14	15	16	17	18	19	20	21	22	23	24
Marks in Statistics :	14	9	8	13	10	13	11	11	12	18	9	7
Marks in Mathematics :	11	3	5	4	10	11	14	7	18	15	15	3

Prepare a correlation table taking the magnitude of each class interval as four marks and the first class interval as "equal to 0 and less than 4". Calculate Karl Pearson's coefficient of correlation between the marks in Statistics and marks in Mathematics from the correlation table.

Ans. 0·5544.

2. An employment bureau asked applicants their weekly wages on jobs last held. The actual wages were obtained for 54 of them; and are recorded in the table below; x represents reported wage, y actual wage, and the entry in the table represents frequency. Find the correlation coefficient and comment on the significance of the computed value. [Four figure log table may be used].

$x \downarrow$ \ $y \rightarrow$	15	20	25	30	35	40
40						2
35				3	5	
30			4	15		
25			20			
20		3	1			
15	1					

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1986*]

3. Calculate the correlation coefficient from the following table :—

$y \downarrow$ \ $x \rightarrow$	0—10	10—20	20—30	30—40
0—5	1	3	2	0
5—10	7	10	8	1
10—15	10	13	10	8
15—20	5	8	10	7
20—25	0	1	5	4

4. (*a*) Find the correlation coefficient between age and salary of 50 workers in a factory :

Age (in years) ↓	*Daily pay in rupees*				
	160—169	170—179	180—189	190—199	200—209
20—30	5	3	1	...	...
30—40	2	6	2	1	...
40—50	1	2	4	2	2
50—60	...	1	3	6	2
60—70	...	...	1	1	5

(*b*) Fnd the coefficient of correlation between the ages of 100 mothers and daughters :

Age of mothers in years (X)	*Age of daughters in years (Y)*					*Total*
	5—10	*10—15*	*15—20*	*20—25*	*25—30*	
15—25	6	3				9
25—35	3	16	10			29
35—45		10	15	7		32
45—55			7	10	4	21
55—65				4	5	9
Total	9	29	32	21	9	100

[***Madras Univ. B.Sc. (Main Maths.), 1991***]

5. Given the following frequency distribution of (X, Y) :

$Y \downarrow$ \ $X \rightarrow$	5	10	Total
10	30	20	50
20	20	30	50
Total	50	50	100

find the frequency distribution of (U, V), where

$$U = \frac{X - 7{\cdot}5}{2{\cdot}5}, \; V = \frac{Y - 15}{-5}.$$

What shall be the relationship between the correlation coefficients between X, Y, and U, V ?

6. (*a*) Find the coefficient of correlation between X and Y for the following table :

$X \downarrow$ \ $Y \rightarrow$	y_1	y_2	Total
x_1	p_{11}	p_{12}	P
x_2	p_{21}	p_{22}	Q
Total	P'	Q'	1

(*b*) Consider the following probability distribution :

$X \downarrow$ \ $Y \rightarrow$	0	1	2
0	0·1	0·2	0·1
1	0·2	0·3	0·1

Calculate $E(X)$, Var (X), Cov(X, Y) and $r\,(X, Y)$.

[***Delhi Univ. M.A. (Eco.), 1991***]

(*c*) Let (X, Y) have the p.m.f.

$$p(0, 1) = p(1, 0) = \frac{1}{3};\ p(0, -1) = p(-1, 0) = \frac{1}{6}.$$

Find $r(X, Y)$. Are X and Y independent ? For what values of k, $X + kY$ and $kX + Y$ are uncorrelated ?

10·5. Probable Error of Correlation Coefficient. If r is the correlation coefficient in a sample of n pairs of observations, then its *standard error* is given by

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

Probable error of correlation coefficient is given by

$$\text{P.E.}(r) = 0{\cdot}6745 \times \text{S.E.} = 0{\cdot}6745 \frac{(1 - r^2)}{\sqrt{n}} \qquad \ldots(10{\cdot}6)$$

Probable error is an old measure for testing the reliability of an observed correlation coefficient. The reason for taking the factor 0·6745 is that in a normal distribution, the range $\mu \pm 0{\cdot}6745\ \sigma$ covers 50% of the total area. According to Secrist, "*The probable error of the correlation co-efficient is an amount which if added to and substracted from the mean correlation coefficient, produces amounts within which the chances are even that a coefficient of correlation from a series selected at random will fall.*"

If $r <$ P.E.(r), correlation is not at all significant. If $r > 6$ P.E.(r), it is definitely significant. A rigorous method (t-test) of testing the significance of an observed correlation coefficient will be discussed later in "*tests of significance*" in sampling [c.f. § 14·4·11].

Probable error also enables us to find the limits within which the population correlation coefficient can be expected to vary. The limits are $r \pm$ P.E.(r).

10·6. 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. Let (x_i, y_i); $i = 1, 2, \ldots, n$ be the ranks of the ith individual in two characteristics A and B respectively. Pearsonian coefficient of correlation between the ranks x_i's and y_i's is called the rank correlation coefficient between A and B for that group of individuals.

Assuming that no two individuals are bracketed equal in either classification, each of the variables X and Y takes the values $1, 2, \ldots, n$.

Hence $$\bar{x} = \bar{y} = \frac{1}{n}(1 + 2 + 3 + \ldots + n) = \frac{n+1}{2}$$

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

$$= \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}$$

$$\therefore \qquad \sigma_X^2 = \frac{n^2-1}{12} = \sigma_Y^2$$

In general $x_i \neq y_i$. Let $d_i = x_i - y_i$

$$\therefore \qquad d_i = (x_i - \bar{x}) - (y_i - \bar{y}) \qquad (\because \bar{x} = \bar{y})$$

Squaring and summing over i from 1 to n, we get

$$\Sigma d_i^2 = \Sigma\{(x_i - \bar{x}) - (y_i - \bar{y})\}^2$$

$$= \Sigma(x_i - \bar{x})^2 + \Sigma(y_i - \bar{y})^2 - 2\Sigma(x_i - \bar{x})(y_i - \bar{y})$$

Dividing both sides by n, we get

$$\frac{1}{n}\Sigma d_i^2 = \sigma_X^2 + \sigma_Y^2 - 2\,\text{Cov}\,(X, Y) = \sigma_X^2 + \sigma_Y^2 - 2\rho\,\sigma_X\sigma_Y,$$

where ρ is the rank correlation coefficient between A and B.

$$\therefore \qquad \frac{1}{n}\Sigma d_i^2 = 2\sigma_X^2 - 2\rho\sigma_X^2 \;\Rightarrow\; 1 - \rho = \frac{\Sigma d_i^2}{2n\sigma_X^2}$$

$$\Rightarrow \qquad \rho = 1 - \frac{\sum_{i=1}^{n} d_i^2}{2n\sigma_X^2} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad \ldots(10\cdot7)$$

which is the *Spearman's formula for the rank correlation coefficient.*

Remark. We always have

$$\Sigma d_i = \Sigma(x_i - y_i) = \Sigma x_i - \Sigma y_i = n(\bar{x} - \bar{y}) = 0 \qquad (\because \bar{x} = \bar{y})$$

This serves as a check on the calculations.

10·6·1. Tied Ranks. If some of the individuals receive the same rank in a ranking of merit, they are said to be tied. Let us suppose that m of the individuals, say, $(k+1)^{th}$, $(k+2)^{th}$, ..., $(k+m)^{th}$ are tied. Then each of these m individuals is assigned a common rank, which is the arithmetic mean of the ranks $k+1, k+2, ..., k+m$.

Derivation of $\rho(X, Y)$. We have :

$$\rho(X, Y) = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{[\Sigma (X - \bar{X})^2 . \Sigma (Y - \bar{Y})^2]^{1/2}} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}} \quad ...(*)$$

where $x = X - \bar{X}, y = Y - \bar{Y}$.

If X and Y each takes the values $1, 2, ..., n$, then we have

$$\bar{X} = (n+1)/2 = \bar{Y}$$

and $$n\sigma_X^2 = \Sigma x^2 = \frac{n(n^2-1)}{12} = \text{ and } n\sigma_Y^2 = \Sigma y^2 = \frac{n(n^2-1)}{12} \quad ...(**)$$

Also $$\Sigma d^2 = \Sigma (X - Y)^2 = \Sigma [(X - \bar{X}) - (Y - \bar{Y}]^2 = \Sigma (x - y)^2$$

$$\Rightarrow \quad \Sigma d^2 = \Sigma x^2 + \Sigma y^2 - 2\Sigma xy$$

$$\Rightarrow \quad \Sigma xy = \frac{1}{2}[\Sigma x^2 + \Sigma y^2 - \Sigma d^2] \quad ...(***)$$

We shall now investigate the effect of common ranking, (in case of ties), on the sum of squares of the ranks. Let S^2 and S_1^2 denote the sum of the squares of untied and tied ranks respectively.

Then we have :

$$S^2 = (k+1)^2 + (k+2)^2 + ... + (k+m)^2$$
$$= mk^2 + (1^2 + 2^2 + ... + m^2) + 2k.(1 + 2 + ... + m)$$
$$= mk^2 + \frac{m(m+1)(2m+1)}{6} + mk(m+1)$$

$$S_1^2 = m \text{ (Average rank)}^2$$
$$= m\left[\frac{(k+1) + (k+2) + ... + (k+m)}{m}\right]^2$$
$$= m\left(k + \frac{m+1}{2}\right)^2 = mk^2 + \frac{m(m+1)^2}{4} + mk(m+1)$$

$$\therefore \quad S^2 - S_1^2 = \frac{m(m+1)}{12}[2(2m+1) - 3(m+1)] = \frac{m(m^2-1)}{12}$$

Thus the effect of tying m individuals (ranks) is to reduce the sum of the squares by $m(m^2-1)/12$, though the mean value of the ranks remains the same, *viz.*, $(n+1)/2$.

Suppose that there are s such sets of ranks to be tied in the X-series so that the total sum of squares due to them is

$$\frac{1}{12}\sum_{i=1}^{s} m_i(m_i^2 - 1) = \frac{1}{12}\sum_{i=1}^{s}(m_i^3 - m_i) = T_X, \text{ (say)} \quad ... (10·7a)$$

Similarly suppose that there are t such sets of ranks to be tied with respect to the other series Y so that sum of squares due to them is :

$$\frac{1}{12}\sum_{j=1}^{t} m_j' \cdot (m_j'^2 - 1) = \frac{1}{12}\sum_{j=1}^{t} (m_j'^3 - m_j') = T_Y \text{ , (say)} \qquad \ldots(10\cdot7b)$$

Thus, in the case of ties, the new sums of squares are given by :

$$n \text{ Var}'(X) = \Sigma x^2 - T_X = \frac{n(n^2-1)}{12} - T_X$$

$$n \text{ Var}'(Y) = \Sigma y^2 - T_Y = \frac{n(n^2-1)}{12} - T_Y$$

and $$n \text{ Cov}'(X, Y) = \frac{1}{2}[\Sigma x^2 - T_X + \Sigma y^2 - T_Y - \Sigma d^2] \qquad \text{[From (***)]}$$

$$= \frac{1}{2}\left[\frac{n(n^2-1)}{12} - T_X + \frac{n(n^2-1)}{12} - T_Y - \Sigma d^2\right]$$

$$= \frac{n(n^2-1)}{12} - \frac{1}{2}\left[(T_X + T_Y) + \Sigma d^2\right]$$

$$\rho(X, Y) = \frac{\frac{n(n^2-1)}{12} - \frac{1}{2}\left[T_X + T_Y + \Sigma d^2\right]}{\left[\frac{n(n^2-1)}{12} - T_X\right]^{1/2}\left[\frac{n(n^2-1)}{12} - T_Y\right]^{1/2}}$$

$$= \frac{\frac{n(n^2-1)}{6} - \left[\Sigma d^2 + T_X + T_Y\right]}{\left[\frac{n(n^2-1)}{6} - 2T_X\right]^{1/2}\left[\frac{n(n^2-1)}{6} - 2\,T_Y\right]^{1/2}} \qquad \ldots(10\cdot7c)$$

where T_X and T_Y are given by (10·7a) and (10·7b).

Remark. If we adjust only the covariance term *i.e.*, Σxy and not the variances σ_X^2 (or Σx^2) and σ_Y^2 (or Σy^2) for ties, then the formula (10·7c) reduces to :

$$\rho(X, Y) = \frac{\frac{n(n^2-1)}{6} - (\Sigma d^2 + T_X + T_Y)}{n(n^2-1)/6}$$

$$= 1 - \frac{6\,[\Sigma d^2 + T_X + T_Y]}{n(n^2-1)}, \qquad \ldots(10\cdot7d)$$

a formula which is commonly used in practice for numerical problems. For illustration, see Example 10·18.

Example 10·16. *The ranks of same 16 students in Mathematics and Physics are as follows. Two numbers within brackets denote the ranks of the students in Mathematics and Physics.*

(1, 1) (2, 10) (3, 3) (4, 4) (5, 5) (6, 7) (7, 2) (8, 6) (9, 8) (10, 11) (11, 15) (12, 9) (13, 14) (14, 12) (15, 16) (16, 13).

Calculate the rank correlation coefficient for proficiencies of this group in Mathematics and Physics.

Solution.

Ranks in Maths. (X)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Ranks in Physics (Y)	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13	
$d = X - Y$	0	−8	0	0	0	−1	5	2	1	−1	−4	3	−1	2	−1	3	0
d^2	0	64	0	0	0	1	25	4	1	1	16	9	1	4	1	9	136

Rank correlation coefficient is given by

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 136}{16 \times 255} = 1 - \frac{1}{5} = \frac{4}{5} = 0{\cdot}8$$

Example 10·17. *Ten competitors in a musical test were ranked by the three judges A, B and C in the following order :*

Ranks by A :	1	6	5	10	3	2	4	9	7	8
Ranks by B :	3	5	8	4	7	10	2	1	6	9
Ranks by C :	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 in music.

Solution. Here $n = 10$

Ranks by A (X)	*Ranks by B (Y)*	*Ranks by C (Z)*	$d_1 = X-Y$	$d_2 = X-Z$	$d_3 = Y-Z$	d_1^2	d_2^2	d_3^2
1	3	6	−2	−5	−3	4	25	9
6	5	4	1	2	1	1	4	1
5	8	9	−3	−4	−1	9	16	1
10	4	8	6	−2	−4	36	4	16
3	7	1	−4	2	6	16	4	36
2	10	2	−8	0	8	64	0	64
4	2	3	2	1	−1	4	1	1
9	1	10	8	−1	−9	64	1	81
7	6	5	1	2	1	1	4	1
8	9	7	−1	1	2	1	1	4
Total			$\Sigma d_1 = 0$	$\Sigma d_2 = 0$	$\Sigma d_3 = 0$	$\Sigma d_1^2 = 200$	$\Sigma d_2^2 = 60$	$\Sigma d_3^2 = 214$

$$\rho(X, Y) = 1 - \frac{6 \Sigma d_1^2}{n(n^2 - 1)} = 1 - \frac{6 \times 200}{10 \times 99} = 1 - \frac{40}{33} = -\frac{7}{33}$$

$$\rho(X, Z) = 1 - \frac{6 \Sigma d_2^2}{n(n^2 - 1)} = 1 - \frac{6 \times 60}{10 \times 99} = 1 - \frac{4}{11} = \frac{7}{11}$$

$$\rho(Y, Z) = 1 - \frac{6\sum d_3^2}{n(n^2-1)} = 1 - \frac{6 \times 214}{10 \times 99} = -\frac{49}{165}$$

Since $\rho(X, Z)$ is maximum, we conclude that the pair of judges A and C has the nearest approach to common likings in music.

10·6·2. Repeated Ranks (Continued). If any two or more individuals are bracketed equal in any classification with respect to characteristics A and B, or if there is more than one item with the same value in the series, then the Spearman's formula (10·7) for calculating the rank correlation coefficient breaks down, since in this case each of the variables X and Y does not assume the values 1, 2, ..., n and consequently, $\bar{x} \neq \bar{y}$.

