Study Material
on
Bivariate Sampling
Paper: MTM .C204 A
Course: Mathematics (CBCS) PG 2nd Sem.
Prepared by Dr. Ajoy Kumar Maiti

Bivariate Sample:

The population (existent and hypothetical) of a single random variable x has been defined earlier and such population has been called a univariate population. Often we come across situations in which our focus is simultaneously on two or more variables and invariably, we observe that movements in one variable are accompained by movements in other variable.

For example, husband's age and wife's age, Studies in income and expenditure on households or price and demand of commodities.

Let X, Y be the random variables defined on the event spaces S of a random experiment E. Then (X, Y) is a two-dimensional random variable which a mapping of S to R×R, where R is the set of all real numbers. Let E be performed once and let the outcome $\omega \in S$ be obtained.

If $(X,Y)(\omega) = (X(\omega), Y(\omega)) = (x, y)$ where $x \in R$, $y \in R$, then (x, y) is called an observed value of the two-dimensional random variable (X, Y). The totality of all such observed values of (X, Y) obtained by repeating E under uniform conditions infinite number of times, is called a bivariate population of X and Y.

The joint distribution function $F(x, y)$ of x and y is said to determine the distribution of the bivariate population of x and y.

Bivariate Sample: The observed values of (X, Y) are obtained as $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ then the ordered n-tuple of pairs of observed values $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$ is called a bivariate random sample or simply a bivariate sample of size n drawn from the bivariate population of x and y.

Distribution of bivariate sample:

Let $((x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n))$ be a bivariate sample drawn from the bivariate population of X and Y. We define two fake random variables $\hat{x}, \hat{y}$ as follows:

$$\hat{x}(x_i,y_i) = x_i, \quad \hat{y}(x_i,y_i) = y_i, \quad i = 1,2,\ldots,n$$

The joint distribution of $\hat{x}$ and $\hat{y}$ defined by

$P(\hat{x}=x_i, \hat{y}=y_i) = \frac{1}{n}$ for $i = 1,2,\ldots,n$. is called the distribution of the bivariate sample $((x_1,y_1), (x_2,y_2), \ldots (x_n,y_n))$.

Now $P(\hat{x}=x_i) = \frac{1}{n}$ for $i = 1,2,\ldots,n$, is the marginal distribution of $\hat{x}$.

Similarly, $P(\hat{y}=y_i) = \frac{1}{n}$ for $i = 1,2,\ldots,n$, is the marginal distribution of $\hat{y}$.

Now the mean $\bar{x}$, variance S_x, the 2nd order moment $a_{20} = a_{x_2}$ of $\hat{x}$ are given by

$$\bar{x} = E(\hat{x}) = \sum_{i=1}^{n} x_i P(\hat{x}=x_i, \hat{y}=y_i) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\therefore \boxed{\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

$$a_{20} = a_{x_2} = E(\hat{x}^2) = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$S_x^2 = E[(\hat{x}-\bar{x})^2] = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\therefore \boxed{S_x = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = a_{x_2} - \bar{x}^2}$$

Similarly, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $a_{y_2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2$, $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$.

For the bivariate sample $((x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n))$, the sample correlation coefficient r (or ρ) is defined as the correlation coefficient between $\bar{x}$ & $\bar{y}$ and so it is given by

$$r = \frac{Cov(\bar{x},\bar{y})}{S_x S_y}$$

$$= \frac{E[(\bar{x}-\bar{x})(\bar{y}-\bar{y})]}{S_x S_y}, \quad S_x > 0, \ S_y > 0$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\sum x_i^2 - \bar{x}^2\right)\left(\frac{1}{n}\sum y_i^2 - \bar{y}^2\right)}}$$

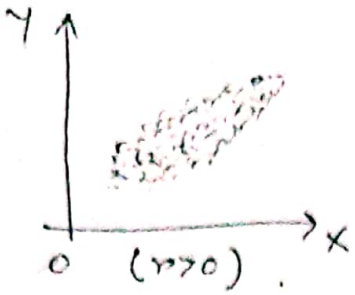
Correlation:

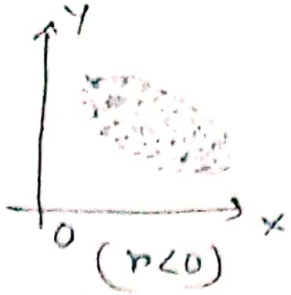
Meaning of Correlation:

In a bivariate distribution, there is any correlation or covariance 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.

