A Comprehensive Guide to Correlation Coefficients: Pearson, Spearman, and Kendall using Stata

Authored by

Mohammed looti

November 8, 2025

In the realm of statistics and data analysis, the concept of correlation is absolutely fundamental. It quantifies the statistical relationship between two variables, specifically detailing both the strength and the direction of that association. This relationship is summarized by a correlation coefficient, a standardized metric that always ranges between **-1** and **1**. A coefficient of **-1** indicates a perfect negative relationship, meaning that as one variable increases, the other decreases proportionally. Conversely, **1** denotes a perfect positive relationship where both variables increase together. A coefficient of **0** signifies the complete absence of a linear relationship between the variables under observation.

Selecting the correct correlation method is a critical step in any statistical analysis, depending heavily on the scale and distribution of the data. Statisticians primarily rely on three distinct types of correlation coefficients to handle various data scenarios:

Pearson Correlation (r): As the most common parametric measure, Pearson's *r* is specifically designed to assess the linear association between two **continuous variables** that are assumed to be approximately normally distributed. A typical application involves examining the relationship between physical measurements like height and weight.
Spearman Correlation (rho): This powerful non-parametric measure calculates the correlation based on the **ranks** of the observations, rather than their raw values. It is the ideal choice when dealing with ordinal data or when continuous data violates the strict normality assumptions required for Pearson's correlation.
Kendall's Correlation (tau): An alternative non-parametric measure, Kendall's tau (τ) is often preferred over Spearman's rho, particularly when the sample size is small or when the dataset includes a high number of tied ranks. It assesses association by counting the concordance and discordance of paired observations.

This comprehensive tutorial will guide you step-by-step through the process of calculating, interpreting, and applying all three major correlation coefficients using **Stata**, one of the leading statistical software packages utilized across academic and professional research environments.

Preparing the Data in Stata

Before diving into the correlation calculations, we must first prepare the data within the **Stata** environment. For demonstration purposes throughout this article, we will utilize a widely accessible, built-in dataset known as *auto*. This dataset is standard in Stata documentation and provides rich information concerning 74 different automobiles, including characteristics such as weight, length, and engine displacement.

To load this demonstration dataset into your current Stata session, execute the following command directly in the Command box. This action retrieves the dataset from the official Stata Press repository, ensuring immediate access to the necessary variables for analysis:

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

A crucial preliminary step after loading any dataset is performing a rapid data inspection. We achieve this overview by running the standard **summarize** command. The output provides essential descriptive statistics, including the total number of observations, the mean, the standard deviation, and the minimum and maximum values for every variable present in the dataset. This quick assessment helps verify data integrity and scale.

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

As confirmed by the summary output displayed above, the *auto* dataset contains 74 complete observations across 12 distinct variables. Understanding this initial structure is vital, as it highlights the presence of any potential missing values or initial distributional characteristics that may influence whether we proceed with a Pearson (parametric) test or opt for one of the rank-based (non-parametric) alternatives like Spearman or Kendall's tau.

Calculating and Interpreting Pearson Correlation

The Pearson Correlation Coefficient (*r*), also known as the Pearson product-moment correlation coefficient, is the gold standard for measuring the strength and direction of the linear association between two variables that are **continuous** and satisfy the assumptions of normality. In **Stata**, the primary command used for calculating pairwise Pearson correlations is **pwcorr**.

To illustrate, we will calculate the correlation between the vehicle characteristics *weight* and *length*.

By listing these variables after the **pwcorr** command, Stata immediately returns the correlation coefficient, which quantifies the relationship:

pwcorr weight length

```
. pwcorr weight length
```

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

The coefficient alone only tells us the observed strength. To determine if this relationship is genuine--that is, whether it is likely to hold true for the larger population rather than being a result of random sampling--we must calculate the associated **p-value**. To request the significance test alongside the coefficient, we must include the **sig** option in the command syntax:

pwcorr weight length, sig

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

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

Reviewing the output, the p-value is displayed as **0.000**. Since this value is substantially lower than the widely accepted alpha threshold of 0.05, we confidently conclude that the correlation between vehicle weight and length is **statistically significant**. Furthermore, the positive coefficient (0.9460) indicates a very strong, near-perfect positive linear relationship: heavier cars are almost invariably longer cars in this dataset.

For analyses requiring simultaneous assessment of multiple relationships, **pwcorr** can easily generate a correlation matrix. By listing several variables, and again including the **sig** option, we obtain a comprehensive overview of all pairwise associations, such as comparing *weight*, *length*, and *displacement*:

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 interpretation of this matrix reinforces the initial findings and extends them across three dimensions:

The Pearson Correlation between *weight* and *length* (0.9460, p=0.000) shows an extremely strong, highly significant positive correlation.
The Pearson Correlation between *weight* and *displacement* (0.8949, p=0.000) also demonstrates a very strong, highly significant positive correlation.
The Pearson Correlation between *displacement* and *length* (0.8351, p=0.000) confirms a strong, highly significant positive correlation.

