How to Choose the Right Correlation for Your Data

Authored by

stats writer

December 28, 2025

Understanding the relationship between two variables is fundamental to statistical analysis. Correlation serves as a powerful statistical measure used to quantify both the strength and direction of such a relationship. However, the world of correlation is not one-size-fits-all; analysts must select the appropriate coefficient based on the nature and distribution of the data being examined. Choosing the correct method is critical for accurate inference.

This guide explores the three most common correlation coefficients: Pearson's correlation coefficient (for linear relationships in normally distributed data), Spearman's rank correlation coefficient, and Kendall's Tau correlation coefficient (both non-parametric measures). Each method possesses distinct assumptions, advantages, and limitations. By understanding these nuances, researchers can ensure they select the most suitable correlation test for their specific analytical needs, leading to robust and reliable conclusions.

The Fundamentals of Correlation Analysis

In the field of statistics, the concept of correlation encapsulates both the magnitude and the trajectory of the association between two or more variables. This relationship is quantified by a correlation coefficient, a standardized value that facilitates easy interpretation.

The range of a correlation coefficient is strictly defined, spanning from **-1** to **+1**. A value of **-1** signifies a perfect negative relationship, meaning that as one variable increases, the other decreases consistently. Conversely, a value of **+1** represents a perfect positive relationship, where both variables increase or decrease together proportionally. A coefficient of **0** indicates no linear relationship whatsoever between the variables under observation. It is vital to remember that correlation measures association, but does not imply causation.

Selecting the Appropriate Correlation Measure

The decision regarding which correlation coefficient to employ is driven primarily by the type of data involved and the assumptions regarding its distribution. Researchers must first determine if their variables are continuous, ordinal (ranked), or if the relationship is linear or monotonic. While all correlation coefficients aim to describe bivariate relationships, their underlying mathematical foundations differ significantly.

We typically rely on three primary methodologies to measure correlation, each suited for specific data characteristics:

Pearson Correlation: This parametric test is used exclusively to measure the linear relationship between two continuous variables (e.g., relating a person's **height** to their **weight**). It requires the data to be approximately normally distributed and free from extreme outliers.

Spearman Correlation: A non-parametric alternative, the Spearman coefficient is designed to measure the monotonic relationship between two ranked variables, or when the assumption of normality for continuous data is violated. It assesses how well the relationship between the variables can be described using a monotonic function. A common example is correlating the **rank of a student's math exam score** versus the **rank of their science exam score** in a class.

Kendall's Correlation: This non-parametric test is often preferred over Spearman's when dealing with smaller sample sizes or when the data contains a large number of **tied ranks**. It measures the probability that two variables are in the same order (concordant) versus the probability that they are in different orders (discordant).

The subsequent sections of this tutorial detail the practical application of these three types of correlation analysis using the statistical software package, Stata.

Setting Up Your Analysis Environment in Stata

To demonstrate the calculation of these correlation coefficients, we will utilize a built-in dataset commonly used for introductory examples in Stata. This dataset, named *auto*, provides various characteristics of automobiles.

To load this dataset into your active Stata session, execute the following command in the Command window:

use http://www.stata-press.com/data/r13/auto

Once the data is successfully loaded, it is good practice to perform a quick summary check to familiarize ourselves with the dataset's structure, variables, and data completeness. This is achieved by typing the following command:

summarize

```
. use http://www.stata-press.com/data/r13/auto
(1978 Automobile Data)

. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
make	0				
price	74	6165.257	2949.496	3291	15906
mpg	74	21.2973	5.785503	12	41
rep78	69	3.405797	.9899323	1	5
headroom	74	2.993243	.8459948	1.5	5
trunk	74	13.75676	4.277404	5	23
weight	74	3019.459	777.1936	1760	4840
length	74	187.9324	22.26634	142	233
turn	74	39.64865	4.399354	31	51
displacement	74	197.2973	91.83722	79	425
gear_ratio	74	3.014865	.4562871	2.19	3.89
foreign	74	.2972973	.4601885	0	1

The output confirms that the *auto* dataset contains 74 total observations across 12 different variables, ready for our subsequent correlation analyses.

Calculating Pearson Correlation in Stata

The Pearson correlation coefficient (r) measures the linear association between two continuous, normally distributed variables. In Stata, the primary command for calculating pairwise correlations is **pwcorr**.

To find the Pearson correlation between the variables *weight* and *length*, which are both continuous measures, we use the **pwcorr** command followed by the names of the variables:

pwcorr weight length