Positive or direct correlation: If the two variables deviate in the same direction ie if the increase (or decrease) in one results in a corresponding increase (or decrease) in the other, correlation is said to be direct positive.

Ex (i) The heights and weights of a group of persons
(ii) The income and expenditure are the examples of positive correlation.





Negative or diverse Correlation: If the increase (or decrease) in one results in corresponding decrease (or increase) in the other, the Correlation is said to be diverse or negative.

Ex: (i) Price and demand of a commodity

(ii) ~~and~~ Volume and pressure of a perfect gas

are the examples are negative correlation.

The ~~Corrde~~ Correlation is said to be perfect if the deviation in one variable is followed by a corresponding and proportional deviation in the other.

What is correlation?

Defn: The word 'correlation' is used to denote the degree of association between variables. If two variables x and y are so related that variations in the magnitude of one variable tend to be accompanied by variations in the magnitude of the other variable, they are said to be correlated.

Correlation may be linear or non-linear. If the ~~amount~~ amount of change in one variable tends to bear a constant ratio to the amount of change in the other variable, then the Correlation is said to be linear, because the scatter diagram would show a linear path.

Here, we shall be concerned with linear correlation or simple correlation. This is measured by 'correlation Coefficient'.

Scatter Diagram:

When statistical data relating to the simultaneous measurement on two variables are available, each pair of observations can be geometrically represented by a point on the graph paper — the values of one variable being shown along the X-axis and those of the other variable along Y-axis. If there are n pairs of observations, finally the graph paper will contain n points.

The diagrammatic representation of bivariate data is known as scatter diagram.

A scatter diagram indicates the nature of association between the two variables, ie the type of correlation between them.

Covariance:

Given a set of n pairs of observations $(x_1,y_1), (x_2,y_2), \ldots (x_n,y_n)$ relating to two variables x & y, the Covariance of x & y denoted by $\text{Cov}(x,y)$, is defined as

$$\text{Cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

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

$$\therefore S_x^2 = \frac{1}{n}\Sigma x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$S_y^2 = \frac{1}{n}\Sigma y_i^2 - \bar{y}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

where $\bar{x} = \frac{1}{n}\Sigma x_i$ & $\bar{y} = \frac{1}{n}\Sigma y_i$.

Correlation Coefficient:

Let $(x_1,y_1), (x_2,y_2) \ldots (x_n,y_n)$ be a given set of n pairs of observations on two variables x and y. The correlation coefficient between x & y is denoted by

$$r = \frac{\text{Cov}(x,y)}{S_x S_y}$$

where S_x & S_y are sample variance of x and y respectively.

& $Cov(x,y)$ denotes the covariance of x and y.

Properties of Correlation Coefficient:

(i) The correlation coefficient r is independent of the choice of both origin and scale of observations (measurement).

(ii) The correlation coefficient lies between -1 and $+1$.

i.e. $-1 \le r \le 1$.

(iii) The correlation coefficient r is a pure number and is independent of the units of measurement.

Proof (i): If $u = \frac{x-a}{c}$, $v = \frac{y-b}{d}$ then prove that $r_{xy} = r_{uv}$ when c and d are of same signs and $r_{xy} = -r_{uv}$ when c and d are of opposite signs.

Here $u = \frac{x-a}{c} \Rightarrow x = a + cu$ $\therefore \bar{x} = \frac{\sum x}{n} = \frac{1}{n}\sum(a+cu) = \frac{1}{n}(na + c\sum u)$

$= a + c\frac{\sum u}{n} = a + c\bar{u}$

Similarly, $\bar{y} = b + d\bar{v}$

$\sigma_x^2 = \frac{1}{n}\sum(x-\bar{x})^2 = \frac{1}{n}\left[\sum\{a + cu - (a + c\bar{u})\}^2\right] = c^2 \cdot \frac{1}{n}\sum(u-\bar{u})^2$

$\therefore \sigma_x^2 = c^2\sigma_u^2 \Rightarrow \sigma_x = |c|\sigma_u$.

Similarly, $\sigma_y = |d|\sigma_v$.

$\therefore Cov(x,y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y}) = \frac{1}{n}\sum[c(u-\bar{u})\,d(v-\bar{v})]$

$= cd \cdot \frac{1}{n}\sum(u-\bar{u})(v-\bar{v}) = cd\,Cov(u,v)$.

Now, $r_{xy} = \frac{Cov(x,y)}{\sigma_x\sigma_y} = \frac{cd\,Cov(u,v)}{|c||d|\,\sigma_u\sigma_v} = \frac{cd}{|c||d|}\,r_{uv}$.

