Covariance and Correlation:

When we consider the joint distribution of two random variables, it is useful to have a numerical summary that enables us to measure the association between the two variables. The covariance and correlation are the attempts to measure that association or dependence.

Def^n: (Covariance)

The covariance of two random variables X and Y having finite expectations $E[X] = \mu_x$ and $E[Y] = \mu_y$, denoted by $Cov(X, Y)$, is defined as

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] \quad \text{——} \; (1)$$

It is provided that the expectation in eq^n (1) exists.
The value of $Cov(X, Y)$ may be positive, zero or negative.
It follows from (1) that

$$Cov(X,Y) = E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y]$$

$$= E[XY] - \mu_X E[Y] - \mu_Y E[X] + E[\mu_X \mu_Y]$$

$$= E[XY] - \cancel{\mu_X \mu_Y} - \mu_Y \mu_X + \cancel{\mu_X \mu_Y}$$

$$\therefore Cov(X,Y) = E[XY] - \mu_X \mu_Y$$

CamScanner

Note that if X and Y are independent, then

$$E[XY] = E[X]\,E[Y]$$

Therefore,

$$Cov(X,Y) = E[X]\,E[Y] - \mu_X \mu_Y$$

$$= \mu_X \mu_Y - \mu_X \mu_Y$$

$$\therefore Cov(X,Y) = 0$$

Properties of Covariance:

For any random variables X, Y & Z and constants a, b

(1) $Cov(X,X) = V(X)$

(2) $Cov(X,Y) = Cov(Y,X)$

(3) $Cov(aX, bY) = ab\, Cov(X,Y)$

(4) $Cov(X+Y, Z) = Cov(X,Z) + Cov(Y,Z)$

Proof of (1):

By defn $Cov(X,X) = E\left[(X-\mu_X)(X-\mu_X)\right]$

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

$$= V(X)$$

Proof of (2):

By def^n, $Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$

$= E[(Y-\mu_Y)(X-\mu_X)]$

$= Cov(Y,X)$.

Proof of (3):

By def^n $Cov(aX, bY) = E[\{aX - E(aX)\}\{bY - E(bY)\}]$

$= E[\{aX - aE[X]\}\{bY - bE[Y]\}]$

$= E[a\{X - E[X]\} b\{Y - E[Y]\}]$

$= abE[(X-\mu_X)(Y-\mu_Y)]$

$\therefore Cov(aX, bY) = ab\,Cov(X,Y)$

Proof of (4):

By def^n of covariance, we know that

$Cov(X,Y) = E[XY] - E[X].E[Y]$

$\therefore Cov(X+Y, Z) = E[(X+Y)Z] - E[X+Y]E[Z]$

$$= E[XZ+YZ] - (E[X]+E[Y])E[Z]$$

$$= E[XZ] + E[YZ] - E[X]\cdot E[Z] - E[Y]\cdot E[Z]$$

$$= \{E[XZ] - E[X]\cdot E[Z]\} + \{E[YZ] - E[Y]\cdot E[Z]\}$$

$$\therefore Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)$$

Thm:

If $X_1, X_2, X_3, \ldots, X_n$ are random variables with finite means, then

$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) + 2\sum_{i<j}^{n}\sum Cov(X_i, X_j)$$

Proof:

We know that,

$$Cov(X,X) = V(X)$$

$$\therefore V\left(\sum_{i=1}^{n} X_i\right) = Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j)$$

$$= \sum_{i=1}^{n} Cov(X_i, X_i) + \sum\sum_{i \neq j} Cov(X_i, X_j)$$

$$= \sum_{i=1}^{n} V(X_i) + 2\sum\sum_{i<j} Cov(X_i, X_j)$$

It is noted here that if $X_1, X_2, X_3, \ldots, X_n$ are independent, then $Cov(X_i, X_j) = 0$ for all $i \neq j$.

Then,
$$V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i).$$

(Proved)

Problem:

Consider the following joint PDF of X & Y

$$f(x,y) = \begin{cases} \frac{6}{5}(x^2 + 2xy); & 0 \leq x \leq 1,\ 0 \leq y \leq 1 \\ 0; & \text{otherwise} \end{cases}$$

Find the covariance between X and Y.

Sol^n:

The marginal density function of X is

$$f_1(x) = g(x) = \begin{cases} \frac{6}{5} x(x+1); & 0 \leq x \leq 1 \\ 0; & \text{otherwise} \end{cases}$$

The marginal density function of Y is

$$f_2(y) = h(y) = \begin{cases} \frac{2}{5}(1+3y); & 0 \le y \le 1 \\ 0; & \text{otherwise} \end{cases}$$

Now,

$$\mu_X = E[X] = \int_0^1 x \cdot \frac{6}{5} x(x+1)\, dx$$

$$= \frac{6}{5}\int_0^1 x^2(x+1)\, dx$$

$$= 7/10$$

$$\mu_Y = E[Y] = \int_0^1 y\, \frac{2}{5}(1+3y)\, dy$$

$$= \frac{2}{5}\int_0^1 y(3y+1)\, dy$$

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

$$E[XY] = \int_0^1\int_0^1 xy\,\frac{6}{5}(x^2+2xy)\, dx\, dy$$

$$= \frac{6}{5}\int_0^1\int_0^1 (x^3y + 2x^2y^2)\, dx\, dy$$

$$= 5/12$$

CS CamScanner

Thus,

$$Cov(X,Y) = E[XY] - E[X] \cdot E[Y]$$

$$= \frac{5}{12} - \left(\frac{7}{10}\right)\left(\frac{3}{5}\right)$$

$$= -\frac{1}{300}$$

Ans

Correlation:

Correlation means association—more preciously it is a measure of the extent to which two variables are related. There are three possible results of a correlation study: a positive correlation, a negative correlation, and no correlation.