In this case, common ranks are given to the repeated items. This common rank is the average of the ranks which these items would have assumed if they were sightly 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 rank correlation formula [c.f. (10·7*c*) and (10·7*d*)].

In the formula, we add the factor $\frac{m(m^2-1)}{12}$ to $\sum d^2$, where m is the number of times an item is repeated. This correction factor is to be added for each repeated value in both the X-series and Y-series.

Example 10·18. *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.

CALCULATIONS FOR RANK CORRELATION

X	Y	Rank X (x)	Rank Y (y)	$d = x - y$	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
				$\sum d = 0$	$\sum d^2 = 72$

In the X-series we see that the value 75 occurs 2 times. The common rank given to these values is 2·5 which is the average of 2 and 3, the ranks which these values would have taken if they were different. The next value 68, then gets the next rank which is 4. Again we see that value 64 occurs thrice. The common rank given to it is 6 which is the average of 5, 6 and 7. Similarly in

the Y-series, the value 68 occurs twice and its common rank is 3·5 which is the average of 3 and 4. As a result of these common rankings, the formula for 'ρ' has to be corrected. To Σd^2 we add $\frac{m(m^2-1)}{12}$ for each value repeated, where m is the number of times a value occurs. In the X-series the correction is to be applied twice, once for the value 75 which occurs twice ($m = 2$) and then for the value 64 which occurs thrice ($m = 3$). The total correction for the X-series is

$$\frac{2(4-1)}{12} + \frac{3(9-1)}{12} = \frac{5}{2}$$

Similarly, this correction for the Y-series is $\frac{2(4-1)}{12} = \frac{1}{2}$, as the value 68 occurs twice.

Thus $$\rho = 1 - \frac{6\left[\Sigma d^2 + \frac{5}{2} + \frac{1}{2}\right]}{n(n^2-1)} = 1 - \frac{6(72+3)}{10 \times 99} = 0{\cdot}545$$

10·6·3. Limits for the Rank Correlation Coefficient. Spearman's rank correlation coefficient is given by

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

'ρ' is maximum, if $\sum_{i=1}^{n} d_i^2$ is minimum, *i.e.*, if each of the deviations d_i is minimum. But the minimum value of d_i is zero in the particular case $x_i = y_i$, *i.e.*, if the ranks of the ith individual in the two characteristic are equal. Hence the maximum value of ρ is + 1, *i.e.*, $\rho \leq 1$.

'ρ' is minimum , if $\sum_{i=1}^{n} d_i^2$ is maximum, *i.e.*, if each of the deviations d_i is maximum which is so if the ranks of the n individuals in the two characteristics are in the opposite directions as given below :

x	1	2	3	...	...	$n-1$	n	
y	n	$n-1$	$n-2$	...	...	2	1	...(*)

Case 1. Suppose n is odd and equal to $(2m + 1)$ then the values of d are :

d : $2m, 2m-2, 2m-4, \ldots, 2, 0, -2, -4, \ldots -(2m-2), -2m.$

$$\therefore \sum_{i=1}^{n} d_i^2 = 2\{(2m)^2 + (2m-2)^2 + \ldots + 4^2 + 2^2\}$$

$$= 8\{m^2 + (m-1)^2 + \ldots + 1^2\} = \frac{8m\,(m+1)\,(2m+1)}{6}$$

$$\text{Hence} \quad \rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{8m(m+1)(2m+1)}{(2m+1)\{(2m+1)^2-1\}}$$

$$= \frac{8m(m+1)}{(4m2+4m)} = 1 - \frac{8m(m+1)}{4m(m+1)} = -1$$

Case II. Let n be even and equal to $2m$, (say).

Then the values of d are

$(2m-1), (2m-3), \ldots, 1, -1, -3, \ldots, -(2m-3), -(2m-1)$

$$\therefore \quad \Sigma d_i^2 = 2\{(2m-1)^2 + (2m-3)^2 + \ldots + 1^2\}$$

$$= 2[\{(2m)^2 + (2m-1)^2 + (2m-2)^2 + \ldots + 2^2 + 1^2\} - \{(2m)^2 + (2m-2)^2 + \ldots + 4^2 + 2^2\}]$$

$$= 2[1^2 + 2^2 + \ldots + (2m)^2 - \{2^2m^2 + 2^2(m-1)^2 + \ldots + 2^2\}]$$

$$= 2\left[\frac{2m\,(2m+1)(4m+1)}{6} - 4\,\frac{m(m+1)(2m+1)}{6}\right]$$

$$= \frac{2m}{3}\left[(2m+1)\,(4m+1) - 2(m+1)(2m+1)\right]$$

$$= \frac{2m}{3}\left[(2m+1)(4m+1-2m-2)\right]$$

$$= \frac{2m}{3}(2m+1)(2m-1) = \frac{2m(4m^2-1)}{3}$$

$$\therefore \quad \rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{4m(4m^2-1)}{2m(4m^2-1)} = -1$$

Thus the limits for rank correlation coefficient are given by $-1 \le \rho \le 1$.

Aliter. For an alternate and simpler proof for obtaining the minimum value of ρ, from (*) onward, proceed as in Hint to Question Number 9 of Exercise 10(*c*).

Remarks on Spearman's Rank Correlation Coefficient.

1. $\Sigma d = \Sigma x - \Sigma y = n(\bar{x} - \bar{y}) = 0$, which provides a check for numerical calculations.

2. Since Spearman's rank correlation coefficient ρ is nothing but Pearsonian correlation coefficient between the ranks, it can be interpreted in the same way as the Karl Pearson's correlation coefficient.

3. Karl Pearson's correlation coefficient assume that the parent population from which sample observations are drawn is normal. If this assumption is violated then we need a measure which is distribution free (or non-parametric). A distribution-free measure is one which does not make any assumptions about the parameters of the population. Spearman's ρ is such a measure (*i.e.*, distribution-free), since no strict assumptions are made about the form of the population from which sample observations are drawn.

4. Spearman' formula is easy to understand and apply as compared with Karl Pearson's formula. The value obtained by the two formulae, *viz.*, Pearsonian r and Spearman's ρ, are generally different. The difference arises due to the fact that when ranking is used instead of full set of observations, there is

always some loss of information. Unless many ties exist, the coefficient of rank correlation should be only slightly lower than the Pearsonian coefficient.

5. Spearman's formula is the only formula to be used for finding correlation coefficient if we are dealing with qualitative characteristics which cannot be measured quantitatively but can be arranged serially. It can also be used where actual data are given. In case of extreme observations, Spearman's formula is preferred to Pearson's formula.

6. Spearman's formula has its limitations also. It is not practicable in the case of bivariate frequency distribution (Correlation Table). For $n > 30$, this formula should not be used unless the ranks are given, since in the contrary case the calculations are quite time-consuming.

EXERCISE 10(c)

1. Prove that Spearman's rank correlation coefficient is given by

$1 - \dfrac{6 \sum d_i^2}{n^3 - n}$, where d_i denotes the difference between the ranks of ith individual.

2. (*a*) Explain the difference between product moment correlation coefficient and rank correlation coefficient.

(*b*) 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 correlation coefficient.

3. (*a*) Calculate the coefficient of correlation for ranks from the following data :

(X, Y) : (5, 8), (10, 3), (6, 2), (3, 9), (19, 12), (5, 3), (6, 17), (12, 18), (8, 22), (2, 12), (10, 17), (19, 20).

[*Calicut Univ. B.Sc. (Subs. Stat.), Oct. 1991*]

(*b*) Ten recruits were subjected to a selection test to ascertain their suitability for a certain course of training. At the end of training they were given a proficiency test.

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	10
X :	10	15	12	17	13	16	24	14	22	20
Y :	30	42	45	46	33	34	40	35	39	38

Calculate product moment correlation coefficient and rank correlation coefficient. Why are two coefficients different ?

4. (*a*) The I.Q.'s of a group of 6 persons were measured, and they then sat for a certain examination. Their I.Q.'s and examination marks were as follows :

Person :	A	B	C	D	E	F
I.Q. :	110	100	140	120	80	90
Exam. marks :	70	60	80	60	10	20

Compute the coefficients of correlation and rank correlation. Why are the correlation figures obtained different ?

Ans. 0·882 and 0·9.

The difference arises due to the fact that when ranking is used instead of the full set of observations, there is always some loss of information.

(*b*) The value of ordinary correlation (r) for the following data is 0·636 :—

X :	·05	·14	·24	·30	·47	·52	·57	·61	·67	·72
Y :	1·08	1·15	1·27	1·33	1·41	1·46	1·54	2·72	4·01	9·63

(*i*) Calculate Spearman's rank-correlation (ρ) for this data.

(*ii*) What advantage of ρ was brought out in this example ?

4. Ten competitors in a beauty contest are ranked by three judges as follows :

	Competitors									
Judges :	1	2	3	4	5	6	7	8	9	10
A :	6	5	3	10	2	4	9	7	8	1
B :	5	8	4	7	10	2	1	6	9	3
C :	4	9	8	1	2	3	10	5	7	6

Discuss which pair of judges has the nearest approach to common tastes of beauty.

5. A sample of 12 fathers and their eldest sons gave the following data about their height in inches :

Father :	65	63	67	64	68	62	70	66	68	67	69	71
Son :	68	66	68	65	69	66	68	65	71	67	68	70

Calculate coefficient of rank correlation. (**Ans.** 0·7220)

6. The coefficient of rank correlation between marks in Statistics and marks in Mathematics obtained by a certain group of students is 0·8. If the sum of the squares of the difference in ranks is given to be 33, find the number of student in the group (**Ans.** 10). [***Madras Univ. B.Sc., 1990***]

7. The coefficient of rank correlation of the marks obtained by 10 students in Maths and Statistics was found to be 0·5. It was later discovered that the difference in ranks in two subjects obtained by one of the students was wrongly taken as 3 instead of 7. Find the correct coefficient of rank correlation.

Hint. $$0{\cdot}5 = 1 - \frac{6\Sigma d^2}{10 \times 99}$$

$$\Rightarrow \qquad \Sigma d^2 = \frac{990}{6 \times 2} = 82{\cdot}5$$

Since one difference was wrongly taken as 3 instead of 7, the correct value of Σd^2 is given by

$$\text{Corrected } \Sigma d^2 = 82{\cdot}5 - (3)^2 + (7)^2 = 122{\cdot}5$$

$$\therefore \qquad \text{Corrected} \quad \rho = 1 - \frac{6 \times 122{\cdot}5}{10 \times 99} = 0{\cdot}2576$$

8. If d_i be the difference in the ranks of the ith individual in two different characteristics, then show that the maximum value of $\sum_{i=1}^{n} d_i^2$ is $\frac{1}{3}(n^3 - n)$. Hence or otherwise, show that rank correlation coefficient lies between -1 and $+1$. [*Delhi Univ. B.Sc. (Maths. Hons.), 1986*]

9. Let $x_1, x_2, \ldots, x_n$ be the ranks of n individuals according to a character A and $y_1, y_2, \ldots, y_n$ be the ranks of the same individuals according to other character B. Obviously $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$ are permutations of 1, 2, ..., n. It is given that $x_i + y_i = 1 + n$, for $i = 1, 2, \ldots, n$. Show that the value of the rank correlation coefficient ρ between the characters A and B is -1.

Hint. We are given $\quad x_i + y_i = n + 1 \; \forall i = 1, 2, \ldots, n$

Also $\quad x_i - y_i = d_i$

$$\therefore \quad 2x_i = n + 1 + d_i \Rightarrow d_i = 2x_i - (n + 1)$$

$$\therefore \quad \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} [4x_i^2 + (n + 1)^2 - 2(n + 1)2x_i]$$

$$= 4\frac{n(n + 1)(2n + 1)}{6} + n(n + 1)^2 - \frac{4(n + 1)n(n + 1)}{2}$$

$$= \frac{n(n^2 - 1)}{3}$$

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

Remark. From Spearmans' formula we note that ρ will be minimum if Σd_i^2 is maximum, which will be so if the ranks X and Y are in opposite directions as given below :

x	1	2	3	...	n
y	n	$n - 1$	$n - 2$	...	1

This gives us

$$x_i + y_i = n + 1, i = 1, 2, \ldots, n.$$

Hence the value of $\rho = -1$ obtained above is minimum value of ρ.

10. Show that in a ranked bivariate distribution in which no ties occur and in which the variables are independent

(*a*) $\sum_i d_i^2$ is always even, and

(*b*) there are not more than $\frac{1}{6}(n^3 - n) + 1$ possible values of r.

11. Show that if X, Y be identically distributed with common probability mass function : $P(X = k) = \frac{1}{N}$, for $k = 1, 2, \ldots, N; N > 1$,

then $\rho_{X,Y}$, the correlation coefficient between X and Y, is given by

$$1 - \frac{6E(X - Y)^2}{N^2 - 1}$$

[*Delhi Univ. B.Sc. (Maths Hons.), 1992*]

10·7. Regression. The term "*regression*" literally means "*stepping back towards the average*". It was first used by a British biometrician Sir Francis Galton (1822—1911), in connection with the inheritance of stature. Galton found that the offsprings of abnormally tall or short parents tend to "regress" or "step back" to the average population height. But the term "regression" as now used in Statistics is only a convenient term without having any reference to biometry.

Definition. *Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data.*

In regression analysis there are two types of variables. The variable whose value is influenced or is to be predicted is called *dependent variable* and the variable which influences the values or is used for prediction, is called *independent variable*. In regression analysis independent variable is also known as *regressor or predictor* or *explanatory variable* while the dependent variable is also known as *regressed* or *explained variable*.

10·7·1. Lines of Regression. If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster round some curve called the "*curve of regression*". If the curve is a straight line, it is called the line of regression and there is said to be *linear regression* between the variables, otherwise regression is said to be *curvilinear*.

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable. Thus the line of regression is the line of "*best fit*" and is obtained by the *principles of least squares*.

Let us suppose that in the bivariate distribution (x_i, y_i); $i = 1, 2, \ldots, n$; Y is dependent variable and X is independent variable. Let the line of regression of Y on X be $Y = a + bX$.

According to the principle of least squares, the normal equations for estimating a and b are {*c.f.* (9·2*a*)}

$$\sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \qquad \ldots(10\cdot8)$$

and
$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \qquad \ldots(10\cdot9)$$

From (10·8) on dividing by n, we get

$$\bar{y} = a + b\bar{x} \qquad \ldots(10\cdot10)$$

Thus the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$.

Now

$$\mu_{11} = \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\,\bar{y} \Rightarrow \frac{1}{n} \sum_{i=1}^{n} x_i y_i = \mu_{11} + \bar{x}\,\bar{y} \qquad \ldots(10\cdot11)$$

Also
$$\sigma_X^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \Rightarrow \frac{1}{n} \sum_{i=1}^{n} x_i^2 = \sigma_X^2 + \bar{x}^2 \qquad \ldots(10\cdot11a)$$

Dividing (10·9) by n and using (10·11) and (10·11a), we get

$$\mu_{11} + \bar{x}\,\bar{y} = a\bar{x} + b(\sigma_X^2 + \bar{x}^2) \qquad \ldots(10\cdot12)$$

Multiplying (10·10) by $\bar{x}$ and then subtracting from (10·12), we get

$$\mu_{11} = b\sigma_X^2 \Rightarrow b = \frac{\mu_{11}}{\sigma_X^2} \qquad \ldots(10\cdot13)$$

Since 'b' is the slope of the line of regression of Y on X and since the line of regression passes through the point $(\bar{x}, \bar{y})$, its equation is

$$Y - \bar{y} = b(X - \bar{x}) = \frac{\mu_{11}}{\sigma_X^2}(X - \bar{x}) \qquad \ldots(10\cdot14)$$

$$\Rightarrow \qquad Y - \bar{y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \qquad \ldots(10\cdot14a)$$

Starting with the equation $X = A + BY$ and proceeding similarly or by simply interchanging the variables X and Y in (10·14) and (10·14a), the equation of the line of regression of X on Y becomes

$$X - \bar{x} = \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{y}) \qquad \ldots(10\cdot15)$$

$$\Rightarrow \qquad X - \bar{x} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{y}) \qquad \ldots(10\cdot15a)$$

Aliter. The straight line defined by

$$Y = a + bX \qquad \ldots(i)$$

and satisfying the residual (least square) condition

$$S = E[(Y - a - bX)^2] = \text{Minimum} \qquad \ldots(ii)$$

for variations in a and b, is called the line of regression of Y on X.

