

1.9: Correlation

Introduction to Correlation

Correlation measures the extent to which two or more variables are related to each other and is typically expressed as a correlation coefficient. The correlation test is a basic but commonly used method for examining the relationship between variables by assessing how two variables change together. However, there's a common misconception among students who interpret the warning about correlation: "correlation does not equal causation." Some students mistakenly believe that two correlated variables can never be causally linked. This is an erroneous conclusion. A more accurate understanding of the warning is that correlation does not necessarily imply causation. In other words, while correlation between variables is necessary for causation, it is not always sufficient (Vogt & Johnson, 2011).

Pearson Product-Moment Correlation Coefficient

Correlation measures the strength and direction of the linear relationship between two variables. There are various types of correlations. However, this chapter will focus on a Pearson product-moment correlation coefficient (aka Pearson correlation coefficient, Pearson's r , or Pearson's correlation) designed to evaluate the relationship between continuous variables (or one dichotomous variable and one continuous variable). It is calculated by dividing the covariance (a measure of how two variables covary together) by the product of their standard deviations).

The Pearson correlation coefficient ranges from -1 to +1, indicating the strength and direction of the linear relationship between two variables. A positive correlation (value closer to +1) suggests that as one variable increases, the other also tends to increase. On the other hand, a negative correlation (value closer to -1) indicates that as one variable increases, the other tends to decrease. The absolute value of the correlation coefficient indicates the strength of the relationship between the variables, with larger absolute values representing stronger relationships.

What standards should we use to determine whether a relationship is strong or weak? Many people use Cohen's (1988) guideline to interpret the magnitude of Pearson correlation coefficients. A small correlation, falling within the $r = 0.1$ to 0.29 range, suggests a relatively weak relationship between the variables under consideration. When the correlation coefficient falls between $r = 0.3$ and 0.49 , it is classified as a medium correlation, indicating a moderate association between the variables. Conversely, a large correlation, defined as $r \geq 0.5$, signifies a robust relationship between the variables. Finally, a perfect correlation of 1 or -1 indicates that the value of one variable can be precisely determined by knowing the value of the other variable. Conversely, a correlation of 0 indicates no discernible relationship between the two variables.

Computing Correlation Using the USArrests Dataset

We will use the data that we used in the first chapter to estimate correlation. As reviewed in Chapter 1, the built-in 'USArrests' dataset includes information on the number of arrests per 100,000 residents for assault, murder, and rape in each of the 50 states in the United States in 1973 and the percentage of the population residing in urban areas. Each row in the dataset represents a US state. Specifically, we will be evaluating if the number of arrests per 100,000 residents for murder correlates with the number of arrests per 100,000 residents for assault.

Correlation can be obtained through the following functions:

```
# Load the dataset
data("USArrests")
library(tidyverse)
USArrests %>%
  summarize(cor.murder.assault = cor(x = Murder, y = Assault, use = "complete"))
```

The Pearson's product-moment correlation coefficient obtained is 0.8018733. The number of arrests per 100,000 residents for murder was positively correlated with the number of arrests per 100,000 residents for assault. This means that, in the US in 1973, as the number of arrests per 100,000 residents for murder went up, so did the number of arrests per 100,000 residents for assault also increase. According to Cohen's (1988) guideline, this value indicates a very strong correlation between the variables of interest.

However, a crucial aspect is missing: the assessment of statistical significance. Inferential statistics play a vital role here, enabling us to draw conclusions from sample data by inferring insights about a population. Thus, determining whether the observed correlation is statistically significant is essential for meaningful interpretation in inferential statistics. Technically speaking, in our dataset, we do not need to conduct inferential statistics because the data sampled all 50 states instead of only 25 states out of 50 states. But, for demonstration purposes, I will show how we can conduct a statistical test for correlation coefficients, following the steps for hypothesis testing.

NHST Steps for Pearson's R Correlation Coefficient

Step 1: Formulate the Null and Alternative Hypotheses.

- H_0 : There is no relationship between the number of arrests per 100,000 residents for murder and the number of arrests per 100,000 residents for assault.
- H_A : There is a relationship between the number of arrests per 100,000 residents for murder and the number of arrests per 100,000 residents for assault.