```
. pwcorr weight length
```

	weight	length
weight	1.0000	
length	0.9460	1.0000

The initial output provides the correlation value itself. However, in rigorous statistical analysis, determining the statistical significance of the coefficient is paramount. We can easily obtain the

corresponding p-value by adding the **sig** option to the command:

pwcorr weight length, sig

```
. pwcorr weigh length, sig
```

	weight	length
weight	**1.0000**	
length	**0.9460** **0.0000**	**1.0000**

Upon reviewing the output, we observe that the p-value is reported as **0.000**. Since this value is considerably less than the conventional significance level of 0.05 (alpha), we conclude that the correlation observed between *weight* and *length* is highly statistically significant. This suggests that the relationship is unlikely to have occurred by random chance.

If the goal is to examine the relationships among several variables simultaneously, the **pwcorr** command efficiently handles lists of variables. To obtain the Pearson Correlation Coefficients and associated p-values for *weight*, *length*, and *displacement*, simply list them after the command, retaining the **sig** option:

pwcorr weight length displacement, sig

```
. pwcorr weight length displacement, sig
```

	weight	length	displa~t
weight	**1.0000**		
length	**0.9460** **0.0000**	**1.0000**	
displacement	**0.8949** **0.0000**	**0.8351** **0.0000**	**1.0000**

The resulting matrix provides the pairwise correlation coefficients and their significance levels for all combinations of the selected variables. Here is a detailed interpretation of this comprehensive

output:

The Pearson Correlation between **weight** and **length** is 0.9460, indicating a very strong positive association. The corresponding p-value is 0.000.
The Pearson Correlation between **weight** and **displacement** is 0.8949, also showing a strong positive relationship. The p-value is 0.000.
The Pearson Correlation between **displacement** and **length** is 0.8351, representing a strong positive link. The p-value is 0.000.

Executing Spearman Rank Correlation in Stata

When dealing with ordinal data, non-normally distributed continuous data, or when seeking a relationship that is merely monotonic (not strictly linear), the Spearman's rank correlation coefficient (rho) is the appropriate non-parametric test. This method works by ranking the data for each variable and then applying the Pearson formula to the ranks.

In Stata, the dedicated command for this procedure is **spearman**. Let's calculate the Spearman coefficient between *trunk* size and the *rep78* (repair record 1978) variable:

spearman trunk rep78

```
. spearman trunk rep78

 Number of obs =      69
Spearman's rho =      -0.2235

Test of Ho: trunk and rep78 are independent
    Prob > |t| =      0.0649
```

The output generated by the **spearman** command provides key metrics necessary for interpretation:

Number of obs: This crucial statistic indicates the number of paired observations used in the calculation. Since the variable *rep78* had missing values, Stata employed only 69 pairwise observations out of the total 74 records, utilizing pairwise deletion.
Spearman's rho: This is the Spearman correlation coefficient itself. In this instance, rho is -0.2235, which suggests a weak negative monotonic relationship. As the rank of one variable increases, the rank of the other tends to decrease slightly.
Prob > |t|: This represents the p-value associated with the test of the hypothesis that the true correlation is zero. With a p-value of 0.0649, which is greater than the standard significance

threshold ($\alpha = 0.05$), we conclude that there is not a statistically significant correlation between these two variables.

To calculate Spearman Correlation Coefficients for multiple variables, list them following the command. Furthermore, to structure the output neatly into a matrix that includes both the correlation coefficient (rho) and the p-value, we employ the **stats(rho p)** option:

spearman trunk rep78 gear_ratio, stats(rho p)