The necessary and sufficient conditions for a minima of S, subject to variations in a and b are :

(*i*) $\dfrac{\partial S}{\partial a} = 0$, $\dfrac{\partial S}{\partial b} = 0$ and ...(*)

(*ii*) $\Delta = \begin{vmatrix} \dfrac{\partial^2 S}{\partial a^2} & \dfrac{\partial^2 S}{\partial a \partial b} \\ \dfrac{\partial^2 S}{\partial b\, \partial a} & \dfrac{\partial^2 S}{\partial b^2} \end{vmatrix} > 0$ and $\dfrac{\partial^2 S}{\partial a^2} > 0$ ... (**)

Using (*), we get

$$\frac{\partial S}{\partial a} = -2\, E\,[Y - a - bX] = 0 \qquad \text{...(iii)}$$

$$\frac{\partial S}{\partial b} = -2\, E\,[X(Y - a - bX)] = 0 \qquad \text{...(iv)}$$

$\Rightarrow \quad E(Y) = a + bE(X)$ ...(*v*) and $E(XY) = aE(X) + bE(X^2)$ ...(*vi*)

Equation (*v*) implies that the line (*i*) of regression of Y on X passes through the mean value $[E(X), E(Y)]$.

Multiplying (*v*) by $E(X)$ and substracting from (*vi*), we get

$$E(XY) - E(X)E(Y) = b[E(X^2) - \{E(X)\}^2]$$

$$\Rightarrow \quad \text{Cov}\,(X, Y) = b\,\sigma_X^2 \quad \Rightarrow \quad b = \frac{\text{Cov}\,(X, Y)}{\sigma_X^2} = \frac{r\sigma_Y}{\sigma_X} \qquad \text{...(vii)}$$

Subtracting (*v*) from (*i*) and using (*vii*), we obtain the equation of line of regression of Y on X as :

$$Y - E(Y) = \frac{\text{Cov}\,(X, Y)}{\sigma_X^2}[X - E(X)] \Rightarrow Y - E(Y) = \frac{r\sigma_Y}{\sigma_X}[X - E(X)]$$

Similarly, the straight line defined by $X = A + BY$
and satisfying the residual condition

$$E[X - A - BY]^2 = \text{Minimum},$$

is called the line of regression of X on Y.

Remarks 1. We note that

$$\frac{\partial^2 S}{\partial a^2} = 2 > 0, \text{ and}$$

$$\frac{\partial^2 S}{\partial b^2} = 2E(X^2) \text{ and } \frac{\partial^2 S}{\partial a \partial b} = 2E(X)$$

Substituting in (**), we have

$$\Delta = \frac{\partial^2 S}{\partial a^2} \cdot \frac{\partial^2 S}{\partial b^2} - \left(\frac{\partial^2 S}{\partial a \partial b}\right)^2$$

$$= 4\,[E(X^2) - (E(X))^2] = 4.\,\sigma_X^2 > 0$$

Hence the solution of the least square equations (*iii*) and (*iv*), in fact, provides a minima of S.

2. The regression equation (10·14*a*) implies that the line of regression of Y on X passes through the point $(\bar{x}, \bar{y})$. Similarly (10·15*a*) implies that the line of regression of X on Y also passes through the point $(\bar{x}, \bar{y})$. Hence both the lines of regression pass through the point $(\bar{x}, \bar{y})$. In other words, the mean

values ($\bar{x}, \bar{y}$) can be obtained as the point of intersection of the two regression lines.

3. *Why two lines of Regression ?* There are always two lines of regression, one of Y on X and the other of X on Y. The line of regression of Y on X (10·14*a*) is used to estimate or predict the value of Y for any given value of X, *i.e.*, when Y is a dependent variable and X is an independent variable. The estimate so obtained will be best in the sense that it will have the minimum possible error as defined by the principle of least squares. We can also obtain an estimate of X for any given value of Y by using equation (10·14*a*) but the estimate so obtained will not be best since (10·14*a*) is obtained on minimising the sum of the squares of errors of estimates in Y and not in X. Hence to estimate or predict X for any given value of Y, we use the regression equation of X on Y (10·15*a*) which is derived on minimising the sum of the squares of errors of estimates in X. Here X is a dependent variable and Y is an independent variable. The two regression equations are not reversible or interchangeable because of the simple reason that the basis and assumptions for deriving these equations are quite different. The regression equation of Y on X is obtained on minimising the sum of the squares of the errors parallel to the Y-axis while the regression equation of X on Y is obtained on minimising the sum of squares of the errors parallel to the X-axis.

In a particular case of perfect correlation, positive or negative, *i.e.*, $r \pm 1$, the equation of line of regression of Y on X becomes :

$$Y - \bar{y} = \pm \frac{\sigma_Y}{\sigma_X}(X - \bar{x})$$

$$\Rightarrow \qquad \frac{Y - \bar{y}}{\sigma_Y} = \pm \left(\frac{X - \bar{x}}{\sigma_X}\right) \qquad \ldots(10\cdot16)$$

Similarly, the equation of the line of regression of X on Y becomes :

$$X - \bar{x} = \pm \frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

$$\Rightarrow \qquad \frac{Y - \bar{y}}{\sigma_Y} = \pm \left(\frac{X - \bar{x}}{\sigma_X}\right),$$

which is same as (10·16).

Hence in case of perfect correlation, ($r = \pm 1$), both the lines of regression coincide. Therefore, in general, we always have two lines of regression except in the particular case of perfect correlation when both the lines coincide and we get only one line.

10·7·2. Regression Curves. In modern terminology, the conditional mean $E(Y \mid X = x)$ for a continuous distribution is called the regression function of Y on X and the graph of this function of x is known as the regression curve of Y on X or sometimes the regression curve for the mean of Y. Geometrically, the regression function represents the y co-ordinate of the centre of mass of the bivariate probability mass in the infinitesimal vertical strip bounded by x and $x + dx$.

Similarly, the regression function of X on Y is $E(X \mid Y = y)$ and the graph of this function of y is called the regression curve (of the mean) of X on Y.

In case a regression curve is a straight line, the corresponding regression is said to be *linear*. If one of the regressions is linear, it does not however follow that the other is also linear. For illustration, See Example 10·21.

Theorem 10·4. *Let (X, Y) be a two-dimensional random variable with $E(X) = \bar{X}$, $E(Y) = \bar{Y}$, $V(X) = \sigma_X^2$, $V(Y) = \sigma_Y^2$ and let $r = r(X, Y)$ be the correlation coefficient between X and Y. If the regression of Y on X is linear then*

$$E(Y \mid X) = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X}) \qquad \ldots(10\cdot16a)$$

Similarly, if the regression of X on Y is linear, then

$$E(X \mid Y) = \bar{X} + r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y}) \qquad \ldots(10\cdot16b)$$

Proof. Let the regression equation of Y on X be

$$E(Y \mid x) = a + bx \qquad \ldots(1)$$

But by definition,

$$E(Y \mid x) = \int_{-\infty}^{\infty} y\, f(y \mid x)\, dy = \int_{-\infty}^{\infty} y\, \frac{f(x, y)}{f_X(x)}\, dy$$

$$\therefore \qquad \frac{1}{f_X(x)} \int_{-\infty}^{\infty} y\, f(x, y)\, dy = a + bx \qquad \ldots(2)$$

Multiplying both sides of (2) by $f_X(x)$ and integrating w.r.t. x, we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x, y)\, dy dx = a\int_{-\infty}^{\infty} f_X(x)\, dx + b\int_{-\infty}^{\infty} x\, f_X(x) dx$$

$$\Rightarrow \qquad \int_{-\infty}^{\infty} y\, \left[\int_{-\infty}^{\infty} f(x, y)\, dx\right] dy = a + bE(X)$$

$$\Rightarrow \qquad \int_{-\infty}^{\infty} y\, f_Y(y) dy = a + bE(X)$$

i.e., $\qquad E(Y) = a + bE(X) \quad \Rightarrow \quad \bar{Y} = a + b\bar{X} \qquad \ldots(3)$

Multiplying both sides of (2) by $x\, f_X(x)$ and integrating w.r.t. x, we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x, y)\, ay\, dx = a\int_{-\infty}^{\infty} x\, f_X(x)\, dx + b\int_{-\infty}^{\infty} x^2 f_X(x)\, dx$$

$$\Rightarrow \qquad E(XY) = a\, E(X) + b\, E(X^2)$$

$$\Rightarrow \qquad \mu_{11} + \bar{X}\,\bar{Y} = a\bar{X} + b(\sigma_X^2 + \bar{X}^2) \qquad \ldots(4)$$

$$[\because \mu_{11} = E(XY) - E(X)E(Y) \quad = E(XY) - \bar{X}\,\bar{Y}\ ;$$

$$\sigma_X^2 = E(X^2) - \{E(X)\}^2 = E(X^2) - \bar{X}^2$$

Solving (3) and (4) simultaneously, we get

$$b = \frac{\mu_{11}}{\sigma_X^2} \quad \text{and} \quad a = \bar{Y} - \frac{\mu_{11}}{\sigma_X^2}\bar{X}$$

Substituting in (1) and simplifying, we get the required equation of the line of regression of Y on X as

$$E(Y \mid x) = \bar{Y} + \frac{\mu_{11}}{\sigma_X^2}(x - \bar{X})$$

$$\Rightarrow \qquad E(Y \mid X) = \bar{Y} + \frac{\mu_{11}}{\sigma_X^2}(X - \bar{X})$$

$$\Rightarrow \qquad E(Y \mid X) = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

By starting with the line $E(X \mid y) = A + By$ and proceeding similarly we shall obtain the equation of the line of regression of X on Y as

$$E(X \mid Y) = \bar{X} + \frac{\mu_{11}}{\sigma_Y^2}(Y - \bar{Y}) = \bar{X} + r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

Example 10·19. *Given*

$$f(x, y) = xe^{-x(y+1)};\ x \geq 0,\ y \geq 0,$$

find the regression curve of Y on X. **[B.H. Univ. M.Sc., 1989]**

Solution. Marginal p.d.f. of X is given by

$$f_1(x) = \int_0^\infty f(x, y)\,dy = \int_0^\infty xe^{-x(y+1)}\,dy$$

$$= xe^{-x}\int_0^\infty e^{-xy}\,dy = xe^{-x}\left|\frac{e^{-xy}}{-x}\right|_0^\infty$$

$$= e^{-x},\ x \geq 0$$

Conditional p.d.f. of Y on X is given by

$$f(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{xe^{-x(y+1)}}{e^{-x}} = xe^{-xy},\ y \geq 0.$$

The regression curve of Y on X is given by

$$y = E(Y \mid X = x) = \int_0^\infty y\,f(y \mid x)\,dy = \int_0^\infty yxe^{-xy}\,dy$$

$$= x\left[\left| \frac{ye^{-xy}}{-x} \right|_0^\infty + \int_0^\infty \frac{e^{-xy}}{x} dy\right] = 0 + \left| \frac{e^{-xy}}{-x} \right|_0^\infty = \frac{1}{x}$$

i.e., $$y = \frac{1}{x} \quad \Rightarrow \quad xy = 1.$$

which is the equation of a rectangular hyperbola. Hence the regression of Y on X is not linear.

Example 10·20. *Obtain the regression equation of Y on X for the following distribution* :

$$f(x, y) = \frac{y}{(1+x)^4} \exp\left(-\frac{y}{1+x}\right); x, y \geq 0$$

Solution. Marginal p.d.f. of X is given by

$$f_1(x) = \int_0^\infty f(x, y)\, dy = \frac{1}{(1+x)^4} \int_0^\infty y e^{-y/(1+x)} dy$$

$$= \frac{1}{(1+x)^4} . \Gamma 2 . (1+x)^2 \qquad \text{(Using Gamma Integral)}$$

$$= \frac{1}{(1+x)^2}; \; x \geq 0$$

The conditional p.d.f. of Y (for given X) is

$$f(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{y}{(1+x)^2} \exp\left(-\frac{y}{1+x}\right); y \geq 0$$

Regression equation of Y on X is given by

$$y = E(Y \mid X) = \int_0^\infty y f(y \mid x)\, dy = \frac{1}{(1+x)^2} \int_0^\infty y^2 e^{-y/(1+x)} dx$$

$$= \frac{1}{(1+x)^2} . \; \Gamma 3 . \; (1+x)^3 \qquad \text{[Using Gamma Integral]}$$

$$\Rightarrow \qquad y = 2(1+x) \qquad [\because \Gamma 3 = 2! = 2]$$

Hence the regression of Y on X is linear.

Example 10·21. *Let (X, Y) have the joint p.d.f. given by*

$$f(x, y) = \begin{cases} 1, \text{if } |y| < x, \; 0 < x < 1 \\ 0, \text{ otherwise} \end{cases}$$

Show that the regression of Y on X is linear but regression of X on Y is not linear.

Solution. $|y| < x \Rightarrow -x < y < x$ and $x > |y|$.

The marginal p.d.f's $f_1(.)$ of X and $f_2(.)$ of Y are given by :

$$f_1(x) = \int_{-x}^{x} f(x, y)\, dy = \int_{-x}^{x} 1 . dy = 2x \; ; 0 < x < 1$$

$$f_2(y) = \int_{|y|}^{1} f(x, y)\, dx = \int_{|y|}^{1} 1 . dx = 1 - |y| \; ; -1 < y < 1$$

$$\therefore \quad f_1(x \mid y) = \frac{f(x, y)}{f_2(y)} = \frac{1}{1 - |y|}; -1 \le y < 1, 0 < x < 1$$

$$= \begin{cases} \dfrac{1}{1-y}, & 0 < y < 1 \; ; \; 0 < x < 1 \\ \dfrac{1}{1+y}, & -1 < y < 0 \; ; \; 0 < x < 1 \end{cases}$$

$$f_2(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{1}{2x}, 0 < x < 1; |y| < x$$

$$E(Y \mid X = x) = \int_{-x}^{x} y \cdot f_2(y \mid x)\, dy = \int_{-x}^{x} \frac{y}{2x} dy = \frac{1}{4x} \cdot \left| y^2 \right|_{-x}^{x} = 0$$

Hence the curve of regression of Y on X is $y = 0$, which is a straight line.

$$E(X|Y = y) = \int x f_1(x \mid y) dx$$

$$\therefore \quad E(X \mid Y = y) = \int_0^1 x \left(\frac{1}{1-y} \right) dx = \frac{1}{2(1-y)}, 0 < y < 1$$

and $$E(X|Y = y) = \int_0^1 x \left(\frac{1}{1+y} \right) dx = \frac{1}{2(1+y)}, -1 < y < 0$$

Hence the curve of regression of X on Y is

$$x = \begin{cases} \dfrac{1}{2(1-y)}, & 0 < y < 1 \\ \dfrac{1}{2(1+y)}, & -1 < y < 0, \end{cases}$$

which is not a straight line.

Example 10·22. *Variables X and Y have the joint p.d.f.*

$$f(x, y) = \tfrac{1}{3}(x + y), \; 0 \le x \le 1, \; 0 \le y \le 2.$$

Find :

(i) $r(X, Y)$

(ii) The two lines of regression

(iii) The two regression curves for the means.