In conclusion, for this set of vehicle dimensions, all three pairs exhibit robust, statistically significant positive linear relationships, indicating that increases in any one measure (weight, length, or engine displacement) are strongly associated with increases in the others.

Calculating and Interpreting Spearman Correlation

When the data scale is **ordinal** or when the critical assumption of normality is violated for continuous variables, the **Spearman Correlation Coefficient** (rho) provides a robust non-parametric alternative to Pearson's *r*. Spearman's method works by assessing the correlation between the **ranks** assigned to the observations, effectively measuring monotonic relationships--whether variables tend to move together, even if the relationship isn't perfectly linear.

In **Stata**, calculating Spearman's rho is straightforward using the command **spearman**. We will analyze the relationship between *trunk* (trunk space, a continuous measure) and *rep78* (the repair record in 1978, which is an ordinal rating).

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 is highly informative, providing all necessary components for interpretation:

Number of obs: This indicates that only 69 pairwise observations were used, signifying that 5 vehicles had missing data for *rep78*. Spearman only uses complete pairs.
Spearman's rho: The calculated coefficient is **-0.2235**. This suggests a weak negative relationship; specifically, vehicles with higher trunk space ranks tend to have slightly lower repair record ranks.
Prob > |t|: This two-sided p-value tests the null hypothesis that rho equals zero. Since the p-value of **0.0649** is just above the conventional significance level of 0.05, we lack sufficient evidence to declare this correlation **statistically significant**.

To efficiently evaluate multiple rank correlations, the **spearman** command supports matrix output. By utilizing the **stats()** option, we can customize which statistics appear in the matrix. Using **stats(rho p)** displays both the Spearman's rho coefficient and the corresponding p-value for every pairing:

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**

Analyzing the expanded rank correlation matrix reveals clearer patterns among the variables:

The correlation between *trunk* and *rep78* remains weak and non-significant (rho = -0.2235, p = 0.0649).
The Spearman Correlation between *trunk* and *gear_ratio* is **-0.5187** (p = 0.0000). This indicates a strong, highly significant inverse relationship: cars ranked higher in trunk size tend to have lower gear ratios.
The Spearman Correlation between *gear_ratio* and *rep78* is **0.4275** (p = 0.0002). This shows a moderate, highly significant positive relationship, suggesting that vehicles with higher gear ratio ranks are associated with better repair records.

Calculating and Interpreting Kendall's Correlation

The **Kendall's Correlation Coefficient** (τ), or Kendall's tau, offers a third approach to assessing monotonic association, serving as a primary non-parametric alternative to Spearman's rho. Kendall's tau is particularly valuable in situations involving small datasets or when the data exhibits a high frequency of tied ranks across observations. Its calculation relies on counting the relative number of concordant (in agreement) versus discordant (in disagreement) pairs.

To calculate Kendall's tau in **Stata**, we use the command **ktau**. We will use the same variables as the previous section, *trunk* and *rep78*, which allows for a direct comparison between the Spearman and Kendall rank correlation results:

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)
```

The output from the **ktau** command provides nuanced information about the rank association:

Number of obs: Like the Spearman test, the analysis utilized 69 valid observations, confirming the impact of missing data in the *rep78* variable.
Kendall's tau-b: Stata typically reports tau-b, which includes adjustments specifically designed to handle tied ranks, making it highly appropriate for discrete or ordinal data. A coefficient of **-0.1752** suggests a weak negative correlation.
Prob > |z|: The p-value associated with the Z-test statistic is **0.0662**. Since this value marginally exceeds the conventional 0.05 threshold, we again conclude that this rank correlation is not **statistically significant** at the 5% level.

If a comprehensive overview is required, the **ktau** command can also produce a full correlation matrix. When generating this matrix, it is necessary to specify the required statistics using the **stats()** option. By entering **stats(taub p)**, we instruct Stata to display Kendall's tau-b coefficient and its corresponding p-value for all requested pairings:

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**

A detailed analysis of the Kendall's tau-b matrix reveals consistent patterns compared to Spearman's analysis:

The Kendall's Correlation between *trunk* and *rep78* is -0.1752 (p-value = 0.0662), confirming a weak, non-significant negative association.
The Kendall's Correlation between *trunk* and *gear_ratio* is **-0.3753** (p-value = 0.0000), indicating a moderate and highly significant negative relationship.
The Kendall's Correlation between *gear_ratio* and *rep78* is **0.3206** (p-value = 0.0006), demonstrating a moderate and highly significant positive relationship.

In summary, both non-parametric methods--Spearman's rho and Kendall's tau--arrive at the same qualitative conclusions regarding the significance and direction of the relationships. While Kendall's tau coefficients are often numerically smaller than Spearman's rho for the same data, Kendall's tau-b remains a highly reliable measure, especially favored by researchers when dealing with complex ranking structures or a smaller number of observations.