possible mathematical relation between the two. Very naturally, this relation depends on the assumption of independence and dependence of variables. That is, if x is taken as independent variable, y as dependent variable, then we get one relation, called the *regression equation of* y *on* x, and similarly if y is taken as independent variable and x as an dependent variable, then we get another relation, called the *regression equation* x *on* y. It is to be noted that under some stringent conditions the two relations may be identical.

Correlation analysis

Consider the following bivariate data:

$$\begin{array}{llllll} x: & x_1 & x_2 & x_3 & \dots & x_n \\ y: & y_1 & y_2 & y_3 & \dots & y_n \end{array}$$

Then the covariance of the two variables x and y is denoted by $\mathrm{Cov}(x, y)$ and defined by

$$\mathrm{Cov}(x.y) = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

where,

$$\bar{x} = \frac{1}{n}\Sigma_{i=1}^{n}x_i \quad \text{and} \quad \bar{y} = \frac{1}{n}\Sigma_{i=1}^{n}y_i.$$

Another form of the covariance formula is

$$\mathrm{Cov}(x, y) = \frac{1}{n}\Sigma_{i=1}^{n}x_iy_i - \bar{x}\bar{y}$$

which one can deduce from the above covariance formula.

In fact

$$\begin{aligned} \mathrm{Cov}(x.y) &= \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{n}\Sigma_{i=1}^{n}(x_iy_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) \\ &= \frac{1}{n}\Sigma_{i=1}^{n}x_iy_i - \bar{y}\frac{1}{n}\Sigma_{i=1}^{n}x_i - \bar{x}\frac{1}{n}\Sigma_{i=1}^{n}y_i + \frac{1}{n}\Sigma_{i=1}^{n}\bar{x}\bar{y} \\ &= \frac{1}{n}\Sigma_{i=1}^{n}x_iy_i - \bar{y}\bar{x} - \bar{x}\bar{y} + \frac{1}{n}n\bar{x}\bar{y} \\ &= \frac{1}{n}\Sigma_{i=1}^{n}x_iy_i - \bar{x}\bar{y} \end{aligned}$$

Karl Pearson correlation coefficient

The correlation coefficient of the two variables x and y is denoted by r and is defined by

$$r_{xy} = \frac{\mathrm{Cov}(x, y)}{\sigma_x\sigma_y}$$

where $\sigma_x = \frac{1}{n}\Sigma_{i=1}^n x_i^2 - (\bar{x})^2$ and $\sigma_y = \frac{1}{n}\Sigma_{i=1}^n y_i^2 - (\bar{y})^2$.

Hence,

$$r_{xy} = \frac{\text{Cov}(x,y)}{\sigma_x\sigma_y} = \frac{\frac{1}{n}\Sigma_{i=1}^n x_i y_i - \bar{x}\bar{y}}{\sqrt{\left(\frac{1}{n}\Sigma_{i=1}^n x_i^2 - (\bar{x})^2\right)}\left(\frac{1}{n}\Sigma_{i=1}^n y_i^2 - (\bar{y})^2\right)}$$

Property 1. The correlation coefficient r is a pure number and is independent of units of measurement, i.e., it has no unit.

Property 2. The correlation of coefficient r is independent of the choice of origin.

Proof: Let (x, y) and (u, v) be the two sets of bivariate data such that $u = x - a$ and $v = y - b$ where a and b are constants.

$$\therefore \bar{u} = \bar{x} - a \text{ and } \bar{v} = \bar{y} - b.$$

$$\therefore \bar{u} = \bar{x} - a \text{ and } \bar{v} = \bar{y} - b.$$

$$\therefore \text{var}(u) = \sigma_u^2 = \frac{1}{n}\Sigma_{i=1}^n(u_i - \bar{u})^2 = \frac{1}{n}\Sigma_{i=1}^n(x_i - a - \bar{x} + a)^2 = \frac{1}{n}\Sigma_{i=1}^n(x_i - \bar{x})^2 = \sigma_x^2.$$

Similarly $\text{var}(v) = \sigma_v^2 = \sigma_y^2$.

$$\begin{aligned}
\text{Cov}(u.v) &= \frac{1}{n}\Sigma_{i=1}^n(u_i - \bar{u})(v_i - \bar{v}) \\
&= \frac{1}{n}\Sigma_{i=1}^n(x_i - a - \bar{x} + a)(y_i - b - \bar{y} + b) \\
&= \frac{1}{n}\Sigma_{i=1}^n(x_i - \bar{x})(y_i - \bar{y}) \\
&= \text{Cov}(x,y)
\end{aligned}$$

$$\therefore \quad r_{uv} = \frac{\text{Cov}(u,v)}{\sigma_u\sigma_v} = \frac{\text{Cov}(x,y)}{\sigma_x\sigma_y} = r_{xy}.$$

Property 3. let (x, y) and (u, v) be such that $u = ax + b$ and $v = cy + d$, where a, b, c, d are constants; then

$$r_{uv} = \frac{ac}{|a||c|}r_{xy}$$

$$= \begin{cases} r_{xy} & \text{when } a \text{ and } c \text{ have the same sign} \\ -r_{xy} & \text{when } a \text{ and } c \text{ have opposite sign} \end{cases}$$

Proof: Since $u = ax + b$ and $v = cy + d$, then $\bar{u} = a\bar{x} + b$ and $\bar{v} = c\bar{y} + d$.

$$\begin{aligned}
\text{var}(u) &= \sigma_u^2 = \frac{1}{n}\Sigma_{i=1}^n(u_i - \bar{u})^2 = \frac{1}{n}\Sigma_{i=1}^n(ax_i + b - a\bar{x} - b)^2 \\
&= a^2\frac{1}{n}\Sigma_{i=1}^n(x_i - \bar{x})^2 = a^2\sigma_x^2.
\end{aligned}$$

$\therefore \sigma_u = |a|\ \sigma_x$. Similarly, $\sigma_v = |c|\ \sigma_y$.

Now,

$$\begin{aligned}
\text{Cov}(u,v) &= \frac{1}{n}\Sigma_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v}) && (1.10)\\
&= \frac{1}{n}\Sigma_{i=1}^{n}(ax_i+b-a\bar{x}-b)(cy_i+d-c\bar{y}-d) && (1.11)\\
&= \frac{1}{n}\ ac\ \Sigma_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) && (1.12)\\
&= ac\ \text{Cov}(x,y). && (1.13)
\end{aligned}$$

$$\therefore r_{uv} = \frac{\text{Cov}(u,v)}{\sigma_u\sigma_v} = ac\ \frac{\text{Cov}(x,y)}{|a|\ \sigma_x|c|\ \sigma_y} = r_{uv} = \frac{ac}{|a||c|}r_{xy}$$

$$= \begin{cases} r_{xy} & \text{when } a \text{ and } c \text{ have the same sign} \\ -r_{xy} & \text{when } a \text{ and } c \text{ have opposite signs.} \end{cases}$$

Property 4. The value of r lies between -1 and 1, i.e., $-1 \le r \le 1$.

Proof: Let u_i and v_i be the two sets of two variables such that

$$u_i = \frac{x_i-\bar{x}}{\sigma_x} \text{ and } v_i = \frac{y_i-\bar{y}}{\sigma_y}$$

where the symbols on the r.h.s have usual meaning.

$$\therefore \Sigma u_i^2 = \Sigma\frac{(x_i-\bar{x})}{\sigma_x^2} = \frac{1}{\sigma_x^2}\Sigma(x_i-\bar{x})^2 = \frac{1}{\sigma_x^2}n\sigma_x^2 = n.$$

Similarly, $\Sigma v_i^2 = n$.

Again,

$$\begin{aligned}
\Sigma u_i v_i &= \Sigma\frac{(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} = \frac{n}{\sigma_x\sigma_y}\Sigma\frac{(x_i-\bar{x})(y_i-\bar{y})}{n}\\
&= \frac{n\ \text{Cov}(x,y)}{\sigma_x\sigma_y} = nr_{xy}.
\end{aligned}$$

Now $(u_i \pm v_i)^2$ cannot be negative.

$$\begin{aligned}
\therefore \Sigma(u_i \pm v_i)^2 &\ge 0\\
\implies \Sigma(u_i^2+v_i^2 \pm 2u_iv_i) &\ge 0\\
\implies \Sigma u_i^2 + \Sigma v_i^2 \pm 2\Sigma u_iv_i &\ge 0\\
\implies n+n+\pm 2nr_{xy} &\ge 0\\
\implies 2n(1\pm r_{xy}) &\ge 0 \implies (1\pm r_{xy}) \ge 0.
\end{aligned}$$

Hence, $-1 \leq r_{xy} \leq 1$.