$\therefore$ If c and d are of same signs, then $\frac{cd}{|c||d|} = 1 \therefore r_{xy} = r_{uv}$.

If c and d are of opposite signs then $\frac{cd}{|c||d|} = -1 \therefore r_{xy} = -r_{uv}$.

Thus, the absolute value of the correlation coefficient is independent of the choice of origin and the scale.

(ii) If r be the sample correlation coefficient of a bivariate sample $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ then $-1 \le r \le 1$.

Prof

Ans: The sample correlation coefficient r is given by

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

& $s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

Now since $x_1, x_2, \ldots, x_n$; $y_1, y_2, \ldots y_n$ are all real numbers

we have $\left(\frac{x_i - \bar{x}}{s_x} \pm \frac{y_i - \bar{y}}{s_y}\right)^2 \ge 0$ for $i = 1, 2, \ldots, n$

So we have $\left(\frac{x_i - \bar{x}}{s_x}\right)^2 + \left(\frac{y_i - \bar{y}}{s_y}\right)^2 \pm \frac{2(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \ge 0$

Hence we get

$$\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^2 + \sum_{i=1}^{n}\left(\frac{y_i - \bar{y}}{s_y}\right)^2 \pm 2\sum_{i=1}^{n}\frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \ge 0$$

$$\Rightarrow \frac{1}{s_x^2} n s_x^2 + \frac{1}{s_y^2} n s_y^2 \pm 2nr \ge 0$$

$$\Rightarrow 2 \pm 2r \ge 0 \Rightarrow 1 \pm r \ge 0 \Rightarrow -1 \le r \le 1.$$

Regression:

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

If the variables in a bivariate distribution are related, we will find that the points in the scatter diagram will cluster round

Some curve is 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.

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 principle of least squres.

Regression equation of y on x:

The regression equation of y on x is the equation of the best fitting straight line in the form $y = a + bx$, obtained by the method of Least Squres, from the set of n pairs of observations $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$.

According to the principle of least squres, we have to determine a and b so that

$$E = \frac{1}{n}\sum_{i=1}^{n} (y_i - a - bx_i)^2 \text{ is minimum.}$$

From the principle of maxima & minima, $\frac{\partial E}{\partial a} = 0 = -2\sum_{i=1}^{n} (y_i - a - bx_i)$

$$\Rightarrow \Sigma y_i = na + b\Sigma x_i \quad \text{---(1)}$$

and $$\frac{\partial E}{\partial b} = 0 = -2\sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0$$

$$\Rightarrow \Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 \quad \text{---(ii)}$$

Dividing both sides of (1) by n, we get

$$\bar{y} = a + b\bar{x} \Rightarrow a = \bar{y} - b\bar{x}.$$

Substituting this in equation $y = a + bx$

$$\therefore y - \bar{y} = b(x - \bar{x})$$

Again multiplying (i) by Σx_i and (ii) by n we have

$$\Sigma x_i \Sigma y_i = na\Sigma x_i + b(\Sigma x_i)^2$$

$$\& \quad n\Sigma x_i y_i = na\Sigma x_i + nb\Sigma x_i^2$$

& Subtracting 1st from 2nd, we have

$$n\Sigma x_i y_i - \Sigma x_i \Sigma y_i = b\{n\Sigma x_i^2 - (\Sigma x_i)^2\}$$

$$\therefore b = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$= \frac{\frac{1}{n}\Sigma x_i y_i - \frac{\Sigma x_i}{n}\frac{\Sigma y_i}{n}}{\frac{1}{n}\Sigma x_i^2 - \left(\frac{\Sigma x_i}{n}\right)^2} = \frac{Cov(x,y)}{S_x^2} = b_{yx}.$$

∴ The required regression equation of y on x is

$$\boxed{y - \bar{y} = b_{yx}(x - \bar{x})} \quad \text{where } b_{yx} = \frac{Cov(x,y)}{S_x^2}$$

Again $r = \frac{Cov(x,y)}{S_x S_y} \Rightarrow Cov(x,y) = r S_x S_y$

$$\therefore \boxed{b_{yx} = \frac{r S_x S_y}{S_x^2} = r\frac{S_y}{S_x}}$$

- Similarly, the regression equation of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y}) \quad \text{where } b_{xy} = \frac{Cov(x,y)}{S_y^2}$$

$$= r\frac{S_x}{S_y}$$

* Find the angle between the regression lines and comments when $r = 0, \pm 1$.