Positive Correlation:

A positive correlation is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases.

An example of positive correlation would be height and weight. Taller people tend to be heavier.

CS CamScanner

Negative Correlation:

A negative correlation is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example of negative correlation would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).

Zero Correlation:

A zero correlation exists when there is no relationship between two variables. For example there is no relationship between the amount of tea drunk, and the level of intelligence.

Scattergrams:

A correlation can be expressed visually. This is done by drawing a scattergram (also known as a scatterplot, scatter graph, scatter chart, or scatter diagram)

A scattergram is a graphical display that shows the relationships or association between two numerical variables (or co-variables), which are represented as points (or dots) for each pair of score

A scattergraph indicates the strength and direction of the correlation between the co-variables.

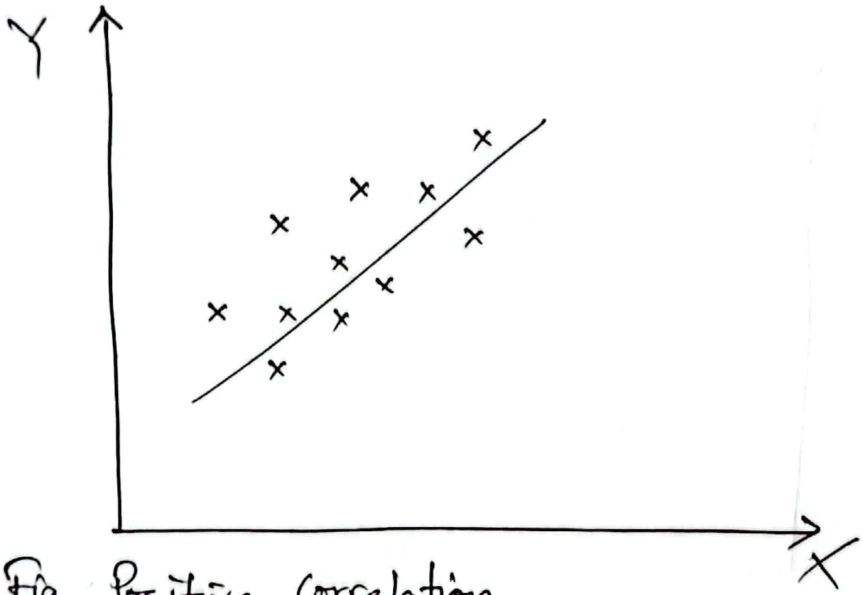


Fig. Positive Correlation

The points lie close to a straight line, which has a positive gradient.

This shows that as one variable increases, the other increases too.

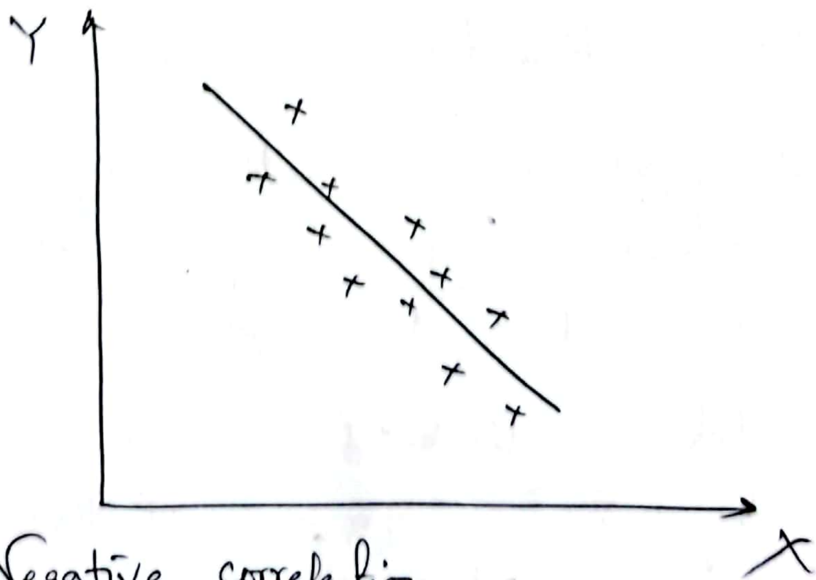
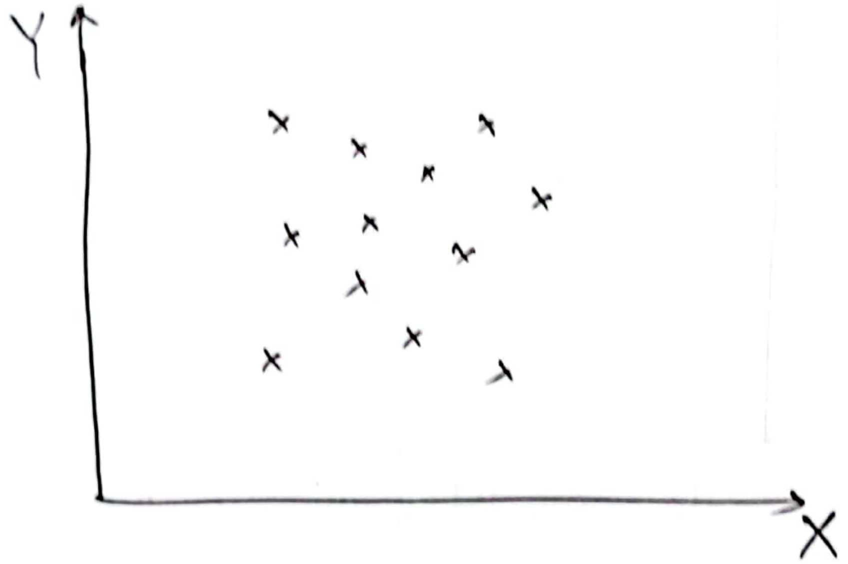