Property 5. Let (x, y) represent bivariate data for the two variables x and y. Then, $\text{var}(x \pm y) = \sigma_x^2 + \sigma_y^2 + \pm 2 r_{xy}\sigma_x\sigma_y$.

Proof: By definition $\text{var}(x \pm y) = \frac{1}{n}\Sigma_{i=1}^n[(x_i \pm y_i) - (\bar{x} \pm \bar{y})]^2$

$$\begin{aligned}
&= \frac{1}{n}\Sigma_{i=1}^n[(x_i - \bar{x}) \pm (y_i - \bar{y})]^2 \\
&= \frac{1}{n}\Sigma_{i=1}^n[(x_i - \bar{x})^2 + (y_i - \bar{y})^2 \pm 2(x_i - \bar{x})(y_i - \bar{y})] \\
&= \frac{1}{n}\Sigma_{i=1}^n(x_i - \bar{x})^2 + \frac{1}{n}\Sigma_{i=1}^n(y_i - \bar{y})^2 \pm 2\frac{1}{n}\Sigma_{i=1}^n(x_i - \bar{x})(y_i - \bar{y}) \\
&= \sigma_x^2 + \sigma_y^2 \pm 2\ \text{Cov}(x, y) \\
&= \sigma_x^2 + \sigma_y^2 \pm 2\ r_{xy}\sigma_x\sigma_y
\end{aligned}$$

Notes:

1. The standard error of correlation coefficient is given by $\frac{1-r_{xy}^2}{\sqrt{n}}$.
2. If two variables x and y are uncorrelated, then $r_{xy} = 0$.
3. Probable error $= 0.67485 \times \frac{1-r_{xy}^2}{\sqrt{n}}$.

Example. Find the correlation coefficient of the following data:

x :	65	63	67	64	68	62	70	66
y :	68	66	68	65	69	66	68	65

Solution: Since the correlation coefficient is unaffected by change of origin, let us change the origin of x and y to 65 and 687, respectively.

Then, we write $u = x - 65$ and $v = y - 67$.

x	y	u	v	u^2	v^2	uv
65	68	0 ‘	1	0	1	0
63	66	-2	-1	4	1	2
67	68	2	1	4	1	2
64	65	-1	-2	1	4	2
68	69	3	2	9	4	6
62	66	-3	-1	9	1	3
70	68	5	1	25	1	5
66	65	1	-2	4	4	-2
		$\Sigma u = 5$	$\Sigma v = -1$	$\Sigma u^2 = 53$	$\Sigma v^2 = 17$	$\Sigma uv = 18$

$$\therefore \sigma_u^2 = \frac{1}{n}\Sigma u^2 - (\bar{u})^2 = \frac{1}{8}\ 53 - \left(\frac{5}{8}\right)^2 = \frac{399}{64} \tag{1.14}$$

$$\sigma_v^2 = \frac{1}{n}\Sigma v^2 - (\bar{v})^2 = \frac{1}{8}\ - \left(\frac{-1}{8}\right)^2 = \frac{135}{64}. \tag{1.15}$$

$$\mathrm{Cov}(u,v) = \frac{1}{n}\Sigma uv - \bar{u}\bar{v} = \frac{1}{8}\ 18 - \left(\frac{5}{8}\right)\left(\frac{-1}{8}\right) = \frac{149}{64} \tag{1.16}$$

$$r_{xy} = r_{uv} = \frac{\mathrm{Cov}(u,v)}{\sigma_u \sigma_v} = \frac{\frac{149}{64}}{\sqrt{\frac{399}{64}}\sqrt{\frac{135}{64}}} = \frac{149}{\sqrt{399 \times 136}} = 0.64. \tag{1.17}$$

hence the required correlation coefficient is 0.64.

Example. Find the correlation coefficient of the following data:

x :	23.3	17.5	17.8	20.7	18.1	20.9	22.9	20.8
y :	4.2	3.8	4.6	3.2	5.2	4.7	1.1	5.6

Solution: We know that the correlation coefficient is unaffected by the change of origin and scale. Therefore, we assume

$$u = \frac{x - 20.7}{1} \quad \text{and} \quad v = \frac{y - 4.4}{1}.$$

x	y	u	v	u^2	v^2	uv
23.5	4.2	26 ‘	-2	676	4	-52
17.5	3.8	-32	–6	1024	36	192
17.8	4.6	-29	2	841	4	-58
20.7	3.2	-0	-12	0	144	0
18.1	5.2	-26	8	676	64	-208
20.9	4.7	2	3	4	9	6
22.9	4.4	22	0	484	0	0
20.8	5.6	1	12	1	144	12
		$\Sigma u = -36$	$\Sigma v = 5$	$\Sigma u^2 = 3076$	$\Sigma v^2 = 405$	$\Sigma uv = -108$

We know

$$r_{xy} = r_{uv} = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{[n\Sigma u^2 - (\Sigma u)^2]\,[n\Sigma v^2 - (\Sigma v)^2]}} \tag{1.18}$$

$$= \frac{8 \times (-108) - (-36) \times 5}{\sqrt{[8 \times 3706 - (-36)^2]\,[8 \times 405 - 5^2]}} \tag{1.19}$$

$$= -\frac{684}{6547.34} = -0.0716. \tag{1.20}$$

$$
\begin{aligned}
\text{Probability of error is given by P.E.} &= 0.6745 \times \frac{1-r^2}{\sqrt{n}} \\
&= 0.6745 \times \frac{1-(-0.072)^2}{\sqrt{8}} \\
&= 0.6745 \times \frac{0.9948}{2 \times 1.414} \\
&= 0.2372.
\end{aligned}
$$

Example. While calculating the correlation coefficient between variables x and y, the following results are found:

$$\Sigma_{i=1}^{25} x_i = 125, \Sigma_{i=1}^{25} = 100, \Sigma_{i=1}^{25} x_i^2 = 650, \Sigma_{i=1}^{25} y_i^2 = 460 \text{ and } \Sigma_{i=1}^{25} x_i y_i = 508.$$

Later it was found that at the time of checking two pairs of observations (x, y) were copied wrongly as $(6, 14)$ and $(8, 6)$ while the correct values were $(8, 12)$ and $(6, 8)$ respectively. Determine the correlation coefficient between x and y.

Solution: Now

$$
\begin{aligned}
\text{Corrected } \Sigma x_i &= 125 - (6+8) + (8+6) = 125 \\
\text{Corrected } \Sigma y_i &= 10 - (14+6) + (12+8) = 100 \\
\text{Corrected } \Sigma x_i^2 &= 650 - (6^2+8^2) + (8^2+6^2) = 650 \\
\text{Corrected } \Sigma y_i^2 &= 460 - (14^2+6^2) + (12^2+8^2) = 436 \\
\text{Corrected } \Sigma x_i y_i &= 508 - (6 \times 14 + 8 \times 6) + (8 \times 12 + 6 \times 8) = 520
\end{aligned}
$$

$$
\begin{aligned}
\therefore \text{ Corrected Cov}(x, y) &= \frac{1}{n}\Sigma x_i y_i - \bar{x}\bar{y} \\
&= \frac{1}{25} \times 520 - \frac{125}{25}\frac{100}{25} = \frac{104}{5} - 20 = \frac{4}{5}.
\end{aligned}
$$

$$
\begin{aligned}
\therefore \text{ Corrected } \sigma_x^2 &= \frac{1}{n}\Sigma x_i^2 - (\bar{x})^2 \\
&= \frac{1}{25} \times 650 - \left(\frac{125}{25}\right)^2 = 26 - 25 = 1.
\end{aligned}
$$

and

$$
\begin{aligned}
\therefore \text{ Corrected } \sigma_y^2 &= \frac{1}{n}\Sigma y_i^2 - (\bar{y})^2 \\
&= \frac{1}{25} \times 436 - \left(\frac{100}{25}\right)^2 = \frac{436}{25} - 16 = \frac{36}{25}.
\end{aligned}
$$

$\therefore$ Corrected Correlation coefficient is

$$r_{xy} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \frac{\frac{4}{5}}{\sqrt{1}\sqrt{\frac{36}{25}}} = \frac{4}{6} = \frac{2}{3}.$$

Example. If var$(x+y) = 81$, var(x)=36 and var(y)=25, then find the correlation coefficient between x and y.

