ASSIGNMENT TITTLE:

Correlation and Covariance Solutions

Q1. Define Covariance and explain how it differs from Correlation in terms of scale and interpretation.

Covariance is a statistical measure that describes the directional relationship between two variables (X and Y). A positive covariance indicates that X and Y tend to increase or decrease together, while a negative covariance indicates that as one increases, the other tends to decrease.

Formula: $Cov(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$ (for a sample)

Feature	Covariance	Correlation
Scale	Unbounded (Can range from $-\infy$ to $+\infy$).	Bounded (Ranges strictly from -1 to $+1$).
Interpretation	Difficult to interpret the strength of the relationship,as the magnitude depends on the original units of measurement.	Easy to interpret the strength and direction. A value closer to ± 1 indicates a stronger relationship.
Unit	Has units (e.g., if X is in meters and Y is in kilograms, covariance is in meter-kilograms).	Unitless (A standardized measure).

Q2. What does a positive, negative, and zero covariance indicate about the relationship between two variables?

Positive Covariance ($Cov > 0$): Indicates a positive linear relationship. When the values of one variable (X) are above their mean, the values of the other variables (Y) tend to be above their mean as well. The variables move in the same direction.

Negative Covariance ($Cov < 0$): Indicates a negative (inverse) linear relationship. When the values of X are above their mean, the values of Y tends to be below their mean. The variables move in opposite Directions.

Zero Covariance ($Cov \approx 0$): Indicates that there is no linear relationship between the variables. The values of X and Y are independent of each other in a linear sense.

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

The main limitation is the scale-dependency and lack of standardisation.
Scale Dependence: The magnitude of covariance is directly affected by the units of the variables. If you change the units (e.g., from meters to centimetres), the covariance value will change dramatically, even though The relationship between the variables remains the same.

Lack of Standardization: Because it is not bounded, it's impossible to tell if a covariance of, say, 100 indicates a weak or strong relationship without knowing the variables' standard deviations.

Why Correlation is Preferred
Correlation (specifically the Pearson correlation coefficient, ρ or r) is

preferred because it standardizes the measure of the relationship:
Standardisation: Correlation is calculated by dividing the covariance by the product of the standard deviations of the two variables, effectively removing the units.

$$r = \frac{Cov(X, Y)}{s_X s_Y}$$
Ease of Interpretation: The resulting value is always between -1 and $+1$, allowing for immediate and universal interpretation of the strength and direction of the linear relationship, regardless of the variables' original units.

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

Feature	Pearson's Correlation Coefficient (r)	Spearman's Rank Correlation Coefficient (ρ or rs)
Measures	Linear relationship between the actual data values.	Monotonic relationship (consistency of direction)between the ranks of the data.
Assumptions	Assumes data is approximately normally distributed and the relationship is linear (Parametric test).	Makes no assumptions about the distribution of the data (Non-parametric test).
Data Type	Interval or Ratio scale data.	Ordinal, Interval, or Rato scale data.

When to Prefer Spearman's Correlation:
You would prefer to use Spearman's rank correlation when:

1. The relationship is non-linear but monotonic: The variables consistently move in the same direction, but not at a constant rate (e.g., Y increases slowly, then quickly, as X increases).

2. The data is ordinal: The data consists of ranks (e.g., quality ratings, finish order in a race).

3. The data contains outliers: Pearson's r is highly sensitive to outliers, while Spearman's r_s is more robust since it uses ranks, which reduces the impact of extreme values.

4. Assumptions for Pearson's r are violated (e.g., data is not normally distributed).

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?**

Interpretation of $r = 0.85$
A correlation coefficient of $r = 0.85$ indicates a strong, positive linear relationship between variables X and Y.
Positive: As the values of X increase, the values of Y also tend to increase.
Strong: The data points cluster closely around a straight line, suggesting that the linear model is a good fit for the relationship.
Can you infer causation?
No, you cannot infer causation from a correlation value alone.

Why not? Correlation only measures the degree to which two variables are linearly associated (they move together). It does not tell us *why* they are associated.

The relationship could be due to a third, unobserved variable (a confounding variable) that influences both X and Y.
This concept is summarized by the phrase: "Correlation does not imply causation."

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

X	2	4	6	8
Y	7	5	10	3

$n = 4$.

1. Calculate Means ($\bar{X}$ and $\bar{Y}$):

$\sum X = 2 + 4 + 6 + 8 = 20 \implies \bar{X} = 20 / 4 = 5$

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

2. Calculate the Sum of Products of Deviations ($\sum (X_i - \bar{X})(Y_i - \bar{Y})$):

Xi	**Yi**	**Xi−X¯**	**Yi−Y¯**	**(Xi−X¯)(Yi−Y¯)**
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**

3. Calculate Covariance ($Cov(X, Y)$):

$$Cov(X, Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{19.00}{4-1} = \frac{19}{3} \approx **6.333$$

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

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

$n = 5$.
Notice that the values in B are approximately 0.6 times the values in A (e.g., $10 \times 0.6 + 2 \approx 8$, $50 \times 0.6 - 2 \approx 28$). When a relationship is perfectly linear, the Pearson correlation coefficient is $+1$ or -1.
Let's check if B is a perfect linear transformation of A. Let $B = 0.5A + 3$:

