COVARIANCE

Q. What is covariance?

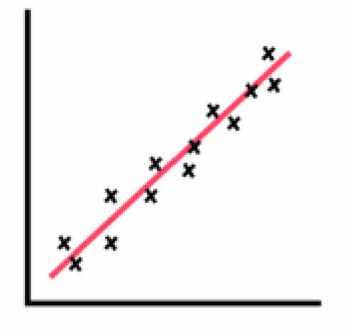
➞Covariance is a measure that tells us **how two variables change together**. Theoretically,

Covariance is a **statistical measure** that indicates the **degree and direction of the linear relationship** between two variables. It shows how two variables **vary together**.

Moving forward, there are three types of correlation:

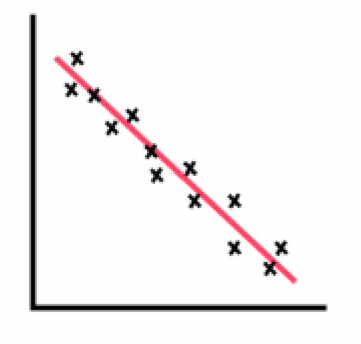
1)Positive covariance (r>0) 📈

➞If both variables increase or decrease at the same time, the covariance is **positive**.



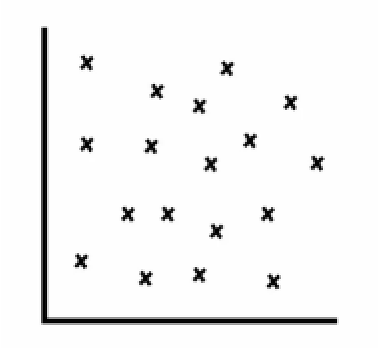
2)Negative covariance (r<0) 📉

If one increases while the other decreases, the covariance is **negative**.



3)Zero covariance(r≈ 0)

If there is no clear pattern, the covariance is **close to zero**.



Let us understand types of covariance through the following example where,

X- No. of hours studied

Y-Exam score

i)POSITIVE COVARIANCE

When more number of hours is spent on studying there are high chances of scoring good marks!

ii) NEGATIVE COVARIANCE

When less number of hours is spent on studying, then there are chances where lower marks are expected.

iii)ZERO COVARIANCE

Studying more does not necessarily mean getting higher or lower marks :)

FORMULAE

i) Population Sample

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

Here,

- Xi is the values of the X-variable
- Yi is the values of the Y-variable
- $\bar{X}$ is the mean of the X-variable
- $\bar{Y}$ is the mean of the Y-variable
- n is the number of data points

ii)Sample covariance

$$\text{Cov (X, Y) = } = \frac{\left\{\sum_{\{i=1\}}^{\{n\}}(X_i-\bar{X})(Y_i-\bar{Y}\{\ \ \})\right\}}{N-1}$$

Example 1: Calculate the sample covariance for the following data.

Student (i)	**Xi (Hours)**	**Yi (Score)**
1	2	4
2	4	8
3	6	6

➡ Calculation

1. **Calculate the Means ($\bar{X}$ and $\bar{Y}$)**
   - **$\bar{X}$ = (2 + 4 + 6) / 3 = 12 / 3 = 4**
   - **$\bar{Y}$ = (4 + 8 + 6) / 3 = 18 / 3 = 6**
2. **Calculate Deviations and Products (The Numerator $\sum(X_i - \bar{X})(Y_i - \bar{Y})$**

Xi	**Yi**	**(Xi - $\bar{X}$)**	**(Yi–$\overline{Y)}$**	**Product (Xi - $\bar{X}$) (Yi –$\overline{Y)}$**
2	**4**	**2 - 4 = -2**	**4 – 6 = -2**	**(-2)x(-2)= 4**
4	**8**	**4 – 4= 0**	**8 – 6 = 2**	**(0)x(2)= 0**
6	**6**	**6 – 4 = 2**	**6 – 6 = 0**	**(2)x(0)=0**

Xi	Yi	(Xi - $\overline{X}$)	(Yi$-\overline{Y})$	Product (Xi - $\overline{X}$) (Yi $-\overline{Y})$
Sum ($\sum$)				4

3. **Final Covariance Calculation:**
    - **Numerator ($\sum$)=4**
    - **Denominator ($n-1$): 3 - 1 = 2**

$$Cov(X,Y) = \frac{4}{2} = 2$$

Example 2: Calculate Population covariance for the following data.

