On

Note:- If $f(x)$ is the p.d.f of a random variable 'X' which is defined in the interval (a, b), then

(1) Mean of $X = \int_a^b x\, f(x)\, dx$

(2) Harmonic mean $= \int_a^b \frac{1}{x} f(x)\, dx$

(3) Geometric Mean G is given by

$$\log G = \int_a^b \log x\, f(x)\, dx$$

(4) Moments about origin $\mu_r' = \int_a^b x^r f(x)\, dx$

(5) Moment about any point A

$$\mu_r' = \int_a^b (x-A)^r f(x)\, dx$$

(6) Moment about mean $\mu_r = \int_a^b (x-\text{mean})^r f(x)\, dx$

(7) Variance $= \int_{-\infty}^{\infty} (x-\text{mean})^2 f(x)\, dx$

(8) Relation between Pdf and distribution function

$$f(x) = \frac{dF(x)}{dx}.$$

Continuous distribution function:-

If $f(x)$ is a p.d.f of a continuous r.v 'X' then the function $F_X(x) = F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\, dx$, $-\infty < x < \infty$ is called the distribution function or cumulative distribution function of the r.v. X.

properties of c.d.f of a r.v. 'X'.

(i) $0 \le F(x) \le 1$

(ii) $\lim_{x \to -\infty} F(x) = 0$

(iii) $\lim_{x \to \infty} F(x) = 1$

(iv) $P(a \le x \le b) = \int_a^b f(x)\, dx$

$= F(b) - F(a)$

(v) $F'(x) = \frac{dF(x)}{dx} = f(x) \ge 0$

$dF(x)$ is called the Probability differential of the r.v. X.

Expectation of a Linear combination of r.vs:-

Let $X_1, X_2, \cdots X_n$ be any 'n' r.vs and if $a_1, a_2, \cdots a_n$ are constants then $E[a_1X_1 + a_2X_2 + \cdots + a_nX_n] = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n)$

Note:- Let X and Y be two r.v such that $Y \leq X$ then $E(Y) \leq E(X)$.

Note:- If X is a r.v, then

$Var(aX+b) = a^2 var(X)$ where 'a' and 'b' are constants.

Covariance:- If X and Y are r.v then the covariance between them is defined as

$$cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}$$
$$= E\{XY - XE(Y) - E(X)Y + E(X)E(Y)\}$$
$$= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y)$$
$$\therefore cov(X,Y) = E(XY) - E(X)\cdot E(Y).$$

If X and Y are independent, then

$$E(XY) = E(X)E(Y)$$
$$cov(X,Y) = 0$$

Note:-

1. $cov(aX, bY) = ab\, cov(X,Y)$
2. $cov(cX+a, Y+b) = cov(X,Y)$
3. $cov(aX+b, cY+d) = ac\, cov(X,Y)$
4. $Var(X_1+X_2) = var(X_1) + var(X_2) + 2cov(X_1,X_2)$
5. $Var(X_1-X_2) = var(X_1) + var(X_2) - 2cov(X_1,X_2)$
6. If X_1, X_2 are independent

$$Var(X_1 \pm X_2) = var(X_1) \pm var(X_2).$$

Expected values of a two-dimensional r.v:-

Let X and Y be two-dimensional discrete r.v's with joint probability mass function $P(x_i, y_j) = P_{ij}$. Then the mathematical expectation of a function $g(X,Y)$ is defined by

$$E[g(X,Y)] = \sum_i \sum_j g(x_i, y_j) \cdot P(x_i, y_j)$$

If (X, Y) is a two-dimensional continuous r.v with pdf $f(x,y)$ then, the mathematical expectation of the function $g(x,y)$ is defined by $E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\cdot f(x,y)\,dx\,dy$.

