Ben-Gurion University - School of Neuroscience and Cognition - 197.2.006

Lecture 1: Correlation Measures

Fall 2025

Lecturer: Yoav Noah

In this tutorial, we will explore key correlation measures, Pearson's r, Spearman's ρ, and Kendall's τ, to understand relationships between variables in statistical analysis. We'll begin by examining the concept of covariance and how it relates to correlation and z-scores, setting the foundation for Pearson's r as a measure of linear association between continuous variables. Next, we will introduce Spearman's rho, which is ideal for assessing monotonic relationships with ordinal or non-linear data, and Kendall's τ a robust measure for ordinal association especially useful with small datasets or ties.

1 Introduction

In statistical analysis, understanding relationships between variables often begins with *covariance*, which measures how two variables change together. For two jointly distributed real valued random variables X and Y with finite second moments, the covariance is defined as the expected value (or mean) of the product of their deviations from their individual expected values. The formula for covariance is:

$$C(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{1}$$

where $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ is the expected value of X and Y, respectively. If the (real) random variable pair (X,Y) can take on the values (X_i, Y_i) for $i = 1, \ldots, N$, with equal probabilities $P_i = \frac{1}{N-1}$, then the covariance can be equivalently written as:

$$C(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N-1} \tag{2}$$

where X_i and Y_i are individual values, $\bar{X}$ and $\bar{Y}$ are the means of X and Y, and N is the number of observations. The units of measurement of the covariance are those of X times those of Y. While covariance indicates the direction of a relationship, its magnitude depends on the units of the variables. *Correlation* standardizing covariance to measure the strength and direction of the relationship between variables, ranging from -1 to 1, making it unit-free and easier to interpret.
A related concept, the *z-score*, expresses how far a value X_i deviates from the mean in terms of standard deviations, calculated as:

$$z_i = \frac{X_i - \bar{X}}{\sigma_X} \tag{3}$$

where σ_X is the standard deviation of X, and $\bar{X}$ is the mean of X. The z-score standardizes values, allowing for direct comparison across variables.

2 Pearson's Correlation Coefficient (r)

Pearson's correlation coefficient, denoted as r, measures the strength and direction of the linear relationship between two continuous variables.

It standardizes covariance, providing a value between -1 and 1, where $r = 1$ indicates a perfect positive linear relationship, $r = -1$ indicates a perfect negative linear relationship, and $r = 0$ suggests no linear relationship. The formula for r is:

$$r_{XY} = \frac{C(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \sum_{i=1}^{N}(Y_i - \bar{Y})^2}} \tag{4}$$

where $C(X,Y)$ is the covariance matrix between X and Y, and $\bar{X}$, $\bar{Y}$, and σ_X, σ_Y are their means and standard deviations, respectively. An equivalent equation for equation 4 is:

$$r_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(\frac{X_i - \bar{X}}{S_X})(\frac{Y_i - \bar{Y}}{S_Y}) = \frac{\sum_{i=1}^{N} X_i Y_i - N\bar{X}\bar{Y}}{(N-1)S_X S_Y} \tag{5}$$

where $S_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}$ and $S_Y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$ are the sample standard deviation for X and Y.

Note: Pearson's r captures only linear associations, meaning it may not accurately reflect relationships that are non-linear.

Exercise 1:

A neuroscience research team is investigating the relationship between brainwave activity and stress levels in students. They measure alpha brainwave activity (variable X) during a mindfulness exercise and cortisol levels (variable Y) right after the session. The researchers hypothesize that higher alpha brainwave activity, often linked to relaxation, might correspond to lower cortisol, indicating reduced stress. The data they collected is as follows: $X = [3, 5, 7, 9, 13]$, $Y = [10, 14, 18, 22, 28]$. Calculate Pearson's correlation coefficient r to determine the relationship between alpha brainwave activity and cortisol levels. Interpret the result: is there a positive or negative relationship, and is it strong or weak?

1. Calculate the means and standard deviation of X and Y
2. Calculate the covariance between X and Y
3. Calculate Pearson's correlation coefficients (r)
4. Interpret the correlation result

Solution:

To calculate Pearson's r, we use equation 4. First, we need to calculate the means and standard deviation:

$\bar{X} = \frac{3+5+7+9+13}{5} = 7.4$, and $\bar{Y} = \frac{10+14+18+22+28}{5} = 18.4$.