Employee (i)	Xi (Years Experience)	Yi (Salary in $'000s)
1	2	4
2	4	8
3	6	6
4	8	10
Sum ($\sum$)	20	28

➡Calculation

1. Find the Population Means (μ_X and μ_Y)

The population mean is found by dividing the sum by the population size, N=4

- **Mean of X $(\boldsymbol{\mu_X})$:**

  $\boldsymbol{\mu_X} = \{\sum \boldsymbol{X_i}\}/\{\boldsymbol{N}\}$

  = 20/4 = 5

- **Mean of Y$(\boldsymbol{\mu_Y})$**

  $\boldsymbol{\mu_Y}$ **=** $\{\sum \boldsymbol{Y_i}\}/\{\boldsymbol{N}\}$

  =28/4 = 7

---

2. Calculate Deviations and Products (The Numerator)

Now we calculate how far each point deviates from its *population* mean and multiply those deviations.

Xi	Yi	(Xi−μX) (μX =5)	(Yi−μY) (μY=7)	Product (Xi−μX)(Yi −μY)
2	4	2 – 5 = -3	4 - 7 = -3	(-3)x(-3)= 9
4	8	4 - 5 = -1	8 - 7 = 1	(-1)x(1) =1
6	6	6 - 5 = 1	6 – 7 = -1	(1)x(-1) = -1
8	10	8 - 5 = 3	10 - 3 = 7	(3)x(3) = 9
Sum $(\sum)$				9+(-1)+(-1)+9= 16

The numerator $\sum (\boldsymbol{X_i} - \boldsymbol{\mu_X})(\boldsymbol{Yi} - \boldsymbol{\mu_Y})$ **is 16.**

3. Calculate Population Covariance $(\boldsymbol{\sigma_{XY}})$

We use the population formula, which divides the sum by N(the population size, 4)

$$\sigma_{XY} = \sum_{\{i=1\}}^{\{N\}} (X_i - \mu_X)(Y_i - \mu_Y)\}/\{N\}$$

$$\sigma_{XY} = 16/4$$

$$= 4$$

COVARIANCE QUIZ QUESTIONS

1.What primary feature of the relationship between two variables does the sign (+or -) of the covariance value indicate?

A. The strength of the relationship.

B. The cause-and-effect relationship.

C. The average value of the data set.

D. The direction of the linear relationship (direct or inverse).

ANS- D

Explanation- As we know that the sign of the covariance value indicates the direction of the linear relationship between two variables (X and Y)

- **Positive Covariance ($+$): Means the variables move in the same direction (a direct relationship). As X increases, Y tends to increase.**
- **Negative Covariance ($-$): Means the variables move in opposite directions (an inverse relationship). As X increases, Y tends to decrease.**
- **Zero Covariance (0): Means there is no linear relationship.**

2. What does a Positive Covariance mean for two variables ($X\ and\ Y$)?

A. The variables have no linear relationship.

B. The variables tend to move in the same direction (e.g., both increase or both decrease).

C. The strength of the relationship is very weak.

D. The variables move in opposite directions.

ANS-B

Explanation-When the covariance is positive, the variables tend to have same relationship to their respective means:

i) **If X is above its mean $\overline{X}$ then Y is likely to be above its mean $\overline{Y}$.**
ii) **If X is below its mean $\overline{X}$ then Y is likely to be below its mean $\overline{Y}$.**

3.Which formula uses the denominator (the total population size)?

A. The Population Covariance formula.

B. Both Sample and Population formulas.

C. The Correlation Coefficient formula.

D. The Sample Covariance formula.

ANS-A

Explanation- The total population size N in the denominator is because it is calculating the true average (the arithmetic mean) of the product of the deviations for the *entire* data set.

4. Which of these pairs would most likely have a Negative Covariance?

A. Years of Education and Annual Salary.

B. Number of Pets Owned and Favorite Color.

C. Height and Weight of Adults.

D. Temperature and Heating Costs.

ANS- D

Explanation- We know that Negative Covariance indicates an inverse relationship, meaning as one variable increases, the other tends to decrease.

> **As the Temperature (outside) increases, the need to run the heater decreases, leading to lower Heating Costs. They move in opposite directions.**

5. If you are using the Population Covariance formula, what is the correct notation for the mean of ?

A. μ_X

B. X_i

C. n

D. None

ANS- A

Explanation- It is the standard symbol used in statistics to represent the mean (average) of a variable X when you are dealing with the entire population.