Ex:- Given the following probability distribution of X compute (i) E(X) (ii) $E(X^2)$ (iii) $E(2X\pm3)$ (iv) $Var(2X\pm3)$.

x	-3	-2	-1	0	1	2	3
P(x)	0.05	0.10	0.30	0	0.30	0.15	0.10

Soln:- We know that for a discrete r.v 'x',

(i) $E(x) = \sum_{i=1}^{n} x_i p(x_i)$

$= (-3)(0.05) - 2(0.1) - 1(0.30) + 0 + 1(0.30)$
$+ 2(0.15) + 3(0.10)$

$= 0.25$

(ii) $E(X^2) = (-3)^2(0.05) + (-2)^2(0.1) + (-1)^2(0.30) + 0 + (1)^2(0.3)$
$+ (2)^2(0.15) + 3^2(0.10)$

$= 2.95$

(iii) $E(2X\pm3) = 2E(X) \pm 3$

$= 2(0.25) \pm 3$

$= 0.5 \pm 3$

(iv) $Var(2X\pm3) = 2^2 Var(X)$

$= 4\left[E(X^2) - (E(X))^2\right]$

$= 4\left[2.95 - (0.25)^2\right] = 2.8875 \times 4$

$= 11.55$

Let X be a r.v with $E(X) = 10$, $V(X) = 25$. Find the +ve values of 'a' and 'b' such that $Y = aX - b$ has expectation 0 and variance 1.

Soln:- Given $E(X) = 10$, $V(X) = 25$

Now, $E(Y) = E(aX - b)$

$= aE(X) - b$

$= a(10) - b = 0$

$10a - b = 0$ —— (1)

$Var(Y) = V(aX - b) = a^2 Var(X)$

$= a^2(25)$

$= 25a^2 = 1$

$25a^2 = 1$ —— (2)

from (1) and (2)

$a = 1/5$, $b = 2$.

Mathematical expectations:-

Let 'x' be a r.v with probability density function (or pmf) $f(x)$ (or $p(x)$). Then the mathematical expectation of 'x' is denoted by $E(x)$ and is given by

$$E(x) = \int_{-\infty}^{\infty} x f(x)\,dx \quad \text{for a continuous r.v}$$

$$= \sum_{x} x p(x) \quad \text{for a discrete r.v.}$$

r^{th} moment (about origin):-

Consider a continuous r.v 'x' with pdf $f(x)$ then the r^{th} moment (about origin) of the probability distribution is defined as

$$E(x^r) = \int_{-\infty}^{\infty} x^r f(x)\,dx$$

It is denoted by $\mu_r' = \int_{-\infty}^{\infty} x^r f(x)\,dx$

Thus $\mu_1' = E(x)$

$\mu_2' = E(x^2)$

$\therefore$ Mean $= \bar{x} = \mu_1' = E(x)$

Variance $= \mu_2 = \mu_2' - \mu_1'^2$

Variance $= E(x^2) - (E(x))^2$

NOTE: Now $E\{x - E(x)\}^r = \int_{-\infty}^{\infty} \{x - E(x)\}^r f(x)\,dx$

$$= \int_{-\infty}^{\infty} (x - \bar{x})^r f(x)\,dx$$

Thus $\mu_r = \int_{-\infty}^{\infty} (x - \bar{x})^r f(x)\,dx$ this gives the r^{th} momentum about mean and is denoted by μ_r

Put $r = 1$, we get $\int_{-\infty}^{\infty} (x - \bar{x}) f(x)\,dx$

$$= \int_{-\infty}^{\infty} x f(x)\,dx - \int_{-\infty}^{\infty} \bar{x} f(x)\,dx$$

$$= \bar{x} - \bar{x} \int_{-\infty}^{\infty} f(x)\,dx = \bar{x} - \bar{x} = 0$$

Put $r = 2$, we get Variance $= \int_{-\infty}^{\infty} (x - \bar{x})^2 f(x)\,dx = E\{x - E(x)\}^2$

Note:- for discrete r.v. 'x',

$$E(x^r) = \sum_x x^r p(x)$$

(ii) $\mu_r' = E(x^r) = \sum_x x^r p(x)$

Put $r=1$, we get

Mean $= \mu_1' = \sum x\, p(x)$

Variance $= \mu_2 = \mu_2' - \mu_1^2 = E(x^2) - (E(x))^2$

The r^{th} moment about mean

$$\mu_r = E\left[(x - E(x))^r\right]$$

$$= \sum_x (x - E(x))^r p(x), \quad E(x) = \bar{x}$$

$$= \sum_x (x - \bar{x})^r p(x).$$

Put $r = 2$, we get

Variance $= \mu_2 = \sum_x (x-\bar{x})^2 p(x)$.

Addition theorem:- If X and Y are two continuous r.vs with pdf $f_x(x)$ and $f_y(y)$ then

$$E(X+Y) = E(X) + E(Y)$$

Multiplication theorem:-

If X and Y are independent r.v. then $E(XY) = E(X) \cdot E(Y)$.

