Lecture Notes 6: Correlation Fundamentals & Statistical Test Selection

Dr. Ratnesh Srivastava, CSIT, Guru Ghasidas Viswavidyalaya, Bilaspur

19.07.2025

Contents

1 Theoretical Foundations

1.1 Covariance: Direction of Relationship

Measures joint variability between two variables:

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

Interpretation:

Positive: Variables move in same direction **Negative**: Variables move in opposite directions **Zero**: No linear relationship (but possible non-linear)

1.2 Pearson Correlation: Strength & Direction

Standardizes covariance to [-1, 1] range:

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

Interpretation:

$|r| > 0.7$: Strong relationship $0.3 < |r| < 0.7$: Moderate relationship $|r| < 0.3$: Weak relationship

1.3 Parametric vs. Non-Parametric Tests

Key Idea: Assumptions About Population Distribution

The fundamental difference lies in assumptions about the underlying population distribution:

Parametric Tests: Assume specific distribution (usually normal) **Non-Parametric Tests**: Make no distributional assumptions

Table 1: Core Differences Between Test Types

Feature	**Parametric Tests**	**Non-Parametric Tests**
Distribution assumption	Normal distribution	None
Data requirements	Interval/Ratio scale	Ordinal/Nominal okay
Handles skewed data?	Poor	Excellent
Power when assumptions met	High	Lower
Power when assumptions violated	Low	High
Example tests	t-test, ANOVA, Pearson	Mann-Whitney, Spearman

Illustrative Example: Drug Effectiveness

Compare two groups:

Group A (Drug A): [5, 6, 7, 8, 9]   Group B (Drug B): [2, 3, 4, 5, 6]

Parametric Approach (t-test):

$$\bar{x}_A = 7, \quad \bar{x}_B = 4$$
$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \approx 3.67$$
$$p < 0.05 \Rightarrow \text{significant difference}$$

Caution: Requires normality and equal variance

Non-Parametric Approach (Mann-Whitney U):

Pool and rank data: [2(1), 3(2), 4(3), 5(4.5), 5(4.5), 6(6.5), 6(6.5), 7(8), 8(9), 9(10)] Sum ranks: Group A = 4.5+6.5+8+9+10 = 38, Group B = 1+2+3+4.5+6.5=17 $U = \min(U_A, U_B) = \min(38 - \frac{5(5+1)}{2}, 17 - \frac{5(5+1)}{2}) = \min(23, 2) = 2$ Critical U (5,5) = 2 at $\alpha = 0.05 \Rightarrow$ significant difference

Advantage: No distributional assumptions

1.4 Specialized Correlation Measures

Spearman (ρ): For monotonic relationships (non-linear) **What is a Monotonic Relationship?** A monotonic relationship describes a consistent directional association between two variables. This means that as one variable increases, the other variable either consistently increases (positive monotonic) or consistently decreases (negative monotonic). Unlike a linear relationship, the rate of change does not have to be constant. For example, if studying gets harder as you get closer to an exam, your stress might increase, but not necessarily at a steady rate.

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

Kendall's Tau (τ): For ordinal data with ties

$$\tau = \frac{C - D}{\binom{n}{2}}$$

Point-Biserial (r_{pb}): Binary + Continuous

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{s} \cdot \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

Phi (ϕ): Binary + Binary

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Partial Correlation ($r_{XY.Z}$): Controlling for confounders

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

2 Applied Examples with Student Dataset

2.1 Data Summary

Student	Math (X)	Science (Y)	Rank	CP	Gender	Scholarship (S)	Sports
A	85	88	2	9	M	1	1
B	78	75	5	7	F	0	0
C	92	95	1	10	M	1	0
D	70	72	7	6	F	0	1
E	88	85	3	8	M	1	0
F	65	67	8	5	F	0	1
G	80	82	4	8	M	1	1
H	72	70	6	6	F	0	1

Key: CP = Class Participation (1-10 scale)

