Calculate Spearman Rank Correlation in R

Authored by

Mohammed looti

November 3, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Calculate Spearman Rank Correlation in R*. PSYCHOLOGICAL STATISTICS. Retrieved from https://statistics.arabpsychology.com/?p=9118

In the field of statistics, the concept of **correlation** is fundamental. It quantifies the strength and direction of the linear or monotonic relationship shared between two variables. Understanding correlation is critical for predictive modeling and observational data analysis. The resulting value, known as the correlation coefficient, is strictly confined to the range of -1 to 1, providing clear interpretations of the relationship:

-1: This signifies a **perfect negative relationship** between two variables. As one variable increases, the other decreases consistently.
0: A coefficient of zero indicates that there is **no linear relationship** detected between the two variables.
1: This represents a **perfect positive relationship**, meaning that the variables increase or decrease together in lockstep.

While the standard Pearson correlation measures linear relationships between interval or ratio data, many real-world datasets involve ordinal data or non-linear associations. For these scenarios, the **Spearman Rank Correlation** coefficient, often denoted as Rho (ρ), becomes the statistical tool of choice. This non-parametric measure assesses the monotonic relationship between two variables by operating on the ranks of the data rather than the raw scores themselves. This approach makes it robust against outliers and suitable for ordinal data, such as comparing the rank order of student performance in two different subjects.

Understanding Correlation and Rank Statistics

The choice of correlation method depends entirely on the nature of the data and the underlying assumptions about its distribution. Pearson correlation assumes that the data is normally distributed and that the relationship is strictly linear. However, when these assumptions are violated, relying on rank statistics provides a powerful alternative that retains much of the relationship information without demanding strict distributional characteristics.

By converting raw scores into their respective ranks (e.g., the lowest score receives rank 1, the next lowest rank 2, and so on), the Spearman method focuses solely on the consistency of the ordering. It determines whether a high rank in one variable corresponds consistently to a high rank in the second variable, regardless of the precise distance between the scores. This monotonic assessment is often more informative when dealing with data derived from surveys, subjective assessments, or non-linear biological phenomena.

The robustness of Spearman's Rho is a significant advantage. If you were comparing, for instance, a team's spending budget against their final league ranking, a single outlier budget (an extraordinarily high spend) would skew a Pearson coefficient drastically. Spearman's Rho, working on the ranks, minimizes the impact of this outlier, providing a more stable and representative measure of the overall trend.

What is Spearman Rank Correlation (Rho)?

Spearman's Rho is mathematically defined as the Pearson correlation coefficient calculated on the rank-transformed data. This transformation is key to its utility. If there are no repeated data values (ties), a simpler calculation based on the difference between ranks can be used. However, most modern statistical packages, including R, handle ties automatically by assigning the average rank to the tied observations, ensuring accuracy.

The resulting Rho coefficient measures the degree to which the relationship between the variables can be described using a monotonic function. A monotonic relationship is one where the variables tend to move in the same general direction (positive) or opposite directions (negative), but not necessarily at a constant rate (which would be required for linearity).

For researchers and analysts, Spearman's Rho provides a straightforward, interpretable statistic. A high positive Rho (close to +1) suggests that as the rank of one variable increases, the rank of the other variable tends to increase as well. Conversely, a value close to -1 implies an inverse ranking pattern. Understanding this distinction is vital for making correct inferences about the relationship between observed phenomena.

The R Function: Using cor.test() for Spearman's Rho

The R programming language offers a highly efficient and comprehensive suite of functions for statistical testing. To calculate the Spearman rank correlation coefficient, we utilize the built-in function `cor.test()`. While `cor()` provides only the coefficient, `cor.test()` is preferred because it performs a formal hypothesis test, providing the Rho value, the test statistic (S), and the corresponding **p-value**.

The null hypothesis for this test asserts that the true correlation coefficient (Rho) is zero, meaning there is no monotonic relationship between the population variables. The alternative hypothesis, typically two-sided by default, suggests that the true Rho is not equal to zero. Rejecting the null hypothesis depends on the calculated **p-value**.

To specify that we require the Spearman method, a simple argument is added to the function call. The following basic syntax demonstrates how to execute this test in R, requiring only two vectors (`x` and `y`) containing the data points:

corr <- cor.test(x, y, method = 'spearman')

This command returns a comprehensive object containing the test results. The next sections illustrate how to apply this function effectively, first using simple vectors and then applying it to columns within a structured data frame, which is a common scenario in data analysis projects.