V. IMP

Ans: The regression line of y on x is $y-\bar{y} = b_{yx}(x-\bar{x})$

Let m_1 = slope of this line = $b_{yx} = r\frac{s_y}{s_x}$.

The regression line of x on y is $x-\bar{x} = b_{xy}(y-\bar{y})$.

Let m_2 = slope of this line = $\frac{1}{b_{xy}} = \frac{1}{r}\frac{s_y}{s_x}$.

Let θ be the angle between these lines then

$$\tan\theta = \frac{m_1 - m_2}{1+m_1m_2} = \frac{\frac{1}{r}\frac{s_y}{s_x} - r\frac{s_y}{s_x}}{1 + r\cdot\frac{1}{r}\cdot\frac{s_y^2}{s_x^2}}$$

$$= \frac{s_y}{s_x}\left(\frac{1-r^2}{r}\right)\frac{s_x^2}{(s_y^2+s_x^2)}$$

$$= \frac{1-r^2}{r}\cdot\frac{s_x s_y}{s_x^2+s_y^2}$$

When $r=0$, $\tan\theta = \infty$, $\Rightarrow \theta = 90^\circ$.

The angle between them is 90° i.e the regression lines are $\perp^r$.

when $r = \pm 1$, $\tan\theta = 0 \Rightarrow \theta = 0 \Rightarrow$ these lines are coincident.

(Ex-1) Using the method of least squares, find the values of x & y from the following $x+y=3$, $2x-y=0.03$, $x+3y=7.03$ and $3x+y=4.97$.

Ans:

$$S = (x+y-3)^2 + (2x-y-0.03)^2 + (x+3y-7.03)^2 + (3x+y-4.97)^2$$

We choose x & y in such a way that S is minimum.

$\therefore \frac{ds}{dx} = 0$ & $\frac{ds}{dy} = 0$

$$\Rightarrow \left.\begin{aligned} 3x+y-5 &= 0 \\ \text{&}\quad 5x+12y &= 29.03 \end{aligned}\right\} \Rightarrow \begin{aligned} x &= .9990322 \\ y &= 2.0029032 \end{aligned}$$

The estimated values of x & y are 0.9990322 & 2.0029032 respectively.

Ex-2 In a partially destroyed ~~~~ laboratory record of an analysis of correlation data is given below.

The following results are only legible

Variance of $x = 9$

Regression equations $8x - 10y + 66 = 0$

$40x - 18y = 214$.

Find (i) The mean values of x & y.

(ii) Correlation coefficient between x & y.

(iii) The standard deviation of y.

Ans (i) The means of x and y are given by

$$8\bar{x} - 10\bar{y} + 66 = 0$$
$$40\bar{x} - 18\bar{y} - 214 = 0$$

Solving these two equations, $\bar{x} = 13$ & $\bar{y} = 17$.

$\therefore$ The means are 13 and 17 respectively.

(ii) The regression line y on x is

$$8x - 10y + 66 = 0$$
$$\Rightarrow y = \frac{8}{10}x + \frac{66}{10}$$
$$\therefore b_{yx} = \frac{8}{10}$$

Also the regression line of x on y is

$$40x - 18y = 214$$
$$x = \frac{18}{40}y + \frac{214}{40}$$
$$\therefore b_{xy} = \frac{18}{40} = \frac{9}{20}$$

$\therefore$ The regression coefficient are $b_{yx} = \frac{4}{5}$ and $b_{xy} = \frac{9}{20}$.

$$\therefore r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{\frac{9}{20} \cdot \frac{4}{5}} = \frac{3}{5}$$

(iii) We know that $b_{yx} = r \cdot \frac{S_y}{S_x}$

$$\Rightarrow \frac{4}{5} = \frac{3}{5} \cdot \frac{S_y}{3}$$

$S_x^2 = 9$

$\Rightarrow S_x = 3$

$\Rightarrow S_y = 1.$

(Ex-3) Calculate the correlation coefficient and determine the regression line of y on x and x on y for the sample

x	8	10	5	8	9
y	1	3	1	2	3

C.H-90, V.H-94

Ans: The calculation are shown in the following table.

x	y	x^2	y^2	xy
8	1	64	1	8
10	3	100	9	30
5	1	25	1	5
8	2	64	4	16
9	3	81	9	27
$\Sigma x = 40$	$\Sigma y = 10$	$\Sigma x^2 = 334$	$\Sigma y^2 = 24$	$\Sigma xy = 86.$

$\therefore \bar{x} = \frac{\Sigma x}{n} = \frac{40}{5} = 8, \quad \bar{y} = \frac{1}{n}\Sigma y = \frac{10}{5} = 2$

