Introduction to Stochastics
Exercise 6

Part 1: Variance and Standard Deviation

Dataset A

Student	Hours Studied (X)	Test Score (Y)
A	2	65
B	4	70
C	6	75
D	8	85
E	10	95

Tasks

- Compute the variance and standard deviation for both variables:

$$\text{Variance:} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \text{Standard Deviation:} \quad S = \sqrt{S^2}$$

- Compute the (population) covariance of both variables:

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

Part 2: Pearson Correlation Coefficient

0.1 Short introduction to the equation

Definition of the pearson correlation coefficient via covariance and standard deviations:

$$r = \frac{\mathrm{Cov}(X,Y)}{S_x \cdot S_y}$$

Where the covariance is defined as:

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

Standard deviations:

$$S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad S_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Putting it all together:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n}\sum(y_i - \bar{y})^2}}$$

Simplified (used in this exercise):

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Tasks

- Using the data from Dataset A, manually calculate the Pearson correlation coefficient.
- Interpret the value of r:
  - Is the correlation weak, moderate, or strong?
  - Is it positive or negative?
  - What does it imply about the relationship between hours studied and test scores?

Part 3: Interpretation of Correlation Values

Below are five values for Pearson's r. Interpret the strength and direction of each.

Value of r
-0.95
0.25
0.00
0.72
-0.40

Tasks

For each value:

- Indicate the strength (none, small, medium, strong)
- Indicate the direction (positive or negative)
- Give a brief interpretation (e.g., as variable X increases, what happens to Y?)

Part 4: Spearman's Rank Correlation

Dataset B

The dataset contains the ratings of two movie critics. Calculate Spearman's rank correlation!

Movie	**Critic A**	**Critic B**	**Rank A**	**Rank B**	d	d^2
M1	1	4				
M2	9	8				
M3	3	2				
M4	7	7				
M5	10	9				

Tasks

- Compute the difference in ranks d and d^2 for each movie.
- Use the Spearman rank correlation formula, that only applies when there aren't any ties:

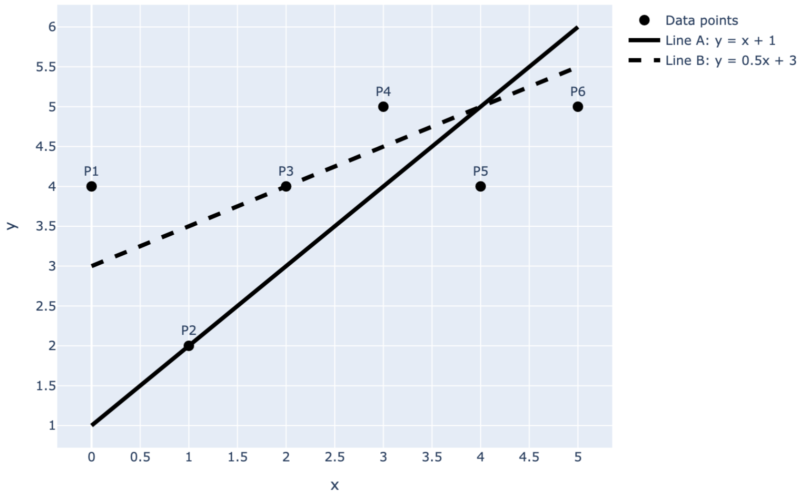
  $$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

  where n is the number of data points.
- Interpret the value of ρ.

Part 5: Linear Regression

Task 1: Model Comparison Using Sum of Squared Residuals (SSR)

Below you see a plot with data points and two different regression lines (Line A and Line B).



Tasks:

- For each line (A and B), calculate the Sum of Squared Residuals (SSR):

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  where $\hat{y}_i$ is the predicted value for x_i on the line.

- Based on the SSR values, determine which line fits the data better.
- Briefly explain why SSR is a good metric for model quality.

Task 2: Compute the Linear Regression Line Step-by-Step

Use the dataset below to manually compute the regression line. Use the formulas provided and fill in the helper table.

X **(Hours Studied)**	Y **(Test Score)**
1	45
2	65
3	70
4	80

Step 1: Compute the following values

- Mean (average):

$$\bar{x} = \frac{1}{n}\sum x_i, \qquad \bar{y} = \frac{1}{n}\sum y_i$$

- Standard deviation:

$$S_x = \sqrt{\frac{1}{n}\sum (x_i - \bar{x})^2}, \qquad S_y = \sqrt{\frac{1}{n}\sum (y_i - \bar{y})^2}$$

- Pearson correlation coefficient:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \cdot S_x \cdot S_y}$$

Helper Table:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$

Step 2: Compute the regression coefficients

- Slope:

$$b_1 = r \cdot \left(\frac{S_y}{S_x}\right)$$

- Intercept:

$$b_0 = \bar{y} - b_1 \cdot \bar{x}$$

- Final regression equation:

$$\hat{y} = b_0 + b_1 x$$

Step 3: Application

Use the regression equation to predict the test score of a student who studied for **5 hours**.