```
. spearman trunk rep78 gear_ratio, stats(rho p)
(obs=69)
```

Key
rho *Sig. level*

	trunk	rep78	gear_r~o
trunk	**1.0000**		
rep78	**-0.2235** **0.0649**	**1.0000**	
gear_ratio	**-0.5187** **0.0000**	**0.4275** **0.0002**	**1.0000**

Interpretation of the multiple variable output:

Spearman Correlation between **trunk** and **rep78** = -0.2235, with a p-value of 0.0649 (not significant).
Spearman Correlation between **trunk** and **gear_ratio** = -0.5187, indicating a moderate negative relationship. The p-value of 0.0000 confirms high statistical significance.
Spearman Correlation between **gear_ratio** and **rep78** = 0.4275, suggesting a moderate positive relationship. The p-value of 0.0002 confirms statistical significance.

Using Kendall's Tau for Small Samples and Ties

The third major method for measuring rank correlation is Kendall's Tau correlation coefficient (τ). This non-parametric statistic is especially useful when the sample size is small, or when the data features a significant number of identical values, known as "tied ranks," which can potentially bias the Spearman coefficient. Kendall's Tau often provides a more robust estimate of the population

correlation.

In Stata, we use the **ktau** command to perform this analysis. We will calculate the Kendall's Tau coefficient between *trunk* and *rep78* again for comparison:

ktau trunk rep78

```
. ktau trunk rep78

  Number of obs =      69
Kendall's tau-a =      -0.1424
Kendall's tau-b =      -0.1752
Kendall's score =    -334
    SE of score =     181.254   (corrected for ties)

Test of Ho: trunk and rep78 are independent
     Prob > |z| =       0.0662  (continuity corrected)
```

Interpreting the **ktau** command output:

Number of obs: Consistent with the Spearman calculation, 69 pairwise observations were used due to missing data in the *rep78* variable.
Kendall's tau-b: This value, τb, is the standard measure reported for Kendall's rank correlation when ties are present in the data (as opposed to tau-a, which assumes no ties). In this case, tau-b = -0.1752, indicating a weak negative correlation between the two variables, similar in direction to the Spearman result, but typically lower in magnitude.
Prob > |z|: This is the p-value derived from the hypothesis test. With a p-value of 0.0662, which is marginally greater than the alpha level of 0.05, we fail to reject the null hypothesis, concluding that the correlation is not statistically significant.

To calculate Kendall's Correlation Coefficient for multiple variable pairs, list them after the **ktau** command. To display the coefficient and the corresponding p-value in a clean matrix format, use the **stats(taub p)** option, ensuring the use of tau-b for handling ties:

ktau trunk rep78 gear_ratio, stats(taub p)

```
Test of Ho: trunk and rep78 are independent
     Prob > |z| =       0.0662  (continuity corrected)

. ktau trunk rep78 gear_ratio, stats(taub p)
(obs=69)
```

Key
tau_b *Sig. level*

	trunk	rep78	gear_r~o
trunk	**1.0000**		
rep78	**-0.1752** **0.0662**	**1.0000**	
gear_ratio	**-0.3753** **0.0000**	**0.3206** **0.0006**	**1.0000**

The pairwise results for the multiple variable Kendall's Tau test are as follows:

Kendall's Correlation between **trunk** and **rep78** = -0.1752, with a p-value of 0.0662 (not significant).
Kendall's Correlation between **trunk** and **gear_ratio** = -0.3753, indicating a moderate negative rank correlation. The p-value of 0.0000 confirms strong statistical significance.
Kendall's Correlation between **gear_ratio** and **rep78** = 0.3206, showing a moderate positive rank correlation. The p-value of 0.0006 confirms statistical significance.

Summary of Correlation Choice

Selecting the correct correlation method is essential for obtaining meaningful results. If your data consists of two continuous variables that meet parametric assumptions (linearity and normality), **Pearson's r** is preferred. If your data is ordinal, or if the continuous data violates normality assumptions, the non-parametric measures—**Spearman's rho** or Kendall's Tau—are appropriate. Kendall's Tau offers added robustness when ties are frequent or the sample size is limited, making it a powerful tool for reliable rank correlation estimation.