Note:- If $X_1, X_2, \ldots X_n$ are 'n' independent r.vs, then

$$E(X_1, X_2, \ldots X_n) = E(X_1) \cdot E(X_2) \cdot \ldots E(X_n).$$

Note:- (1) If 'X' is a r.v with pdf $f(x)$ and 'a', 'b' are constants, then $E(ax+b) = aE(x) + b$

(2) If $b=0$, then $E(ax) = aE(x)$

(3) $E\left(\frac{1}{x}\right] \neq \frac{1}{E(x)}$

(4) $E(\log(x)) \neq \log E(x)$

$E(x^2) \neq (E(x))^2$.

Ex:- Two r.v's X and Y have joint pdf $f(x,y) = \begin{cases} \frac{xy}{96}, & 0<x<4,\ 1<y<5 \\ 0, & \text{otherwise} \end{cases}$

find (i) $E(X)$ (ii) $E(Y)$ (iii) $E(XY)$ (iv) $E(2X+3Y)$ (v) $V(X)$ (vi) $V(Y)$ (vii) $Cov(X,Y)$.

Sol:- (i) $E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\, dx\,dy = \int_1^5\int_0^4 x\left(\frac{xy}{96}\right) dx\,dy = 8/3$

(ii) $E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\, dx\,dy = \int_1^5\int_0^4 y\left(\frac{xy}{96}\right) dx\,dy = 31/9$

(iii) $E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\, dx\,dy = \int_1^5\int_0^4 xy\cdot\left(\frac{xy}{96}\right) dx\,dy = \frac{248}{27}$

(iv) $E(2X+3Y) = 2E(X) + 3E(Y)$

$= 2(8/3) + 3(31/9) = 47/3$

(v) $Var(X) = E(X^2) - [E(X)]^2$

now $E(X^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 f(x,y)\, dx\,dy = \int_1^5\int_0^4 x^2\left(\frac{xy}{96}\right) dx\,dy = 8$

$\therefore Var(X) = 8 - (8/3)^2 = 8/9$

(vi) $V(Y) = E(Y^2) - (E(Y))^2$

$V(Y^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2 f(x,y)\, dx\,dy = \int_1^5\int_0^4 y^2\left(\frac{xy}{96}\right) dx\,dy = 13$

$\therefore Var(Y) = E(Y^2) - (E(Y))^2 = 13 - \left(\frac{31}{9}\right)^2 = \frac{92}{81}$

(vii) $Cov(X,Y) = E(XY) - E(X)E(Y)$

$= \frac{248}{27} - \left(\frac{8}{3}\right)\left(\frac{31}{9}\right) = 0.$

Ex:- If the joint pdf of (X,Y) is given by $f(x,y) = 24y(1-x)$, $0 \le x \le y \le 1$ find $E(XY)$

Sol:- $E(XY) = \int_0^1\int_y^1 24y(1-x)\, dx\,dy = 1/15$

Ex:- If X and Y is a two-dimensional r.v uniformly distributed over the triangular region R bounded by $y=0$, $x=3$ and $y = \frac{1}{3}x$. find $E(X)$, $E(Y)$, $V(X)$, $V(Y)$.

Sol:- Given X and Y are uniformly distributed.

$\therefore f(x,y) = k$ a constant

we know that $\iint f(x,y)\, dx\,dy = 1$

$\int_0^3\int_{x/3}^{1} k\, dx\,dy = 1 \Rightarrow k = 1/6$

$\therefore f(x,y) = 1/6$, $0<y<1$, $0<x<3$

Correlation:-

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 result in a corresponding increase (or decrease) in the other, correlation is said to be direct or positive. But if they constantly deviate in opposite directions i.e., if increase (or decrease) in one results in corresponding decrease (or increase) in the other, correlation is said to be -ve.

Correlation Coefficient:-

Correlation coefficient between two r.v. X and Y usually denoted by $r(X,Y)$ or $\rho(X,Y)$ and is defined as

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$$

where $Cov(x,y) = \frac{1}{n}\sum xy - \bar{x}\cdot\bar{y}$

$$\sigma_x = \sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2}$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum y^2 - \bar{y}^2}$$

n is the no. of items in given data.

Note:- The range of correlation coefficient is $|r| \le 1$ (or) $-1 \le r \le 1$.

Note: ① correlation coefficient does not exceed unity

② when $r = 1$ the correlation is perfect and positive

③ Two independent variables are uncorrelated.

