Correlation

Correlation is a mathematical measure of linear relationship between two variables. The simplest way to get an idea whether the variables are correlated is to obtain the diagram of the points (x_i, y_i), $i = 1, 2, 3, \ldots n$ called the scatter diagram.

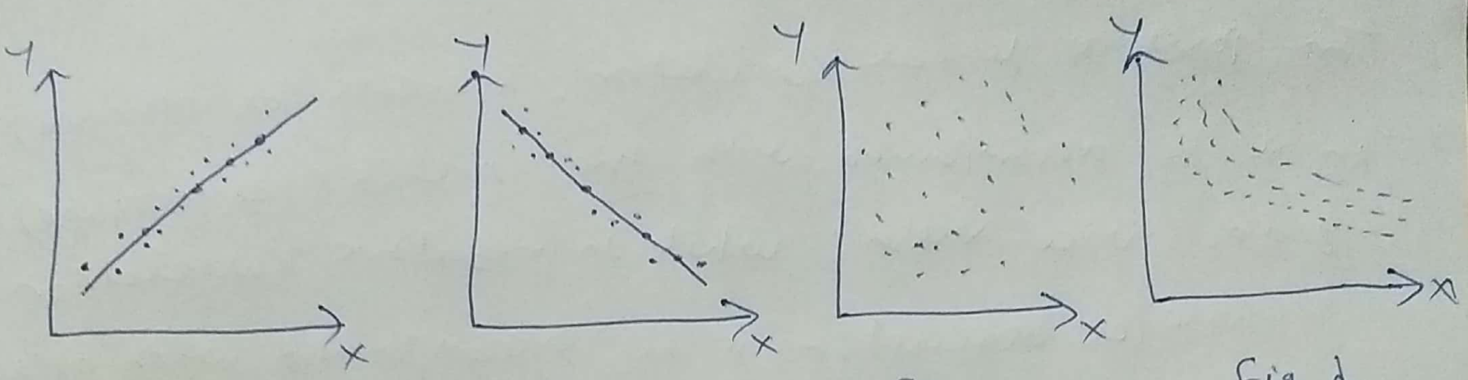

Fig. a   Fig. b   Fig. c   Fig. d

If all the points on the scatter diagram lie on a straight line, the perfect linear correlation is said to exist. But In general, there is no perfect linear correlation exist.

When the points tend to concentrate about a straight line with positive slope, then positive linear correlation is said to exist. refer Fig a.

When the points tend to concentrate about a straight line with negative slope, the negative linear correlation is said to exist, refer Fig. b.

2

In case the points are widely scattered as shown in Figure c, then a poor correlation or no correlation is expected. However when the points in the scatter diagram tend to concentrate around some specific curve, as shown in fig. d, then curvilinear correlation is said to exist. Here we shall study only linear correlation.

In positive linear correlation, increase (or decrease) in x is accompanied with the increase (or decrease) in y and vice-versa; while in negative correlation increase (or decrease) in x is accompanied with decrease (or increase) in y and vice-versa.

Covariance

If X and Y are two random variables, then Covariance between X and Y is defined as

$$\text{Cov.}(X,Y) = E\{(X - E(X))(Y - E(Y))\} \quad \text{—— (1)}$$

By doing some easy calculations,

we have $$\text{Cov.}(X,Y) = E(XY) - E(X)E(Y) \quad \text{—— (2)}$$

Scanned with CamScanner

Note:

If X and Y are independent, then

$Cov.(X,Y) = 0$ $\quad [\because E(XY) = E(X)E(Y)]$

Correlation Coefficient

or

Karl Pearson's Coefficient of Linear Correlation

The correlation coefficient between X and Y is defined as

$$\rho = \frac{Cov.(X,Y)}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are standard deviations of X and Y respectively

$$\therefore \rho = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - (E(X))^2}\sqrt{E(Y^2) - (E(Y))^2}}$$

Note:

(i) The value of ρ lies between -1 and 1

$\therefore -1 \leq \rho \leq 1$

(ii) Correlation is perfect and positive for $\rho = 1$

(iii) Correlation is perfect and negative for $\rho = -1$

(iv) If $\rho = 0$, then there is no linear correlation between x and y and variables are said to be uncorrelated

Q. (1) Let X and Y have joint p.m.f. given as

(x,y)	$(1,1)$	$(1,2)$	$(1,3)$	$(2,1)$	$(2,2)$	$(2,3)$
$p(x,y)$	$2/15$	$4/15$	$3/15$	$1/15$	$1/15$	$4/15$

And zero elsewhere. Find the correlation coefficient between X and Y.

Solⁿ

The Marginal p.m.f.s can be obtained as.

X \ Y	1	2	3	$P_1(x)$
1	$2/15$	$4/15$	$3/15$	$9/15 = 3/5$
2	$1/15$	$1/15$	$4/15$	$6/15 = 2/5$
$P_2(y)$	$3/15 = 1/5$	$5/15 = 1/3$	$7/15$	$\Sigma P_1(x) = 1$, $\Sigma P_2(y) = 1$

Now correlation coefficient $\rho = \dfrac{Cov.(x,y)}{\sigma_x \sigma_y}$

Now, $E(X) = \sum_{x=1}^{2} x P_1(x) = (1)\left(\frac{3}{5}\right) + (2)\left(\frac{2}{5}\right) = 7/5$

$E(x^2) = \sum_{x=1}^{2} x^2 P_1(x) = (1)^2(3/5) + (2)^2(2/5) = 11/5$

$V(X) = E(x^2) - \{E(x)\}^2 = \frac{11}{5} - \frac{49}{25} = \frac{6}{25}$

$\sigma_x = \sqrt{V(x)} = \sqrt{\frac{6}{25}} = \frac{1}{5}\sqrt{6}$.