Step 2: Calculate the Test Statistic.

When testing the null hypothesis for the correlation coefficient, we employ a t-statistic to compare the observed correlation coefficient (r) to a hypothesized value of 0. This t-statistic helps us determine whether the observed correlation is statistically significant or if it could have occurred by chance. We will use R codes to perform this task.

```
cor.test(x = USArrests$Murder,  
y = USArrests$Assault)
```

You will see that the correlation coefficient of 0.8018733.

Step 3: Determine the Probability (P-Value) of Obtaining a Test Statistic at Least as Extreme as the Observed Value, Assuming no Relationship Exists.

The p-value, the probability that the t statistic (of 9.2981 in this case) would occur by sampling error, was 2.596e-12, which was essentially much smaller than 0.05.

Steps 4 & 5: If the P-Value is Very Small, Typically Less Than 5%, Reject the Null Hypothesis, but if the P-Value is Not Small, Typically 5% or Greater, Retain the Null Hypothesis.

The very small p-value (much smaller than 0.05) indicates that a very strong positive relationship between the number of arrests per 100,000 residents for assault and murder is very unlikely if the null hypothesis were true.

Reporting the Results for Pearson's Product-Moment Correlation Coefficient

The number of arrests per 100,000 residents for murder is statistically significant and very strongly positively correlated with the number of arrests per 100,000 residents for assault in 50 US states in 1973 [$r = .80$; $t(48) = 9.30$; $p < .05$]. As the number of arrests per 100,000 residents for murder goes up, the number of arrests per 100,000 residents for assault goes up. While the correlation is .80 in this sample, the correlation is probable between .68 and .88 in the population (95% CI: .68 –.88).

Assumptions That Need To Be Met To Perform Correlation Analysis

It is important to highlight that correlation analysis has specific conditions and assumptions that must be met for accurate interpretation. While these assumptions are crucial for sound statistical analysis, I have not delved deeply into them in this book. This decision stems from the complexity of these assumptions, which could potentially overwhelm those learning the field. Instead, the aim of this book is to provide a broad overview of statistics and analysis, focusing on fundamental concepts rather than intricate technical details. Ensuring that certain assumptions are satisfied before conducting correlation analysis is crucial. Five key assumptions must be met for reliable results:

- The observations need to be independent of each other. This means that one observation's value should not influence another's value.
- Both variables being analyzed should be continuous. This ensures that the correlation analysis is applicable and meaningful.
- Both variables need to follow a normal distribution. This implies that the data points are evenly distributed around the mean in a bell-shaped curve.

- The relationship between the two variables should be linear. In other words, as one variable increases, the other should either increase or decrease consistently.
- The variance between the two variables should be constant, meaning that the spread of data points around the line of best fit remains consistent throughout the range of values.

It's worth noting that many advanced statistical techniques have been developed precisely because data often fail to meet one or more of these assumptions, highlighting the importance of understanding and addressing these issues in statistical analysis.

Scatter Plot

A scatter plot visually represents the relationship between two variables by displaying individual points on a graph. Each point on the plot corresponds to a unique data point or observation, formed by the intersection of the values of the two variables being studied. By examining the pattern of these points, we can discern the strength and direction of the correlation between the two variables (Vogt & Johnson, 2011). If you want to see how to create a scatter plot, refer to Chapter 1.

Conclusion

In this chapter, we covered correlation and how to compute the Pearson's product-moment correlation coefficient. Chapter 10 will delve into regression analysis, a powerful statistical technique used to understand the relationship between variables. Specifically, we will explore what regression entails and how to calculate regression coefficients, which are essential in quantifying the strength and direction of these relationships.

References

Cohen, J. (1988). Statistical power analysis for the behavioral science. Erlbaum Associates.

Vogt, W. P., & Johnson, R. B. (2011). Dictionary of statistics & methodology: A nontechnical guide for the social sciences (4th ed.). Sage.

This page titled [1.9: Correlation](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Jaeyong Choi](#) ([The Pennsylvania Alliance for Design of Open Textbooks \(PA-ADOPT\)](#)) via [source content](#) that was edited to the style and standards of the LibreTexts platform.