Solution. The marginal p.d.f.'s of X and Y are given by :

$$f_1(x) = \int_0^2 f(x, y)\, dy = \frac{1}{3} \int_0^2 (x + y)\, dy = \frac{1}{3} \left| xy + \frac{y^2}{2} \right|_0^2$$

$$\Rightarrow \quad f_1(x) = \frac{2}{3}(1 + x) \; ; \; 0 \le x \le 1 \qquad \ldots(1)$$

$$f_2(y) = \int_0^1 f(x, y)\, dx = \frac{1}{3} \int_0^1 (x + y)\, dx = \frac{1}{3} \left| \frac{x^2}{2} + xy \right|_0^1$$

$$\Rightarrow \quad f_2(y) = \frac{1}{3}\left(\frac{1}{2} + y \right); 0 \le y \le 2 \qquad \ldots(2)$$

The conditional distributions are given by :

$$f_3(y \mid x) = \frac{f(x, y)}{f_1(x)} = \frac{1}{2}\left(\frac{x + y}{1 + x}\right)$$

$$f_4(x \mid y) = \frac{f(x, y)}{f_2(y)} = \frac{2(x+ y)}{1 + 2y} \qquad \ldots(3)$$

$$E(Y|x) = \int_0^2 y \,.\, f_3(y \mid x)\, dy = \frac{1}{2(1 + x)} \int_0^2 y(x + y)dy$$

$$= \frac{1}{2(1 + x)} \left|\frac{xy^2}{2} + \frac{y^3}{3}\right|_{y=0}^{y=2} = \frac{3x + 4}{3(x + 1)}$$

Similarly, we shall get

$$E(X \mid y) = \int_0^1 x f_4(x \mid y)\, dx = \frac{2}{1 + 2y} \int_0^1 (x^2 + xy)\, dx = \frac{2 + 3y}{3(1 + 2y)}$$

(*iii*) Hence the regression curves for means are :

$$y = E(Y|x) = \frac{3x + 4}{3(x + 1)} \quad \text{and} \quad x = E(X|y) = \frac{2 + 3y}{3(1 + 2y)}.$$

From the marginal distributions we shall get

$$E(X) = \int_0^1 x\, f_1(x)\, dx = \frac{5}{9}, \quad E(X^2) = \int_0^1 x^2 f_1(x)dx = \frac{7}{18}$$

$$\Rightarrow \quad \text{Var}(X) = \sigma_X^2 = \frac{7}{18} - \left(\frac{5}{9}\right)^2 = \frac{13}{162}$$

Similarly, we shall get

$$E(Y) = \frac{11}{9}, E(Y^2) = \frac{16}{9}; \sigma_Y^2 = \frac{16}{9} - \left(\frac{11}{9}\right)^2 = \frac{23}{81}$$

Also $$E(XY) = \int_0^1 \int_0^2 xy\, f(x, y)\, dxdy = \frac{1}{3} \int_0^1 \int_0^2 (x^2y + xy^2)\, dxdy$$

$$= \frac{1}{3}\left\{\left(\int_0^1 x^2dx\right)\left(\int_0^2 y\, dy\right) + \left(\int_0^1 x\, dx\right)\left(\int_0^2 y^2dy\right)\right\}$$

$$= \frac{1}{3}\left[\frac{1}{3} \times 2 + \frac{1}{2} \times \frac{8}{3}\right] = \frac{2}{3}$$

$$\therefore \quad \text{Cov}(X, Y) = E(XY) - E(X)\, E(Y) = \frac{2}{3} - \frac{5}{9} \times \frac{11}{9} = -\frac{1}{81}$$

(*i*) $$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \,.\, \sigma_Y} = \frac{-\frac{1}{81}}{\sqrt{\frac{13}{162} \times \frac{23}{81}}} = -\left(\frac{2}{299}\right)^{1/2}$$

(*ii*) The two lines of regression are :

$$Y - E(Y) = \frac{\text{Cov}(X, Y)}{\sigma_X^2}[X - E(X)] \quad \Rightarrow \quad Y - \frac{11}{9} = -\frac{2}{13}\left(X - \frac{5}{9}\right)$$

and $X - E(X) = \frac{\text{Cov}(X, Y)}{\sigma_Y^2}[Y - E(Y)] \quad \Rightarrow \quad X - \frac{5}{9} = -\frac{1}{23}\left(Y - \frac{11}{9}\right)$

10·7·3. Regression Coefficients. 'b', the slope of the line of regression of Y on X is also called the coefficient of regression of Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X. More precisely, we write

$$b_{YX} = \text{Regression coefficient of } Y \text{ on } X = \frac{\mu_{11}}{\sigma_X^2} = r\frac{\sigma_Y}{\sigma_X} \qquad \ldots(10\cdot17)$$

Similarly, the coefficient of regression of X on Y indicates the change in the value of variable X corresponding to a unit change in the value of variable Y and is given by

$$b_{XY} = \text{Regression coefficient of } X \text{ on } Y = \frac{\mu_{11}}{\sigma_Y^2} = r\frac{\sigma_X}{\sigma_Y} \qquad \ldots(10\cdot17a)$$

10·7·4. Properties of Regression Coefficients.

(*a*) *Correlation coefficient is the geometric mean between the regression coefficients.*

Proof. Multiplying (10·17) and (10·17*a*), we get

$$b_{XY} \times b_{YX} = r\frac{\sigma_X}{\sigma_Y} \times r\frac{\sigma_Y}{\sigma_X} = r^2$$

$$\therefore \qquad r = \pm\sqrt{b_{XY} \times b_{YX}} \qquad \ldots(10\cdot18)$$

Remark. We have

$$r = \frac{\mu_{11}}{\sigma_X \cdot \sigma_Y}, \quad b_{YX} = \frac{\mu_{11}}{\sigma_X^2} \quad \text{and} \quad b_{XY} = \frac{\mu_{11}}{\sigma_Y^2}$$

It may be noted that the sign of correlation coefficient is the same as that of regression coefficients, since the sign of each depends upon the co-variance term μ_{11}. Thus if the regression coefficients are positive, 'r' is positive and if the regression coefficients are negative 'r' is negative.

From (10·18), we have

$$r = \pm\sqrt{b_{XY} \times b_{YX}}$$

the sign to be taken before the square root is that of the regression coefficients.

(*b*) *If one of the regression coefficients is greater than unity, the other must be less than unity.*

Proof. Let one of the regression coefficients (say) b_{YX} be greater than unity, then we have to show that $b_{XY} < 1$.

Now $\quad b_{YX} > 1 \Rightarrow \frac{1}{b_{YX}} < 1 \qquad \ldots(*)$

Also $\quad r^2 \le 1 \quad \Rightarrow \quad b_{YX} \cdot b_{XY} \le 1$

Hence $\quad b_{XY} \le \frac{1}{b_{YX}} < 1 \qquad$ [From (*)]

(*c*) *Arithmetic mean of the regression coefficients is greater than the correlation coefficient r, provided r > 0.*

Proof. We have to prove that $\frac{1}{2}(b_{YX} + b_{XY}) \geq r$

or $$\frac{1}{2}\left(r\frac{\sigma_Y}{\sigma_X} + r\frac{\sigma_X}{\sigma_Y}\right) \geq r \quad \text{or} \quad \frac{\sigma_Y}{\sigma_X} + \frac{\sigma_X}{\sigma_Y} \geq 2 \qquad (\because r > 0)$$

$$\Rightarrow \quad \sigma_Y^2 + \sigma_X^2 - 2\sigma_X\sigma_Y \geq 0 \quad i.e., \quad (\sigma_Y - \sigma_X)^2 \geq 0$$

which is always true, since the square of a real quantity is ≥ 0.

(d) Regression coefficients are independent of the change of origin but not of scale.

Proof. Let $U = \dfrac{X-a}{h}$, $V = \dfrac{Y-b}{k} \Rightarrow X = a + hU, Y = b + kV$,

where $a, b, h\ (> 0)$ and $k\ (> 0)$ are constants.

Then Cov $(X, Y) = hk$ Cov (U, V), $\sigma_X^2 = h^2\sigma_U^2$ and $\sigma_Y^2 = k^2\sigma_V^2$

$$b_{YX} = \frac{\mu_{11}}{\sigma_X^2} = \frac{hk \text{ cov } (U, V)}{h^2\sigma_U^2}$$

$$= \frac{k}{h} \cdot \frac{\text{cov } (U, V)}{\sigma_U^2} = \frac{k}{h} b_{VU}$$

Similarly, we can prove that

$$b_{XY} = (h/k)\, b_{UV}$$

10·7·5. Angle Between Two Lines of Regression. Equations of the lines of regression of Y on X, and X on Y are

$$Y - \bar{y} = r \cdot \frac{\sigma_Y}{\sigma_X}(X - \bar{x}) \text{ and } X - \bar{x} = r \cdot \frac{\sigma_X}{\sigma_Y}(Y - \bar{y})$$

Slopes of these lines are $r \cdot \dfrac{\sigma_Y}{\sigma_X}$ and $\dfrac{\sigma_Y}{r\sigma_X}$ respectively. If θ is the angle between the two lines of regression then

$$\tan\theta = \frac{r\dfrac{\sigma_Y}{\sigma_X} \sim \dfrac{\sigma_Y}{r\sigma_X}}{1 + r\dfrac{\sigma_Y}{\sigma_X} \cdot \dfrac{\sigma_Y}{r\sigma_X}} = \frac{r^2 \sim 1}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)$$

$$= \frac{1 - r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right) \qquad (\because r^2 \leq 1)$$

$$\therefore \quad \theta = \tan^{-1}\left\{\frac{1 - r^2}{r}\left(\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2}\right)\right\} \qquad \ldots(10\cdot19)$$

Case (i). $(r = 0)$. If $r = 0$, $\tan\theta = \infty \Rightarrow \theta = \dfrac{\pi}{2}$

Thus if the two variables are uncorrelated, the lines of regression become perpendicular to each other.

Case (ii). $(r = \pm 1)$. If $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0$ or π.

In this case the two lines of regression either coincide or they are parallel to each other. But since both the lines of regression pass through the point

$(\bar{x}, \bar{y})$, they cannot be parallel. Hence in the case of perfect correlation, positive or negative, the two lines of regression coincide.

Remarks 1. Whenever two lines intersect, there are two angles between them, one acute angle and the other obtuse angle. Further $\tan\theta > 0$ if $0 < \theta < \pi/2$, *i.e.*, θ is an acute angle and $\tan\theta < 0$ if $\pi/2 < \theta < \pi$, *i.e.*, θ is an obtuse angle and since $0 < r^2 < 1$, the acute angle (θ_1) and obtuse angle θ_2 between the two lines of regression are given by

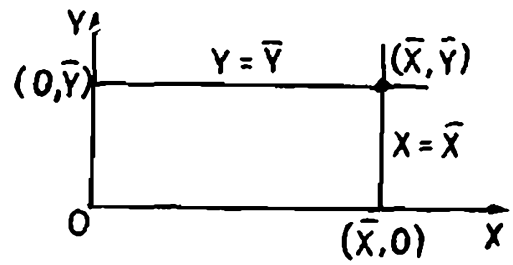
$$\theta_1 = \text{Acute angle} = \tan^{-1}\left\{\frac{\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2} \cdot \frac{1 - r^2}{r}\right\}, r > 0$$

and
$$\theta_2 = \text{Obtuse angle} = \tan^{-1}\left\{\frac{\sigma_X \cdot \sigma_Y}{\sigma_X^2 + \sigma_Y^2} \cdot \frac{r^2 - 1}{r}\right\}, r > 0$$

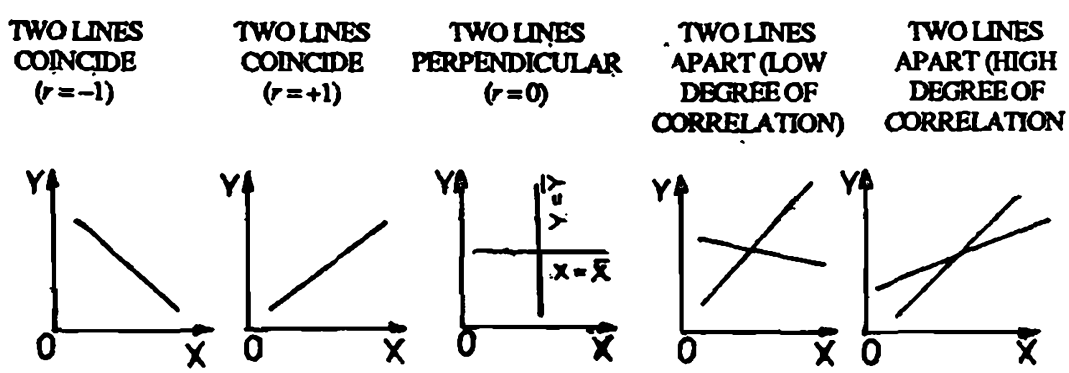
2. When $r = 0$, *i.e.*, variables X and Y are uncorrelated, then the lines of regressions of Y on X and X on Y are given respectively by : [From (10·14*a*) and (10·15*a*)]

$$Y = \bar{Y} \text{ and } X = \bar{X},$$

as shown in the adjoining diagram. Hence, in this case ($r = 0$), the lines of regression are perpendicular to each other and are parallel to X- axis and Y-axis respectively.



3. The fact that if $r = 0$ (variables uncorrelated), the two lines of regression are perpendicular to each and if $r = \pm 1$, $\theta = 0$, *i.e.*, the two lines coincide, leads us to the conclusion that for higher degree of correlation between the variables, the angle between the lines is smaller, *i.e.*, the two lines of regression are nearer to each other. On the other hand, if the lines of regression make a larger angle, they indicate a poor degree of correlation between the variables and ultimately for $\theta = \pi/2$, $r = 0$, *i.e.*, the lines become perpendicular if no correlation exists between the variables. Thus by plotting the lines of regression on a graph paper, we can have an approximate idea about the degree of correlation between the two variables under study. Consider the following illustrations :



10·7·6. Standard Error of Estimate or Residual Variance. The equation of the line of regression of Y on X is

$$Y = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$\Rightarrow \qquad \frac{Y - \bar{Y}}{\sigma_Y} = r\left(\frac{X - \bar{X}}{\sigma_X}\right)$$

The residual variance s_Y^2 is the expected value of the squares of deviations of the observed values of Y from the expected values as given by the line of regression of Y on X. Thus

$$s_Y^2 = E[Y - \{\bar{Y} + (r\sigma_Y (X - \bar{X})/\sigma_X)\}]^2$$

$$= \sigma_Y^2 E\left[\frac{Y - \bar{Y}}{\sigma_Y} - r\left(\frac{X - \bar{X}}{\sigma_X}\right)\right]^2 = \sigma_Y^2 E(Y^* - rX^*)^2$$

where X^* and Y^* are standardised variates so that

$$E(X^{*2}) = 1 = E(Y^{*2}) \text{ and } E(X^* Y^*) = r.$$

$$\therefore \qquad s_Y^2 = \sigma_Y^2[E(Y^{*2}) + r^2 E(X^{*2}) - 2r E(X^*Y^*)] = \sigma_Y^2 (1 - r^2)$$

$$\Rightarrow \qquad s_Y = \sigma_Y (1 - r^2)^{1/2}$$

Similarly, the standard error of estimate of X is given by

$$s_X = \sigma_X (1 - r^2)^{1/2}$$

Remarks 1. Since s_X^2 or $s_Y^2 \geq 0$, it follows that

$$(1 - r^2) \geq 0 \quad \Rightarrow \quad |r| \leq 1$$

Hence $\qquad -1 \leq r(X, Y) \leq 1$

2. If $r = \pm 1$, $s_X = s_Y = 0$ so that each deviation is zero, and the two lines of regression are coincident.

3. Since, as $r^2 \to 1$, s_X and $s_Y \to 0$, it follows that departure of the value r^2 from unity indicates the departure of the relationship between the variables X and Y from linearity.