Solution: We know that
$$\begin{aligned} \text{var}(x+y) &= \text{var}(x) + \text{var}(y) + 2\ \text{Cov}(x,y) \\ &= \text{var}(x) + \text{var}(y) + 2\ r_{xy}\ \sigma_x\sigma_y \end{aligned}$$
or, $\quad 81 = 36 + 25 + 2.6.5.r_{xy}$
or, $\quad r_{xy} = \frac{81-61}{60} = \frac{20}{60} = \frac{1}{3}.$

Example. If $\Sigma xy = 60$, $\sigma_y = 2.5, \Sigma x^2 = 90$ and $r_{xy} = 0.8$, them find the number of items where $\Sigma x = \Sigma y = 0$.

Solution: Now, $\text{Cov}(x,y) = \frac{1}{n}\Sigma(x-\bar{})(y-\bar{y}) = \frac{1}{n}\Sigma xy = \frac{60}{n}$, where n is the number of terms and
$$\sigma_x^2 = \frac{1}{n}\Sigma x^2 - (\bar{x})^2 = \frac{90}{n}.$$
$$\text{We know,} \quad r_{xy} = \frac{\text{Cov}(x,y)}{\sigma_x\sigma_y} = \frac{\frac{60}{n}}{\sqrt{\frac{90}{n}} \times 2.5}$$
$$\text{or,} \qquad 0.8 = \frac{60}{\sqrt{90}\times\sqrt{n}\times 2.5}$$
$$\begin{aligned} &\text{or,} \qquad \sqrt{90}\times\sqrt{n}\times 2 = 60 \\ &\text{or,} \qquad 90n = 30\times 30 \\ &\text{or,} \qquad n = 10. \end{aligned}$$

Example. If $u - 7x = 5$ and $v - 5y = 11$ and the correlation coefficient of x and y is 0.23, then find the correlation coefficient of u and v.

Solution: From the given relations, we get $u = 7x+5$ and $v = 5y+11$ which are linear functions of x and y then $r_{xy} = r_{uv}$, since the coefficients of x and y have same sign.
$$\therefore r_{uv} = 0.23.$$

Example. Two variables x and y have n pair of values. The variance of x, y and $x-y$ are given by σ_x^2, σ_y^2 and σ_{x-y}^2 respectively. Prove that correlation coefficient r_{xy} between x and y is given by
$$r_{xy} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\ \sigma_x\sigma_y}.$$

Solution: Let $u_i = x_i - y_i, \quad i = 1, 2, ..., n$, then $\bar{u} = \bar{x} - \bar{y}$

and

$$\begin{aligned}
\sigma_u^2 &= \frac{1}{n}\Sigma(u_i - \bar{u})^2 = \frac{1}{n}\Sigma[(x_i - y_i) - (\bar{x} - \bar{y})]^2 \\
&= \frac{1}{n}\Sigma[(x_i - \bar{x}) - (y_i - \bar{y})]^2 \\
&= \frac{1}{n}\Sigma[(x_i - \bar{x})^2 + (y_i - \bar{y})^2 - 2(x_i - \bar{x})(y_i - \bar{y})] \\
&= \frac{1}{n}\Sigma(x_i - \bar{x})^2 + \frac{1}{n}\Sigma(y_i - \bar{y})^2 - 2\frac{1}{n}\Sigma(x_i - \bar{x})(y_i - \bar{y}) \\
&= \sigma_x^2 + \sigma_y^2 - 2\ \mathrm{Cov}(x, y) \\
\text{or,} \quad \sigma_{x-y}^2 &= \sigma_x^2 + \sigma_y^2 - 2r_{xy}\sigma_x\sigma_y \\
\text{or,} \quad 2r_{xy}\sigma_x\sigma_y &= \sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2 \\
\text{or,} \quad r_{xy} &= \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\ \sigma_x\sigma_y}.
\end{aligned}$$

Regression Analysis

The word *regression* refers to the method of finding the most suitable equation for *predicting* or *estimating* one variable for a given value of other. It also refers to the method of finding the *error* in such prediction.

Let us suppose that the variables are x and y where x is independent and y is depends on x.

Linear regression: If the dependence can be expressed in the form $y = a + bx$, then the regression that is studied is known as *linear regression*, because the above equation represents straight line.

Curvillinear regression: If the dependence is given by an equation representing a curve then the regression is known as curvillinear regression, *e.g.*, $y = ax^2 + bx + c$. We shall discuss here linear regression only.

Normal equation: The equations

$$\begin{aligned}
\Sigma y &= na + b\Sigma x \\
\Sigma xy &= a\Sigma x + b\Sigma x^2.
\end{aligned}$$

are called *normal equation* for the regression equation $y = a + bx$.

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 square.
Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ be a set of n pair of observations and let us fit a straight line in the form

$$y = a + bx$$

to these data . Applying method of last squares, the constants a and b are obtained by solving the normal equations, *i.e.,*

$$\begin{aligned} \Sigma y &= \quad na + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2. \end{aligned} \tag{1.21}$$

We now solve the normal equations in a and b. Multiplying Σx with first equation of (1.21) and second equation by n and then subtracting, we get

$$\begin{aligned} \Sigma x\Sigma y - n\Sigma xy &= b(\Sigma x)^2 - nb\Sigma x^2 \\ \text{or,} \qquad b &= \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{\frac{1}{n}\Sigma xy - \frac{\Sigma x}{n}\frac{\Sigma y}{n}}{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2} \\ &= \frac{\frac{1}{n}\Sigma xy - \bar{x}\bar{y}}{\frac{1}{n}\Sigma x^2 - (\bar{x})^2} \\ &= \frac{\mu_{11}}{\sigma_x^2}. \end{aligned}$$

where $\mu_{11} = \text{Cov}(x, y) = \frac{1}{n}\Sigma xy - \bar{x}\bar{y}$.

Putting the value of b in (1.21), we obtain

$$\Sigma y = na + \frac{\mu_{11}}{\sigma_x^2}\Sigma x \tag{1.22}$$

$$\text{or,} \quad \frac{1}{n}\Sigma y = a + \frac{\mu_{11}}{\sigma_x^2}\frac{\Sigma x}{n} \tag{1.23}$$

$$\text{or,} \quad \bar{y} = a + \frac{\mu_{11}}{\sigma_x^2}\bar{x} \tag{1.24}$$

$$\text{or,} \quad a = \bar{y} - \frac{\mu_{11}}{\sigma_x^2}\bar{x} \tag{1.25}$$

Substituting these values a and b in the regression equation,

$$y = \left(\bar{y} - \frac{\mu_{11}}{\sigma_x^2}\bar{x}\right) + \frac{\mu_{11}}{\sigma_x^2}x$$

$$\text{or,} \quad (y - \bar{y}) = \frac{\mu_{11}}{\sigma_x^2}(x - \bar{x})$$

which is the equation of the line of regression of y on x.

Regression coefficient of y on x

The coefficient b i.e., $\frac{\mu_{11}}{\sigma_x^2}$ or $\frac{\text{Cov}(x,y)}{\sigma_x^2}$ is called the *regression coefficient of y on x* and is denoted by b_{yx}.

The regression equation of y on x is, therefore, written as

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

Regression equation of x on y

The best fitting straight line of bivariate distribution representing a regression equation of the form

$$x = c + dy$$

where y is the independent variable and x is the dependent variable, known as the *line of regression of x on y.*

Proceeding exactly in the same manner as before, we obtain the *regression equation of x on y* as

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

where $b_{xy} = \frac{\mu_{11}}{\sigma_y^2}$ or $\frac{\text{Cov}(x,y)}{\sigma_y^2}$ which is known as the *regression coefficient of x on y.*

Properties of regression coefficient

Property 1. Regression coefficients are unaffected by the change of origin.

Proof. Let $u_i = x_i - a$ and $v_i = y_i - b$.

$$\begin{aligned}
\text{Now} \qquad b_{yx} &= \frac{\text{Cov}(x,y)}{\sigma_x^2} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \\
&= \frac{\Sigma\{(u_i + a) - (\bar{u} + a)\}\{(v_i + b) - (\bar{v} + b)\}}{\Sigma[(u_i + a) - (\bar{u} + a)]^2} \\
&= \frac{\Sigma(u_i - \bar{u})(v_i - \bar{v})}{\Sigma(u_i - \bar{u})^2} = \frac{\text{Cov}(u,v)}{\sigma_u^2} = b_{uv}.
\end{aligned}$$

which is the regression coefficient of v on u. It can be similarly proved that $b_{xy} = b_{uv}$.

