$cov(x,y) = \frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})$, where $\bar{x}$ and $\bar{y}$ are means of the variables x and respectively

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

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

and similarly $Var(y) = \frac{1}{n}\sum y_i^2 - \bar{y}^2$

So, we can write

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

7.4 Properties of Correlation coefficients :-

(a) Correlation coefficient of any two variables is a pure number, i.e., is independent of the units of measurement. This follows from the definition of $r(=r_{xy})$

(b) $r_{xy} = r_{yx}$. It also follows from the definition

(c) The numerical value of Correlation coefficient is independent of the change of origins and scales of the variables.

Suppose we are given n pair of values (x_i, y_i), $i=1,2,\ldots,n$, of the variables x and y. Let us introduce $u = \frac{x-a}{c}$, $v = \frac{y-b}{d}$, where a, b, c and d are arbitrary constants and $c \neq 0, d \neq 0$. Then corresponding to each pair of values (x_i, y_i) of x and y, we have a pair of values (u_i, v_i) of u and v where $u_i = \frac{x_i-a}{c}$, $v_i = \frac{y_i-b}{d}$

Scanned with CamScanner

So, $x_i = a + cu_i$, $y_i = b + dv_i$

$\therefore \quad x_i - \bar{x} = c(u_i - \bar{u})$

Similarly, $y_i - \bar{y} = d(v_i - \bar{v})$

$Var(x) = c^2 var(u)$

$var(y) = d^2 var(v)$

$Cov(x,y) = cd\ cov(u,v)$

Hence $r_{xy} = \dfrac{cd\ Cov(u,v)}{\sqrt{c^2 var(u)}\sqrt{d^2 var(v)}} = \dfrac{cd}{|c||d|} r_{uv}$

So, $r_{xy} = \begin{cases} r_{uv} & \text{when } c, d \text{ are of same sign} \\ -r_{uv} & \text{when } c, d \text{ are of opposite sign.} \end{cases}$

(d) $-1 \le r \le 1$

Suppose we are given n pair of values (x_i, y_i), $i = 1, 2, \ldots, n$, of variable x and y. Then $r = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$

$= \frac{1}{n} \sum_i \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{1}{n} \sum_i p_i q_i$

where $p_i = \dfrac{x_i - \bar{x}}{s_x}$, $q_i = \dfrac{y_i - \bar{y}}{s_y}$

So, $\sum p_i q_i = nr$

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

& similarly $\sum q_i^2 = n$

Now since $\sum_{i=1}^{n} (p_i + q_i)^2 \ge 0$ (since squares of real quantities are non-negative)

or, $\sum p_i^2 + \sum q_i^2 + 2 \sum_i p_i q_i \ge 0$

or, $n + n + 2nr \ge 0$

or, $r \ge -1 \quad \ldots (i)$

Scanned with CamScanner

Again $\sum_{i=1}^{n}(p_i q - q_i)^2 \geq 0$ or $\sum p_i^2 + \sum q_i^2 - 2\sum p_i q_i \geq 0$

or, $n + n - 2nr \geq 0$ or $r \leq 1$ ... (ii)

from (i) and (ii), we have $-1 \leq r \leq 1$

Remark: When $r = -1$ then, for each i $(p_i + q_i)^2 = 0$

or, $q_i = -p_i$ or, $\frac{y_i - \bar{y}}{\sigma_y} = -\frac{x_i - \bar{x}}{\sigma_x}$

or, $y_i = \bar{y} - \frac{\sigma_y}{\sigma_x}(x_i - \bar{x})$. Hence all pair of values of x, y lies on the straight line $y = \bar{y} - \frac{\sigma_y}{\sigma_x}(x - \bar{x})$, slope of the line being negative

Similarly when $r = 1$, for each i, $(p_i - q_i)^2 = 0$

or, $y_i = \bar{y} + \frac{\sigma_y}{\sigma_x}(x_i - \bar{x})$

So, all pair of values of x, y lies on the straight line $y = \bar{y} + \frac{\sigma_y}{\sigma_x}(x - \bar{x})$, the slope of the line is positive.

Example 7.1 $n = 10$, $\sum x = 140$, $\sum y = 150$, $\sum (x-10)^2 = 180$, $\sum (y-15)^2 = 215$, $\sum (x-10)(y-15) = 60$. Find the correlation coefficient between x and y.

Solution: Here $\bar{x} = \frac{\sum x}{n} = \frac{140}{10} = 14$, $\bar{y} = \frac{\sum y}{n} = \frac{150}{10} = 15$

Now $\sum (x-10)^2 = 180$

or, $\sum (x - 14 + 4)^2 = 180$

or, $\sum (x-14)^2 + 8\sum (x-14) + 16 \times 10 = 180$ [Since $n = 10$]

or, $\sum (x-14)^2 + 160 = 180$ [$\because \sum (x-14) = \sum (x - \bar{x}) = 0$]

or, $\sum (x-14)^2 = 20$ i.e $\sum (x - \bar{x})^2 = 20$

Scanned with CamScanner

$\sum(y-15)^2 = 215$ or, $\sum(y-\bar{y})^2 = 215$

Again $\sum(x-10)(y-15) = 60$

or, $\sum(x-14+4)(y-15) = 60$

or, $\sum(x-14)(y-15) + 4\sum(y-15) = 60$

or, $\sum(x-14)(y-15) = 60$ [Since $\sum(y-15) = \sum(y-\bar{y}) = 0$]

i.e., $\sum(x-\bar{x})(y-\bar{y}) = 60$

Hence $r = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} = \dfrac{60}{\sqrt{20\times215}} = \dfrac{60}{\sqrt{4300}} = 0.92$

Example 7.2 Let x_0 and y be uncorrelated variables with standard deviations s_1 and s_2. Show that the correlation coefficient between x and xy is $\dfrac{s_1}{\sqrt{s_1^2+s_2^2}}$

Solution: $r_{xy} = 0$ so that $cov(x,y) = 0$

Let $u = x+y$ or $\bar{u} = \bar{x}+\bar{y}$ or $u-\bar{u} = (x-\bar{x}) + (y-\bar{y})$

Now, $Cov(x, xy) = cov(x,u) = \frac{1}{n}\sum(x-\bar{x})(u-\bar{u})$

$= \frac{1}{n}\sum(x-\bar{x})\{(x-\bar{x})+(y-\bar{y})\}$

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

$= var(x) + cov(x,y) = s_1^2$

$Var(u) = Var(x+y) = Var(x) + Var(y) + 2cov(x,y) = s_1^2 + s_2^2$

So, $r_{xu} = \dfrac{cov(x,u)}{\sqrt{Var(x)Var(u)}}$

$= \dfrac{s_1^2}{s_1\sqrt{s_1^2+s_2^2}} = \dfrac{s_1}{\sqrt{s_1^2+s_2^2}}$

Scanned with CamScanner