$S_x = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2} = \sqrt{\frac{1}{5} \times 334 - 64} = \sqrt{2.8}$

$S_y = \sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2} = \sqrt{\frac{1}{5} \times 24 - 4} = \sqrt{.8}$

$Cov(x,y) = \frac{1}{n}\Sigma xy - \bar{x}\bar{y} = \frac{1}{5} \times 86 - 8 \times 2 = 1.2$

$\therefore r = \frac{Cov(x,y)}{S_x S_y} = \frac{1.2}{\sqrt{2.8}\sqrt{.8}} = .8017837$

$\therefore$ The regression line of x on y is

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

$$\Rightarrow x - 8 = .8\sqrt{\frac{2.8}{0.8}} \cdot (y - 2)$$

Similarly, for the regression line of y on x.

Ex-4 Two regression lines $10x+3y-16=0$ and $6x+5y-16=0$ have been calculated from certain data on two variables x and y. Find the correlation coefficient between the variables.

Ans: Let the regression line of y on x is

$$y = \frac{16-6x}{5}$$

& the regression line of x on y is $x = \frac{16-3y}{10}$

∴ The regression coefficients are $b_{xy} = -\frac{3}{10}$

& $b_{yx} = -6/5$.

∴ $r = -.6$.

Ex-5 The following results were obtained in the analysis of the data on yeild of dry bark (y) and age in years (x) of 200 plants. correlation coefficient = 0.84.

	x	y
Average	9.2	16.5
S.d	2.1	4.2

Construct the two lines of regression and estimate the yield dry bark of a plant of age 8 years. V. H '89.

Ans: The regression line of age (x) on dry bark (y) is

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

$$\Rightarrow x-9.2 = r.\frac{\sigma_x}{\sigma_y}(y-16.5)$$

$$\Rightarrow x = 9.2 + .84 \times \frac{2.1}{4.2}(y-16.5)$$

$$\therefore x = .42y + 2.27.$$

Now the regression line of the dry bark (y) on the age (x) is

$$y-\bar{y} = b_{yx}(x-\bar{x}) \Rightarrow y-16.5 = .84 \times \frac{4.2}{2.1}(x-9.2)$$

$\Rightarrow y = 1.68x + 1.044$.

∴ The duty bank (y) of a plant of age 8 years is

$$y = 1.68 \times 8 + 1.044 = 14.484.$$

Ex 6 Find the most likely price in Bombay corresponding to the price of 70 at Calcutta from the following

	Calcutta	Bombay
Average	65	67
S.d	2.5	3.5

Correlation coefficient between the price commodities in the two cities is 0.8. C.H-94

Ans: Let the price of Calcutta and Bombay be x & y resp.

The regression line of y on x is

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

$$\Rightarrow y - 67 = 0.8 \times \frac{3.5}{2.5}(x - 65)$$

When $x = 70$ at Calcutta, then

$$y = 72.6.$$

∴ Price in Bombay corresponding to the price of Rs. 70 at Calcutta is Rs. 72.6.

Ex 7 If a linear relation exists between the variables x and y Prove that $r = \pm 1$. [V.H-93, 89]

Ans: Let the relation between x and y be $ax + by + c = 0$

Now, $\text{cov}(x, y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$

Now, $\text{Cov}(x,y) = \frac{1}{n}\sum (x_i-\bar{x})\left(-\frac{ax_i+c}{b} + \frac{a\bar{x}+c}{b}\right)$.

$= \frac{1}{n}\sum \left(-\frac{a}{b}\right)(x_i-\bar{x})(x_i-\bar{x})$

$= -\frac{a}{b}\frac{1}{n}\sum (x_i-\bar{x})^2 = -\frac{a}{b} s_x^2$

Also, $s_y^2 = \frac{1}{n}\sum (y_i-\bar{y})^2$

$= \frac{1}{n}\sum \left(-\frac{ax_i+c}{b} + \frac{a\bar{x}+c}{b}\right)^2$

$= \left(\frac{a}{b}\right)^2 \frac{1}{n}\sum (x_i-\bar{x})^2 = \frac{a^2}{b^2} s_x^2$.

Now, $r = \frac{\text{cov}(x,y)}{s_x s_y} = \frac{-\frac{a}{b}\cdot s_x^2}{s_x \cdot \left|\frac{a}{b}\right| s_x}$

$= \frac{\left(-\frac{a}{b}\right)}{\left|\frac{a}{b}\right|}$

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

Ex-8 Two random variables X and Y are connected by the relation $3X+4Y+5=0$. A sample (x_i, y_i), $i=1,2,\cdots,n$ is taken from the bivariate population (X,Y); obtain the correlation coefficient of the sample. V.H -97, 2002, 09

Ans: Given relation is $3x+4y+5=0$.