Property 2. Regression coefficient is affected by the change of scale.

Proof. Let $u_i = \frac{x_i - a}{c}$ and $v_i = \frac{y_i - b}{d}$.

$$\begin{aligned}
\therefore \qquad x_i = a + cu_i \qquad &\text{and} \qquad y_i = b + dv_i \\
\therefore \qquad \bar{x} = a + c\bar{u} \qquad &\text{and} \qquad \bar{y} = b + d\bar{v}
\end{aligned}$$

$$\begin{aligned}
\text{Hence} \quad b_{yx} &= \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \\
&= \frac{\Sigma\{(a + cu_i) - (a + c\bar{u})\}\{(b + dv_i) - (b + d\bar{v})\}}{\Sigma[(a + cu_i) - (a + c\bar{u})]^2} \\
&= \frac{c.d\ \Sigma(u_i - \bar{u})(v_i - \bar{v})}{c^2\Sigma(u_i - \bar{u})^2} \\
&= \frac{d}{c}\ b_{vu}.
\end{aligned}$$

It can be similarly be shown that $b_{xy} = \frac{c}{d}\, b_{uv}$.

Relation between regression coefficient and between regression coefficient and correlation coefficient

1. It is known that $r = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}$ where r is the correlation coefficient

$$\begin{aligned} \text{and} \quad b_{yx} &= \frac{\text{Cov}(x,y)}{\sigma_x^2} = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y} \frac{\sigma_y}{\sigma_x} \\ &= r\frac{\sigma_y}{\sigma_x}. \end{aligned}$$

Similarly, $b_{xy} = r\frac{\sigma_x}{\sigma_y}$. Hence $b_{yx} b_{xy} = r^2$.

In other words, r is the geometric mean of the regression coefficients.

2. Both the regression coefficients must have the same algebraic signs. If b_{yx} and b_{xy} are positive the r is positive and if b_{yx} and b_{xy} are negative, then r is negative.

3. Since $-1 \leq r \leq 1$, both the regression coefficients cannot be greater than 1.

4. Arithmetic mean of two regression coefficients is either equal to or greater than the correlation coefficient

$$\text{i.e.,} \quad \frac{b_{yx} + b_{xy}}{2} \geq r.$$

5. The regression lines are usually different. But since they always pass through $(\bar{x}, \bar{y})$, therefore, they become identical if their slopes become equal, *i.e.,* if $b_{yx} = \frac{1}{b_{xy}}$ or if $b_{xy} b_{yx} = 1$. In such a case

$$\left(r\frac{\sigma_y}{\sigma_x}\right)\left(r\frac{\sigma_x}{\sigma_y}\right) = 1 \quad \Longrightarrow r^2 = 1 \quad \Longrightarrow r = \pm 1.$$

6. If $r = +1$, then both regression equation take the form

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

7. If $r = -1$, then both regression equation take the form

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

8. Correlation is said to be of high degree if $\frac{3}{4} \leq |r| \leq 1$, of moderate degree if $\frac{1}{4} \leq |r| < \frac{3}{4}$ and of low degree if $0 \leq |r| < \frac{1}{4}$.

9. The acute angle between two regression line is given by

$$tan\theta = \left|\frac{1 - r^2}{b_{xy} + b_{yx}}\right|.$$

$\therefore$ Two lines coincide, iff $\theta = 0$. *i.e.*, iff $r = \pm 1$.

Example. Given the following bivariate data:

$$\begin{array}{lcccccccc} x: & 1 & 5 & 3 & 2 & 1 & 1 & 7 & 3 \\ y: & 6 & 1 & 0 & 0 & 1 & 2 & 1 & 5 \end{array}$$

Fit the regression line of y on x and that of x on y. Predict y when $x = 10$ and x when $y = 2.5$.

Solution: we are to find the equations

$$\begin{aligned} y - \bar{y} &= b_{yx}(x - \bar{x}) \\ x - \bar{x} &= b_{xy}(y - \bar{y}) \end{aligned}$$

We shall assume $u = x - 3$ and $v = y - 3$ and use the formula

$$b_{yx} = b_{vu} = \frac{n\Sigma uv - \Sigma u \Sigma v}{n\Sigma u^2 - (\Sigma u)^2}$$

$$\text{and} \quad b_{xy} = b_{uv} = \frac{n\Sigma uv - \Sigma u \Sigma v}{n\Sigma v^2 - (\Sigma v)^2}$$

x	y	u	v	u^2	v^2	uv
1	6	-2 ‘	-3	4	9	-6
5	1	2	-2	4	4	-4
3	0	-0	-3	0	9	0
2	0	-1	-3	1	9	3
1	1	-2	-2	4	4	4
1	2	-2	-1	4	1	2
7	1	4	-2	16	4	-8
3	5	0	2	0	4	0
		$\Sigma u = -1$	$\Sigma v = -8$	$\Sigma u^2 = 33$	$\Sigma v^2 = 44$	$\Sigma uv = -9$

$$\therefore \quad \bar{x} = \bar{u} + 3 = -\frac{1}{8} + 3 = 2.875 \tag{1.26}$$

$$\bar{y} = \bar{v} + 3 = -\frac{8}{8} + 3 = 2 \tag{1.27}$$

$$b_{yx} = \frac{8 \times (-9) - (-1) \times (-8)}{8 \times 33 - (-1^2)} = \frac{-72 - 8}{264 - 1} = -0.304 \tag{1.28}$$

$$b_{xy} = \frac{8 \times (-9) - (-1) \times (-8)}{8 \times 44 - (-8^2)} = \frac{-80}{352 - 64} = -0.278. \tag{1.29}$$

The regression line of y on x is

$$(y - 2) = -0.304\ (x - 2.875),\ \text{ or } y = -0.304x + 2.874.$$

Value of y when $x = 10$ is $y = -0.304 \times 10 + 2.874 = -0.166$.

The regression line of x on y is

$$(x - 2.875) = -0.278(y - 2), \quad \text{or} \quad x = -0.278x + 3.431.$$

Value of x when $y = 2.5$ is $x = -0.278 \times 2.5 + 3.431 = 2.736$.

Example. Find the equation of regression line x on y for the following bivariate data:

x :	1	1.5	2	2.5	3	3.5	4
y :	5.3	5.7	6.3	7.2	8.2	8.7	8.4

Solution: For simplifying the calculations, let us make a change of origin and scale for both the variables as follows:

$$u = \frac{x - 2.5}{0.5},\ v = \frac{y - 7.0}{0.1}.$$

x	y	u	v	v^2	uv
1	5.3	-3 ‘	-17	289	51
1.5	5.7	-2	-13	169	26
2	6.3	-1	-7	49	7
2.5	7.2	0	2	4	0
3	8.2	1	12	144	12
3.5	8.7	2	17	289	34
4	8.4	3	14	196	42
17.5	49.8	$\Sigma u = 0$	$\Sigma v = 8$	$\Sigma u^2 = 1140$	$\Sigma uv = 172$

We know that

$$\begin{aligned} b_{xy} &= \frac{c}{d}\frac{n\Sigma uv - \Sigma u \Sigma v}{n\Sigma v^2 - (\Sigma v)^2} \quad \text{where} \quad c = 0.5 \ \text{ and } \ d = 0.1 \\ &= \frac{0.5}{0.1} \times \frac{172 \times 7 - 0 \times 8}{7 \times 1140 - (8)^2} = \frac{5 \times 1204}{7916} = 0.76. \\ \bar{x} &= \frac{17.5}{7} = 2.5 \text{ and } \ \bar{y} = \frac{49.8}{7} = 7.11. \end{aligned}$$

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

$$\begin{aligned} x - 2.5 &= 0.76\ (y - 7.11) \\ x &= 0.76y - 2.90. \end{aligned}$$

Example. Let the line of regression concerning two variables x and y be given by $y = 32 - x$ and $x = 13 - 0.25y$. Obtain the values of the means and correlation coefficient.

Solution: Since the regression lines intersects at $(\bar{x}, \bar{y})$, the means will be obtained by solving the two equations. Solving $y = 32 - x$ and $x = 13 - 0.25y$, we get $x = 6.7$ and $y = 25.3$. So $\bar{x} = 6.7$ and $\bar{y} = 25.3$.