2.2 Pearson Correlation: Math vs Science

$$\begin{aligned}
\bar{x} &= 78.75, \quad \bar{y} = 79.25 \\
r &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}} \\
&= \frac{54.69 + \cdots}{\sqrt{(37.56 + \cdots)(43.56 + \cdots)}} \approx 0.99
\end{aligned}$$

Interpretation: Near-perfect positive linear relationship

2.3 Spearman Rank Correlation: Rank vs CP

$$\begin{aligned}
d_i &= \text{Rank}_{CP} - \text{Rank}_{Academic} \\
\sum d_i^2 &= 20 \\
\rho &= 1 - \frac{6 \times 20}{8(64 - 1)} = 1 - \frac{120}{504} \approx 0.76
\end{aligned}$$

Interpretation: Strong monotonic relationship

2.4 Kendall's Tau: Rank vs CP

$$\begin{aligned} C &= 22 \text{ (concordant pairs)} \\ D &= 6 \text{ (discordant pairs)} \\ \tau &= \frac{22-6}{\binom{8}{2}} = \frac{16}{28} \approx 0.57 \end{aligned}$$

Interpretation: Moderate rank agreement

2.5 Point-Biserial: Math vs Scholarship

$$\begin{aligned} \bar{X}_1 &= 86.25, \quad \bar{X}_0 = 71.25 \\ s &\approx 9.75 \\ r_{pb} &= \frac{86.25 - 71.25}{9.75} \cdot \sqrt{\frac{4 \times 4}{8 \times 7}} \approx 0.77 \end{aligned}$$

Interpretation: Scholarship students score significantly higher

2.6 Phi Coefficient: Scholarship vs Sports

	Sports=1	Sports=0
Scholarship=1	2	2
Scholarship=0	3	1

$$\begin{aligned} \phi &= \frac{(2)(1) - (2)(3)}{\sqrt{4 \times 5 \times 3 \times 4}} \\ &= \frac{-4}{\sqrt{240}} \approx -0.26 \end{aligned}$$

Interpretation: Weak negative association

2.7 Partial Correlation: Math vs Science | CP

$$\begin{aligned} r_{XY} &= 0.99, \quad r_{XZ} = 0.89, \quad r_{YZ} = 0.88 \\ r_{XY.Z} &= \frac{0.99 - (0.89)(0.88)}{\sqrt{(1 - 0.89^2)(1 - 0.88^2)}} \\ &\approx \frac{0.2068}{0.216} \approx 0.96 \end{aligned}$$

Interpretation: Strong relationship persists after controlling for participation

3 Test Selection Guide

3.1 When to Use Parametric Tests

Data is normally distributed (check with Shapiro-Wilk or Q-Q plot) Interval/ratio scale data Homogeneity of variance (for group comparisons) Sample size ¿ 30 (Central Limit Theorem applies) Examples: Pearson correlation, t-test, ANOVA

3.2 When to Use Non-Parametric Tests

Data is skewed or has outliers Ordinal data (e.g., Likert scales, rankings) Small sample sizes (n ¡ 30) Violated parametric assumptions Examples: Spearman/Kendall correlation, Mann-Whitney, Wilcoxon

3.3 Decision Framework for Correlation

Table 2: Correlation Method Selection Guide

Data Type 1	**Data Type 2**	**Recommended Method**
Continuous	Continuous	Pearson (if linear/normal)
Ordinal	Ordinal	Spearman or Kendall
Binary	Continuous	Point-Biserial
Binary	Binary	Phi Coefficient
Continuous	Continuous (with control variable)	Partial correlation (controls confounders)

4 Summary & Interpretation Guide

Method	**Variable Pair**	**Value**	**Interpretation**
Pearson	Math vs Science	0.99	Strong positive linear
Spearman	Rank vs CP	0.76	Strong monotonic
Kendall's Tau	Rank vs CP	0.57	Moderate agreement
Point-Biserial	Math vs Scholarship	0.77	Large effect size
Phi	Scholarship vs Sports	-0.26	Weak negative
Partial Corr	Math vs Science | CP	0.96	Strong controlled relationship

When to Use Each Measure

Measure	When to Use
Pearson	Continuous + continuous variables, for linear relationships.
Spearman	Ordinal data or for assessing monotonic (non-linear but consistently directional) relationships.
Kendall's Tau	Small samples with tied ranks, for ordinal data.
Point-Biserial	When one variable is binary (dichotomous) and the other is continuous.
Phi	When both variables are binary (dichotomous).
Partial	When you need to control for the influence of one or more confounding variables on the relationship between two others.