Now, $E(Y) = \sum_{y=1}^{3} y P_2(y) = (1)\left(\frac{1}{15}\right) + (2)\left(\frac{1}{3}\right) + (3)\left(7/15\right) = \frac{34}{15}$

$E(y^2) = \sum_{y=1}^{3} y^2 P_2(y) = (1)^2\left(\frac{1}{15}\right) + (2)^2\left(\frac{1}{3}\right) + (3)^2(7/15) = 86/15$

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{86}{15} - \left(\frac{34}{15}\right)^2$$

$$= \frac{86}{15} - \frac{1156}{225} = \frac{1290 - 1156}{225}$$

$$= \frac{134}{225}$$

$$\therefore \sigma_Y = \sqrt{V(Y)} = \sqrt{\frac{134}{225}} = \frac{1}{15}\sqrt{134}$$

Now, $$E(XY) = \sum_{x=1,2}^{2} \sum_{y=1}^{3} xy\, p(x,y)$$

$$= (1)(1)\,p(1,1) + (1)(2)\,p(1,2) + (1)(3)\,p(1,3)$$
$$+ (2)(1)\,p(2,1) + (2)(2)\,p(2,2) + (2)(3)\,p(2,3)$$

$$= (1)(2/15) + (2)(4/15) + (3)(3/15)$$
$$+ (2)(1/15) + (4)(1/15) + (6)(4/15)$$

$$= 2/15 + \frac{8}{15} + 9/15 + 2/15 + 4/15 + 24/15$$

$$= \frac{49}{15}$$

Now, $$\rho = \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y}$$

$$\Rightarrow \rho = \frac{49/15 - (7/5)(34/15)}{\left(\frac{1}{5}\sqrt{6}\right)\left(\frac{1}{15}\sqrt{134}\right)} = \frac{49 - 47.6}{5.671}$$

$$\Rightarrow \rho = \cancel{\frac{49 - 47.6}{3\sqrt{6}\sqrt{134}}} = \cancel{11.5758}$$

$$\Rightarrow \cancel{\rho =} \Rightarrow \rho = 0.246$$

Hence, there is positive linear correlation between X and Y.

Q.(2) Let X and Y are jointly distributed as

$$f(x,y) = \begin{cases} 2-x-y & , 0 \le x \le 1, 0 \le y \le 1 \\ 0 & , \text{elsewhere} \end{cases}$$

(a) Find the covariance between X and Y

(b) Find the correlation of X and Y

Soln (a) Marginal pdf of X is

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_0^1 (2-x-y)\,dy$$

$$\Rightarrow f_1(x) = \left[2y - xy - \frac{y^2}{2}\right]_0^1 = 2 - x - \frac{1}{2}$$

$$\Rightarrow f_1(x) = \frac{3}{2} - x, \quad \text{for } 0 \le x \le 1$$

Marginal pdf of Y is

$$f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_0^1 (2-x-y)\,dx$$

$$\Rightarrow f_2(y) = \left[2x - \frac{x^2}{2} - xy\right]_0^1 = 2 - \frac{1}{2} - y$$

$$\Rightarrow f_2(y) = \frac{3}{2} - y, \quad \text{for } 0 \le y \le 1$$

Scanned with CamScanner

Now, $E(X) = \int_0^1 x f_1(x)\,dx$

$= \int_0^1 x\left\{\left(\frac{3}{2}\right) - x\right\} dx = \left[\frac{3}{2}\frac{x^2}{2} - \frac{x^3}{3}\right]_0^1$

$= \frac{3}{4} - \frac{1}{3} = \frac{9-4}{12} = \frac{5}{12}$

$E(X^2) = \int_0^1 x^2 f_1(x)\,dx = \int_0^1 x^2\left(\frac{3}{2} - x\right) dx$

$= \left[\frac{3}{2}\cdot\frac{x^3}{3} - \frac{x^4}{4}\right]_0^1 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

$V(X) = E(X^2) - (E(X))^2 = \frac{1}{4} - \frac{25}{144} = \frac{11}{144}$

$\sigma_X = \sqrt{V(X)} = \sqrt{\frac{11}{144}} = \frac{1}{12}\sqrt{11}$

Now,

$E(Y) = \int_0^1 y f_2(y)\,dy = \int_0^1 y\left(\frac{3}{2} - y\right) dy = \frac{5}{12}$

$E(Y^2) = \int_0^1 y^2 f_2(y)\,dy = \frac{1}{4}$.

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

$\sigma_Y = \sqrt{V(Y)} = \frac{1}{12}\sqrt{11}$

Now, $E(XY) = \int_0^1\int_0^1 xy\, f(x,y)\,dx\,dy$

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

$= \int_0^1\left\{\int_0^1 (2xy - x^2y - xy^2)\,dx\right\} dy$

$$= \int_0^1 \left(y - y/3 - y^2/2 \right) dy$$

$$= \int_0^1 \left(\frac{2y}{3} - y^2/2 \right) dy$$

$$= \left[\frac{2y^2}{6} - y^3/6 \right]_0^1 = 2/6 - 1/6$$

$$= 1/6$$

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

$$= \left(\frac{1}{6}\right) - \left(\frac{5}{12}\right)\left(\frac{5}{12}\right) = \frac{1}{6} - \frac{25}{144}$$

$$= -\frac{1}{144}$$

(b) Correlation coefficient is

$$\rho = \frac{Cov.(X, Y)}{\sigma_X \sigma_Y}$$

$$\Rightarrow \rho = \frac{-1/144}{\frac{1}{12}\sqrt{11} \cdot \frac{1}{12}\sqrt{11}} = \frac{-1/144}{11/144}$$

$$\Rightarrow \rho = -\frac{1}{11}.$$