- $10 \ times 0.5 + 3 = 8$ (Correct)
- $20 \ times 0.5 + 3 = 13$ (Close to 14)
- $30 \ times 0.5 + 3 = 18$ (Correct)
- $40 \ times 0.5 + 3 = 23$ (Close to 24)
- $50 \ times 0.5 + 3 = 28$ (Correct)

Let's check $B = 0.6A + 2$:

- 10 \ times 0.6 + 2 = 8$ (Correct)
- $20 \ times 0.6 + 2 = 14$ (Correct)
- $30 \ times 0.6 + 2 = 20$ (Close to 18)
- $40 \ times 0.6 + 2 = 26$ (Close to 24)
- $50 \ times 0.6 + 2 = 32$ (Close to 28)

Since the values of B are not a perfect linear transformation of A, we must calculate r using the full formula:
$$r = \frac{\sum (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum (A_i - \bar{A})^2

\sum (B_i - \bar{B})^2}}$$

1.Calculate Means:

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

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

2. Calculate Deviations and Products:

Ai	Bi	Ai−A¯	Bi−B¯	(Ai−A¯)2	Bi−B¯)2	(Ai−A¯)(Bi−B¯)
10	8	-20	-10.4	400	108.16	208.0
20	14	-10	-4.4	100	19.36	44.0
30	18	0	-0.4	0	0.16	0.0
40	24	10	5.6	100	31.36	56.0
50	28	20	9.6	400	92.16	192.0
Sum				$\mathbf{1000}$	$\mathbf{251.2}$	$\mathbf{500.0}$

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

This also requires the Pearson correlation coefficient (r). $n=5$.

1. Calculate Means:

$\sum H = 150 + 160 + 165 + 170 + 180 = 825 \implies \bar{H} = 825 / 5 = 165$

$\sum W = 50 + 55 + 58 + 62 + 70 = 295 \implies \bar{W} = 295 / 5 = 59$

2. Calculate Deviations and Products:

Hi	Wi	Hi-H¯	Wi−W¯	(Hi−H¯2	(Wi−W)2	(Hi−H¯)(−W¯)
150	50	-15	-9	225	81	135
160	55	-5	-4	25	16	20
165	58	0	-1	0	1	0
170	62	5	3	25	9	15
180	70	15	11	225	121	165
Sum				$\mathbf{50	$\mathbf{22	$\mathbf{33

3. Calculate r:

$$r = \frac{335}{\sqrt{500 \tmes 228}} = \frac{335}{\sqrt{114000}} = \frac{335}{337.638} \approx **0.9922$$

There is an extremely strong positive linear correlation between the students' Height and Weight.

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

Reasoning:

1. Analyze the trend in X: The values of X are strictly increasing (1, 2, 3,4,5).

2.Analyze the trend in Y: The corresponding values of Y are strictly decreasing (15, 12, 9, 7, 3).

3.Conclusion: As one variable (X) increases, the other variable (Y) consistently decreases. This indicates that the variables move in opposite directions.

Therefore, there is a negative correlation between X and Y.

Q10. Two investment portfolios having 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

This problem requires calculating the sample covariance and Pearson correlation coefficient (r) for the returns of Portfolio A (A) and Portfolio B (B) over $n=5$ years.

1. Calculate the Means ($\bar{A}$ and $\bar{B}$)

Year	Portfolio A (Ai)	Portfolio B (Bi)
1	8	6
2	10	9
3	12	11
4	9	8
5	11	10
um ($\sum$)	50	44

Ai	Bi	Ai−A¯	Bi−B¯	(Ai−A¯)2	(Bi−B¯)2	Ai−A¯)(Bi−B¯)
8	6	-2	-2.8	4	7.84	5.6
10	9	0	0.2	0	0.04	0.0
12	11	2	2.2	4	4.84	4.4
9	8	-1	0.8	1	0.64	0.8
11	10	1	1.2	1	1.44	1.2
Sum				$\mathbf{10}$	$\mathbf{14.8}$	$\mathbf{12.0}$

$$\bar{A} = \frac{\sum A_i}{n} = \frac{50}{5} = \mathbf{10.0}

$$$$\bar{B} = \frac{\sum B_i}{n} = \frac{44}{5} = \mathbf{8.8}$$

2. Calculate Deviations and Products

We need the sum of the products of deviations ($\sum (A_i - \bar{A})(B_i - \bar{B})$) for the covariance and the sum of squared deviations for the correlation coefficient.

3. Compute Covariance

The formula for sample covariance is:

$$Cov(A, B) = \frac{\sum (A_i - \bar{A})(B_i - \bar{B})}{n-1}$$

$$Cov(A, B) = \frac{12.0}{5-1} = \frac{12.0}{4} = \mathbf{3.0}$$

4. Compute Correlation Coefficient (r)

The formula for the Pearson correlation coefficient is:

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

$$r = \frac{12.0}{\sqrt{10 \tmes 14.8}} = \frac{12.0}{\sqrt{148}}

$$$$r = \frac{12.0}{12.1655} \approx \mathbf{0.9864}$$

5. Interpretation

Covariance (3.0): Since the covariance is positive (> 0), it indicates that when the return of Portfolio A is above its average, the return of Portfolio B tends to be above its average as well. The portfolios tend to move in the same direction.

Correlaton Coefcient (0.9864): Since the correlaton coefcient is very close to $+1$, it indicates an extremely strong, positive linear relationship between the returns of the two portfolios.

Conclusion: Both the covariance and correlation indicate that the two portfolios move closely together and in the same direction. From an investment perspective, this means they offer very little diversification benefit.