In real, $3x+4y+5=0$

Replacing a by 3 and b by 4 in the above example-7. we get the correlation coefficient of the sample,

$$r = \frac{-\frac{3}{4}}{\left|\frac{3}{4}\right|} = -1$$

$\therefore r = -1$. [Proceeding same as Ex-7]

Ex-9: Two regression lines are of the form $5x+12y=7$ and $3x+8y=11$. Identify the regression lines.

Ans: Let us assume that the regression line of y on x is $5x+12y=7$ and so $b_{yx} = -\frac{5}{12}$

Then obviously, the regression line of x on y will be $3x+8y=11$ and so $b_{xy} = -8/3$.

Now, b_{xy} and b_{yx} are of the same sign. So, these two lines are possible regression lines. Again, $r^2 = b_{xy} \times b_{yx}$

$$= -\frac{8}{3} \times \left(-\frac{5}{12}\right)$$

$$= \frac{10}{9} > 1 \text{ which is not possible.}$$

So, our assumption is wrong.

Hence, the regression line of y on x is $3x+8y=11$

and the regression line of x on y is $5x+12y=7$.

Ex-10: The following results were obtained from records of age(x) and systolic blood pressure(y) of a group of 10 women.

	x	y
Mean	53	142
Variance	130	165

$\Sigma (x-\bar{x})(y-\bar{y}) = 1220$.

Find the appropriate regression equation and use it to estimate the blood pressure of a women whose age is 45.

Ans: Here, the appropriate regression line will be y on x and it is given by

$$y-\bar{y} = b_{yx}(x-\bar{x}) \text{ where } b_{yx} = r\frac{s_y}{s_x}.$$

Now, $\bar{x}=53$, $\bar{y}=142$, $s_x^2=130$, $s_y^2=165$, $n=10$.

$$\therefore r = \frac{\frac{1}{n}\Sigma (x-\bar{x})(y-\bar{y})}{s_x s_y}$$

$$= \frac{1}{10} \times \frac{1120}{\sqrt{130} \times \sqrt{165}}$$

Therefore, $b_{yx} = \frac{1}{10} \times \frac{1120}{\sqrt{130} \times \sqrt{165}} \times \frac{\sqrt{165}}{\sqrt{130}}$

$$= \frac{112}{130} = \frac{61}{65}$$

So, the regression line will be

$$y - 142 = \frac{61}{65}(x - 53)$$

$$\Rightarrow 61x - 65y + 5997 = 0$$

Now, when age is 45, $x = 45$

$$\therefore 65y = 61 \times 45 + 5997 = 8742$$

$$\therefore y = \frac{8742}{65} = 134.49.$$

$\therefore$ Blood pressure would be 134.49.

(Ex) A computer while calculating correlation coefficient between two variables x and y from 25 pairs of observation, obtain 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 it had copied two pairs as

x	y
8	14
8	6

while the correct values were

x	y
8	12
6	8

Obtain the correct values of correlation coefficient.

[V.H-90], C.H-02

Ans

Let X and Y denotes the correct values of x and y. So, the correct correlation coefficient is

$$r = \frac{\frac{1}{n}\Sigma XY - \frac{\Sigma X}{n} \cdot \frac{\Sigma Y}{n}}{\sqrt{\frac{1}{n}\Sigma X^2 - \left(\frac{\Sigma X}{n}\right)^2} \cdot \sqrt{\frac{1}{n}\Sigma Y^2 - \left(\frac{\Sigma Y}{n}\right)^2}}$$

—(1)

So, $\Sigma x = 125 - 6 - 8 + 8 + 6 = 125$

$\Sigma y = 100 - 14 - 6 + 12 + 8 = 100$

$\Sigma x^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

$\Sigma y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$.

$\Sigma xy = 508 - 6\times 14 - 8\times 6 + 8\times 12 + 6\times 8 = 520$.