Example 1: Analyzing Correlation Between Two Vectors

In this first example, we define two simple vectors, `x` and `y`, perhaps representing the scores of ten students in two different, non-standardized tests. We want to determine if there is a monotonic relationship in their performance--that is, whether students who rank highly in test X also tend to rank highly in test Y.

We first define the datasets and then execute the `cor.test()` function, explicitly setting the method to 'spearman'. Note that the output provides not only the Rho estimate but also the test statistic (S) and the critical **p-value** for hypothesis testing.

#define data
x <- c(70, 78, 90, 87, 84, 86, 91, 74, 83, 85)
y <- c(90, 94, 79, 86, 84, 83, 88, 92, 76, 75)

#calculate Spearman rank correlation between x and y
cor.test(x, y, method = 'spearman')

Spearman's rank correlation rho

data: x and y
S = 234, p-value = 0.2324
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
-0.4181818

The output clearly shows that the Spearman rank correlation coefficient (rho) is **-0.41818**. This value indicates a moderate negative correlation between the two sets of scores. A negative value suggests that students ranking higher in test `x` tend to rank slightly lower in test `y`, although the relationship is not strong.

Crucially, we must examine the **p-value**, which in this case is **0.2324**. If we use the standard threshold for **statistical significance** (alpha = 0.05), since 0.2324 is significantly greater than 0.05, we fail to reject the null hypothesis. Therefore, while a negative relationship exists in the sample data, we do not have sufficient evidence to conclude that a statistically significant monotonic relationship exists between these two variables in the broader population.

Example 2: Assessing Correlation in a Data Frame

In real-world data science, variables are typically stored within data structures such as data frames. This example demonstrates how to calculate the Spearman correlation between two specific columns--`points` scored and `assists` made--within a data frame representing basketball team performance metrics. Analyzing rank correlation here is appropriate as we might suspect the relationship is monotonic but not perfectly linear (e.g., a team with slightly more points might have a huge jump in assists).

We first construct the data frame, ensuring the variables we wish to correlate (`points` and `assists`) are correctly defined. To execute the test, we reference the columns using the standard R data frame notation: `df$column_name`.

#define data frame
df <- data.frame(team=c('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J'),
points=c(67, 70, 75, 78, 73, 89, 84, 99, 90, 91),
assists=c(22, 27, 30, 23, 25, 31, 38, 35, 34, 32))

#calculate Spearman rank correlation between x and y
cor.test(df$points, df$assists, method = 'spearman')

Spearman's rank correlation rho

data: df$points and df$assists
S = 36, p-value = 0.01165
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.7818182

The resulting Spearman rank correlation coefficient is **0.7818**. This value indicates a strong, positive monotonic relationship between the number of points a team scores and the number of assists they accumulate. Teams ranking higher in points generally rank higher in assists.

Furthermore, the associated **p-value** is **0.01165**. Since this value is less than the standard significance level of 0.05, we reject the null hypothesis. This means the strong positive correlation observed is highly unlikely to have occurred by random chance, confirming that there is a **statistically significant** monotonic relationship between points and assists in this dataset.

Interpreting the Results: Rho and P-value

Interpreting the output of `cor.test(method='spearman')` requires attention to two primary elements: the estimated Rho value and the **p-value**. Both are essential for drawing accurate conclusions from the analysis.

First, the Rho value (sample estimates) provides the magnitude and direction of the monotonic association. Values closer to 1 or -1 indicate a stronger relationship. A Rho of 0.78, as seen in Example 2, suggests a substantial positive association, implying that knowledge of one variable's rank helps predict the rank of the other. In contrast, a Rho near zero suggests ranks are independent.

Second, the **p-value** determines if the observed relationship is likely due to chance. The p-value tests the likelihood of observing a correlation as extreme as the one calculated, assuming that the true population correlation is zero. If the p-value is less than the predetermined significance level (alpha, usually 0.05), the relationship is deemed **statistically significant**. This significance indicates confidence that the relationship observed in the sample data reflects a true association in the population from which the sample was drawn.

It is important to remember that significance does not imply causality. Even a perfect **Spearman Rank Correlation** (Rho = 1) only confirms a strong association in ranking, but does not prove that one variable causes the other to increase or decrease. Proper interpretation requires considering the context and design of the study.

The flexibility and non-parametric nature of the **Spearman Rank Correlation** make it an invaluable tool when analyzing complex data where normality assumptions cannot be met, providing reliable insights into the ordered relationships between variables.

Additional Resources