Practical Considerations

Always visualize relationships first (scatterplots) Check assumptions (linearity, normality for Pearson) Correlation $\neq$ causation (especially in observational data) Effect size matters more than statistical significance

Final Thought: Measurement Analogy

Parametric tests are like precise laser measures - highly accurate in ideal conditions but fail with irregular surfaces **Non-parametric tests** are like flexible tape measures - work in diverse conditions but with slightly less precision

Classroom Activity: Test Selection Challenge

Scenario 1: Compare salaries (skewed) between graduates of 3 universities
Solution: Kruskal-Wallis (non-parametric alternative to ANOVA)

2. **Scenario 2**: Examine relationship between temperature (continuous) and ice cream sales (continuous) with normal distribution
Solution: Pearson correlation

3. **Scenario 3**: Assess agreement between two judges' rankings of 15 contestants
Solution: Kendall's Tau (handles ties better than Spearman)

5 Questions and Answers

Question 1

What is the primary difference in interpretation between a positive covariance and a negative covariance?

Answer 1

A **positive covariance** indicates that two variables tend to move in the same direction; as one increases, the other tends to increase, and vice versa. A **negative covariance** indicates that two variables tend to move in opposite directions; as one increases, the other tends to decrease, and vice versa.

Question 2

Why is Pearson correlation often preferred over covariance when comparing the strength of relationships between different pairs of variables?

Answer 2

Pearson correlation is preferred because it standardizes the covariance by dividing it by the product of the standard deviations of the two variables. This standardization results in a coefficient that ranges from -1 to 1, making it a unitless measure of the strength and direction of a linear relationship. Covariance, on the other hand, is not standardized and its magnitude depends on the units of the variables, making it difficult to compare across different datasets or variable pairs.

Question 3

You are analyzing the relationship between a student's "satisfaction level" (measured on an ordinal scale from 1 to 5) and their "ranking in a debate competition." Which correlation coefficient would be most appropriate to use, and why?

Answer 3

For analyzing the relationship between a student's "satisfaction level" (ordinal data) and their "ranking in a debate competition" (ordinal data), the **Spearman Rank Correlation** (ρ) or **Kendall's Tau** (τ) would be most appropriate. This is because both are non-parametric measures designed for ordinal data or for assessing monotonic (not necessarily linear) relationships. Kendall's Tau is often preferred for smaller sample sizes or when there are many tied ranks.

Question 4

Under what conditions would you choose a non-parametric test over a parametric test? Provide at least two scenarios.

Answer 4

You would choose a non-parametric test over a parametric test under the following conditions:

- When the data is **not normally distributed** (e.g., highly skewed or has significant outliers), and the sample size is small, making the Central Limit Theorem less applicable.
- When the data is measured on an **ordinal or nominal scale**, as parametric tests typically require interval or ratio scale data.
- When the assumption of **homogeneity of variance** (for group comparisons) is violated.

Question 5

Explain the concept of "controlling for confounders" in the context of partial correlation.

Answer 5

In partial correlation, "controlling for confounders" means removing the influence of one or more third variables (confounders) on the relationship between two other variables. For instance, in $r_{XY.Z}$, we are examining the correlation between X and Y, but only after the linear effects of Z on both X and Y have been accounted for. This helps to isolate the true, direct relationship between X and Y, reducing the risk of spurious correlations that might arise due to a shared relationship with the confounding variable Z.

References

[1] Agresti, A. (2018). *Statistical Methods for the Social Sciences*. Pearson.

[2] Pearson, K. (1895). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society.*

[3] Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.