$E(x)$ = mean

variance = $E(x^2) - (E(x))^2$

$E(X+Y) = E(X) + E(Y)$ with $f_X(x)$ and $f_Y(y)$

$E(XY) = E(X)\cdot E(Y)$ if X and Y are Independent

$E(ax+b) = aE(X) + b$

$Var(ax+b) = a^2 Var(X)$

$Cov(XY) = E(XY) - E(X)\cdot E(Y)$.

$E\left(\frac{1}{X}\right) \neq \frac{1}{E(X)}$

$E(\log X) \neq \log(E(X))$

Ex:- 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	71	69	71

Sol:-

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

Now $\bar{x} = \frac{544}{8} = 68$, $\bar{y} = \frac{552}{8} = 69$

$\bar{x}\bar{y} = 68 \times 69 = 4692$

$\sigma_x = \sqrt{\frac{1}{n}\sum x^2 - \bar{x}^2} = \sqrt{\frac{37028}{8} - 4624}$

$= 2.121$

$\sigma_y = \sqrt{\frac{1}{n}\sum y^2 - \bar{y}^2} = \sqrt{\frac{38132}{8} - 4761} = 2.345$

$\therefore r(x,y) = \frac{\text{Cov}(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\sigma_x \cdot \sigma_y} = \frac{\frac{1}{8} \times 37560 - 4692}{2.121 \times 2.345}$

$= 0.6030.$

Ex:- Find the correlation for the following data

X	10	14	18	22	26	30
Y	18	12	14	6	30	36

Sol:- $\rho(X,Y) = ?$

Ex:- If the joint pdf of (X,Y) is given by $f(x,y) = x+y$, $0 \le x \le 1$, $0 \le y \le 1$ find $\rho(X,Y)$.

Sol:- $$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{E(XY) - E(X)\cdot E(Y)}{\sigma_X \cdot \sigma_Y}$$

Now $$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\, dx\,dy$$

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

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

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

$$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\, dx\,dy = \frac{7}{12}$$

$$E(Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\, dx\,dy = \frac{7}{12}$$

$$E(X^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2 f(x,y)\,dx\,dy = \int_0^1\int_0^1 x^2 f(x,y)\,dx\,dy = 5/12$$

$$E(Y^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y^2 f(x,y)\,dx\,dy = \int_0^1\int_0^1 y^2(x+y)\,dx\,dy = 5/12$$

$$Var(X) = E(X^2) - (E(X))^2 = \frac{5}{12} - \left(\frac{7}{12}\right)^2 = \frac{11}{144}$$

$$Var(Y) = E(Y^2) - (E(Y))^2 = \frac{11}{144}$$

$$\sigma_X = \sigma_Y = \frac{\sqrt{11}}{12}$$

$$\therefore \rho(X,Y) = \frac{\frac{1}{3} - \frac{7}{12}\cdot\frac{7}{12}}{\frac{\sqrt{11}}{12}\cdot\frac{\sqrt{11}}{12}} = \frac{-1}{11}$$

Ex:- The independent variables X and Y have pdf's given by

$f_X(x) = \begin{cases} 4x, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$, $f_Y(y) =$. Find the correlation coefficient between X and Y.

Rank correlation :-

Let us suppose that a group of n individuals are arranged in order of merit or proficiency in possession of two characteristics A and B.

Let (x_i, y_i), $i = 1, 2, \ldots n$ be the n ranks of 'n' individual in two characteristics A and B respectively. pearson co-efficient of correlation between the ranks x_i's and y_i's is called the rank correlation coefficient between the characteristics A and B for that group of individuals and is given by

$$\rho = r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \quad \text{where } d_i = x_i - y_i$$

This formula is called Spearman's formula for rank correlation.

Ex:- Find the rank correlation coefficient from the following data.

Rank in x	1	2	3	4	5	6	7
Rank in y	4	3	1	2	6	5	7

Sol:-

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

Rank correlation coefficient,

$$r(x, y) = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 20}{7(49-1)}$$

$$= 1 - \frac{120}{336}$$

$$= 0.6429$$

Ex:- The ranks of some 16 students in mathematics and physics are as follows. Calculate Rank correlation coefficient for proficiency in mathematics and physics.