$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\bar{X})^2}{N-1}} = \sqrt{\frac{(3-7.4)^2+(5-7.4)^2+(7-7.4)^2+(9-7.4)^2+(13-7.4)^2}{4}} = 3.847$

$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{N}(Y_i-\bar{Y})}{N-1}} = \sqrt{\frac{(10-18.4)^2+(14-18.4)^2+(18-18.4)^2+(22-18.4)^2+(28-18.4)^2}{4}} = 6.986.$

Next, we need to calculate the covariance using equation 2:

$C(X,Y) = \frac{(3-7.4)(10-18.4)+(5-7.4)(14-18.4)+(7-7.4)(18-18.4)+(9-7.4)(22-18.4)+(13-7.4)(28-18.4)}{4} = \frac{107.2}{4} = 26.8$

Now we can calculate Pearson's correlation, $r = \frac{26.8}{3.847 \times 6.986} = 0.997$

In this example, the value of $r = 0.997$ indicates a very strong positive linear relationship between the two variables (alpha brainwave activity and cortisol levels among the neuroscience students).

3 Spearman's Rank Correlation Coefficient

Spearman's Rank Correlation Coefficient is a non-parametric measure that assesses the strength and direction of the association between two ranked variables. It is particularly useful when the data do not meet the assumptions required for Pearson's correlation, such as normality or linearity. By converting raw data into ranks, Spearman's method evaluates how well the relationship between two variables can be described by a monotonic function. The coefficient ranges from -1 to 1, with 1 indicating a perfect positive correlation, -1 indicating a perfect negative correlation, and 0 suggesting no correlation. The equation for Spearman's rank correlation is given by:

$$\rho = 1 - \frac{\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \tag{6}$$

where $d_i = rank(X_i) - rank(Y_i)$ represents the difference between the ranks of each pair of observations, and N is the number of paired observations. This equation highlights how the degree of correlation decreases as the differences in ranks increase, providing a robust measure for analyzing ordinal data.

Exercise 2:

Students collected data on the relationship between the number of neuronal connections in the prefrontal cortex and performance scores on a cognitive flexibility test. The data for six participants is as follows:

Participant	**Neural Connections (X)**	**Performance Score (Y)**
1	150	80
2	120	70
3	180	85
4	130	75
5	160	90
6	140	78

Table 1: Dataset of neural connections and performance score taken from six participants.

1. Rank the values of X and Y
2. Calculate the Spearman's rank correlation coefficient (ρ)
3. Interpret the correlation result

Solution:

First, we will rank the values of X and Y from Table. 1, as shown in Table. 2. Next, we will calculate the sum of squares of the rank differences.

$\sum_{i=1}^{6} d_i^2 = \sum_{i=1}^{6} (rank(X_i) - rank(Y_i))^2 = (4-4)^2 + (1-1)^2 + (6-5)^2 + (2-2)^2 + (5-6)^2 + (3-3)^2 = 2.$

Now we can substituting the values into the equation, $\rho = 1 - \frac{6*2}{6*(6^2-1)} = 1 - \frac{12}{210} = 1 - 0.0571 = \frac{33}{35} \approx 0.9429$.
The calculated Spearman's rank correlation coefficient of $\rho \approx 0.9429$ indicates a very strong positive association between the number of neuronal connections in the prefrontal cortex and performance scores on the cognitive flexibility test. This suggests that as the number of neuronal connections increases, participants tend to achieve higher performance scores, reinforcing the idea that greater neuronal connectivity may enhance cognitive flexibility. This strong correlation supports the hypothesis that the structural properties of the brain can significantly influence cognitive performance.

Participant	**Neural Connections** (X)	**Rank of** X	**Performance Score** (Y)	**Rank of** Y
1	150	4	80	4
2	120	1	70	1
3	180	6	85	5
4	130	2	75	2
5	160	5	90	6
6	140	3	78	3

Table 2: Dataset of six participants with ranks.