Thus, from (1), we have,

$$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

$$= \frac{25\times 520 - 125\times 100}{\sqrt{(25\times 650 - 125\times 125)(25\times 436 - 100\times 100)}}$$

$$= \frac{25\times 20}{\sqrt{25\times 25\times 25\times 36}} = \frac{25\times 20}{25\times 6\times 5}$$

$$= \frac{2}{3}$$

(Ex) A bivariate sample of size 11 give the results $\bar{x} = 7$, $S_x = 2$, $\bar{y} = 9$, $S_y = 4$, $r = 0.5$. It was later found that 1 pair of the sample values $x = 7$ & $y = 9$ was inaccurate and was rejected. How would the original value of r be effected by rejection? v. H-04

Ans: As the one pair be rejected, so, the remaining sample size is $11 - 1 = 10$.

Given that $n = 11$, $\bar{x} = 7$, $S_x = 2$, $\bar{y} = 9$, $S_y = 4$, $r = 0.5$.

$\therefore \bar{x} = 7 \Rightarrow \frac{1}{n}\Sigma x = 7 \Rightarrow \Sigma x = 77$.

Also, $\bar{y} = 9 \Rightarrow \frac{1}{n}\Sigma y = 9 \Rightarrow \Sigma y = 99$.

$S_x = 2$

$\Rightarrow \sqrt{\frac{1}{n}\Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2} = 2$

$\Rightarrow \frac{\Sigma x^2}{n} = 4 + \left(\frac{77}{11}\right)^2 = (4 + 7^2) = 4 + 49 = 53$

$\Rightarrow \Sigma x^2 = 583.$

Again, $S_y = 4 \Rightarrow \frac{1}{n}\Sigma y^2 = 16 + \left(\frac{\Sigma y}{11}\right)^2$

$= 16 + 81 = 97.$

$\therefore \Sigma y^2 = 1067.$

Also, $r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$

$= \frac{11 \times \Sigma xy - 77 \times 99}{\sqrt{11 \times 583 - 77 \times 77}\sqrt{11 \times 1067 - 99 \times 99}}$

$= \frac{11 \times (\Sigma xy - 693)}{\sqrt{11 \times 44 \times 11 \times 11 \times 16}} = \frac{11 \times (\Sigma xy - 693)}{11 \times 11 \times 2 \times 4}$

$= \frac{\Sigma xy - 693}{88}$

$\Rightarrow 88 \times 0.5 = \Sigma xy - 693$

$\Rightarrow \Sigma xy = 693 + 44 = 737.$

Let x and y denote the correct values of x & y.

So, $\Sigma x = 77 - 7 = 70$

$\Sigma y = 99 - 9 = 90$

$\Sigma x^2 = 583 - 49 = 534$

$\Sigma y^2 = 1067 - 81 = 986$

$\Sigma xy = 737 - 63 = 674.$

Thus, the correct Correlation coefficient is

$$r = \frac{10\times 674 - 70\times 90}{\sqrt{10\times 534 - 70\times 70}\ \sqrt{10\times 986 - 90\times 90}}$$

$$= \frac{10\times 44}{10\times 2\times 22\times 2} = \frac{1}{2} = 0.5.$$

Thus, there is no effect due to rejection.

Ex: In order to find the correlation coefficient between two variables x and y from 12 pairs of observations, the following calculation were made $\Sigma x = 30$, $\Sigma y = 5$, $\Sigma x^2 = 670$, $\Sigma y^2 = 285$ and $\Sigma xy = -334$. On subsequent verification, it was found that the pair $(x=11, y=4)$ was copied ~~one~~ wrongly, the correct value being $(x=10, y=14)$.

Find the correct value of correlation coefficient.

Ans: Let X and Y be the correct values of x and y resp.

$\therefore$ $\Sigma x = 30 - 11 + 10 = 29$

$\Sigma Y = 5 - 4 + 14 = 15$

$\Sigma x^2 = 670 - 11^2 + 10^2 = 649$

$\Sigma Y^2 = 285 - 4^2 + 14^2 = 465$

$\Sigma XY = -334 - 11\times 4 + 10\times 14 = -238$

$n = 12$.

The Correlation coefficient $(r) = \dfrac{n\Sigma XY - \Sigma x \Sigma Y}{\sqrt{\{n\Sigma x^2 - (\Sigma x)^2\}\{n\Sigma Y^2 - (\Sigma Y)^2\}}}$

$$= \frac{12\times(-238) - 29\times 15}{\sqrt{12\times 649 - 29^2}\ \sqrt{12\times 465 - (15)^2}}$$

$$= -0.539$$

So, the correct value of the correlation coefficient is -0.539.

Ex: If two variables x and y are connected by the relation $6x-y-10$, find the correlation coefficient between x and y.