Now $y = 32 - x$ is the regression equation of y on x,

$$\therefore \quad b_{yx} = -1$$

and $x = 13 - 0.25y$ being the regression equation of x on y,

$$\therefore \quad b_{xy} = -0.25$$

$$\therefore \quad r^2 = b_{yx} \times b_{xy} = (-1) \times (-0.25) = 0.25$$

$$\therefore \quad r = \pm\sqrt{0.25} = \pm 0.5.$$

Bur, since both regression coefficients are negative (note that both must have same sign), the correlation coefficient must be negative, *i.e.*, $r = -0.5$.

Example. For the variables x and y, the equations of the regression lines are $4x - 5y + 33 = 0$ and $20x - 9y = 107$. Identify the regression line of y on x and that of x on y. What is the correlation coefficient? If the variance of x is 9 find the standard deviation of y. Also find $\bar{x}, \bar{y}$. What is the estimate value of y at $x = 10$? If this estimate be y_0, find the estimated value of x when $y = y_0$.

Solution: Let the regression line of y on x be the $4x - 5y + 33 = 0$, then

$$5y = 4x + 33 \quad \text{or} \quad y = \frac{4}{5}x + \frac{33}{5}.$$

$\therefore$ The regression coefficient of y on x is given by $b_{yx} = \frac{4}{5}$.

Let the regression line of x on y be the $20x - 9y = 107$, then

$$20x = 9x + 107 \quad \text{or} \quad x = \frac{9}{20}y + \frac{107}{20}.$$

$\therefore$ The regression coefficient of x on y is given by $b_{xy} = \frac{9}{20}$.

$$\therefore \quad r^2 = b_{yx} \times b_{xy} = \frac{4}{5} \times \frac{9}{20} = \frac{9}{25}$$

$$\therefore r = \pm\,\frac{3}{5} = \pm\,0.6.$$

Since $b_{xy} and b_{yx}$ are both positive, then $r = 0.6$. So our hypothesis is correct.

Hence the regression line of y on x and of x on y are given by

$$4x - 5y + 33 = 0 \text{ and } 20x - 9y = 107, \quad \text{respectively.}$$

Again the variance of $x = 9$. So $\sigma_x = 3$.

$$\text{Now,} \quad b_{yx} = r\frac{\sigma_y}{\sigma_x} \tag{1.30}$$

$$\text{or,} \quad \frac{4}{5} = \frac{3}{5}\frac{\sigma_y}{\sigma_3} \tag{1.31}$$

$$\text{or,} \quad \sigma_y = 4. \tag{1.32}$$

$\therefore$ The standard deviation of y is 4.

We know that the two regression line s intersects at the point $(\bar{x}, \bar{y})$, where $\bar{x}$ and $\bar{y}$ are the mean of x and y respectively.

$\therefore \quad 4\bar{x} = 5\bar{y} + 33 = 0$ and $20\bar{x} - 9\bar{y} - 107 = 0$. Solving, we get $\bar{x} = 13$ and $\bar{y} = 17$.

Also when $x = 10$, $y_0 = \frac{4}{5}x + 6.6 = \frac{4}{5} \times 10 + 6.6 = 8 + 6.6 = 14.6$.

For $y = y_0 = 14.6$, $x_0 = \frac{9}{20}y + \frac{107}{20} = \frac{9}{20} \times 14.6 + \frac{107}{20} = \frac{238.4}{20} = 11.92$.

Example. If $x = 4y + 5$ and $y = Kx + 4$ be two regression lines of x on y and of y on x respectively, find the interval in which K lies.

Solution: Since $x = 4y + 5$ and $y = Kx + 4$ be two regression lines of x on y and of y on x respectively, then the regression coefficients of x on y and y on x are given by

$$b_{xy} = 4 \text{ and } b_{yx} = K$$

Since

$$r_{xy}^2 = b_{xy}\, b_{yx} \quad \therefore \quad r_{xy}^2 = 4K.$$

As $-1 \leq r_{xy} \leq 1$, $0 \leq r_{xy}^2 \leq 1 \qquad \therefore 0 \leq 4K \leq 1$, or $0 \leq K \leq \frac{1}{4}$.

Example. The relationship between travel expenses (y) and the duration of travel (x) is found to be linear. A summary of data for 102 pairs is given below:

$\Sigma x = 510$, $\Sigma y = 7140, \Sigma x^2 = 4150, \Sigma xy = 54900$ and $\Sigma y^2 = 7,40,200$.

1. Find the two regression coefficients.
2. Find the two regression line.
3. A given trip has to take seven days. How much money should a salesman be allowed so that he will not run short of money?

Solution: Here $\bar{x} = \frac{1}{n}\Sigma x = \frac{510}{102} = 5$ where $n = 102$ and $\bar{y} = \frac{1}{n}\Sigma y = \frac{7140}{102} = 70$.

$$\begin{aligned}
\text{Cov } (x, y) &= \frac{1}{n}\Sigma xy - \bar{x}\bar{y} = \frac{54900}{102} - 5 \times 70 = \frac{9150}{17} - 350 = 188.24 \\
\sigma_x^2 &= \frac{1}{n}\Sigma x^2 - (\bar{x})^2 = \frac{4150}{102} - 25 = \frac{2075}{51} - 25 = 15.686 \\
\sigma_y^2 &= \frac{1}{n}\Sigma y^2 - (\bar{y})^2 = \frac{740200}{12} - 4900 = 2356.863
\end{aligned}$$

1.

$$b_{xy} = r_{xy}\ \frac{\sigma_x}{\sigma_y} = \frac{\sigma_x}{\sigma_y}\ \frac{\text{Cov}(x, y)}{\sigma_x\sigma_y} = \frac{\text{Cov}(x, y)}{\sigma_y^2} = \frac{188.24}{2356.863} = 0.08.$$

and

$$b_{yx} = \frac{\text{Cov}(x, y)}{\sigma_x^2} = \frac{188.24}{15.686} = 12.$$

2. The regression line of y on x is $y - \bar{y} = b_{yx}\ (x - \bar{x})$ or $y - 70 = 12(x - 5)$ or $y = 12x + 10$.

   The regression line of x on y is $x - \bar{x} = b_{xy}\ (y - \bar{y})$ or $x - 5 = 0.08(y - 70)$ or $x = 0.08y - 0.6$.

3. For $x = 7$, $y = 12x + 10 = 12 \times 7 + 10 = 94$.

Example. If $\text{var}(x) = 4,\ \text{var}(y) = 9$ and $r_{xy} = \frac{2}{3}$, then find $\text{var}(2x - 3y)$.

Solution: Now

$$\begin{aligned}
\text{var}(2x - 3y) &= \text{var}(2x) + \text{var}(3y) - 2\sqrt{\text{var}(2x)}\sqrt{\text{var}(3y)}\ r_{2x,3y} \\
&= 2^2\text{var}(x) + 3^2\text{var}(y) - 2\sqrt{2^2\text{var}(x)}\sqrt{3^2\text{var}(y)}\ r_{xy} \\
&= 4 \times 4 + 9 \times 9 - 2 \times 2 \times 3 \times \sqrt{4} \times \sqrt{9} \times \frac{2}{3} \\
&= 16 + 81 - 48 = 49.
\end{aligned}$$

Example. If x and y are two correlated variables with same variance and the correlation coefficient is r, find the regression coefficient of x on $(x+y)$ and that of $(x+y)$ on x. Hence find the correlation coefficient between x and $(x + y)$.

Solution: Let var(x)= var(y)= σ^2.

We know that $r = \frac{\text{Cov}(x,y)}{\sigma_x\sigma_y} = \frac{\text{Cov}(x,y)}{\sigma\sigma} = \frac{\text{Cov}(x,y)}{\sigma^2}$.

$\therefore \text{Cov}(x, y) = r\sigma^2$.

