Covariance and Correlation: Comprehensive Notes

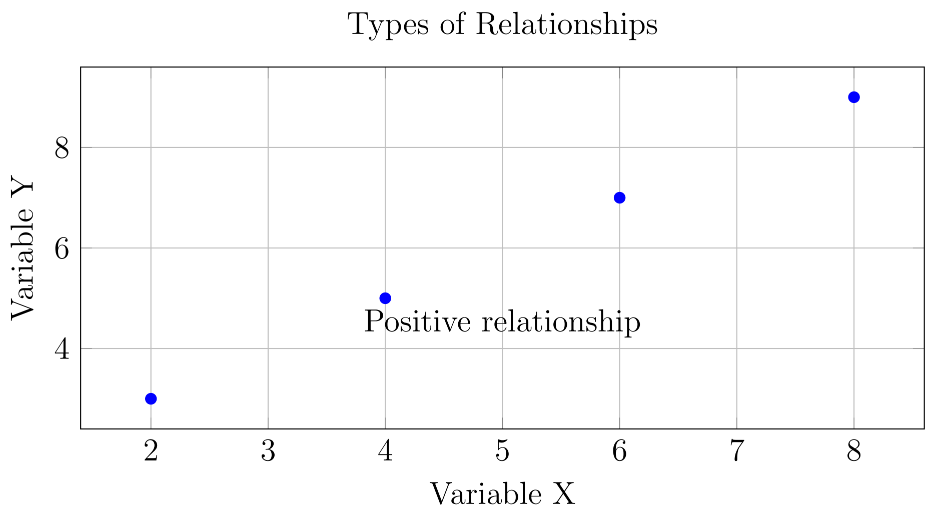
Your Name

June 16, 2025

1 Introduction

Covariance and correlation measure the relationship between two **continuous variables**. Both quantify:

- How changes in one variable relate to changes in another
- The direction (positive/negative) and strength of association



2 Covariance

2.1 Definition

Covariance measures how two random variables change together:

- **Positive**: Both increase or decrease together ($X \uparrow, Y \uparrow$ or $X \downarrow, Y \downarrow$)
- **Negative**: One increases when the other decreases ($X \uparrow, Y \downarrow$ or $X \downarrow, Y \uparrow$)

2.2 Formula

The sample covariance formula:

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

Where:

- n = number of data points
- $\bar{x}, \bar{y}$ = sample means

2.3 Calculation Example

Hours studied (X) vs exam scores (Y):

X (hours)	Y (scores)
2	50
3	60
4	70
5	80
6	90

Steps: 1. Calculate means: $\bar{x} = 4$, $\bar{y} = 70$
2. Compute covariance:

$$\begin{aligned}\mathrm{cov}(X,Y) &= \frac{1}{4}[(2-4)(50-70) + (3-4)(60-70) \\ &\quad + (4-4)(70-70) + (5-4)(80-70) + (6-4)(90-70)] \\ &= \frac{1}{4}[(-2)(-20) + (-1)(-10) + (0)(0) + (1)(10) + (2)(20)] \\ &= \frac{1}{4}[40 + 10 + 0 + 10 + 40] = \frac{100}{4} = 25\end{aligned}$$

Interpretation: Positive covariance (25) confirms that as study hours increase, exam scores increase.

2.4 Properties

Advantages:

- Quantifies direction of relationship
- Foundation for correlation calculations

Disadvantages:

- No standardized range (values from $-\infty$ to $+\infty$)
- Cannot compare strength across different datasets
- Sensitive to measurement units

3 Correlation

3.1 Pearson Correlation Coefficient

Pearson correlation (ρ) standardizes covariance to range [-1, 1]:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

Where σ_X, σ_Y are standard deviations.

figurePearson correlation examples (Source: Course Video)

3.1.1 Interpretation

- $\rho = 1$: Perfect positive linear relationship
- $\rho = -1$: Perfect negative linear relationship
- $\rho = 0$: No linear relationship
- $|\rho| > 0.7$: Strong correlation

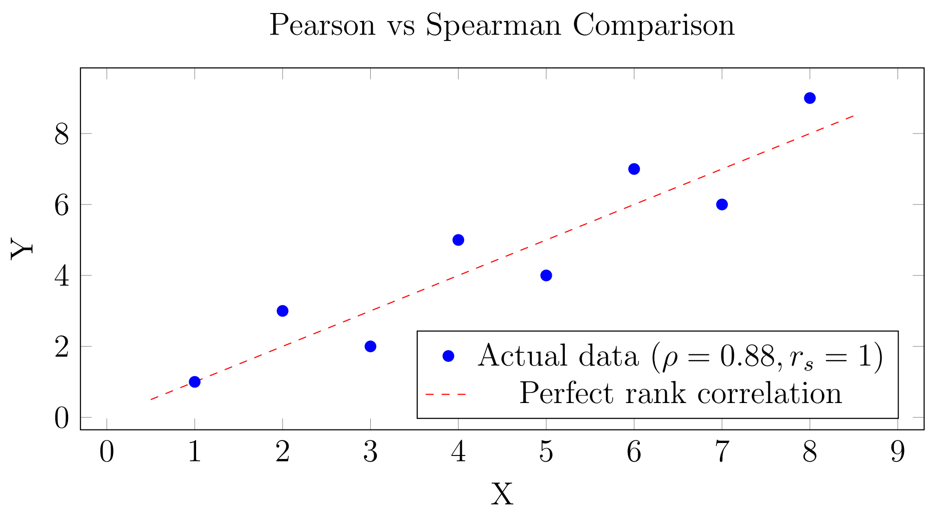
3.2 Spearman Rank Correlation

Spearman correlation (r_s) measures monotonic relationships (linear or non-linear) using **ranks**:

$$r_s = \frac{\text{cov}(\text{rank}(X), \text{rank}(Y))}{\sigma_{\text{rank}(X)} \sigma_{\text{rank}(Y)}}$$

Advantages over Pearson:

- Handles non-linear relationships
- Robust to outliers
- Works with ordinal data



4 Applications in Data Science

4.1 Feature Selection

Correlation helps identify relevant features for predictive modeling:

- High $|\rho|$ with target variable = important feature
- Near-zero ρ = potential feature removal

House price prediction:

Feature	Correlation with Price	Decision
Size	$\rho = 0.85$	Keep
Number of rooms	$\rho = 0.78$	Keep
Haunted status	$\rho = -0.65$	Keep
Residents count	$\rho = 0.05$	Remove

4.2 Key Differences Summary

Covariance	**Correlation**
Measures direction only	Measures direction and strength
Range: $-\infty$ to $+\infty$	Range: -1 to 1
Unit-dependent	Unitless
Not scalable	Scalable for comparison

5 Conclusion

- **Covariance**: Direction of linear relationship (positive/negative)
- **Pearson**: Strength of *linear* relationships (range -1 to 1)
- **Spearman**: Strength of *monotonic* relationships (handles non-linearity)
- Applications: Feature selection, EDA, dimensionality reduction