Ans: Given that $6x - y = 10$

$$\Rightarrow y = 6x - 10$$

$$\Rightarrow \bar{y} = 6\bar{x} - 10$$

$$\Rightarrow y - \bar{y} = 6(x - \bar{x}) \quad \text{---(1)}$$

$$\Rightarrow (y-\bar{y})^2 = 36(x-\bar{x})^2$$

$$\Rightarrow \frac{1}{n}\sum (y-\bar{y})^2 = \frac{36}{n}\sum (x-\bar{x})^2$$

$$\Rightarrow \sigma_y^2 = 36\,\sigma_x^2$$

$$\Rightarrow \sigma_y = 6\sigma_x. \quad \text{---(2)}$$

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

$$= \frac{1}{n}\sum (x-\bar{x})\,6\,(x-\bar{x}) \quad [\text{using (1)}]$$

$$= 6 \cdot \frac{1}{n}\sum (x-\bar{x})^2 = 6\sigma_x^2$$

$\therefore$ Correlation coefficient $= \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{6\sigma_x^2}{\sigma_x \cdot 6\sigma_x}$ [using (2)]

$$= 1.$$

Ex: Fit a straight line by the method of least squares to the following

x:	60	61	62	63	64
y:	40	42	48	52	55.

Ans: Let the equation of the best fitted straight line $y = a + bx$ ---(1).

Let $u = x - 62$, $v = y - 48$ then (1) takes the form

$$v = A + Bu \quad \text{---(2)}$$

Here $n = 5$

The normal equations are $\Sigma v = nA + B\Sigma u$ ---(3)

and $\Sigma uv = A\Sigma u + B\Sigma u^2$ ---(4)

Calculation for fitting straight line

x	y	$u=x-62$	$v=y-48$	uv	u^2
60	40	−2	−8	16	4
61	42	−1	−6	6	1
62	48	0	0	0	0
63	52	1	4	4	1
64	55	2	7	14	4
Total		$\Sigma u=0$	$\Sigma v=-3$	$\Sigma uv=40$	$\Sigma u^2=10$

From (3) and (4) we have

$$-3 = 5A + B\cdot 0 \Rightarrow A = -3/5$$

and $$40 = A\cdot 0 + B\cdot 10 \Rightarrow B = 4.$$

$\because u = x-62$, $v = y-48$.

The straight line is $y - 48 = -3/5 + 4(x-62)$

$$\Rightarrow 20x - 5y = 1003.$$

This is the required best fitted straight line.

Ex: Fit a straight line $y = a+bx$ by the method of least squres to the following data:

x:	0	5	10	15	20	25	30
y:	10	14	19	25	31	36	39

Ans Try yourself

Ex: Fit a 2nd degree parabola $y = a+bx+cx^2$ to the following data:

x:	0	1	2	3	4
y:	1	5	10	22	38

Ans: Given that the 2nd degree parabola $y = a+bx+cx^2$ ... (1)

Let $u = x - 2$ then (1) takes the form $y = A + Bu + Cu^2$ —(2).

x	y	$u = x-2$	u^2	u^3	u^4	uy	u^2y
0	1	−2	4	−8	16	−2	4
1	5	−1	1	−1	1	−5	5
2	10	0	0	0	0	0	6
3	22	1	1	1	1	22	22
4	38	2	4	8	16	76	152
Total	$\Sigma y = 76$	$\Sigma u = 0$	$\Sigma u^2 = 10$	$\Sigma u^3 = 0$	$\Sigma u^4 = 34$	$\Sigma uy = 91$	$\Sigma u^2y = 183$

Here $n = 5$

The normal equations are $\Sigma y = nA + B\Sigma u + C\Sigma u^2$

i.e. $76 = 5A + 10C$ —(3)

$$\Sigma uy = A\Sigma u + B\Sigma u^2 + C\Sigma u^3$$

$$\Rightarrow 91 = A \cdot 0 + B \cdot 10 + C \cdot 0 \Rightarrow B = 9.1$$

and $\Sigma u^2y = A\Sigma u^2 + B\Sigma u^3 + C\Sigma u^4$

$$\Rightarrow 183 = A \times 10 + B \cdot 0 + C \cdot 34$$

$$\Rightarrow A + 3.4C = 18.3 \quad \text{—(4)}.$$

Solving (3) and (4) we have $C = 2.2$ and $A = 10.82$

also $B = 9.1$.

$\therefore$ (2) becomes $y = 10.82 + 9.1u + 2.2 \cdot u^2$

$$\Rightarrow y = 10.82 + 9.1(x-2) + 2.2(x-2)^2$$

$$= 1.42 + 0.3x + 2.2x^2.$$