Covariance And Correlation

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.

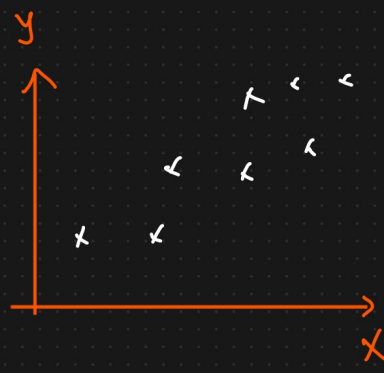
[Quantify the Relationship between X and Y]

X	Y
2	3
4	5
6	7
8	9

X↑	Y↑
X↓	Y↑
X↑	Y↓
X↓	Y↓

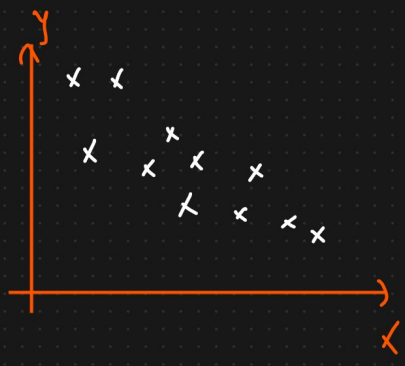
Dataset

↓ ↑ Size of House	Price ↑ ↓
1200	45 lakhs
1300	50 lakh
1500	75 lakh



X↑	Y↑
X↓	Y↓

⇒ +ve Covariance ⇒ +ve value



X↓	Y↑
X↑	Y↓

X	Y
7	10
6	12
5	14
4	16

⇒ -ve Covariance ⇒ -ve value

Covariance

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

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

$$\boxed{Cov(x,x) = Var(x)}$$

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

$x_i \rightarrow$ Datapoint of random variable x

$\bar{x} \rightarrow$ Sample mean of x

$y_i \rightarrow$ Datapoints of random variable y

$\bar{y} \rightarrow$ Sample mean of y

Students

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

$x\uparrow \; y\uparrow$ $x\downarrow \; y\downarrow$ $\Rightarrow$ +ve covariance

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

① $\bar{x} = \frac{2+3+4+5+6}{5} = 4$ //

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

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

$Cov(x,y) = 20$ //.

$\Rightarrow$ The positive covariance indicates the no. of hours studied increased the exam score also.

X	Y
7	10
6	12
5	14

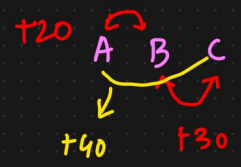
$\Rightarrow$ -Ve //.

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

0.96 Cov(A, B) 0.88 Cov(B, C)

-200 +100 20 = Cov(A, B)

-300 +300 30 = Cov(B, C)



Advantages

-1 to 1

Disadvantage

① Quantify the Relationship between X and Y

① Covariance does not have a specific limit value.

$$Cov(X,Y) \Rightarrow -\infty \text{ to } \infty$$

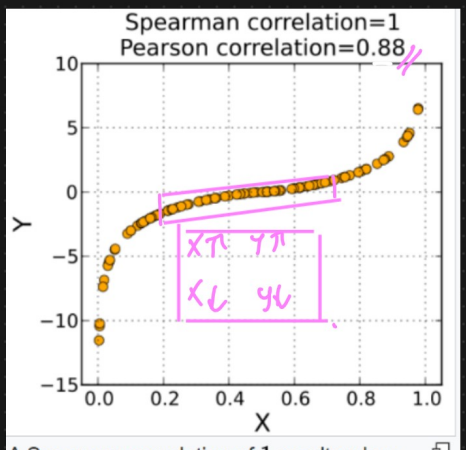
② Correlation
- → Pearson Correlation Coefficient
- → Spearman Rank Correlation

① Pearson Correlation Coefficient $\Rightarrow [-1 \text{ to } 1]$

$$\rho_{x,y} = \frac{Cov(x,Y)}{\sigma_x \cdot \sigma_y} = \frac{20}{\sigma_x \cdot \sigma_y} \Rightarrow 0 \text{ to } 1$$

(Y) The more the value towards +1 the more +ve correlated X & Y is.

(K) The more the value towards -1 the more -ve correlated it is (X,Y)

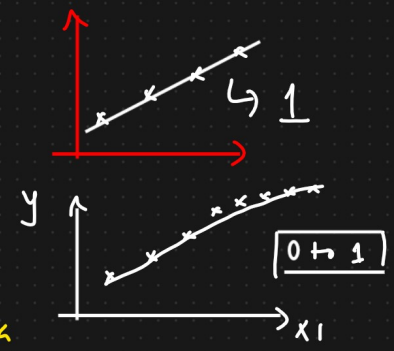


A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data points with greater x values than that of a given data point will have greater y values as well. In contrast, this does not give a perfect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

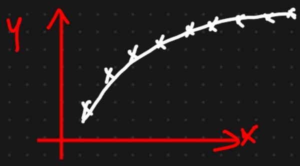
⇒ X↑ Y↑
⇒ X↓ Y↓

Pearson Correlation = 0.88

② Spearman Rank Correlation



Pearson Correlation for non linear data



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

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

Feature Selection

+ve Correlated

Size of House ↑   No. of Room ↑   Location ↑   ~~No. of People staying in the house~~ → Price ↑↑ (o/p)

No

≈ 0

Haunted ↔ Price

-ve Correlation