Let $u = x + y$, then $b_{xu} = \frac{\text{Cov}(x,u)}{\text{var}(u)}$ and $b_{ux} = \frac{\text{Cov}(x,u)}{\text{var}(x)}$.

$$\begin{aligned}
\text{Cov}(x,u) &= \text{Cov}(x,x+y) = \frac{1}{n}\Sigma(x-\bar{x})(x+y-\bar{x}-\bar{y}) && (1.33)\\
&= \frac{1}{n}\Sigma(x-\bar{x})[(x-\bar{x})+(y-\bar{y})] && (1.34)\\
&= \frac{1}{n}\Sigma(x-\bar{x})^2 + \frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y}) && (1.35)\\
&= \text{var}(x) + \text{Cov}(x,y) = \sigma^2 + r\sigma^2 = (1+r)\sigma^2 && (1.36)
\end{aligned}$$

and

$$\begin{aligned}
\text{var}(u) &= \text{var}(x+y) = \text{var}(x) + \text{var}(y) + 2\ \text{Cov}(x,y) && (1.37)\\
&= \sigma^2 + \sigma^2 + 2r\sigma^2 && (1.38)\\
&= 2(1+r)\sigma^2 && (1.39)
\end{aligned}$$

$$\therefore b_{xu} = \frac{\text{Cov}(x,u)}{\text{var}(u)} = \frac{(1+r)\sigma^2}{2(1+r)\sigma^2} = \frac{1}{2}$$

$$b_{ux} = \frac{\text{Cov}(x,u)}{\text{var}(x)} = \frac{(1+r)\sigma^2}{\sigma^2} = 1+r$$

and the correlation coefficient between x and u is given by

$$r_{xu} = \sqrt{b_{xu} \times b_{ux}} = \sqrt{\frac{1}{2}(1+r)} = \sqrt{\frac{1+r}{2}}.$$

Example. For two variables x and y, the two regression lines are $x+4y+3=0$ and $4x+9y+5=0$. Identify which one is of y on x. Find the means of x and y. Find the correlation coefficient between x and y. Estimate the value of x when $y=1.5$.

Solution: Let $x+4y+3=0$ be the regression line of y on x. Then $4x+9y+5=0$ must be the regression line x on y. So if b_{yx} and b_{xy} denote the respective regression coefficients, then we get

$$b_{yx} = -\frac{1}{4} \quad \text{and} \quad b_{xy} = -\frac{9}{4}.$$

$$\therefore r^2 = b_{yx} \times b_{xy} = \frac{9}{16}.$$

Since $0 \le r^2 \le 1$, our assumption is correct, *i.e.*, $x+4y+3=0$ be the regression line of y on x.

Now, $r^2 = \frac{9}{16}$ which gives $r = \pm\frac{3}{4}$.

Since b_{xy} and b_{yx} are both negative, the $r = -\frac{3}{4}$.

Solving the two equations $x+4y+3=0$ and $4x+9y+5=0$, we get

$$x = 1, y = -1.$$

$\therefore \bar{x} = 1$ and $\bar{y} = -1$.

For estimate x when $y = -1.5$, we take the regression line of x on y and putting $y = 1.5$, we get $4x + 9.5 = -5 \implies x = -\frac{18.5}{4} = -4.625$.

Therefore, the estimated value of x is -4.625.

Example. Let (x, y) and (u, v) be two bivariate variables such that $2u = x + 9$ and $3v = 2y + 7$. The regression coefficient of x on y is σ. Then find the regression coefficient of u on v.

Solution: Here $2u = x + 9$ and $3v = 2y + 7$.

$$\therefore u = \frac{x}{2} + \frac{9}{2} \text{ and } v = \frac{2}{3}y + \frac{7}{3}.$$

Now,

$$\begin{aligned}
\bar{u} &= \frac{\bar{x}}{2} + \frac{9}{2} \text{ and } \bar{v} = \frac{2}{3}\bar{y} + \frac{7}{3}. \\
\therefore \sigma_u &= \frac{1}{2}\sigma_x \text{ and } \sigma_v = \frac{2}{3}\sigma_y. \\
\therefore r_{uv} &= \frac{\frac{1}{2}\frac{2}{3}}{\left|\frac{1}{2}\right|\left|\frac{2}{3}\right|} r_{xy} \\
\therefore b_{uv} &= \frac{\sigma_u}{\sigma_v} r_{uv} = \frac{\frac{1}{2}\sigma_x}{\frac{2}{3}\sigma_y} r_{xy} = \frac{3}{4}\frac{\sigma_x}{\sigma_y} r_{xy} = \frac{3}{4} \times \sigma = \frac{3}{4}\sigma.
\end{aligned}$$

Example. The variates x and y are normally correlated and u, v are defined by

$$\begin{aligned}
u &= x \cos\alpha + y \sin \alpha \\
v &= y \cos\alpha - x \sin \alpha
\end{aligned}$$

Show that u and v will be correlated if

$$\tan 2\alpha = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}$$

where r is the correlation coefficient between x and y.

Further show that in the case

$$\sigma_u^2 + \sigma_v^2 = \sigma_x^2 + \sigma_y^2.$$

Solution: Now,

$$\begin{aligned}
\text{Cov}(x,y) &= \frac{1}{n}\Sigma(u_i-\bar{u})(v_i-\bar{v}) \\
&= \frac{1}{n}\Sigma(x_i\cos\alpha+y_i\sin\alpha-\bar{x}\cos\alpha-\bar{y}\sin\alpha) \\
&\quad (y_i\cos\alpha-x_i\sin\alpha-\bar{y}\cos\alpha+\bar{x}\sin\alpha) \\
&= \frac{1}{n}\Sigma[(x_i-\bar{x})\cos\alpha+(y_i-\bar{y})\sin\alpha] \\
&\quad [(y_i-\bar{y})\cos\alpha-(x_i-\bar{x})\sin\alpha] \\
&= (\cos^2\alpha-\sin^2\alpha)\frac{1}{n}\Sigma(x_i-\bar{x})(y_i-\bar{y}) \\
&\quad -\cos\alpha\sin\alpha\left[\frac{1}{n}\Sigma(x_i-\bar{x})^2-\frac{1}{n}\Sigma(y_y-\bar{y})^2\right] \\
&= \cos 2\alpha\ \text{Cov}(x,y)-\frac{1}{2}\sin 2\alpha\ [\sigma_x^2-\sigma_y^2]
\end{aligned}$$

Now u and v will be uncorrelated if $\text{Cov}(u,v)=0$, *i.e.*, if

$$\cos 2\alpha\ \text{Cov}(x,y)-\frac{1}{2}\sin 2\alpha\ [\sigma_x^2-\sigma_y^2]=0$$

$$\textit{i.e.}, \text{ if } \tan 2\alpha=\frac{2r\sigma_x\sigma_y}{\sigma_x^2-\sigma_y^2}.$$

Further

$$\begin{aligned}
\sigma_u^2+\sigma_v^2 &= \frac{1}{n}\Sigma(u_i-\bar{u})^2+\frac{1}{n}\Sigma(v_i-\bar{v})^2 \\
&= \frac{1}{n}\Sigma(x_i\cos\alpha+y_i\sin\alpha-\bar{x}\cos\alpha-\bar{y}\sin\alpha)^2 \\
&\quad +\frac{1}{n}\Sigma(y_i\cos\alpha-x_i\sin\alpha-\bar{y}\cos\alpha+\bar{x}\sin\alpha)^2 \\
&= \frac{1}{n}\Sigma[(x_i-\bar{x})\cos\alpha+(y_i-\bar{y})\sin\alpha]^2 \\
&\quad +\frac{1}{n}\Sigma[(y_i-\bar{y})\cos\alpha-(x_i-\bar{x})\sin\alpha]^2 \\
&= \cos^2\alpha\ \frac{1}{n}\Sigma(x_i-\bar{x})^2+\sin^2\alpha\ \frac{1}{n}\Sigma(y_i-\bar{y})^2 \\
&\quad +2\sin\alpha\cos\alpha\frac{1}{n}\ \Sigma(x_i-\bar{x})(y_i-\bar{y}) \\
&\quad +\sin^2\alpha\ \frac{1}{n}\Sigma(x_i-\bar{x})^2+\cos^2\alpha\ \frac{1}{n}\Sigma(y_i-\bar{y})^2 \\
&\quad -2\sin\alpha\cos\alpha\frac{1}{n}\ \Sigma(x_i-\bar{x})(y_i-\bar{y}) \\
&= \sigma_x^2\cos^2\alpha+\sigma_y^2\sin^2\alpha+\sigma_x^2\sin^2\alpha+\sigma_y^2\cos^2\alpha \\
&= \sigma_x^2(\cos^2\alpha+\sin^2\alpha)+\sigma_y^2(\cos^2\alpha+\sin^2\alpha) \\
&= \sigma_x^2+\sigma_y^2.
\end{aligned}$$

Example. If θ be the acute angle between two regression lines of the variables x nd y, prove that

$$\tan\theta = \frac{1-r_{xy}^2}{r_{xy}} . \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2}$$

where r_{xy} is the correlation coefficient between x and y.