4. From the definition of linear regression, the minima condition implies that s_Y or s_X is the minimum variance.

10·7·7. Correlation Coefficient between Observed and Estimated Value. Here we will find the correlation between Y and

$$\hat{Y} = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

where $\hat{Y}$ is the estimated value of Y as given by the line of regression of Y on X, which is given by

$$r(Y, \hat{Y}) = \frac{\text{Cov}(\hat{Y}, Y)}{\sigma_Y \hat{\sigma}_Y}$$

We have

$$E(\hat{Y}) = E\left[\bar{Y} + r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})\right] = \bar{Y} + r\frac{\sigma_Y}{\sigma_X}E(X - \bar{X}) = \bar{Y}$$

$$\therefore \qquad \text{Var}(\hat{Y}) = E[\hat{Y} - E(\hat{Y})]^2 = E\left[r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})\right]^2 = r^2\sigma_Y^2$$

$$\Rightarrow \qquad \hat{\sigma}_Y = r\,\sigma_Y$$

Also Cov $(Y, \hat{Y}) = E[\{Y - E(Y)\}\ \{\hat{Y} - E(\hat{Y})\}]$

$$= E\left[\{b(X - E(X))\}\left\{r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})\right\}\right]$$

$$= br\frac{\sigma_Y}{\sigma_X}E[\{X - E(X)\}^2] = \left(r\frac{\sigma_Y}{\sigma_X}\right)^2\sigma_X^2 = r^2\sigma_Y^2$$

$$\therefore \qquad r(Y, \hat{Y}) = \frac{r^2\sigma_Y^2}{\sigma_Y\, r\, \sigma_Y} = r = r(X, Y)$$

Hence the correlation coefficient between observed and estimated value of Y is the same as the correlation coefficient between X and Y.

Example 10·23. *Obtain the equations of the lines of regression for the data in Example 10·1. Also obtain the estimate of X for $Y = 70$.*

Solution. Let $U = X - 68$ and $V = Y - 69$, then

$\bar{U} = 0$, $\bar{V} = 0$, $\sigma_U^2 = 4{\cdot}5$, $\sigma_V^2 = 5{\cdot}5$, Cov $(U, V) = 3$ and $r\,(U, V) = 0{\cdot}6$

Since correlation coefficient is independent of change of origin, we get

$$r = r(X, Y) = r(U, V) = 0{\cdot}6$$

We know that if $U = \dfrac{X - a}{h}$, $V = \dfrac{Y - b}{k}$, then

$$\bar{X} = a + h\bar{U},\ \bar{Y} = b + k\bar{V},\ \sigma_X = h\sigma_U \text{ and } \sigma_Y = k\,\sigma_Y$$

In our case $h = k = 1$, $a = 68$ and $b = 69$.

Thus $\quad \bar{X} = 68 + 0 = 68$, $\bar{Y} = 69 + 0 = 69$

$$\sigma_X = \sigma_U = \sqrt{4{\cdot}5} = 2{\cdot}12 \text{ and } \sigma_Y = \sigma_V = \sqrt{5{\cdot}5} = 2{\cdot}35$$

Equation of line of regression of Y on X is

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

i.e., $Y = 69 + 0{\cdot}6 \times \dfrac{2{\cdot}35}{2{\cdot}12}(X - 68) \Rightarrow Y = 0{\cdot}665\,X + 23{\cdot}78$

Equation of line of regression of X on Y is

$$X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

$\Rightarrow \quad X = 68 + 0{\cdot}6 \times \dfrac{2{\cdot}12}{2{\cdot}35}(Y - 69)$ *i.e.*, $X = 0{\cdot}54Y + 30{\cdot}74$

To estimate X for given Y, we use the line of regression of X on Y. If $Y = 70$, estimated value of X is given by

$$\hat{X} = 0{\cdot}54 \times 70 + 30{\cdot}74 = 68{\cdot}54,$$

where $\hat{X}$ is estimate of X.

Example 10·24. *In a partially destroyed laboratory record of an analysis of correlation data, the following results only are legible :*

Variance of X = 9.

Regression equations : 8X – 10Y + 66 = 0, 40X – 18Y = 214.

What were (i) the mean values of X and Y,

(ii) the correlation coefficient between X and Y, and

(iii) the standard deviation of Y ?

[Punjab Univ. B.Sc. (Hons.), 1993]

Solution (*i*) Since both the lines of regression pass through the point $(\bar{X}, \bar{Y})$, we have $8\bar{X} - 10\bar{Y} + 66 = 0$, and $40\bar{X} - 18\bar{Y} = 214$.

Solving, we get $\bar{X} = 13, \bar{Y} = 17$.

(*ii*) Let $8X - 10Y + 66 = 0$ and $40X - 18Y = 214$ be the lines of regression of Y on X and X on Y respectively. These equations can be put in the form :

$$Y = \frac{8}{10}X + \frac{66}{10} \text{ and } X = \frac{18}{40}Y + \frac{214}{40}$$

$\therefore$ b_{YX} = Regression coefficient of Y on $X = \frac{8}{10} = \frac{4}{5}$

and b_{XY} = Regression coefficient of X on $Y = \frac{18}{40} = \frac{9}{20}$

Hence $r^2 = b_{YX} \cdot b_{XY} = \frac{4}{5} \cdot \frac{9}{20} = \frac{9}{25}$

$\therefore$ $r = \pm\frac{3}{5} = \pm 0{\cdot}6$

But since both the regression coefficients are positive, we take $r = +0{\cdot}6$

(*iii*) We have $b_{YX} = r \cdot \frac{\sigma_Y}{\sigma_X} \Rightarrow \frac{4}{5} = \frac{3}{5} \times \frac{\sigma_Y}{3}$ $[\because \sigma_X^2 = 9 \text{ (Given)}]$

Hence $\sigma_Y = 4$

Remarks. 1. It can be verified that the values of $\bar{X} = 13$ and $\bar{Y} = 17$ as obtained in part (*i*) satisfy both the regression equations. In numerical problems of this type, this check should invariably be applied to ascertain the correctness of the answer.

2. If we had assumed that $8X - 10Y + 66 = 0$, is the equation of the line of regression of X on Y and $40X - 18Y = 214$ is the equation of line of regression of Y on X, then we get respectively :

$$8X = 10Y - 66 \text{ and } 18Y = 40X - 214$$

$$\Rightarrow \quad X = \frac{10}{8}Y - \frac{66}{8} \text{ and } Y = \frac{40}{18}X - \frac{214}{18}$$

$$\Rightarrow \quad b_{XY} = \frac{18}{8} \text{ and } b_{YX} = \frac{40}{18}$$

$$\therefore \quad r^2 = b_{XY} \cdot b_{YX} = \frac{10}{8} \times \frac{40}{18} = 2{\cdot}78$$

But since r^2 always lies between 0 and 1, our supposition is wrong.

Example 10·25. *Find the most likely price in Bombay corresponding to the price of Rs. 70 at Calcutta from the following :*

	Calcutta	*Bombay*
Average price	65	67
Standard deviation	2·5	3·5

Correlation coefficient between the prices of commodities in the two cities is 0·8. [***Nagpur Univ. B.Sc., 1993; Sri Venkateswara Univ. B.Sc. (Oct.) 1990***]

Solution. Let the prices, (in Rupees), in Bombay and Calcutta be denoted by Y and X respectively. Then we are given

$\bar{X} = 65$, $\bar{Y} = 67$, $\sigma_X = 2{\cdot}5$, $\sigma_Y = 3{\cdot}5$ and $r = r(X, Y) = 0{\cdot}8$. We want Y for $X = 70$.

Line of regression of Y on X is

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$\Rightarrow \qquad Y = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5}(X - 65)$$

When $X = 70$, $\quad \hat{Y} = 67 + 0{\cdot}8 \times \frac{3{\cdot}5}{2{\cdot}5}(70 - 65) = 72{\cdot}6$

Example 10·26. *Can $Y = 5 + 2{\cdot}8\,X$ and $X = 3 - 0{\cdot}5Y$ be the estimated regression equations of Y on X and X on Y respectively ? Explain your answer with suitable theoretical arguments.* [***Delhi Univ. M.A.(Eco.), 1986***]

Solution. Line of regression of Y on X is :

$$Y = 5 + 2{\cdot}8X \quad \Rightarrow \quad b_{YX} = 2{\cdot}8 \qquad \ldots(*)$$

Line of regression of X on Y is :

$$X = 3 - 0{\cdot}5Y \quad \Rightarrow \quad b_{XY} = -0{\cdot}5 \qquad \ldots(**)$$

This is not possible, since each of the regression coefficients b_{YX} and b_{XY} must have the same sign, which is same as that of Cov (X, Y). If Cov (x, y) is positive, then both the regression coefficients are positive and if Cov (X, Y) is negative, then both the regression coefficients are negative. Hence (*) and (**) cannot be the estimated regression equations of Y on X and X on Y respectively.

EXERCISE 10 (d)

1. (*a*) Explain what are regression lines. Why are there two such lines ? Also derive their equations.

(*b*) Define (*i*) Line of regression, (*ii*) Regression coefficient. Find the equations to the lines of regression and show that the coefficient of correlation is the geometric mean of coefficients of regression.

(*c*) What equation is the equivalent mathematical statement for the following words ?

"If the respective deviations in each series, X and Y, from their means were expressed in units of standard deviations, *i.e.*, if each were divided by the

standard deviation of the series, to which it belongs and plotted to a scale of standard deviations, the slope of a straight line best describing the plotted points would be the correlation coefficient r."

2(*a*) Obtain the equation of the line of regression of Y on X and show that the angle θ, between the two lines of regression is given by

$$\tan\theta = \frac{1-\rho^2}{\rho} \times \frac{\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}$$

where σ_1, σ_2 are the standard deviations of X and Y respectively, and ρ is the correlation coefficient. ***(Delhi Univ. B.Sc. (Maths. Hons.), 1989)***

Interpret the cases when $\rho = 0$ and $\rho = \pm 1$.

(Bangalore Univ. B.Sc. 1990)

(*b*) If θ is the acute angle between the two regression lines with correlation coefficient r, show that $\sin\theta \le 1 - r^2$.

3. (*a*) Explain the term "regression" by giving examples. Assuming that the regression of Y on X is linear, outline a method for the estimation of the coefficients in the regression line based on the random paired sample of X and Y, and show that the variance of the error of the estimate for Y for the regression line is $\sigma_Y^2(1-\rho^2)$, where σ_Y^2 is the variance of Y and ρ is the correlation coefficient between X and Y.

(*b*) Prove that X and Y are linearly related if and only if $\rho_{XY}^2 = 1$. Further show that the slope of the regression line is positive or negative according as $\rho = +1$ or $\rho = -1$.

(*c*) Let X and Y be two variates. Define $X^* = \dfrac{X-a}{b}$, $Y^* = \dfrac{Y-c}{d}$ for some constants a, b, c and d. Show that the regression line (least square) of Y on X can be obtained from that of Y^* on X^*.

(*d*) Show that the coefficient of correlation between the observed and the estimated values of Y obtained from the line of regression of Y on X, is the same as that between X and Y.

4. Two variables X and Y are known to be related to each other by the relation $Y = X/(aX+b)$. How is the theory of linear regression to be employed to estimate the constants a and b from a set of n pairs of observations (x_i, y_i), $i = 1, 2, \ldots, n$?

Hint. $$\frac{1}{Y} = \frac{aX+b}{X} = a + \frac{b}{X}$$

Put $$\frac{1}{X} = U \text{ and } \frac{1}{Y} = V$$

$\therefore$ $$V = a + bU$$

5. Derive the standard error of estimate of Y obtained from the linear regression equation of Y on X. What does this standard error measure ?

6. (*a*) Calculate the coefficient of correlation from the following data :

X :	1	2	3	4	5	6	7	8	9
Y :	9	8	10	12	11	13	14	16	15

Also obtain the equations of the lines of regression and obtain an estimate of Y which should correspond on the average to $X = 6{\cdot}2$.

Ans. $r = 0{\cdot}95$, $Y - 12 = 0{\cdot}95\,(X - 5)$, $X - 5 = 0{\cdot}95\,(Y - 12)$, $13{\cdot}14$

(*b*) Why do we have, in general, two lines of regression ? Obtain the regression of Y on X, and X on Y from the following table and estimate the blood pressure when the age is 45 years :

Age in years (*X*)	*Blood pressure* (*Y*)	*Age in years* (*X*)	*Blood pressure* (*Y*)
56	147	55	150
42	125	49	145
72	160	38	115
36	118	42	140
63	149	68	152
47	128	60	155

Ans. $Y = 1{\cdot}138X + 80{\cdot}778$, $Y = 131{\cdot}988$ for $X = 45$.

(*c*) Suppose the observations on X and Y are given as :

X :	59	65	45	52	60	62	70	55	45	49
Y :	75	70	55	65	60	69	80	65	59	61

where $N = 10$ students, and Y = Marks in Maths, X = Marks in Economics. Compute the least square regression equations of Y on X and of X on Y.

If a student gets 61 marks in Economics, what would you estimate his marks in Maths to be ?

7. (*a*) In a correlation analysis on the ages of wives and husbands, the following data were obtained. Find

(*i*) the value of the correlation coefficient, and (*ii*) the lines of regression.

Estimate the age of husband whose wife's age is 31 years. Estimate the age of wife whose husband is 40 years old.

Age of wife → / Age of Husband ↓	15—25	25—35	35—45	45—55	55—65
15—30	30	6	3	—	—
30—45	18	32	15	12	8
45—60	2	28	40	16	9
60—75	—	4	9	10	8

(*b*) The following table gives the distribution of total cultivable area (X) and area under cultivation (Y) in a district of 69 villages.

Calculate (*i*) the linear regression of Y on X,

(*ii*) the correlation coefficient $r(X, Y)$, and (*iii*) the average area under wheat corresponding to total area of 1,000 Bighas.

		Total area (in Bighas)				
		0—500	500—1000	1000—1500	1500—2000	2000—2500
Area under wheat	0—200	12	6	...	...	...
	200—400	2	18	4	2	1
	400—600	...	4	7	3	...
	600—800	...	1	...	2	1
	800—1000	...	...	1	2	3

Ans. (*i*) $Y = 0{\cdot}7641X - 455{\cdot}3854$, (*ii*) $r(X, Y) = 0{\cdot}756$

(*iii*) $Y = 308{\cdot}7146$ for $X = 1000$

8. (*a*) Compare and contrast the roles of correlation and regression in studying the inter-dependence of two variates.

For 10 observations on price (X) and supply (Y) the following data were obtained (in appropriate units).

$$\Sigma X = 130, \Sigma Y = 220, \Sigma X^2 = 2288, \Sigma Y^2 = 5506 \text{ and } \Sigma XY = 3467$$

Obtain the line of regression of Y on X and estimate the supply when the price is 16 units, and find out the standard error of the estimate.

Ans. $Y = 8{\cdot}8 + 1{\cdot}015X$, $25{\cdot}04$

(*b*) If a number X is chosen at random from among the integers 1, 2, 3, 4 and a number Y is chosen from among those at least as large as X, prove that

$$\text{Cov}(X, Y) = \frac{5}{8}$$

Find also the regression line of X on Y.

(*c*) Calculate the correlation coefficient from the following data :—

$$N = 100, \quad \Sigma X = 12500 \quad \Sigma Y = 8000$$
$$\Sigma X^2 = 1585000, \quad \Sigma Y^2 = 648100 \quad \Sigma XY = 1007425.$$

Also obtain the regression equation of Y on X.

9. (*a*) The means of a bivariate frequency distribution are at (3, 4) and $r = 0{\cdot}4$. The line of regression of Y on X is parallel to the line $Y = X$. Find the two lines of regression and estimate the mean of X when $Y = 1$.