4 Kendall's Tau (τ)

Kendall's Tau is a non-parametric statistic used to measure the strength and direction of the association between two variables. Unlike Pearson's correlation, which assesses linear relationships, Kendall's Tau evaluates the ordinal relationship between the ranks of data points, making it particularly useful for datasets that do not meet the assumptions of normality or homoscedasticity. It is calculated based on the difference between the number of concordant and discordant pairs of observations, providing a robust measure of correlation that is less sensitive to outliers. A pair of observations (X_i, Y_i) and (X_j, Y_j) is concordant pair if ranks for both variables follow the same order, i.e. if $X_i > X_j$ impies $Y_i > Y_j$ or $X_i < X_j$ implies $Y_i < Y_j$. While a pair observations is discordant pair if the ranks for the variables follow opposite orders, i.e. if $X_i > X_j$ implies $Y_i < Y_j$ or $X_i < X_j$ and $Y_i > Y_j$. The equation for Kendall's tau is given by:

$$\tau = \frac{C - D}{0.5 * N(N-1)} \tag{7}$$

where C is the number of concordant pairs, D is the number of discordant pairs, and N is the number of observations. Kendall's Tau ranges from -1 to 1, where values closer to 1 indicate a strong positive association, values near -1 signify a strong negative association, and values around 0 suggest no association. This makes it an ideal choice for analyzing ordinal data, ranks, or when the relationship between variables is not strictly linear.

Exercise 3:
In a study on the effects of sleep quality on memory retention, students collected data on the quality of sleep (measured by a sleep quality index) and memory retention scores from six participants. The data is summarized as follows:

Participant	**Sleep Quality Index** (X)	**Memory Retention Score** (Y)
1	90	85
2	80	70
3	75	75
4	85	80
5	70	90
6	95	95

Table 3: Dataset of sleep quality and memory retention taken from six participants.

1. Calculate Kendall's Tau correlation coefficient (τ)
2. Interpret the correlation result

Solution:
First, we will count the number of concordant and discordant from Table. 3.

- $(90, 85)$ and $(80, 70)$ - Concordant ($90 > 80$ and $85 > 70$)
- $(90, 85)$ and $(75, 75)$ - Concordant ($90 > 70$ and $85 > 75$)
- $(90, 85)$ and $(85, 80)$ - Concordant ($90 > 85$ and $85 > 80$)
- $(90, 85)$ and $(70, 90)$ - Discordant ($90 > 70$ and $85 < 90$)
- $(90, 85)$ and $(95, 95)$ - Concordant ($90 < 95$ and $85 < 95$)
- $(80, 70)$ and $(75, 75)$ - Discordant ($80 > 75$ and $70 < 75$)
- $(80, 70)$ and $(85, 80)$ - Concordant ($80 < 85$ and $70 < 80$)
- $(80, 70)$ and $(70, 90)$ - Discordant ($80 > 70$ and $70 < 90$)
- $(80, 70)$ and $(95, 95)$ - Concordant ($80 < 95$ and $70 < 95$)
- $(75, 75)$ and $(85, 80)$ - Concordant ($75 < 85$ and $75 < 80$)
- $(75, 75)$ and $(70, 90)$ - Discordant ($75 > 70$ and $75 < 90$)
- $(75, 75)$ and $(95, 95)$ - Concordant ($75 < 95$ and $75 < 95$)
- $(85, 80)$ and $(70, 90)$ - Discordant ($85 > 70$ and $80 < 90$)
- $(85, 80)$ and $(95, 95)$ - Concordant ($85 < 90$ and $80 < 90$)
- $(70, 90)$ and $(95, 95)$ - Concordant ($70 < 95$ and $90 < 95$)

There are ten concordant pairs ($C = 10$) and five discordant pairs ($D = 5$).
Therefore, $\tau = \frac{10-5}{0.5*6(6-1)} = \frac{5}{0.5*30} = \frac{5}{15} = \frac{1}{3}$. The calculated Kendall Tau correlation coefficient of $\tau = \frac{1}{3}$ indicates a weak positive association between sleep quality and memory retention scores. This suggests that while there may be a slight drift in which higher sleep quality correlates with better memory retention, the correlation is not strong. Other factors may also influence memory retention.