Rank in maths	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Rank in physics	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13

Sol:- $\sum d_i^2 = 136$

$$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 136}{16(256-1)} = 1 - 0.2 = 0.8$$

Repeated ranks:-

If any two or more individuals are equal in any classification with respect to characteristic A or B, then Spearman's formula for calculating the rank correlation coefficient breaks down. In this case the common rank is given to the repeated ranks. This common rank is the average of the ranks.

In the correlation 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.

Ex:- 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

Sol:

x	y	Rank x (x_i)	Rank y (y_i)	$d_i = x_i - y_i$	d_i^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
					$\sum d_i^2 = 72$

Correction factors

In x series 75 repeated twice

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

In x series 64 repeated thrice

$$\therefore C.F. = \frac{3(3^2-1)}{12} = 2$$

In y series 68 repeated twice.

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

∴ Rank correlation coefficient $r = 1 - \frac{6(\sum d^2 + 1/2 + 2 + 1/2)}{10(10^2-1)}$

$= 1 - \frac{6(72 + 0.5 + 2 + 0.5)}{10 \times 99}$

$= 1 - \frac{450}{990} = 0.5454$

Ex:- A sample space of 12 fathers and their eldest sons have the following data about their heights in inches.

Fathers :	65	63	67	64	68	62	70	66	68	67	69	71
Sons :	68	66	68	65	69	66	68	65	71	67	68	70

Sol:- $\sum d_i^2 = 72.5$, In x series 69, 67 are repeated twice

∴ c.f = 1/2, 1/2 respectively.

In y series 68 repeated four times

$c.f = \frac{4(4^2-1)}{12} = 5$

In y series 66, 65 are repeated twice.

∴ c.f = 1/2, 1/2 respectively.

∴ Rank correlation coefficient $= 1 - \frac{6(72.5 + 0.5 + 0.5 + 5 + 0.5 + 0.5)}{12(144-1)}$

$= 1 - 0.277 = 0.722$.

Ex:- Find the rank correlation coefficient for the following distribution

Marks in statistics	48	60	72	62	58	40	39	52	30
Marks in Accountancy	62	78	65	70	38	54	60	32	31

Ans: 0.667

Ex:- Calculate the rank correlation coefficient for the following data.

Marks in statistics	45	58	39	54	45	40	56	60	30	36
Marks in maths	40	36	30	44	36	32	45	42	20	36

Ans: 0.764

Regression:-

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

Lines of Regression:-

If the variables in a bivariate distribution are related we will find that the points in the scattered diagram will cluster around 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 the regression is said to be curvilinear.

(i) The line of regression of Y on X is given by

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

and $r\cdot\frac{\sigma_y}{\sigma_x}$ is regression coefficient of Y on X, $\bar{x}$ - is mean, $\bar{y}$ - is mean

(ii) The line of regression of X on Y is given by

$$x-\bar{x} = r\cdot\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$, $r\cdot\frac{\sigma_x}{\sigma_y}$ is regression coefficient.

Note:- Both lines of regression passes through $(\bar{x},\bar{y})$

Angle between two lines of regression :-

If the equations of lines of regression of Y on X and X on Y are

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

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

The angle 'θ' between the two lines of regression is given by

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

Note:- ① If $r=0$, we get $\theta=\frac{\pi}{2}$

∴ when $r=0$ the lines of regression are perpendicular to each other.

② If $r=\pm1$ then $\tan\theta=0$

$\theta=0$ or π

∴ When $r=\pm1$, the two regression lines are parallel each other or coincide.

(iii) When $r = 0$, the two variables x and y are uncorrelated.

(iv) When $r = \pm 1$, " " " " are said to be perfect.

Regression coefficient :-

Regression coefficient of y on x

$$r \cdot \frac{\sigma_y}{\sigma_x} = b_{yx} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} \quad \text{---------- (1)}$$

Regression coefficient of x on y.

$$r \cdot \frac{\sigma_x}{\sigma_y} = b_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (y-\bar{y})^2} \quad \text{---------- (2)}$$

from (1) and (2), we get

$$r \cdot \frac{\sigma_y}{\sigma_x} \times r \frac{\sigma_x}{\sigma_y} = b_{xy} \cdot b_{yx}$$

$$r^2 = b_{xy} \cdot b_{yx}$$

$$r = \pm \sqrt{b_{xy} \cdot b_{yx}}$$

Eg- From the following data, find

(i) The two regression equations

(ii) The coefficient of correlation between the marks in Economics and Statistics.