(*b*) For certain data, $Y = 1{\cdot}2\,X$ and $X = 0{\cdot}6\,Y$, are the regression lines. Compute $\rho(X, Y)$ and σ_X/σ_Y. Also compute $\rho(X, Z)$, if $Z = Y - X$.

(*c*) The equations of two regression lines obtained in a correlation analysis are as follows :

$$3X + 12Y = 19, \quad 3Y + 9X = 46$$

Obtain (*i*) the value of correlation coefficient,

(*ii*) mean values of X and Y, and

(*iii*) the ratio of the coefficient of variability of X to that of Y.

Ans. (*i*) $-\frac{1}{2}\sqrt{3}$, (*ii*) $\bar{X} = 5$, $\bar{Y} = 1/3$.

(*d*) For an army personnel of strength 25, the regression of weight of kidneys (Y) on weight of heart (X), both measured in ounces is

$$Y - 0{\cdot}399X - 6{\cdot}934 = 0$$

and the regression of weight of heart on weight of kidney is

$$X - 1{\cdot}212Y + 2{\cdot}461 = 0$$

Find the correlation coefficient between X and Y and their mean values. Can you find out the standard deviation of X and Y as well ?

Ans. $r(X, Y) = 0{\cdot}70$, $\bar{X} = 11{\cdot}5086$, $\bar{Y} = 11{\cdot}5261$, No.

(*e*) Find the coefficient of correlation for distribution in which

S.D of $X = 3{\cdot}0$ units

S.D. of $Y = 1{\cdot}4$ units

Coefficient of regression of Y on $X = 0{\cdot}28$.

10. (*a*) Given that $X = 4Y + 5$ and $Y = kX + 4$, are the lines of regression of X on Y and Y on X respectively, show that $0 < 4k < 1$. If $k = \frac{1}{16}$, find the means of the two variables and coefficient of correlation between them.

[*Punjab Univ. B.Sc. (Hons.), 1989*]

Hint.
$$X = 4Y + 5 \Rightarrow b_{XY} = 4$$
$$Y = kx + 4 \Rightarrow b_{YX} = k$$
$$\therefore \quad r^2 = 4k \qquad \ldots(*)$$

But $0 \le r^2 \le 1 \Rightarrow 0 \le 4k \le 1$.

If $k = \frac{1}{16}$, then from (*), we get

$r^2 = 4 \times \frac{1}{16} \Rightarrow r = +\frac{1}{2}$ [Since both the regression coefficient are positive]

For $k = \frac{1}{16}$, the two lines of regression become

$$X = 4Y + 5 \text{ and } Y = \frac{1}{16}X + 4$$

Solving the two equations, we get $\bar{Y} = 5{\cdot}75$, $\bar{X} = 28$.

(*b*) For 50 students of a class the regression equation of marks in Statistics (X) on marks in Mathematics (Y) is $3Y - 5X + 180 = 0$. The mean marks in Mathematics is 44 and variance of marks in Statistics is 9/16th of the variance of marks in Mathematics. Find the mean marks in Statistics and the coefficient of correlation between marks in two subjects.

[*Bangalore Univ. B.Sc., 1989*]

Hint. We are given $n = 50$, $\bar{Y} = 44$

and
$$\sigma_X^2 = \frac{9}{16}\sigma_Y^2 \Rightarrow \frac{\sigma_X}{\sigma_Y} = \frac{3}{4} \qquad \ldots(*)$$

The equation of the line of regression of X on Y is given to be

$$3Y - 5X + 180 = 0 \quad \Rightarrow \quad X = \frac{3}{5}Y + \frac{180}{5}$$

$$\therefore \quad b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{3}{5} \quad \Rightarrow \quad r \cdot \frac{3}{4} = \frac{3}{5} \quad \text{or} \quad r = 0{\cdot}8$$

Since the lines of regression pass through the point $(\bar{X}, \bar{Y})$, we get

$$\bar{X} = \frac{3}{5}\bar{Y} + \frac{180}{5} = \frac{3}{5} \times 44 + 36 = 62{\cdot}4$$

(*c*) Out of the two lines of regression given by

$$X + 2Y - 5 = 0 \text{ and } 2X + 3Y - 8 = 0,$$

which one is the regression line of X on Y?

Use the equations to find the mean of X and the mean of Y. If the variance of X is 12, calculate the variance of Y.

Ans. $\bar{X} = 1, \bar{Y} = 2, \sigma_Y^2 = 4$

(*d*) The lines of regression in a bivariate distribution are :

$$X + 9Y = 7 \text{ and } Y + 4X = \frac{49}{3}$$

Find (*i*) the coefficient of correlation, (*iii*) the ratios $\sigma_X^2 : \sigma_Y^2 : \text{Cov}(X, Y)$, (*iii*) the means of the distribution and (*iv*) $E(X \mid Y = 1)$.

(*e*) Estimate X when $Y = 10$, if the two lines of regression are :

$$X = -\frac{1}{18}Y + \lambda \text{ and } Y = -2x + \mu,$$

(λ, μ) being unknown and the mean of the distribution is at $(-1, 2)$. Also compute r, λ and μ. [*Gujarat Univ. B.Sc., Oct. 1992*]

11. (*a*) The following results were obtained in the analysis of data on yield of dry bark in ounces (Y) and age in years (X) of 200 cinchona plants :

	X	Y
Average	9·2	16·5
Standard deviation	2·1	4·2

Correlation coefficient = +0·84

Construct the two lines of regression and estimate the yield of dry bark of a plant of age 8 years. [*Patna Univ. B.Sc., 1991*]

(*b*) The following data pertain to the marks in subjects A and B in a certain examination :

Mean marks in A = 39·5

Mean marks in B = 47·5

Standard deviation of marks in A = 10·8

Standard deviation of marks in B = 16·8

Coefficient of correlation between marks in A and marks in B = 0·42.

Draw the two lines of regression and explain why there are two regression equations. Give the estimate of marks in B for candidates who secured 50 marks in A.

Ans. $Y = 0{\cdot}65X + 21{\cdot}825$, $X = 0{\cdot}27Y + 26{\cdot}675$ and $Y = 54{\cdot}342$ for $X = 50$

(*c*) You are given the following information about advertising expenditure and sales :

	Advertising Expenditure (X) *(Rs. lakhs)*	*Sales (Y)* *(Rs. lakhs)*
Mean	10	90
s.d.	3	12

Correlation coefficient = 0·8

What should be the advertising budget if the company wants to attain sales target of Rs. 120 lakhs ? [*Delhi Univ. M.C.A., 1990*]

12. Twenty-five pairs of value of variates X and Y led to the following results :

$N = 25$, $\Sigma X = 127$, $\Sigma Y = 100$, $\Sigma X^2 = 760$, $\Sigma Y^2 = 449$ and $\Sigma XY = 500$

A subsequent scrutiny showed that two pairs of values were copied down as :

X	Y
8	14
8	6

X	Y
8	12
6	8

(*i*) Obtain the correct value of the correlation coefficient.

(*ii*) Hence or otherwise, find the correct equations of the two lines of regression.

(*iii*) Find the angle between the regression lines.

Ans. (*i*) $r(X, Y) = -(0{\cdot}64 \times 0{\cdot}15)^{1/2}$,

(*ii*) $X = -0{\cdot}64Y + 7{\cdot}56$, $Y = -0{\cdot}15X + 4{\cdot}75$.

13. Suppose you have n observations :

$$(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$$

on two variables X and Y, and you have fitted a linear regression $Y = a + bX$ by the method of least squares. Denote the 'expected' value of Y by Y^*, and the residual $Y - Y^*$ by e. Find means and variances of Y^* and e, and the correlation co-efficient between (*i*) X and e, (*ii*) Y and e and (*iii*) Y and Y^*. Use these results to bring out the significance and limitations of the correlation coefficient.

Ans. $r(X, e) = 0$, $r(Y, e) = 0$ and $r(Y, Y^*) = r(X, Y)$.

14. (*a*) The regression lines of Y on X and of X on Y are respectively $Y = aX + b$ and $X = cY + d$. Show that

(*i*) Means are $\bar{X} = (bc + d)/(1 - ac)$ and $\bar{Y} = (ad + b)/(1 - ac)$

(*ii*) Correlation coefficient between X and Y is $\sqrt{a\,c}$

(*iii*) The ratio of the standard deviations of X and Y is $\sqrt{c/a}$.

(*b*) For two random variables X and Y with the same mean, the two regression equations are $Y = aX + b$ and $X = \alpha Y + \beta$. Show that $\frac{b}{\beta} = \frac{1 - a}{1 - \alpha}$. Find also the common mean.

[*Punjab Univ.B.Sc. (Maths Hons.), 1992*]

(*c*) If the lines of regression of Y on X and X on Y are respectively $a_1X + b_1Y + c_1 = 0$ and $a_2X + b_2Y + c_2 = 0$, prove that $a_1b_2 \le a_2b_1$.

(*Delhi Univ. B.Sc. (Stat. Hons.), 1989*)

Hint. $r^2 = b_{YX} \cdot b_{XY} \le 1 \Rightarrow \left(-\frac{a_1}{b_1}\right) \times \left(-\frac{b_2}{a_2}\right) = \frac{a_1b_2}{a_2b_1} \le 1$

15. (*a*) By minimising $\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)^2$ for variations in α and p, show that there are two straight lines passing through the mean of the distribution for which the sum of squares of normal deviations has an extreme value. Prove also that their slopes are given by

$$\tan 2\alpha = \frac{2\mu_{11}}{\sigma_X^2 - \sigma_Y^2}$$

Hint. We have to minimize

$$S = \sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)^2 \quad \ldots(1)$$

Equating to zero, the partial derivatives of (1) w.r.t. α and p, we have

$$\frac{\partial S}{\partial \alpha} = 0 = 2 \sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)(-x_i \sin\alpha + y_i \cos\alpha) \quad \ldots(2)$$

$$\frac{\partial S}{\partial p} = 0 = -2 \sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p) \quad \ldots(3)$$

Equation (3) can be written as

$$\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p) = 0 \Rightarrow \bar{x} \cos\alpha + \bar{y} \sin\alpha - p = 0 \quad \ldots(4)$$

From equation (2), we get a quadratic equation which shows that there are two straight lines for extreme values of E.

From equation (4), it becomes clear that both the straight lines pass through the point $(\bar{x}, \bar{y})$.

Again equation (2) can be written as :

$$\sum_{i=1}^{n} f_i (x_i \cos\alpha + y_i \sin\alpha - p)(y_i \cos\alpha - x_i \sin\alpha) = 0$$

$$\Rightarrow \sum_{i=1}^{n} f_i [\cos\alpha (x_i - \bar{x}) + \sin\alpha (y_i - \bar{y})][y_i \cos\alpha - x_i \sin\alpha] = 0$$

[Using (4)]

$$\Rightarrow \cos^2\alpha \sum_{i=1}^{n} f_i y_i (x_i - \bar{x}) - \sin\alpha \cos\alpha \sum_{i=1}^{n} f_i x_i (x_i - \bar{x})$$

$$+ \sin\alpha \cos\alpha \sum_{i=1}^{n} f_i y_i (y_i - \bar{y}) - \sin^2\alpha \sum_{i=1}^{n} f_i x_i (y_i - \bar{y}) = 0 \quad \ldots(5)$$

We have $\mu_{11} = \frac{1}{N} \sum_i f_i (x_i - \bar{x})(y_i - \bar{y})$

$$= \frac{1}{N} \sum_i f_i x_i (y_i - \bar{y}) - \bar{x} . \frac{1}{N} \sum_i f_i (y_i - \bar{y}) = \frac{1}{N} \sum f_i x_i (y_i - \bar{y})$$

Similarly, $\mu_{11} = \frac{1}{N} \sum_i f_i y_i (x_i - \bar{x})$

$$\sigma_X^2 = \frac{1}{N} \sum_i f_i x_i (x_i - \bar{x}) \text{ and } \sigma_Y^2 = \frac{1}{N} \sum_i f_i y_i (y_i - \bar{y})$$

Substituting these values in (5), we get the required result.

(*b*) If the straight line defined by

$$Y = a + bX$$

satisfies the condition $E[(Y - a - bX)^2]$ = minimum, show that the regression line of the random variable Y on the random variable X is

$$Y - \bar{Y} = r \frac{\sigma_Y}{\sigma_X} (X - \bar{X}), \text{ where } \bar{X} = E(X), \bar{Y} = E(Y)$$

16. (*a*) Define Curve of regression of Y on X.

The joint density function of X and Y is given by :

$$f(x, y) = x + y, \; 0 < x < 1, 0 < y < 1$$
$$= 0, \text{ otherwise}$$

Find

(*i*) the correlation coefficient between X and Y,

(*ii*) the regression curve of Y on X, and

(*iii*) the regression curve of X on Y.

Ans. $\rho(X, Y) = -\frac{1}{11}$. [*Madras Univ. B.Sc., Stat. (Main), 1992*]

(*b*) Let $f(x_1, x_2) = \frac{2}{a^2}; 0 < x_1 < x_2, 0 < x_2 < a$

$$= 0, \text{ elsewhere}$$

be the joint p.d.f. of X_1 and X_2.

Find conditional means and variances. Also show that $\rho = \frac{1}{2}$.

17. If the joint density of X and Y is given by

$$f(x, y) = \begin{cases} (x + y)/3, \text{ for } 0 < x < 1, 0 < y < 2 \\ 0, \text{ otherwise} \end{cases}$$

obtain the regressions (*i*) of Y on X and (*ii*) of X on Y.

Are the regressions linear ? Find the correlation coefficient between X and Y. (*Allahabad Univ. B.Sc. 1992*)

Ans. $y = E(Y|x) = \frac{3x + 4}{3(x + 1)}; x = E(X|y) = \frac{2 + 3y}{3(1 + 2y)}$

Corr. $(X, Y) = -\left(\frac{2}{299}\right)^{1/2}$

18. Let the joint density function of X and Y be given by :

$$f(x, y) = 8xy, \; 0 < x < y < 1$$
$$= 0, \text{ otherwise}$$

Find: (*i*) $E(Y \mid X = x)$, (*ii*) $E[XY \mid X = x]$, (*iii*) $\text{Var}[Y \mid X = x]$

[*Delhi Univ. BSc. (Maths Hons.), 1988*]

Ans. (*i*) $E(Y|x) = \frac{2}{3}\left(\frac{1+x+x^2}{1+x}\right)$; $E(XY \mid x) = x\,E(Y \mid x)$, (*iii*) $E(Y^2 \mid x) = \frac{1+x^2}{2}$

19. Give an example to show that it is possible to have the regression of Y on X constant (does not depend on X), but the regression of X on Y is not constant (does depend on Y).

Hint. See Example 10·21

20. Prove or disprove :

$$E(Y \mid X = x) = \text{constant} \Rightarrow r(X,Y) = 0$$

Ans. True

21. If $f(x,y) = \frac{1}{3}x^2 \exp[-y(1+x)]$, $x \geq 0, y \geq 0$ is the joint p.d.f. of (X,Y), obtain the equation of regression of Y on X.

Ans. $y = E(Y \mid x) = 1/(1+x)$.

22. Variables (X,Y) have joint p.d.f.

$$f(x,y) = 6(1-x-y),\ x > 0,\ y > 0,\ x + y < 1.$$
$$= 0, \text{ otherwise.}$$

Find $f_X(x)$, $f_Y(y)$ and Cov (X,Y). Are X and Y independent ? Obtain the regression curves for the means.

[*Calcutta Univ. B.Sc. (Maths Hons.), 1986*]

Ans. $f_1(x) = 3(1-x)^2,\ 0 < x < 1$; $f_2(y) = 3(1-y)^2,\ 0 < y < 1$.

X and Y are not independent.

