

1. Define Covariance and explain how it differs from Correlation in terms of scale and interpretation.

Covariance

Covariance is a statistical measure that indicates the direction of the linear relationship between two variables.

- If covariance is positive → both variables tend to increase or decrease together
- If covariance is negative → one variable increases while the other decreases
- If covariance is zero → no linear relationship

Formula:

$$\text{"Cov"}(X, Y) = (\sum (X_i - \bar{X})(Y_i - \bar{Y})) / (n - 1)$$

Difference Between Covariance and Correlation

Aspect	Covariance	Correlation
Scale	Depends on the units of the variables	Unit-free (standardized)
Range	No fixed range	Always between -1 and +1
Interpretation	Shows only direction of relationship	Shows both direction and strength
Comparability	Cannot be easily compared across datasets	Easy to compare across datasets
Sensitivity to units	Affected by change in units	Not affected by units

Key Interpretation Difference

- Covariance tells *whether* two variables move together or in opposite directions.
- Correlation tells *how strongly* they move together, making it more meaningful for analysis.

2. What does a positive, negative, and zero covariance indicate about the relationship between two variables?

Positive covariance:

Indicates that the two variables tend to move in the same direction. When one variable increases, the other also tends to increase (and when one decreases, the other tends to decrease).

Negative covariance:

Indicates that the two variables tend to move in opposite directions. When one variable increases, the other tends to decrease, and vice versa.

Zero covariance:

Indicates no linear relationship between the two variables. Changes in one variable do not show a consistent linear pattern with changes in the other (though a non-linear relationship may still exist).

3. Discuss the limitations of covariance as a measure of relationship between two variables. Why is correlation preferred in many cases?

Limitations of Covariance

1. Unit-dependent (Scale problem)
Covariance depends on the units of measurement of the variables (e.g., kg, meters).
→ This makes its numerical value hard to interpret and compare across different datasets.
2. No standardized range
Covariance can take any value from $-\infty$ to $+\infty$, so there is no fixed scale to judge whether the relationship is weak or strong.
3. Difficult interpretation of magnitude
While the sign (positive or negative) shows direction, the magnitude has no intuitive meaning.
4. Not suitable for comparison
Covariances from different variable pairs cannot be directly compared because they may be measured in different units.
5. Sensitive to outliers
Extreme values can disproportionately affect covariance, leading to misleading conclusions.

Why Correlation Is Preferred

1. Unit-free and standardized
Correlation removes the effect of units by standardizing covariance.
2. Fixed range (-1 to $+1$)
Makes it easy to interpret the strength and direction of the relationship.
3. Easy comparison
Correlation values can be directly compared across different datasets.
4. Clear interpretation
Values close to ± 1 indicate strong relationships, while values near 0 indicate weak relationships.

4. Explain the difference between Pearson's correlation coefficient and Spearman's rank correlation coefficient. When would you prefer to use Spearman's correlation?

Pearson's Correlation Coefficient vs. Spearman's Rank Correlation Coefficient

Aspect	Pearson's Correlation (r)	Spearman's Rank Correlation (ρ or r_s)
Type of relationship	Measures linear relationship	Measures monotonic (increasing or decreasing) relationship

Aspect	Pearson's Correlation (r)	Spearman's Rank Correlation (ρ or r_s)
Data requirement	Continuous, interval/ratio data	Ordinal, ranked, or continuous data
Based on	Actual data values	Ranks of the data
Distribution assumption	Assumes approximately normal distribution	No normality assumption
Sensitivity to outliers	Highly sensitive	Less sensitive
Range	-1 to +1	-1 to +1

When to Prefer Spearman's Correlation

You should use Spearman's correlation when:

1. Data is ordinal or ranked (e.g., class ranks, satisfaction levels).
2. Relationship is monotonic but not linear.
3. Data contains outliers that may distort Pearson's correlation.
4. Normality assumption is violated.
5. Sample size is small and distribution is unknown.

5. If the correlation coefficient between two variables X and Y is 0.85, interpret this value in context. Can you infer causation from this value? Why or why not?

A correlation coefficient of 0.85 indicates a strong positive linear relationship between variables X and Y.

Interpretation

- As X increases, Y tends to increase as well.
- The value 0.85 is close to +1, which means the relationship is strong and consistent.
- Most of the variation in Y can be associated with changes in X (specifically, $r^2 = [0.85]^2 \approx 0.72$, so about 72% of the variation in Y is explained by X in a linear sense).

Can we infer causation?

No, causation cannot be inferred from correlation alone.

Why not?

1. Correlation does not imply causation: A strong relationship does not mean X causes Y.

2. There may be a third (confounding) variable affecting both X and Y.
3. The direction of influence may be unclear (Y could influence X).
4. The relationship may be coincidental.

