Covariance and correlation are two statistical measures used to determine the relationship between two variables. Both are used to understand how changes in one variable are associated with changes in another variable.

Covariance
Definition: Covariance is a measure of how much two random variables change together. If the variables tend to increase and decrease together, the covariance is positive. If one tends to increase when the other decreases, the covariance is negative.

Quantify the relationship between x and y

$$x \uparrow \, y \uparrow$$

Dataset

↓ Size of House	↓ Price
1200	1M
1300	2M
1500	8M

$$x \uparrow \, y \uparrow \implies \text{The covariance}$$
$$x \downarrow \, y \downarrow \implies \text{The value}$$

x	y
7	6
6	12
6	14
4	16

→ negative
  covariance

Covariance

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

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

$$Cov(x,x) = Var(x)$$

$x_i \to$ Datapoint of random variable x
$\bar{x} \to$ Sample mean of n
$y_i \to$ Datapoints of random variable y
$\bar{y} \to$ Sample mean of y

Students

Hour Studied (X)	Exam Score(Y)
2	50
3	60
4	70
5	80
6	90

$x \uparrow y \uparrow$ Positive
$x \downarrow y \downarrow$ covariance

① $\bar{x} = \frac{2+3+4+5+6}{5} = 4$

② $\bar{y} = \frac{50+60+70+80+90}{5} = 70$

$Cov(x,y) = \frac{(2-4)(50-70)+(3-4)(60-70)+(4-4)(70-70)}{4}$
$\qquad \qquad \frac{+(5-4)(80-70)+(6-4)(90-70)}{4}$

$Cov(x,y) = 20$

The positive covariance indicates the number of hours studied increased the exam score also.

Consider the next data.

x	y
7	10
6	12
5	14

$\sim Vc$

$x \uparrow y \downarrow$
$x \downarrow y \uparrow$

Advantages

1. Quantify the relationship between X and Y

Disadvantage

2. Covariance does not have a specific limit value.

$$Cov(x,y) \implies -\infty \text{ to } \infty$$
$$\underset{20}{Cov(A,B)} \quad \underset{30}{Cov(B,C)}$$
$$<$$

We have two types

1. Pearson Correlation Coefficient
2. Spearman rank correlation

1) Pearson correlation coefficient
Limits the values between [-1, 1]

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y} = \frac{20}{\sigma_x \sigma_y} \implies 0 \to 1$$

1) The more the value towards +1 the more positive correlated is $[0, 1]$
2) The more the value towards -1 the more negative correlated is $[0, -1]$

Examples of scatter diagrams with different values of correlation coefficient (ρ)

Several sets of (x, y) points, with the correlation coefficient of x and y for each set. The correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.

2) SPEARMAN RANK CORRELATION

A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data points with y values greater than that of a given data point will have greater y values as well. In contrast, this does not give a perfect Pearson correlation.

$x \uparrow y \uparrow$
$x \downarrow y \downarrow$ → Pearson correlation
                            = 0.8 ↓

$\rho = +1$

$\rho \, 1$

Pearson correlation for linear relation data, for non-linear data use rank information of relation of no linear relation.

$$r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) \cdot \sigma(R(y))}$$

x	y	R(x)	R(y)
1	8	2	1
3	4	1	2
5	6	4	3
7	5	5	5
0	7	1	4

Feature selection

size of house ↑   No. of room ↑   Location ↑   Hosptial Price ↑ ↑ ↓
                                                       -ve correlation

=   =   =   ↓
                                              People
                          x stay in the house   o correlation