5 Comparing the Measures

When choosing a correlation measure, it is essential to consider the type of data, distribution, and relationship structure. *Pearson's correlation* works well for continuous data normally distributed with a linear relationship, providing insight into the strength and direction of the association. *Spearman's rank correlation* is valuable when the data are ordinal or when relationships are monotonic but not necessarily linear. *Kendall's Tau* is also suitable for ordinal data and tends to be more robust for small datasets or data with many tied ranks. Each measure gives unique insights and adapts to different data types, making it essential to choose based on the specific characteristics of your dataset.

Measure	**Type of Relationship**	**Suitable For**
Pearson's r	Linear	Continuous data, linear relationships
Spearman's ρ	Monotonic	Ordinal or non-linear continuous data
Kendall's τ	Ordinal	Small datasets or data with ties

Table 4: Comparing between the measures.

6 Statistical Tests for Correlation

Statistical tests for correlation are essential for determining whether observed relationships between two variables are statistically significant or simply due to random chance. In correlation analysis, tests such as the parametric t-test for *Pearson's r* assess whether a linear association between two continuous variables is unlikely under the null hypothesis of no relationship. However, for non-linear or non-normal data, non-parametric tests are often preferred. These include random permutation tests and bootstrapping, which are robust methods that generate a distribution of correlation values by randomly shuffling or resampling the data. This distribution serves as a benchmark to estimate the p-value of the observed correlation, providing a flexible approach when data assumptions are violated. Through these statistical tests, researchers can make data-driven conclusions about the strength and reliability of correlations across various contexts and data types.

Exercise 4:
In a study on the relationship between brain activity in the hippocampus and task performance during a memory game. Student measures the average theta wave power in the hippocampus measured while participants try to recall a series of images they saw earlier, and the participants memory performance score, calculated as the percentage of correctly recalled images out of the total presented. $X = [0.75, 0.80, 0.85, 0.70, 0.60, 0.95, 0.90]$, and $Y = [82, 88, 85, 76, 74, 90, 89]$ respectively. Implement the following using *Python*.

1. Calculate the Pearson correlation coefficient r between the power of the hippocampus theta (X) and the memory scores (Y).

2. To test the significance of the observed correlation, shuffle the memory scores (Y) multiple times, each time calculating the correlation with X. Repeat this shuffling process 1000 times to create a null distribution of correlation values. Based on this null distribution, calculate the p-value to determine if the observed correlation is statistically significant.

3. Use bootstrapping to estimate a 95% confidence interval for the correlation coefficient r. Pairs of samples (X, Y) with replacement, calculate the correlation for each sample, and repeat 1000 times to generate a distribution of r values. Use this distribution to determine the 95% confidence interval for the correlation coefficient.

Solution:
You can find the solution in the notebook `exercise_4_solution.ipynb` on Moodle.

7 Summary

In this tutorial, we explore three correlation measures: *Pearson's r*, *Spearman's rank correlation*, and *Kendall's Tau*. We examined each measure formula, assumptions, and application to different data types, from continuous and normally distributed data to ordinal or rank-based data. We also provided exercises using neuroscience-related examples to help solidify the understanding of when to apply each measure.

Questions:

1. **Can Spearman's rank correlation be used for non-monotonic relationships?**

   No, Spearman's rank correlation assumes a monotonic relationship; if your data do not follow a monotonic trend, consider alternative analysis methods.

2. **When would Kendall's Tau be preferred over Spearman's rank correlation?**

   Kendall's Tau is preferred when dealing with small sample sizes or datasets with many tied ranks, as it handles these situations better than Spearman's rank correlation.

Choosing the appropriate correlation measure:

1. **A psychologist analyzing survey responses with a 5-point Likert scale**

   Answer: Spearman's rank correlation or Kendall's Tau is appropriate here, as Likert scale data is ordinal. Both methods account for the rank-based nature of ordinal data. Kendall's Tau might be preferred if there are many tied responses or a small sample size.

2. **A biologist examining the relationship between two temperature measurements over time**

   Answer: Pearson's correlation is suitable in this case, as temperature is a continuous variable and the relationship is likely linear, especially if the temperature measurements are from a consistent source and meet normality assumptions.

3. **A data scientist analyzing sales rank data between different months**

   Answer: Spearman's rank correlation is the best choice here, as sales ranks are ordinal data. Spearman's correlation can effectively measure the strength and direction of the association between ranks without assuming a linear relationship.