6. Using the dataset below, calculate the covariance between X and Y.

Given data:

- X: 2, 4, 6, 8
- Y: 3, 7, 5, 10

Step 1: Find the means

$$X^- = (2 + 4 + 6 + 8)/4 = 5$$

$$Y^- = (3 + 7 + 5 + 10)/4 = 6.25$$

Step 2: Compute deviations and their products

X	Y	$X - X^-$	$Y - Y^-$	Product
2	3	-3	-3.25	9.75
4	7	-1	0.75	-0.75
6	5	1	-1.25	-1.25
8	10	3	3.75	11.25

Sum of products:

$$9.75 - 0.75 - 1.25 + 11.25 = 19$$

Step 3: Calculate covariance

- Sample covariance:

$$"Cov"(X, Y) = 19/(4 - 1) = 19/3 \approx 6.33$$

7. Compute the Pearson correlation coefficient between variables A and B:

Given data:

- A: 10, 20, 30, 40, 50
- B: 8, 14, 18, 24, 28

Step 1: Compute the means

$$\bar{A} = \frac{10 + 20 + 30 + 40 + 50}{5} = 30$$

$$\bar{B} = \frac{8 + 14 + 18 + 24 + 28}{5} = 18.4$$

Step 2: Compute deviations and products

A	B	$A - \bar{A}$	$B - \bar{B}$	Product
10	8	-20	-10.4	208
20	14	-10	-4.4	44
30	18	0	-0.4	0
40	24	10	5.6	56
50	28	20	9.6	192

$$\Sigma(A - \bar{A})(B - \bar{B}) = 500$$

Step 3: Compute squared deviations

$$\Sigma(A - \bar{A})^2 = 1000$$

$$\Sigma(B - \bar{B})^2 = 251.2$$

Step 4: Pearson correlation coefficient

$$r = \frac{\Sigma(A - \bar{A})(B - \bar{B})}{\sqrt{\Sigma(A - \bar{A})^2 \Sigma(B - \bar{B})^2}}$$

$$r = \frac{500}{\sqrt{1000 \times 251.2}} \approx \frac{500}{501.2} \approx 0.998$$

8. The following table shows heights (in cm) and weights (in kg) of 5 students. ^[L]_{SEP}

Find the correlation coefficient between Height and Weight

Given data:

- Height (cm): 150, 160, 165, 170, 180

- Weight (kg): 50, 55, 58, 62, 70

- Step 1: Find the means

$$\bar{X} = \frac{150+160+165+170+180}{5} = \frac{825}{5} = 165$$

$$\bar{Y} = \frac{50 + 55 + 58 + 62 + 70}{5} = \frac{295}{5} = 59$$

- Step 2: Create the calculation table

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
150	50	-15	-9	135	225	81
160	55	-5	-4	20	25	16
165	58	0	-1	0	0	1
170	62	5	3	15	25	9
180	70	15	11	165	225	121

- Step 3: Find the sums
- $\Sigma(X - \bar{X})(Y - \bar{Y}) = 335$

$$\Sigma(X - \bar{X})^2 = 500$$

$$\Sigma(Y - \bar{Y})^2 = 228$$

- Step 4: Apply Pearson's formula

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

$$r = \frac{335}{\sqrt{500 \times 228}}$$

$$r = \frac{335}{\sqrt{114000}}$$

$$r = \frac{335}{337.64}$$

$$r \approx 0.99$$

9. Given the dataset below, determine whether there is a positive or negative correlation between X and Y. (No need for exact calculation, just reasoning.)

There is a negative correlation between X and Y.

Reasoning (no calculation needed):

- As X increases from 1 to 5, Y decreases from 15 to 3.
- This opposite movement indicates a negative relationship.

Conclusion

X and Y are negatively correlated — higher values of X are associated with lower values of Y.

10. Two investment portfolios have the following returns (%) over 5 years. Compute the covariance and correlation coefficient, and interpret whether the portfolios move together.

Given data (returns in %)

Year	Portfolio A (X)	Portfolio B (Y)
1	8	6

Year	Portfolio A (X)	Portfolio B (Y)
2	10	9
3	12	11
4	9	8
5	11	10

Step 1: Calculate the means

$$\bar{X} = \frac{8 + 10 + 12 + 9 + 11}{5} = \frac{50}{5} = 10$$

$$\bar{Y} = \frac{6 + 9 + 11 + 8 + 10}{5} = \frac{44}{5} = 8.8$$

Step 2: Create the calculation table

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	Product
8	6	-2	-2.8	5.6
10	9	0	0.2	0
12	11	2	2.2	4.4
9	8	-1	-0.8	0.8
11	10	1	1.2	1.2

$$\sum(X - \bar{X})(Y - \bar{Y}) = 12$$

Step 3: Covariance

Sample covariance:

$$\text{Cov}(X, Y) = \frac{12}{5 - 1} = \frac{12}{4} = 3$$

Step 4: Correlation coefficient

First, squared deviations:

$$\sum(X - \bar{X})^2 = 4 + 0 + 4 + 1 + 1 = 10$$

$$\sum(Y - \bar{Y})^2 = 7.84 + 0.04 + 4.84 + 0.64 + 1.44 = 14.8$$

Now apply Pearson's formula:

$$r = \frac{12}{\sqrt{10 \times 14.8}}$$

$$r = \frac{12}{\sqrt{148}} = \frac{12}{12.17}$$

$$r \approx 0.99$$

Final Answers

- Covariance = 3
- Correlation coefficient ≈ 0.99

Interpretation

- The positive covariance shows that the two portfolios tend to move in the same direction.
- The very high correlation (~ 0.99) indicates a strong positive relationship.