Correlation and Covariance

Q1. Define Covariance and explain how it differs from Correlation in terms of scale and interpretation.
Answer: Covariance is a statistical measure that indicates the **direction of the linear relationship** between two random variables.

- If covariance is **positive**, the variables tend to increase or decrease together.
- If it is **negative**, one variable tends to increase when the other decreases.
- A covariance of **zero** suggests no linear relationship.

Mathematically, for two variables X and Y:

$$Cov(X,Y) = 1/n \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Difference between Covariance and Correlation

Aspect	**Covariance**	**Correlation**
Scale	Depends on the units of measurement of X and Y	Unit-free (dimensionless)
Range	Unbounded (can take any positive or negative value)	Bounded between −1 and +1
Interpretation	Indicates only the direction of the relationship	Indicates both direction and strength of the relationship
Comparability	Difficult to compare across different datasets	Easy to compare across datasets

- **Covariance tells *which way* variables move together.**
- **Correlation tells *which way* and *how strongly* they move together.**

Q2. What does a positive, negative, and zero covariance indicate about the relationship between two variables?
Answer: Covariance describes the direction of the linear relationship between two variables.

- **Positive covariance:** Indicates that the variables move in the same direction. When one variable increases, the other tends to increase as well.
- **Negative covariance:** Indicates that the variables move in opposite directions. When one variable increases, the other tends to decrease.
- **Zero covariance:** Indicates no linear relationship between the variables. Changes in one variable do not correspond to changes in the other.

Correlation and Covariance

Q3. Discuss the limitations of covariance as a measure of relationship between two variables. Why is correlation preferred in many cases?

Answer: **Limitations of covariance:**

1. **Depends on units of measurement:** The value of covariance changes with the scale or units of the variables, making it difficult to interpret or compare across datasets.
2. **No standardized range:** Covariance has no fixed minimum or maximum value, so the strength of the relationship cannot be easily judged.
3. **Only indicates direction, not strength:** While covariance shows whether the relationship is positive or negative, it does not clearly indicate how strong the relationship is.
4. **Sensitive to outliers:** Extreme values can significantly affect the covariance, leading to misleading results.

Why correlation is preferred:
Correlation overcomes these limitations by standardizing covariance. It is **unit-free**, has a **fixed range from −1 to +1**, and clearly indicates both the **direction and strength** of the linear relationship. This makes correlation easier to interpret and more useful for comparison across different variables and datasets.

Q4. Explain the difference between Pearson's correlation coefficient and Spearman's rank correlation coefficient. When would you prefer to use Spearman's correlation?

Answer: Difference between Pearson's and Spearman's correlation:

Aspect	**Pearson's Correlation (r)**	**Spearman's Rank Correlation (ρ or rs)**
Type of data	Continuous, interval, or ratio	Ordinal or ranked data (can also handle continuous)
Measures	Linear relationship between variables	Monotonic relationship (increasing or decreasing), not necessarily linear
Assumptions	Assumes normality, linearity, and homoscedasticity	Does not assume normality or linearity
Calculation	Based on actual values	Based on ranks of values
Sensitivity	Sensitive to outliers	Less sensitive to outliers

When to use Spearman's correlation:

- When data are ordinal or ranked.
- When the relationship between variables is monotonic but not linear.
- When data violate Pearson's assumptions, such as non-normal distribution or presence of outliers.

Correlation and Covariance

Example: Comparing student ranks in two different exams is better analyzed with Spearman's correlation than Pearson's.

Q5. 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?

Answer: **Interpretation of correlation coefficient (r = 0.85):**

- The value **0.85** is **positive and close to 1**, indicating a **strong positive linear relationship** between X and Y.
- This means that as X increases, Y also tends to increase, and the relationship is fairly consistent.

Causation:

- **No, we cannot infer causation** from this value.
- Correlation **only measures the strength and direction of a linear relationship**, not whether one variable causes the other to change.
- There could be **other factors (confounding variables)** influencing both X and Y, or the association could be coincidental.

Key point: "Correlation does not imply causation."

Q6. Using the dataset below, calculate the covariance between X and Y.

X	2	4	6	8
Y	3	7	5	10

Answer: **Step 1: Calculate the means of XXX and YYY**

$$\bar{X}=(2+4+6+8) / 4=20/4=5$$

$$\bar{Y}=(3+7+5+10) / 4=25/4=6.25$$

Step 2: Calculate the covariance formula

Covariance formula is:

$$Cov(X,Y)=1/n \sum_{i=1}^{n} (Xi-\bar{X})(Yi-\bar{Y})$$

Where n=4.

Correlation and Covariance

Step 3: Calculate each term $(Xi-\bar{X})(Yi-\bar{Y})$

X_i	Y_i	$X_i-\bar{X}$	$Y_i-\bar{Y}$	Product $(X_i-\bar{X})(Y_i-\bar{Y})$
2	3	2-5=-3	3 - 6.25 = -3.25	(-3) * (-3.25) = 9.75
4	7	4-5=-1	7 - 6.25 = 0.75	(-1) * 0.75 = -0.75
6	5	6-5=1	5 - 6.25 = -1.25	1 * (-1.25) = -1.25
8	10	8-5=3	10 - 6.25 = 3.75	3 * 3.75 = 11.25

Step 4: Sum these products and divide by n

$$\sum(Xi-\bar{X})(Yi-\bar{Y})=9.75-0.75-1.25+11.25=19$$

$$Cov(X,Y)=19/4=4.75$$

Final answer: covariance between X and Y = 4.75

Q7. Compute the Pearson correlation coefficient between variables A and B:

