Assignment 22 March

Q1. Pearson correlation coefficient is a measure of the linear relationship between two variables. Suppose you have collected data on the amount of time students spend studying for an exam and their final exam scores. Calculate the Pearson correlation coefficient between these two variables and interpret the result.

Formula:

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Where:

- $\text{Cov}(X, Y)$: Covariance between X and Y.
- σ_X, σ_Y: Standard deviations of X and Y.

Example Dataset:

Study Time (hours)	Exam Score (%)
2	60
3	70
4	80
5	85
6	90

Calculation Steps:

1. Calculate the mean of both variables.

$$\bar{X} = \frac{2+3+4+5+6}{5} = 4$$

$$\bar{Y} = \frac{60+70+80+85+90}{5} = 77$$

2. Compute the covariance:

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

3. Compute the standard deviations of X and Y.

4. Calculate r using the formula.

Interpretation:

- $r = 1$: Perfect positive linear relationship.
- $r = -1$: Perfect negative linear relationship.
- $r = 0$: No linear relationship.

Q2. Spearman's rank correlation is a measure of the monotonic relationship between two variables.

Example Dataset:

Sleep Hours	**Job Satisfaction (1-10)**
6	8
7	9
5	7
8	10
4	6

Steps:

1. Rank the data for both variables.
2. Calculate the difference d in ranks for each pair.
3. Use the Spearman correlation formula:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

4. Interpret the result:
   - $r_s = 1$: Perfect monotonic relationship.
   - $r_s = 0$: No monotonic relationship.

Q3. Compare Pearson and Spearman correlations for exercise hours and BMI.

Dataset Example:

Exercise Hours	**BMI**
1	30
2	28
3	25
4	23
5	22

Calculation:

1. **Pearson Correlation**:
   - Use the covariance and standard deviations.
   - r_p measures linear relationships.
2. **Spearman Correlation**:
   - Rank the data and apply the Spearman formula.
   - r_s measures monotonic relationships.

Comparison:

- Pearson is sensitive to linear trends.
- Spearman captures monotonic but not necessarily linear relationships.

Q4. Calculate Pearson correlation for TV watching hours and physical activity levels.

Dataset Example:

TV Hours	Physical Activity (hours)
1	5
2	4
3	3
4	2
5	1

Steps:

1. Calculate means of both variables.
2. Use the Pearson formula.
3. Interpret:
   - Negative r: Watching more TV correlates with less physical activity.

Q5. Survey on age and soft drink preference.

Data:

Age	Preference

25	Coke
42	Pepsi
37	(Missing)
19	Mountain Dew
31	Coke
28	Pepsi

Steps:

1. Encode soft drink preferences (e.g., Coke = 1, Pepsi = 2, Mountain Dew = 3).
2. Compute Pearson correlation for Age and Preference.

Q6. Relationship between sales calls per day and sales per week.

Dataset Example:

Sales Calls (per day)	**Sales (per week)**
10	5
15	8
20	10
25	15
30	20

Steps:

1. Compute the covariance of sales calls and sales.
2. Compute the standard deviations of both variables.
3. Calculate the Pearson correlation coefficient.
4. Interpret:
    - r>0: Positive relationship (more calls lead to more sales).

Python Implementation:

For all calculations, use the following Python code:

import pandas as pd

```
import numpy as np

# Example dataset
data = pd.DataFrame({
    'X': [2, 3, 4, 5, 6],
    'Y': [60, 70, 80, 85, 90]
})

# Pearson Correlation
pearson_corr = data.corr(method='pearson')

# Spearman Correlation
spearman_corr = data.corr(method='spearman')

print("Pearson Correlation:\n", pearson_corr)
print("Spearman Correlation:\n", spearman_corr)
```