Fig. Negative correlation

• The points lie close to a straight line, which has a negative gradient.

• This shows that as one variable increases, the other decreases.



• There is no pattern among the points.

• This shows that there is no connection between the two variables.

<u>Correlation Coefficient</u>: (Determining correlation strength)

Instead of drawing a scattergram, a correlation can be expressed numerically as a coefficient, ranging from -1 to 1.

The correlation coefficient indicates the extent to which the pairs of numbers of these two variables lie on a straight line. Values over zero indicate a positive correlation, while values under zero indicat a negative correlation.

A Correlation of -1 indicates a perfect negative correlation, meaning that as one variable goes up, the other goes down. A correlation of $+1$ indicates a perfect positive correlation, meaning that as one variable goes up, the other goes up.

Def^n: (Correlation coefficient)

Let X and Y be two random variables with finite variances $V(X)$ and $V(Y)$, respectively. The correlation coefficient $\rho(X,Y)$ is defined to be zero if $V(X)=0$ or $V(Y)=0$, and otherwise

$$\rho(X,Y)=\frac{Cov(X,Y)}{\sqrt{V(X).V(Y)}}$$

Note that $\rho(X,Y)$ remains unaffected by a change of units, and therefore is a dimensionless quantity.

Problem:

Consider the joint PDF of X & Y as follows:

$$f(x,y)=\begin{cases} \frac{6}{5}(x^2+2xy); & \text{if } 0\le x\le 1,\ 0\le y\le 1 \\ 0; & \text{otherwise} \end{cases}$$

CamScanner

Compute the correlation between X & Y

Sol^n:

The marginal density function of X is

$$g(x) = \begin{cases} \frac{6}{5} x(x+1); & 0 \le x \le 1 \\ 0; & \text{otherwise} \end{cases}$$

The marginal density function of Y is

$$h(y) = \begin{cases} \frac{2}{5}(1+3y); & 0 \le y \le 1 \\ 0; & \text{otherwise} \end{cases}$$

Now, $$\mu_X = E[X] = \int_0^1 x \, \frac{6}{5} x(x+1)\, dx$$

$$= 7/10$$

$$\mu_Y = E[Y] = \int_0^1 y \cdot \frac{2}{5}(1+3y)\, dy$$

$$= 3/5$$

~~Cov~~ $$E[XY] = \int_0^1\int_0^1 xy\, f(x,y)\, \cancel{g(x)h(y)}\, dx\, dy$$

$$= \int_0^1\int_0^1 xy \, \frac{6}{5}(x^2+xy)\, dx\, dy$$

$$= 5/12$$

$$\therefore \operatorname{Cov}(X,Y) = E[XY] - E[X].E[Y]$$
$$= \frac{5}{12} - \left(\frac{7}{10} * \frac{3}{5}\right)$$
$$= -\frac{1}{300}$$

Now,
$$E[X^2] = \int_0^1 x^2 \frac{6}{5} x(x+1)\,dx = \frac{27}{50}$$
$$E[Y^2] = \int_0^1 y^2 \frac{2}{5}(1+3y)\,dy = \frac{13}{30}$$
$$\therefore V[X] = E[X^2] - \{E[X]\}^2 = \frac{1}{20}$$
$$V[Y] = E[Y^2] - \{E[Y]\}^2 = \frac{11}{150}.$$

Thus,
$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{V[X]\cdot V[Y]}}$$
$$= \frac{-1/300}{\sqrt{(1/20)\cdot(11/150)}}$$

CS CamScanner

$\Rightarrow \rho(X,Y) = -0.055$

Therefore, the two random variables X and Y are negatively correlated.