A	**10**	**20**	**30**	**40**	**50**
B	**8**	**14**	**18**	**24**	**28**

Answer: Step 1: Calculate the means $\bar{A}$ and $\bar{B}$

$$\bar{A}=(10+20+30+40+50) / 5=150 / 5 =30$$

$$\bar{B}=(8+14+18+24+28) / 5=92 / 5 =18.4$$

Step 2: Calculate the covariance numerator $\sum(Ai-\bar{A})(Bi-\bar{B})$

A_i	B_i	$A_i-\bar{A}$	$B_i-\bar{B}$	Product $(A_i-\bar{A})(B_i-\bar{B})$
10	8	10 - 30 = -20	8 - 18.4 = -10.4	$-20\times-10.4=208$
20	14	20 - 30 = -10	14 - 18.4 = -4.4	$-10\times-4.4=44$
30	18	30 - 30 = 0	18 - 18.4 = -0.4	$0\times-0.4=0$
40	24	40 - 30 = 10	24 - 18.4 = 5.6	$10\times5.6=56$
50	28	50 - 30 = 20	28 - 18.4 = 9.6	$20\times9.6=192$

Sum of products: 208+44+0+56+192=500

Calculate $\sum(Ai-\bar{A})^2$: $(-20)^2+(-10)^2+0^2+10^2+20^2=400+100+0+100+400=1000$

Calculate $\sum(Bi-\bar{B})^2$: $(-10.4)^2+(-4.4)^2+(-0.4)^2+5.6^2+9.6^2 =108.16+19.36+0.16+31.36+92.16$

$=251.2$

Correlation and Covariance

Now,

$S_A = \sqrt{1000/5} = \sqrt{200} \approx 14.142$

$S_B = \sqrt{251.2/5} = \sqrt{50.24} \approx 7.089$

Step 4: Calculate Pearson correlation coefficient

$r = (1\sum(A_i - \bar{A})(B_i - \bar{B}))/S_A * S_B$

$= (500/5) / 14.142 \times 7.089$

$= 100 / 100.261 \approx 0.997$

Answer: $r \approx 0.997$

Q8. The following table shows heights (in cm) and weights (in kg) of 5 students. Find the correlation coefficient between Height and Weight.

Hight	**150**	**160**	**165**	**170**	**180**
Waight	**50**	**55**	**58**	**62**	**70**

Answer: Step 1: Calculate mean of X and Y

$\bar{X} = (150+160+165+170+180)/5 = 165$

$\bar{Y} = (50+55+58+62+70)/5 = 59$

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	120	225	81
160	55	-5	-4	35	25	16
165	58	0	-1	0	0	1
170	62	5	3	15	25	9
180	70	15	11	165	225	121

$\sum(X-\bar{X})(Y-\bar{Y}) = 335$

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

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

Correlation and Covariance

Step 3: Apply Karl Pearson's correlation formula

$$r = \sum(X-\bar{X})(Y-\bar{Y}) / \sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}$$

$$r = 300 / \sqrt{500 \times 228335}$$

$$r = 335 / 337.64 \quad r \approx 0.99$$

Answer: $r \approx 0.99$

Q9. Given the dataset below, determine whether there is a positive or negative correlation between X and Y. (No need for exact calculation, just reasoning.)

X	1	2	3	4	5
Y	15	12	9	7	3

Answer: From the given data, as X increases (1 → 5), the corresponding Y values decrease (15 → 3).

This shows that:

- When one variable increases, the other decreases.
- Therefore, the relationship between X and Y is negative.

Final Answer: There is a negative correlation between X and Y.

Q10. Two investment portfolios have the following returns (%) over 5 years. Compute the covariance and correlation coefficient, and interpret whether the portfolios move together

Year	Portfolio A	Portfolio b
1	8	6
2	10	9
3	12	11
4	9	8
5	11	10

Answer: Step 1: Calculate the mean returns

$$\bar{X} = (8+10+12+9+11) / 5 = 10$$
$$\bar{Y} = (6+9+11+8+10) / 5 = 8.8$$

Correlation and Covariance

Step 2: Prepare calculation table

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$
8	6	-2	-2.8	5.6	4	7.84
10	9	0	0.2	0	0	0.04
12	11	2	2.2	4.4	4	4.84
9	8	-1	-0.8	0.8	1	0.64
11	10	1	1.2	1.2	1	1.44

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

$$\sum(X-\bar{X})^2=10$$

$$\sum(Y-\bar{Y})^2=14.8$$

Step 3: Covariance calculation

$$Cov(X,Y)=\sum(X-\bar{X})(Y-\bar{Y}) / n$$

$$Cov(X,Y)=12 / 5 = 2.4$$

Step 4: Correlation coefficient (Karl Pearson's r)

$$r=\sum(X-\bar{X})(Y-\bar{Y}) / \sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}$$

$$r=12/ \sqrt{10\times14.8}$$

$$r=12/\sqrt{148} \quad = \quad r \approx 0.99$$

Interpretation:

- **Covariance is positive (2.4)** → both portfolios move in the same direction.
- **The correlation coefficient is close to +1** → very strong positive relationship.
- When **Portfolio A's return increases**, **Portfolio B's return also increases**.

Final Answer:

- **Covariance = 2.4**
- **Correlation coefficient ≈ 0.99**
- The portfolios **move together** and show a **strong positive correlation**.