Solution: The regression lines are

$$\begin{aligned} y - \bar{y} = & \; r_{xy} \frac{\sigma_y}{\sigma_x} (x - \bar{x}) \\ x - \bar{x} = & \; r_{xy} \frac{\sigma_x}{\sigma_y} (y - \bar{y}) \end{aligned} \tag{1.40}$$

Let m_1 and m_2 be the slopes of the lines of (1.40) and (1.40). Then

$$m_1 = r_{xy} \frac{\sigma_y}{\sigma_x} \quad \text{and} \quad m_2 = \frac{\sigma_y}{r_{xy}\ \sigma_x}.$$

Now,

$$\begin{aligned} \tan\theta &= \frac{m_2 \sim m_1}{1 + m_1 m_2} = \frac{\frac{\sigma_y}{r_{xy}\ \sigma_x} - r_{xy} \frac{\sigma_y}{\sigma_x}}{1 + \frac{\sigma_y}{r_{xy}\ \sigma_x} r_{xy} \frac{\sigma_y}{\sigma_x}} \\ &= \frac{\frac{\sigma_y}{\sigma_x}\left(\frac{1}{r_{xy}} - r_{xy}\right)}{1 + \frac{\sigma_y^2}{\sigma_x^2}} = \frac{1 - r_{xy}^2}{r_{xy}} . \frac{\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2} \end{aligned}$$

Exercise.

1. Find the correlation coefficient of the following data:

x :	1	3	4	6	8	9	11	14
y :	1	2	4	4	5	7	8	9

2. Find the covariance and correlation coefficient of the two variables x and y of the following data:

x :	50	53	55	57	60	56	62	52
y :	53	55	57	60	56	52	64	54

3. The bivariate data (x, y) has the following results:
$\Sigma x = 200, \Sigma y = 250, \Sigma x^2 = 2000, \Sigma y^2 = 2900, \Sigma xy = 2250, n = 25$. Find the correlation coefficient between x and y.

4. If var$(x+y)$=45, var(x)=9 and var(y)=16, then find Cov(x, y).

5. Calculate the correlation coefficient from the following data:
$n = 10, \Sigma x = 100, \Sigma y = 150, \Sigma(x-10)^2 = 180, \Sigma(y-15)^2 = 215$ and $\Sigma(x-10)(y-15) = 60$.

6. Find the regression lines of the following data:

x :	60	65	72	64	63	75	77	70
y :	45	48	44	47	51	52	54	50

7. Marks of 5 students in mathematics and statistics are given:

Mathematics :	38	48	43	40	41
Statistics :	31	38	43	33	35

Find the regression lines when marks of a student in Mathematics is 42, determine the most likely marks in statistics.

8. If x and y are uncorrelated variables and their standard deviations are 3 and 4 respectively. Find the correlation coefficient between $5x + 2y$ and $2x - 5y$.

9. if (x, y) and (u, v) be the bivariate variables such that $4u = 2x + 7$ and $6v = 2y - 15$ and if the regression coefficient of y on x is 3, then find the regression coefficient of v on u.

10. Find the regression lines from the following data:
$\bar{x} = 90, \bar{y} = 70, n = 10, \Sigma x^2 = 6360, \Sigma y^2 = 2860, \Sigma xy = 3900$.

11. The regression equation of y on x and x on Y are given by $2x + 3y = 26$ and $6x + y = 31$, respectively. Find the regression coefficient b_{yx} and b_{xy}.

12. For the variables x and y, the equation of regression lines on $3x + 12y = 19$ and $3y + 9x = 46$. Identify the regression lines of y on x and x on y. Find the correlation coefficient and ration of standard deviation of x and y. Find the mean of x and y.

13. If $5y - 7x = 11$ be the regression line of y on x, variance of x is 25 and correlation coefficient between x and y is $\frac{1}{7}$, then find the variance of y.

14. If $4x = 3y + 11$ and $3 = 5x + 7$ be the two regression lines of y on x and x on y respectively, find the interval in which K lies.

15. If σ_x and σ_y are the standard deviations of two uncorrelated variables x and y, prove that the standard deviation of $ax + by$ is $\sqrt{a^2\sigma_x^2 + b^2\sigma_y^2}$.

16. Show that $2x + 3y$ and $4x + 9y$ are uncorrelated if

$$8\sigma_x^2 + 30r\sigma_x\sigma_y + 27\sigma_y^2 = 0.$$

17. The regression lines of y on x and x on y are given by $x + 3y = 0, 3x + 2y = 0$. If $\sigma_x = 1$, then find the regression line of v on u where $u = x + y$ and $v = x - y$.

18. If a, b and c are positive constants, show that the correlation coefficient between $ax + by$ and cy is
$$\frac{ar\sigma_x + b\sigma_y}{\sqrt{a^2\sigma_x^2 + b^2\sigma_y^2 + 2abr\sigma_x\sigma_y}}$$

Answer: 1. 0.977 2. 10.85, 0.752 3. 0.625 4. cov(x, y)=10. 5. 0.305 6.$y - 49.5 = 0.34(x - 68.25), x - 68.25 = 1.56(y - 49.5)$ 7. $y = 0.79x - 2.82, x = 0.52y + 23.28$; 36 8. -0.2 9. 2 10. $x = 0.13y + 80.9, y = 106x + 64.6$ 11. $b_yx = -\frac{3}{2}, b_{xy} = \frac{1}{6}$. 12. y on x is $3x + 12y = 19$ and x on y is $3y + 9x = 46$, $r_{xy} = -1, \bar{x} = 5, \bar{y} = \frac{1}{3}$. 13. 49 14. $0 \leq K \leq 4$ 17. $5v - 3u = 0$.

Chebyshev's Inequality

Let X be an arbitrary random variable with mean μ and variance σ^2. What is the probability that X is within t of its average μ? If we knew the exact distribution of and pdf of X, then we could compute this probability $P(|X - \mu| \leq t) = P(\mu - t \leq X \leq \mu + t)$.

But there is another way to find a lower bound for this probability. For instance, we may obtain an expression like $P(|X - \mu| \leq 2) \geq 0.60$. That is, there is at least a 60% chance for an obtained measurement of this X to be within 2 of its mean.

Theorem 1.1. *Let X be a random variable with mean μ and variance σ^2. For all $t > 0$*
$$P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2} \quad \text{and} \quad P(|X - \mu| \leq t) \geq 1 - \frac{\sigma^2}{t^2}.$$

Proof. Consider
$$Y = \begin{cases} t^2, & \text{if } |X - \mu| > t. \\ 0, & \text{otherwise.} \end{cases} \tag{1.41}$$
□

Observe that $Y \leq |X - \mu|^2$. Then
$$t^2 \times P(|X - \mu| > t) = E[Y] \leq E[|X - \mu|^2] = \text{var}(X) = \sigma^2$$
where $E[Y]$ denotes the expectation of Y. Thus
$$P(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}.$$
$\therefore -P(|X - \mu| > t) \geq -\frac{\sigma^2}{t^2}$ which gives
$$P(|X - \mu| \leq t) = 1 - P(|X - \mu| > t) \geq 1 - \frac{\sigma^2}{t^2}.$$

Note. Chebyshev's Inequality is meaningless when $t \leq \sigma$. For instance, when $t = \sigma$ it is simply saying $P(|X - \mu| > t) \leq 1$ or $P(|X - \mu| \leq t) \geq 0$, which are already obvious. So we must use $t > \sigma$ to apply the inequalities.

Generalized form of Chebyshev's ineqiality

Let $g(X)$ be a non-negative function of random variable X. Then for all $K > 0$,

$$P[g(X) \geq K] \leq \frac{E[g(X)]}{K}.$$

Other forms of Chebyshev's ineqiality

If we put $g(X) = (X - \mu)^2$ and $K = K^2\sigma^2$ in the general form, we obtain

$$\begin{aligned} P[(X-\mu)^2 \geq K^2\sigma^2] &\leq \frac{E[(X-\mu)^2]}{K^2\sigma^2} \\ \text{or,} \quad P[|X-\mu| \geq K\sigma] &\leq \frac{\sigma^2}{K^2\sigma^2} \\ \text{or,} \quad P[|X-\mu| \geq K\sigma] &\leq \frac{1}{K^2}. \end{aligned}$$

Example. (a). Let X is Poisson distributed with parameter $\mu = 9$. Give a lower bound for $P(|X - \mu| \leq 5)$.

(b). Let X be normally distributed with $\mu = 100, \sigma = 15$. Give a lower bound for $P(|X - \mu| \leq 20)$.