(iii) The most likely marks in Statistics when marks in Economics are 30.

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

Sol:-

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

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

Coefficient of regression of y on x is

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

Coefficient of regression of x on y is

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

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

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

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

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

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

Equation of the line of regression of y on x is

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

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

$$y = -0.6643x + 38 + 0.6643 \times 32$$

$$y = -0.6643x + 59.2576.$$

(ii) Coefficient of correlation is

$$r = \pm\sqrt{b_{yx} \times b_{xy}}$$

$$= \pm\sqrt{-0.6643 \times (-0.2337)}$$

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

$$= \pm 0.394$$

(iii) Now we have to find the most likely marks in Statistics (Y) when marks in Economics (X) are 30. We use the line of regression y on x

(ii) $y = -0.6643x + 59.2575$

Put $x = 30$, we get

$$y = -0.6643(30) + 59.2575$$

$$= 39.3282$$

$$= 39.$$

Ex:- Heights of fathers and sons are given in centimetry.

X: Height of father	150	152	155	157	160	161	164	166
Y: " " son	154	156	158	159	160	162	161	164

find the two lines of regression and calculate the expected height of the son when the height of the father is 154 cm.

Sol: $\bar{x} = 158.13$, $\bar{y} = 159.25$

$b_{yx} = 0.555$, $b_{xy} = 1.68$

$x = 1.68y - 109.41$, $y = 0.56x + 70.097$ ——— (2)

When $x = 154$, from (2) $y = 156.937$ cm.

Ex:- The two lines of regression are

$8x - 10y + 66 = 0$ ——— (1)

$40x - 18y - 214 = 0$ ——— (2). The variance of x is 9.

Find (i) The mean values of x and y (ii) correlation coefficient between x and y.

Sol:- Let $\bar{x}$, $\bar{y}$ be the means of x and y respectively.

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

i.e, $8\bar{x} - 10\bar{y} = -66$ ——— (3)

$40\bar{x} - 18\bar{y} = 214$ ——— (4)

(3) × 5 ⇒ $40\bar{x} - 50\bar{y} = -330$

(4) × 1 ⇒ $40\bar{x} - 18\bar{y} = 214$

(-) (+) (-)

$-32\bar{y} = -544$

$\bar{y} = \frac{544}{32} = 17$

Substituting $\bar{y} = 17$ in (3), we get

$8\bar{x} - 10 \times 17 = -66$

$\bar{x} = 13$

(ii) Let us suppose that equation ① is the equation of line of regression of y on x and ② is the equation of line of regression of x on y.

① $\Rightarrow$ $10y = 8x + 66 \Rightarrow y = \frac{8}{10}x + \frac{66}{10}$ $\quad \therefore b_{yx} = \frac{8}{10}$

② $\Rightarrow$ $40x = 18y + 214 \Rightarrow x = \frac{18}{40}y + \frac{214}{40}$ $\quad \therefore b_{xy} = \frac{18}{40}$

$$\therefore r = \pm\sqrt{b_{xy} \cdot b_{yx}}$$

$$= \pm \sqrt{\frac{18}{40} \times \frac{8}{10}}$$

$$= \pm \frac{3}{5}$$

$$\therefore r = \pm 0.6$$

Since both regression coefficients are positive, r must be positive

$$r = 0.6$$

Ex:- The following data were available $\bar{x} = 970$, $\bar{y} = 18$, $\sigma_x = 38$, $\sigma_y = 2$. Correlation coefficient $= 0.6$. Find the line of regression and obtain the value of x on $y = 20$.

Sol:- We know that the line of regression of x on y is given by

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

Given $\bar{x} = 970$, $\bar{y} = 18$, $\sigma_x = 38$, $\sigma_y = 2$, $r = 0.6$

$$\therefore x - 970 = 0.6 \times \frac{38}{2}(y - 18)$$

$$= 11.4y - 205.2$$

$$x = 11.4y + 764.8$$

Now when $y = 20$, we get

$$x = 11.4(20) + 764.8$$

$$x = 992.8.$$

Ex:- The regression equation of x and y is $3y - 5x + 108 = 0$. If the mean value of y is 44 and variance of x is $\frac{9}{16}$ th of the variance of y. Find the mean value of x and the correlation coefficient.