Regression curves for the means are :

$$y = E(Y \mid x) = \frac{1}{3}(1-x) \text{ and } x = E(X \mid y) = \frac{1}{3}(1-y)$$

23. For the joint p.d.f.

$$f(x, y) = 3x^2 - 8xy + 6y^2,\ 0 \leq (x, y) \leq 1,$$

find the least square regression lines and the regression curves for the means.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1987*]

Ans. Regression lines :

$$y - \frac{2}{3} = -\frac{10}{67}\left(x - \frac{5}{12}\right) \ ; \ x - \frac{5}{12} = -\frac{25}{32}\left(y - \frac{2}{3}\right)$$

Regression curves for means are :

$$y = E(Y|x) = \frac{9x^2 - 16x + 9}{6(3x^2 - 4x + 2)} \ ; \ x = E(X \mid y) = \frac{36y^2 - 32y - 9}{12(6y^2 - 4y + 1)}$$

24. Let (X, Y) be jointly distributed with p.d.f.

$$f(x, y) = e^{-y},\ 0 < x < y < \infty$$
$$= 0, \text{ otherwise}$$

Prove that :

$$E(Y \mid X = x) = x + 1 \text{ and } E(X \mid Y = y) = y/2.$$

Hence prove that $r(X, Y) = \sqrt{1/2}$.

25. Let $f(x, y) = e^{-y}(1 - e^{-x})$, $0 < x < y$; $0 < y < \infty$
$= e^{-x}(1 - e^{-y})$, $0 < y < x$; $0 < x < \infty$

(a) Show that $f(x, y)$ is a p.d.f.

(b) Find marginal distributions of X and Y.

(c) Find $E(Y \mid X = x)$ for $x > 0$.

(d) Find $P(X \le 2, Y \le 2)$.

(e) Find the correlation coefficient $r(X, Y)$.

(f) Find another joint p.d.f. having the same marginals.

Ans. (b) $f_1(x) = xe^{-x}$, $0 < x < \infty$; $f_2(y) = ye^{-y}$, $0 < y < \infty$.

(c) $E(Y \mid x) = \dfrac{1 - e^x}{x}[x - 1] + \dfrac{1}{x}\left(\dfrac{x^2}{2} + xe^x + e^{-x} - 1\right)$

(d) $1 - \dfrac{1}{e^4} - \dfrac{4}{e^2}$; (e) $r(x, y) = \dfrac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \dfrac{1}{\sqrt{2}\sqrt{2}} = \dfrac{1}{2}$

(f) **Hint.** $f(x, y, \alpha) = f_1(x) f_2(y)[1 + \alpha(2F(x) - 1)(2F(y) - 1)]$, $|\alpha| < 1$, has the same marginals $f_1(x)$ and $f_2(y)$.

26. Obtain regression equation of Y on X for the distributions :

(a) $f(x, y) = \dfrac{9}{2} \cdot \dfrac{1 + x + y}{(1 + x)^4 (1 + y)^4}$; $x, y \ge 0$

(b) $f(x, y) = \dfrac{4}{5}(x + 3y)e^{-x-2y}$; $x, y \ge 0$

[*Sardar Patel Univ. M.Sc., 1992*]

Ans. (a) **Hint.** See Example 5·25, page 5·55, (b) $\dfrac{x + 3}{2x + 3}$.

27. A ball is drawn at random from an urn containing three white balls numbered 0, 1, 2 ; two red balls numbered 0, 1 and one black ball numbered 0. If the colours white, red and black are again numbered 0, 1 and 2 respectively, find the correlation coefficient between the variates X, the colour number and Y the number of the ball. Write down the equation of regression line of Y on X.

[*Calcutta Univ. B.Sc. (Maths. Hons.), 1986*]

OBJECTIVE TYPE QUESTIONS

I. State, giving reasons, whether each of the following statements is true or false.

(i) Both regression lines of Y on X and of X on Y do not intersect at all.

(ii) In a bivariate regression, $b_{YX} = \frac{1}{5}$, $b_{XY} = 10$

(iii) The regression coefficient of Y on X is 3·2 and that of X on Y is 0·8.

(iv) There is no relationship between correlation coefficient and regression coefficient.

(v) Both the regression coefficients cannot exceed unity.

(*vi*) The greater the value of '*r*', the better are the estimates obtained through regression analysis.

(*vii*) If X and Y are negatively correlated variables, and (0, 0) is on the least squares line of Y on X, and if $X = 1$ is the observed value then predicted value of Y must be negative.

(*viii*) Let the correlation between X and Y be perfect and positive. Suppose the points (3, 5) and (1, 4) are on the regression lines. With this knowledge it is possible to determine the least squares line exactly.

(*ix*) If the lines of regression are $Y = \frac{1}{4}X$ and $X = \frac{1}{9}Y + 1$, then $\rho = \frac{1}{6}$ and $E(X \mid Y = 0) = 1$.

(*x*) In a bivariate distribution, $b_{YX} = 2{\cdot}8$ and $b_{XY} = 0{\cdot}3$.

II. Fill in the blanks :

(*i*) The regression analysis measures ... between X and Y.

(*ii*) Lines of regression are ... if $r_{XY} = 0$ and they are ... if $r_{XY} = \pm 1$.

(*iii*) If the regression coefficients of X on Y and Y on X are $-0{\cdot}4$ and $-0{\cdot}9$ respectively then the correlation coefficient between X and Y is ...

(*iv*) If the two regression lines are $X + 3Y - 5 = 0$ and $4X + 3Y - 8 = 0$, then the correlation coefficient between X and Y is ...

(*v*) If one of the regression coefficients is ... unity, the other must be ... unity.

(*vi*) The farther the two regression lines cut each other, the ... will be the degree of correlation.

(*vii*) When one regression coefficient is positive, the other would be ...

(*viii*) The sign of regression coefficient is ... as that of correlation coefficient.

(*ix*) Correlation coefficient is the.... between regression coefficients.

(*x*) Arithmetic mean of regression coefficients is ... correlation coefficient.

(*xi*) When the correlation coefficient is zero, the two regression lines are ... and when it is ± 1, then the regression lines are ...

III. Indicate the correct answer :

(*i*) The regression line of Y on X (*a*) minimises total of the squares of horizontal deviations, (*b*) total of the squares of the vertical deviations, (*c*) both vertical and horizontal deviations, (*d*) none of these.

(*ii*) The regression coefficients are b_2 and b_1. Then the correlation coefficient r is (*a*) b_1/b_2, (*b*) b_2/b_1, (*c*) $b_1 b_2$ (*d*) $\pm \sqrt{b_1 b_2}$.

(*iii*) The farther the two regression lines cut each other (*a*) the greater will be the degree of correlation, (*b*) the lesser will be the degree of correlation, (*c*) does not matter.

(*iv*) If one regression coefficient is greater than unity, then the other must be (*a*) greater than the first one, (*b*) equal to unity, (*c*) less than unity, (*d*) equal to zero.

(*v*) When the correlation coefficient $r = \pm 1$, then the two regression lines (*a*) are perpendicular to each other, (*b*) coincide, (*c*) are parallel to each other, (*d*) do not exist.

(*vi*) The two lines of regression are given as $X + 2Y - 5 = 0$ and $2X + 3Y = 8$. Then the mean values of X and Y respectively are (*a*) 2, 1, (*b*) 1, 2, (*c*) 2, 5, (*d*) 2, 3.

(*vii*) The tangent of the angle between two regression lines is given as 0·6 and the s.d. of Y is known to be twice that of X. Then the value of correlation coefficient between X and Y is (*a*) $-\frac{1}{2}$, (*b*) $\frac{1}{2}$, (*c*) 0·7, (*d*) 0·3.

IV. σ_X and σ_Y are the standard deviations of two correlated variables X an Y respectively in a large sample, and r is the sample correlation coefficient.

(*i*) State the "Standard Error of Estimate" for linear regression of Y on X.

(*ii*) What is the standard error in estimating Y from X if $r = 0$?

(*iii*) By how much is this error reduced if r is increased to 0·30 ?

(*iv*) How large must r be in order to reduce this standard error to one-half its value for $r = 0$?

(*v*) Give your interpretations for the cases $r = 0$ and $r = 1$.

V. Explain why we have two lines of regression.

10·8. Correlation Ratio. As discused earlier, when variables are linearly related, we have the regression lines of one variable on another variable and correlation coefficient can be computed to tell us about the extent of association between them. However, if the variables are not linearly related but some sort of curvilinear relationship exists between them, the use of r which is a measure of the degree to which the relation approaches a straight line "law" will be misleading. We might come across bivariate distributions where r may be very low or even zero but the regression may be strong, or even perfect. *Correlation ratio* 'η' is the appropriate measure of curvilinear relationship between the two variables. Just as r measures the concentration of points about the straight line of best fit, η measures the concentration of points about the curve of best fit. If regression is linear $\eta = r$, otherwise $\eta > r$ (*c.f.* Remark 2, § 10·8·1).

10·8·1. Measure of Correlation Ratio. In the previous articles we have assumed that there is a single observed value Y corresponding to the given value x_i of X but sometimes there are more than one such value of Y.

Suppose corresponding to the values x_i, $(i = 1, 2, \ldots, m)$ of the variable X, the variable Y takes the values y_{ij} with respective frequencies f_{ij}, $j = 1, 2, \ldots, n$.

Though all the x's in the ith vertical array have the same value, the y's are different. A typical pair of values in the ith array is (x_i, y_{ij}), with frequency f_{ij}.

Thus the first suffix i indicates the vertical array while the second suffix j indicates the positions of y in that array. Let

$$\sum_{j=1}^{n} f_{ij} = n_i \quad \text{and} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \sum_{i=1}^{m}\left(\sum_{j=1}^{n} f_{ij}\right) = \sum_{i=1}^{m} n_i = N, \quad \text{(say).}$$

If $\bar{y}_i$ and $\bar{y}$ denote the means of the ith array and the overall mean respectively, then

$$\bar{y}_i = \frac{\sum_j f_{ij}\, y_{ij}}{\sum_j f_{ij}} = \frac{\sum_j f_{ij}\, y_{ij}}{n_i} = \frac{T_i}{n_i} \quad \text{and} \quad \bar{y} = \frac{\sum_i\sum_j f_{ij}\, y_{ij}}{\sum_i\sum_j f_{ij}} = \frac{\sum_i n_i\, \bar{y}_i}{\sum_i n_i} = \frac{T}{N}$$

In other words $\bar{y}$ is the weighted mean of all the array means, the weights being the array frequencies.

Def. *The correlation ratio of Y on X, usually denoted by η_{YX} is given by*

$$\eta_{YX}^2 = 1 - \frac{\sigma_{eY}^2}{\sigma_Y^2} \qquad \ldots(10\cdot21)$$

where σ_{eY}^2 and σ_Y^2 are given by

$$\sigma_{eY}^2 = \frac{1}{N}\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 \quad \text{and} \quad \sigma_Y^2 = \frac{1}{N}\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y})^2$$

A convenient expression for η_{YX} can be obtained in terms of standard deviation σ_{mY} of the means of the vertical arrays, each mean being weighted by the array frequency.

We have

$$N\sigma_Y^2 = \sum\sum f_{ij}\,(y_{ij} - \bar{y})^2 = \sum_i\sum_j f_{ij}\,\{(y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{y})\}^2$$

$$= \sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 + \sum_i\sum_j f_{ij}\,(\bar{y}_i - \bar{y})^2 + 2\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y})$$

The term $2[\sum_i(\bar{y}_i - \bar{y})\{\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)\}]$ vanishes since $\sum f_{ij}\,(y_{ij} - \bar{y}_i) = 0$, being the algebraic sum of the deviations from mean.

$$\therefore \qquad N\sigma_Y^2 = \sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y}_i)^2 + \sum_i n_i(\bar{y}_i - \bar{y})^2$$

$$\Rightarrow \qquad N\sigma_Y^2 = N\sigma_{eY}^2 + N\sigma_{mY}^2 \;\Rightarrow\; \sigma_Y^2 = \sigma_{eY}^2 + \sigma_{mY}^2$$

$$\Rightarrow \qquad 1 - \frac{\sigma_{eY}^2}{\sigma_Y^2} = \frac{\sigma_{mY}^2}{\sigma_Y^2}$$

which on comparison with (10·21) gives

$$\eta_{YX}^2 = \frac{\sigma_{mY}^2}{\sigma_Y^2} = \frac{\sum_i n_i(\bar{y}_i - \bar{y})^2}{\sum_i\sum_j f_{ij}\,(y_{ij} - \bar{y})^2} \qquad \ldots(10\cdot22)$$

We have

$$N\sigma_{mY}^2 = \sum_i n_i (\bar{y}_i - \bar{y})^2 = \sum_i n_i \bar{y}_i^2 - N\bar{y}^2 = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N}$$

$$\therefore \qquad \eta_{YX}^2 = \left[\sum_i \left(\frac{T_i^2}{n_i}\right) - \frac{T^2}{N}\right] \Big/ N\sigma_Y^2, \qquad \ldots(10\cdot23)$$

a formula, much more convenient for computational purposes.

Remarks 1. (10·21) implies that

$$\sigma_{eY}^2 = \sigma_Y^2 (1 - \eta_{YX}^2)$$

Since σ_{eY}^2 and σ_Y^2 are non-negative, we have

$$1 - \eta_{YX}^2 \geq 0 \quad \Rightarrow \quad \eta_{YX}^2 \leq 1 \quad \Rightarrow \quad |\eta_{YX}| \leq 1$$

2. Since the sum of squares of deviations in any array is minimum when measured from its mean, we have

$$\sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 \leq \sum_i \sum_j f_{ij} (y_{ij} - \hat{y}_{ij})^2 \qquad \ldots(*)$$

where $\hat{y}_{ij}$ is the estimate of y_{ij} for given value of $X = x_i$, say, as given by the line of regression of Y on X *i.e.*, $\hat{y}_{ij} = a + bx_i$, $(j = 1, 2, \ldots, n)$.

But $$\sum_i \sum_j f_{ij} (y_{ij} - \bar{y}_i)^2 = N\sigma_{eY}^2 = N\sigma_Y^2 (1 - \eta_{YX}^2)$$

and $$\sum_i \sum_j f_{ij} (y_{ij} - a - bx_i)^2 = N\sigma_Y^2 (1 - r^2) \qquad (c.f.\ \S\ 10\cdot7\cdot6)$$

$\therefore$ (*) $\Rightarrow$ $$1 - \eta_{YX}^2 \leq 1 - r^2$$

i.e., $$\eta_{YX}^2 \geq r^2 \quad \Rightarrow \quad |\eta_{YX}| \geq |r|$$

Thus the absolute value of the correlation ratio can never be less than the absolute of r, the correlation coefficient.

When the regression of Y on X is linear, straight line of means of arrays coincides with the line of regression and $\eta_{YX}^2 = r^2$. Thus $\eta_{YX}^2 - r^2$ is the departure of regression from linearity. It is also clear (from Remark 1) that the more nearly η_{YX}^2 approaches unity, the smaller is σ_{eY}^2 and, therefore, closer are the points to the curve of means of vertical arrays.

When $\eta_{YX}^2 = 1, \sigma_{eY}^2 = 0 \Rightarrow \sum\sum f_{ij} (y_{ij} - \bar{y}_i)^2 = 0$

$\Rightarrow y_{ij} = \bar{y}_i$, $\forall\ j = 1, 2, \ldots, n$, *i.e.*, all the points lie on the curve of means. This implies that there is a functional relationship between X and Y. η_{YX} is, therefore, the measure of the degree to which the association between the variables approaches a functional relationship of the form $Y = F(X)$, where $F(X)$ is a single valued function of X, $[F(X) = a + bX]$.