Solution: (a) Since X is Poisson distributed with $\mu = 9$, so the mean is $\mu = 9$ and variance $= \sigma^2 = 9$.
Then $P(|X - \mu| \leq 5) = P(|X - 9| \leq 5) \geq 1 - \frac{\sigma^2}{5^2} = 1 - \frac{9}{25} = \frac{16}{25} = 0.64$.

(b) Here mean is $\mu = 100$ and $\sigma = 15$.
$\therefore P(|X - \mu| \leq 20) = P(|X - 100| \leq 20) \geq 1 - \frac{\sigma^2}{20^2} = 1 - \frac{15^2}{20^2} = \frac{175}{400} = 0.4375$.

Note: Using a calculator, we obtain $P(|X - 100| \leq 20) \approx 0.817577$. From these examples, we see that the lower bound provided by Chebyshev's Inequality is not very accurate. However, the inequality is very useful when applied to the sample mean $\bar{x}$ from a large random sample.

Example. A random variable has mean 10 and variance 16. Find the lower bound for $P(5 < X < 15)$.

Solution: By Chebyshev's inequality

$$P[|X - \mu| < K\sigma] \geq 1 - \frac{1}{K^2}$$

$$\text{or} \quad P[\mu - K\sigma < X < \mu + K\sigma] \geq 1 - \frac{1}{K^2}$$

In the present case, $\mu = 10$ and $\sigma = 4$.

$\therefore P[10 - 4K < X < \mu + 4K] \geq 1 - \frac{1}{K^2}$.

Substituting $K = \frac{5}{4}$, we get $P(5 < X < 15) \geq 1 - \frac{1}{\frac{25}{16}} = 1 - \frac{16}{25} = \frac{9}{25}$.

Example. If X a random variable with $E(X) = 3$ and $E(X^2) = 13$, find the lower bound for $P(-2 < X < 8)$ using Chebyshev's inequality.

Solution: We have $\text{var}(X) = E(X^2) - [E(X)]^2 = 13 - 9 = 4$.

By Chebyshev's inequality

$$P(\mu - K\sigma < X < \mu + K\sigma) \geq 1 - \frac{1}{K^2}$$

$$\text{or} \quad P(3 - 2K < X < 3 + 2K) \geq 1 - \frac{1}{K^2}.$$

Putting $K = \frac{5}{2}$, we get

$$P(-2 < X < 8) \geq 1 - \frac{4}{25} = \frac{21}{25}.$$

Example. An unbiased coin is tossed 100 times. Show that the probability that the number of heads will lie between 30 and 70 is greater than 0.93.

Solution: Let X be the number of heads. Then X follows Binomial distribution with mean $np = 100 \times \frac{1}{2} = 50$ and standard deviation= $\sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = 5$.

By Chebyshev's inequality,

$$P(\mu - K\sigma < X < \mu + K\sigma) \geq 1 - \frac{1}{K^2}$$

$$\text{or} \quad P(50 - 5K < X < 50 + 5K) \geq 1 - \frac{1}{K^2}.$$

Putting $K = 4$, we get

$$P(30 < X < 80) \geq 1 - \frac{1}{16} = \frac{15}{16} = 0.9375.$$

$$\therefore \quad P(30 < X < 80) > 0.93.$$

Example. If a die is thrown 3,600 times, show that the probability that the number of sixes lies between 550 and 650 is at least $\frac{4}{5}$.

Solution: Let X be the number of sixes. Clearly, X follows Binomial distribution with mean $n = 3600$ and $p = \frac{1}{6}$.

So $\mu = E(X) = np = 3600 \times \frac{1}{6} = 600$ and $\sigma^2 = \text{var}(X) = np(1-p) = 3600 \times \frac{1}{6} \times \frac{5}{6} = 500$.

Hence, by Chebyshev's inequality,

$$P(|X - 600| < 50) \geq 1 - \frac{\text{var}(X)}{50^2} = 1 - \frac{500}{50^2} = 1 - \frac{1}{5} = \frac{4}{5}.$$

$$i.e., \quad P(550 < X < 650) \geq \frac{4}{5}.$$

Example. Use Chebyshev's inequality to show that for $n \geq 36$, the probability that in n throws of a fair die the number of sixes lies between $\frac{1}{6}n - \sqrt{n}$ and $\frac{1}{6}n + \sqrt{n}$ is at least $\frac{31}{36}$.

Solution: Let X denote the number of sixes in n throws of a fair die.

Then clearly X is binomial (n, p) variate with $p = \frac{1}{6}$.

$\therefore E(X) = np = \frac{n}{6}$ and $\text{var}(X) = np(1-p) = n.\frac{1}{6}.\frac{5}{6} = \frac{5n}{36}$.

Now, by Chebyshev's inequality,

$$\begin{aligned} P\left(\frac{1}{6}n - \sqrt{n} < X < \frac{1}{6}n + \sqrt{n}\right) &= P\left(\left|X - \frac{n}{6}\right| < \sqrt{n}\right) \\ &= 1 - P\left(\left|X - \frac{n}{6}\right| \geq \sqrt{n}\right) \\ &\geq 1 - \frac{5}{36} = \frac{31}{36}. \end{aligned}$$

Example. A random variable X has probability density function $f(x) = 12x^2(1-x)$ for $0 < x < 1$. Compute $P\left(|X - E(X)| \geq 2\sqrt{\text{var}(X)}\right)$ and compare it with the limits determined by Chebyshev's inequality.

Solution: Here,

$$\begin{aligned} E(X) &= \int_0^1 xf(x)dx = \int_0^1 x12x^2(1-x)dx = \frac{3}{5} \\ E(X^2) &= \int_0^1 12x^4(1-x)dx = \frac{2}{5}. \end{aligned}$$

and $\text{var}(X) = E(X^2) - [E(X)]^2 = \frac{2}{5} - \frac{9}{25} = \frac{1}{25}$.

$$
\begin{aligned}
\therefore P\left(|X - E(X)| \geq 2\sqrt{\text{var}(X)}\right) &= P\left(\left|X - \frac{3}{5}\right| \geq \frac{2}{5}\right) \\
&= 1 - P\left(\frac{3}{5} - \frac{2}{5} < X < \frac{3}{5} + \frac{2}{5}\right) \\
&= 1 - P\left(\frac{1}{5} < X < 1\right) \\
&= 1 - 12\int_{\frac{1}{5}}^{1} x^2(1-x)dx = \frac{17}{625}.
\end{aligned}
$$

Now, by Chebyshev's inequality

$$P(|X - E(X)| \geq K\sigma) \leq \frac{1}{K^2}$$

and hence $P\left(|X - E(X)| \geq 2\sqrt{\text{var}(X)}\right) \leq \frac{1}{4}$.

Clearly, $\frac{17}{625} < \frac{1}{4}$. Thus, the above result supports the Chebyshev's limits.

Exercise

1. Let X be a random variable such that $E(X) = 2$ and $E(X^2) = 29$; then find the lower bound for $P(-5 < X < 7)$ using Chebyshev's inequality. **Ans.** $\frac{24}{49}$

2. The probability distribution of a discrete random variable X is given by

$$
\begin{array}{rcccc}
X = i: & -1 & 1 & 3 & 5 \\
P(X = i): & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{2}
\end{array}
$$

   Find the upper bound for $P(|X - 3| \geq 1)$ by Chebyshev's inequality. **Ans.** $\frac{3}{16}$

3. A random variable X has mean 3 and variance 2. Using Chebyshev's inequality to find the upper bound for

   (i) $P(|X - 3| \geq 2)$ (ii) $P(|X - 3| \geq 1)$. $\mathbf{Ans. (i) 1, (ii) \frac{1}{4}}$

4. A continuous random variable X follows normal distribution with parameters m and σ. Find $P(|X - m| \geq 1.5\sigma)$ and compare it with the value given by Chebyshev's inequality. $\mathbf{Ans. 0.1336, 0.444}$

5. A coin is tossed 400 times. Show that the probability that the number of heads will be between 150 and 200 is greater than 0.95.

6. If a die is thrown 1800 times, show that the probability that the number of sixes lies between 250 and 350 is at least $\frac{9}{10}$.

7. The probability density function of a continuous variable is given by

$$f(x) = \begin{cases} 6x(1-x), & \text{if } 0 \le x \le 1. \\ 0, & \text{otherwise.} \end{cases}$$

Find the lower bound for $P\left(|X - \frac{1}{2}| < 2\right)$ by Chebyshev's inequality. **Ans.** $\frac{79}{80}$