Sol:- Since the line of regression passes through $(\bar{x}, \bar{y})$, we get

$$3\bar{y} - 5\bar{x} + 108 = 0$$

$$3(44) - 5\bar{x} + 108 = 0$$

$$5\bar{x} = 108 + 132$$

$$\therefore \bar{x} = 48$$

∴ mean value of x is 48.

Given $3y - 5x + 108 = 0$

$$5x = 3y + 108$$

$$x = \frac{3}{5}y + \frac{108}{5}$$ which is the line of regression of x on y.

$$\therefore b_{xy} = \frac{3}{5}$$

$$r \cdot \frac{\sigma_x}{\sigma_y} = \frac{3}{5}$$

Given $\sigma_x^2 = \frac{9}{16}\sigma_y^2$

$$\sigma_x = \frac{3}{4}\sigma_y$$

$$r \cdot \frac{3}{4} \cdot \frac{\sigma_y}{\sigma_y} = \frac{3}{5}$$

$$r = \frac{12}{15} = 0.8$$

∴ correlation coefficient is $r = 0.8$

Ex:- The two regression equations of the variables x and y are $x = 19.93 - 0.87y$ and $y = 11.64 - 0.50x$ find

(i) mean of x (ii) mean of y (iii) correlation coefficient between x and y

Ans:- (i) $\bar{x} = 15.94$, $\bar{y} = 3.67$, $r = -0.66$

(ii)

Ex:- The following regression equations were obtained from a correlation table : $y = 0.516x + 33.73$, $x = 0.512y + 33.52$

Find the value of (i) the r (ii) the mean of x's (iii) mean of y's,

Ans: $r = 0.514$, $\bar{x} = 69.0268$ $\bar{y} = 69.3476$

Ex:- The table bellow gives the respective heights x and y of a sample of 12 fathers and their sons. Find the equation of regression line of y on x.

Height of fathers	65	63	67	64	68	62	70	66	68	67	69	71
Height of sons	68	66	68	65	69	66	68	65	71	67	68	70

Ans: $y = 0.476x + 35.85$

Ex:- Find the regression line of y on x for the data

x	1	4	2	3	5
y	3	1	2	5	4

(Ans: $y = 2.7 + 0.1x$)

Ex:- Find the regression line of y on x if

x	4000	6000	8000	10000
y	2.3	4.1	5.7	6.9

(Ans: $y = -0.64 + 0.00077x$)

Ex:- Find the regression line of y on x if

x	40	70	50	60	80	50	90	40	60	60
y	2.5	6.0	4.5	5.0	4.5	2.0	5.5	3.0	4.5	3.0

(Ans: $7x - 120y + 66 = 0$)

Partical correlation:-

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

If there are three variables x_1, x_2, x_3, there will be three coefficients of partial correlation, each studying the relationship between two variables when the third is held constant. If we denote $\rho_{12.3}$, i.e, the coefficient of partial correlation between x_1 and x_2 keeping x_3 constant, it is calculated as

$$\rho_{12.3} = \frac{\rho_{12} - \rho_{13}\cdot\rho_{23}}{\sqrt{1-\rho_{13}^2}\cdot\sqrt{1-\rho_{23}^2}}$$

Similarly $$\rho_{13.2} = \frac{\rho_{13} - \rho_{12}\cdot\rho_{23}}{\sqrt{1-\rho_{12}^2}\cdot\sqrt{1-\rho_{23}^2}}$$

$$\rho_{23.1} = \frac{\rho_{23} - \rho_{12}\cdot\rho_{13}}{\sqrt{1-\rho_{12}^2}\cdot\sqrt{1-\rho_{13}^2}}$$

Note:- Correlation between two variables only say ρ_{12} or ρ_{xy} are called as zero order coefficients. partial co-efficients such as $\rho_{12.3}$, $\rho_{13.2}$, $\rho_{23.1}$ are referred to as first order coefficients. $\rho_{12.34}$; $\rho_{23.14}$, etc are called second order coefficients.

Ex:- In a trivariate distribution, it is found that $\rho_{12} = 0.7$, $\rho_{13} = 0.61$ and $\rho_{23} = 0.4$. find the partial correlation coefficients.