3. It is worth noting that the value of η_{YX} is not independent of the classification of the data. As the class intervals become narrower η_{YX} approaches unity, since in that case σ_{mY}^2 gets nearer to σ_Y^2. If the grouping is so fine that only one item appears in each row (related to each x-class), that item will constitute the mean of that column and thus in this case σ_{mY}^2 and σ_Y^2 become equal so that $\eta_{YX}^2 = 1$. On the other hand, a very coarse grouping tends to make the value of η_{YX} approach r. "Student" has given a formula for 'the correction'

to be made in the correlation ratio 'for grouping' in Biometrika (Vol IX page 316-320.)

4. It can be easily proved that η_{YX}^2 is independent of change of origin and scale of measurements.

5. η_{XY}^2, the second correlation ratio of X on Y depends upon the scatter of observations about the line of column means.

6. r_{XY} and r_{YX} are same but η_{YX} is, in general, different from η_{XY}.

7. In terms of expectation, correlation ratio is defined as follows :

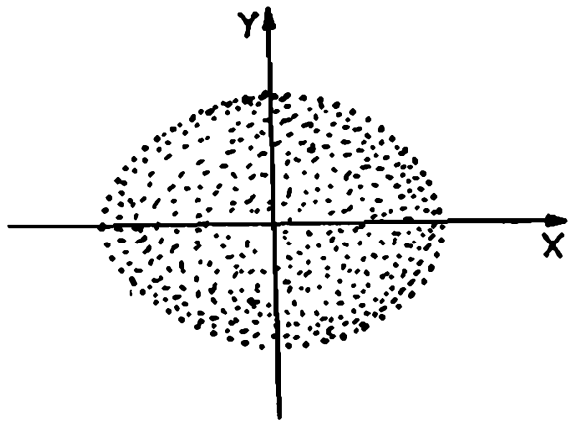
$$\eta_{YX}^2 = \frac{E_X\,[E(Y \mid X) - E(Y)]^2}{E[Y - E(Y)]^2} = \frac{E[E(Y \mid X) - E(Y)]^2}{\sigma_Y^2}$$

and

$$\eta_{XY}^2 = \frac{E_Y\,[E\,(X \mid Y) - E(X)]^2}{E[X - E(X)]^2} = \frac{E[E(X \mid Y) - E(X)]^2}{\sigma_X^2}$$

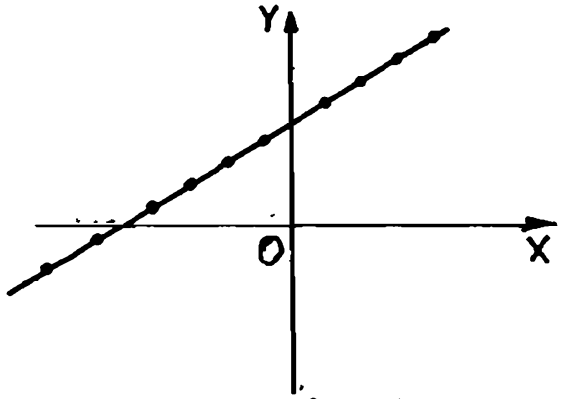
8. We give below some diagrams, exhibiting the relationship between r and η_{YX}.

(i) For completely random scattering of the dots with no trend, both r and η are zero.



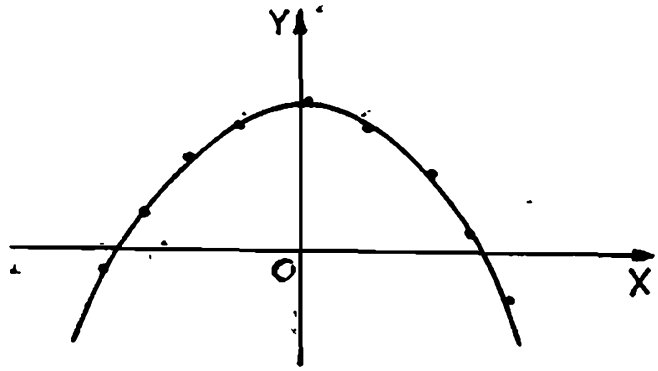
$r = 0, \eta_{YX} = \eta_{XY} = 0$

(ii) If dots lie precisely on a line, $r = 1$ and $\eta = 1$.



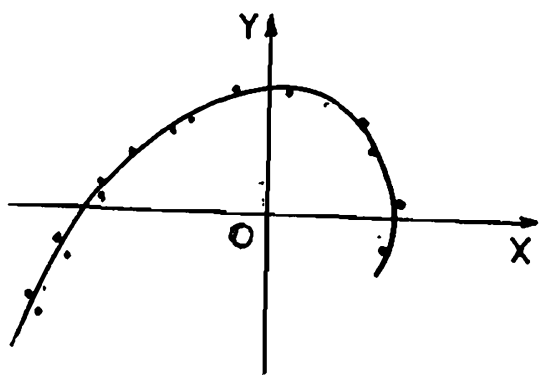
$r = 1, \eta_{YX} = \eta_{XY} = 1$

(iii) If dots lie on a curve, such that no ordinate cuts it more than once, $\eta_{YX} = 1$ and if furthermore, the dots are symmetrically placed about Y-axis, then $\eta_{XY} = 0$, $r = 0$.



$r=0, \eta_{YX}=1, \eta_{XY}=0$

(iv) If $\eta_{YX} > r$, the dots are scattered around a definitely curved trend line.



$\eta_{YX} > r$

EXERCISE 10(e)

1. (*a*) Define correlation coefficient and correlation ratio. When is the latter a more suitable measure of correlation than the former ? Show that the correlation ratio is never less than the correlation coefficient. What do you infer if the two are equal ? Further, show that none of these can exceed one.

[*Delhi Univ. B.Sc. (Stat. Hons.), 1988*]

(*b*) Show that $$1 \geq \eta^2_{YX} \geq r^2_{YX} \geq 0$$

Interpret each of the following statements.

(*i*) $r = 0$, (*ii*) $r^2 = 1$, (*iii*) $\eta^2 = 1$, (*iv*) $\eta^2 = r^2$ and (*v*) $\eta = 0$

(*c*) When the correlation coefficient is equal to unity, show that the two correlation ratios are also equal to unity. Is the converse true ?

(*d*) Define correlation ratio η_{XY} and prove that

$$1 \geq \eta^2_{XY} \geq r^2,$$

where r is the coefficient of correlation between X and Y. Show further that $(\eta^2_{XY} - r^2)$ is a measure of non-linearity of regression.

2. For the joint p.d.f.

$$f(x, y) = \frac{1}{2}x^3 \exp[-x(y+1)], \; y > 0, x > 0$$
$$= 0, \text{ otherwise,}$$

find :

(*i*) Two lines of regression.

(*ii*) The regression curves for the means.

(*iii*) $r(X, Y)$.

(*iv*) η^2_{YX} and η^2_{XY}.

[*Delhi Univ. B.A. (Stat. Hons. Spl. Course), 1987*]

Ans. (*i*) $y = -\frac{1}{6}x + 1$; $x = -\frac{2}{3}y + \frac{10}{3}$

(*ii*) $y = E(Y|x) = \frac{1}{x}$; $x = E(X|y) = \frac{4}{1+y}$

(*iii*) $r(X, Y) = -\frac{1}{3}$ (*iv*) $\eta^2_{YX} = \frac{1}{3}$, $\eta^2_{XY} = \frac{1}{5}$

3. Compute $r(X, Y)$ and η_{YX} for the following data :

X :	0·5 — 1·5	1·5 — 2·5	2·5 — 3·5	3·5 — 4·5	4·5 — 5·5
f :	20	30	35	25	15
$\bar{y}_i$	11·3	12·7	14·7	16·5	19·1

Var $(Y) = 9·61$

Ans. $\eta_{YX} = 0·77$, $r = 0·85$

4. Compute η_{XY} for the following table :

$Y \downarrow$ \ $X \rightarrow$	47	52	57	62	67
57	4	4	2	...	...
62	4	8	8	1	...
67	...	7	12	1	4
72	...	3	1	8	5
77	...	...	3	5	6

10·9. Intra-class Correlation. Intra-class correlation means within class correlation. It is distinguishable from product moment correlation in as much as here both the variables measure the same characteristics. Sometimes specially in biological and agricultural study, it is of interest to know how the members of a family or group are correlated among themselves with respect to some one of their common characteristic. For example, we may require the correlation between the heights of brothers of a family or between yields of plots of an experimental block. In such cases both the variables measure the same characteristic, *e.g.*, height and height or weight and weight. There is

nothing to distinguish one from the other so that one may be treated as X-variable and the other as the Y-variable.

Suppose we have $A_1, A_2, \ldots, A_n$ families with $k_1, k_2, \ldots, k_n$ members, each of which may be represented as

$$\begin{array}{ccccc}
x_{11} & x_{21} & \cdots & x_{i1} & \cdots & x_{n1} \\
x_{12} & x_{22} & & x_{i2} & & x_{n2} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{1j} & x_{2j} & \cdots & x_{ij} & \cdots & x_{nj} \\
\vdots & \vdots & & \vdots & & \vdots \\
x_{1k_1} & x_{2k_2} & \cdots & x_{ik_i} & \cdots & x_{nk_n}
\end{array}$$

and let x_{ij} $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, k_i)$ denote the measurement on the jth member in the ith family.

We shall have $k_i(k_i - 1)$ pairs for the ith family or group like (x_{ij}, x_{il}), $j \neq l$. There will be $\sum_{i=1}^{n} k_i(k_i - 1) = N$ pairs for all the n tamilies or groups. If we prepare a correlation table there will be $k_i(k_i - 1)$ entries for the ith group or family and $\sum_i k_i(k_i - 1) = N$ entries for all the n families or groups. The table is symmetrical about the principal diagonal. Such a table is called an *intra-class correlation table* and the correlation is called *intra-class correlation.*

In the bivariate table x_{i1} occurs $(k_i - 1)$ times, x_{i2} occurs $(k_i - 1)$ times, ..., x_{ik_i} occurs $(k_i - 1)$ times, *i.e.*, from the ith family we have $(k_i - 1)\sum_j x_{ij}$ and hence for all the n families we have $\sum_i (k_i - 1)\sum x_{ij}$ as the marginal frequency, the table being symmetrical about principal diagonal.

$$\therefore \qquad \bar{x} = \bar{y} = \frac{1}{N}\left[\sum_i (k_i - 1) \sum_j x_{ij}\right]$$

Similarly,

$$\sigma_X^2 = \sigma_Y^2 = \frac{1}{N}\left[\sum_i (k_i - 1)\sum_j (x_{ij} - \bar{x})^2\right]$$

Further

$$\text{Cov}(X, Y) = \frac{1}{N}\sum_i\left[\sum_{j,l}(x_{ij} - \bar{x})(x_{il} - \bar{x})\right], j \neq l$$

$$= \frac{1}{N}\sum_i\left[\sum_{j=1}^{k_i}\sum_{l=1}^{k_i}(x_{ij} - \bar{x})(x_{il} - \bar{x}) - \sum_{j=1}^{k_i}(x_{ij} - \bar{x})^2\right]$$

If we write $\bar{x}_i = \sum_j x_{ij} / k_i$, then

$$\sum_i \left[\sum_{j=1}^{k_i} \sum_{l=1}^{k_i} (x_{ij} - \bar{x})(x_{il} - \bar{x}) \right] = \sum_i \left[\sum_j (x_{ij} - \bar{x}) \sum_l (x_{il} - \bar{x}) \right]$$

$$= \sum_i [k_i (\bar{x}_i - \bar{x})\, k_i (\bar{x}_i - \bar{x})]$$

$$= \sum_i k_i^2 (\bar{x}_i - \bar{x})^2$$

Therefore intra-class correlation coefficient is given by

$$r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)\, V(Y)}} = \frac{\sum_i k_i^2 (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{\sum_i \sum_j (k_i - 1)(x_{ij} - \bar{x})^2} \quad \ldots(10\cdot24)$$

If we put $k_i = k$, *i.e.*, if all families have equal members then

$$r = \frac{k^2 \sum_i (\bar{x}_i - \bar{x})^2 - \sum_i \sum_j (x_{ij} - \bar{x})^2}{(k-1) \sum_i \sum_j (x_{ij} - \bar{x})^2}$$

$$= \frac{nk^2 \sigma_m^2 - nk\sigma^2}{(k-1)\, nk\sigma^2} = \frac{1}{(k-1)} \left\{ \frac{k\, \sigma_m^2}{\sigma^2} - 1 \right\} \quad \ldots(10\cdot24a)$$

where σ^2 denotes the variance of X and σ_m^2 the variance of means of families.

Limits. We have from (10·24*a*),

$$1 + (k-1)\, r = \frac{k\sigma_m^2}{\sigma^2} \geq 0 \quad \Rightarrow \quad r \geq -\frac{1}{(k-1)}$$

Also $\quad 1 + (k-1)\, r \leq k$, as the ratio $\dfrac{\sigma_m^2}{\sigma^2} \leq 1 \quad \Rightarrow \quad r \leq 1$

so that $\quad -\dfrac{1}{(k-1)} \leq r \leq 1$

Interpretation. Intraclass correlation cannot be less than $-1/(k-1)$, though it may attain the value $+1$ on the positive side, so that it is a skew coefficient and a negative value has not the same significance as a departure from independence as an equivalent positive value.

EXERCISE 10 (f)

1. If $x_1, x_2, \ldots, x_k$ be k variates with standard deviation σ and m be any number, prove that

$$k^2\sigma^2 = (k-1) \sum_{r=1}^{k} (x_r - m)^2 - \sum_{r=1}^{k} \sum_{s=1}^{k} (x_r - m)(x_s - m),\ r \neq s$$

Hence deduce that the coefficient of intraclass correlation for n families with varying number of members in each family is

$$1 - \frac{\sum_i k_i \sigma_i^2}{\sigma^2 \sum_i k_i(k_i - 1)}$$

where k_i, σ_i^2 denote the number of members and the variance respectively in the $i$tn family and σ^2 is the general variance.

Given $n = 5$, $\sigma_i = i$, $k_i = i + 1$ $(i \leq 5)$, find the least possible intraclass correlation coefficient.

2. What do you understand by intra-class correlation coefficient.

Calculate its value for the following data :

Family No.	*Height of brothers*			
1	60	62	63	65
2	59	60	61	62
3	62	62	64	63
4	65	66	65	66
5	66	67	67	69

3. In four families each containing eight persons, the chest measurements of persons are given below. Calculate the intraclass correlation co-efficient.

Family	1	2	3	4	5	6	7	8
I	43	46	48	42	50	45	45	49
II	33	34	37	39	82	35	37	41
III	56	52	50	51	54	52	39	52
IV	34	37	38	40	40	41	44	44

10·10. Bivariate Normal Distribution. The bivariate normal distribution is a generalization of a normal distribution for a single variate. Let X and Y be two normally correlated variables with correlation coefficient ρ and $E(X) = \mu_1$, Var $(X) = \sigma_1^2$; $E(Y) = \mu_2$, Var $(Y) = \sigma_2^2$. In deriving the bivariate normal distribution we make the following three assumptions.

(*i*) *The regression of Y on X is linear.* Since the mean of each array is on the line of regression $Y = \rho(\sigma_2/\sigma_1)X$, the mean or expected value of Y is $\rho(\sigma_2/\sigma_1)X$, for different values of X.

(*ii*) *The arrays are homoscedastic, i.e.,* variance in each array is same. The common variance of estimate of Y in each array is then given by $\sigma_2^2 (1 - \rho^2)$, ρ being the correlation coefficient between variables X and Y and is independent of X.

(*iii*) *The distribution of Y in different arrays in normal.* Suppose that one of the variates, say X, is distributed normally with mean 0 and standard deviation σ_1 so that the probability that a random value of X will fall in the small interval dx is

$$g(x)\, dx = \frac{1}{\sigma_1\sqrt{(2\pi)}} . \exp\ (-x^2/2\sigma_1^2)\, dx$$

The probability that a value of Y, taken at random in an assigned vertical array will fall in the interval dy is