Sol:- $$\rho_{12.3} = \frac{\rho_{12} - \rho_{13}\cdot\rho_{23}}{\sqrt{1-\rho_{13}^2}\cdot\sqrt{1-\rho_{23}^2}} = \frac{0.7 - (0.61\times 0.4)}{\sqrt{1-(0.6)^2}\sqrt{1-(0.4)^2}} = 0.628$$

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

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

Ex:- Is it possible to get the following from a set of experimental data?
$r_{12} = 0.6$, $r_{23} = 0.8$, $r_{13} = -0.5$

Soln:- $$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{\sqrt{1-r_{13}^2} \cdot \sqrt{1-r_{23}^2}} = \frac{0.6 - (-0.5 \times 0.8)}{\sqrt{1-(-0.5)^2} \cdot \sqrt{1-(0.8)^2}} = 1.923$$

Since the value of $r_{12.3}$ is greater than 1, there is some inconsistency in the given data.

Multiple correlation:-

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

The coefficient of multiple correlation with three variables x_1, x_2, x_3 are $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$. $R_{1.23}$ is the coefficient of multiple correlation related to x_1 as a dependant variable and x_2, x_3 as two independant variables it can be expressed in terms of r_{12}, r_{23}, r_{13} as

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

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

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

Note:- (1) $R_{1.23} = R_{1.32}$, $R_{2.13} = R_{2.31}$ etc.

(2) A coefficient of multiple correlation lies between 0 and 1. If it is equal to 1, the correlation is perfect.

Multiple Regression

If the number of independent variables in a regression model is more than one, then the model is called as multiple regression. In fact, many of the real-world applications demand the use of multiple regression models.

A sample application is as stated below:

$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$

where Y represents the economic growth rate of a country, X_1 represents the time period, X_2 represents the size of the populations of the country, X_3 represents the level of employment in percentage, X_4 represents the percentage of literacy, b_0 is the intercept and b_1, b_2, b_3 and b_4 are the slopes of the variables X_1, X_2, X_3 and X_4 respectively. In this regression model, X_1, X_2, X_3 and X_4 are the independent variables and Y is the dependent variable.

Regression Model with Two Independent Variables using Normal Equations:

Suppose the number of independent variables is two, then $Y = b_0 + b_1X_1 + b_2X_2$.
Normal equations are

$$\Sigma Y = nb_0 + b_1\Sigma X_1 + b_2\Sigma X_2$$
$$\Sigma YX_1 = b_0\Sigma X_1 + b_1\Sigma X_1^2 + b_2\Sigma X_1X_2$$
$$\Sigma YX_2 = b_0\Sigma X_2 + b_1\Sigma X_1X_2 + b_2\Sigma X_2^2$$

where n is the total number of combinations of observations. The solution to the above set of simultaneous equations will form the results for the coefficients b_0, b_1 and b_2 of the regression model.

Example 1: The annual sales revenue (in crores of rupees) of a product as a function of sales force (number of salesmen) and annual advertising expenditure (in lakhs of rupees) for the past 10 years are summarized in the following table. Fit a least squares regression model.

Annual sales revenue Y	20	23	25	27	21	29	22	24	27	35
Sales force X_1	8	13	8	18	23	16	10	12	14	20
Annual advertising expenditures X_2	28	23	38	16	20	28	23	30	26	32

Solution: Let the regression model be $Y = b_0 + b_1X_1 + b_2X_2$
where Y is the annual sales revenue; X_1 is the sales force; X_2 is the annual advertising expenditures.

Y	X_1	X_2	X_1^2	X_2^2	X_1X_2	YX_1	YX_2
20	8	28	64	784	224	160	560

23	13	23	169	529	229	299	529
25	8	38	64	1444	304	200	950
27	18	16	324	256	288	486	432
21	23	20	529	400	460	483	420
29	16	28	256	784	448	464	812
22	10	23	100	529	230	220	506
24	12	30	144	900	360	288	720
27	14	26	196	676	364	378	702
35	20	32	400	1024	640	700	1120
253	**142**	**264**	**2246**	**7326**	**3617**	**3678**	**6751**

Substituting the required values in the normal equations, we get the following simultaneous equations

$$10\,b_0 + 142\,b_1 + 264\,b_2 = 253$$

$$142\,b_0 + 2246\,b_1 + 3617\,b_2 = 3678$$

$$264\,b_0 + 3617\,b_1 + 7326\,b_2 = 6751$$

The solution to the above set of simultaneous equation is $b_0 = 5.1483$, $b_1 = 0.6190$ and $b_2 = 0.4304$.

Therefore, the regression model is $Y = 5.1483 + 0.6190X_1 + 0.4304X_2$.