Chapter 295

Correlation

Introduction

The correlation coefficient, or correlation, is a unit-less measure of the relationship between two variables. The estimation of three correlation types are available in this procedure: the Pearson (product-moment) correlation, the Spearman rank correlation, and Kendall's Tau correlation. Although correlation coefficients are often reported alone, hypothesis tests and confidence intervals are also available in this procedure.

The data for calculation of the sample correlation typically comes from measurements of two variables on a number of individuals or units. The range of correlation is -1 to 1. Correlation values close to -1 indicate a strong negative relationship (high values of one variable generally indicate low values of the other). Correlation values close to 1 indicate a strong positive relationship (high values of one variable generally indicate high values of the other). Correlation values near 0 indicated little relationship among the two variables.

Pearson Correlation Details

The Pearson correlation coefficient is the most common correlation measure. It is sometimes called the simple correlation coefficient, the Pearson product-moment correlation, the sample correlation coefficient, or simple linear correlation. For *n* measurements of two variables, *X* and *Y*, the Pearson correlation coefficient is calculated as

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}}$$

Confidence Intervals

The Pearson Correlation report of this procedure gives two confidence intervals for the correlation.

Correlation Distribution Confidence Interval

Although less commonly used due to the complexity in calculation, the confidence interval for the population correlation may computed directly from the correlation distribution. These calculations are outside the scope of this documentation, but details may be found in Pearson and Hartley (1984) and Guenther (1977). This confidence interval is preferred, even though it is not as commonly described in correlation chapters of texts. An assumption of this formulation is that the two variables are characterized by a bivariate Normal distribution.

Normal Approximation Confidence Interval

An approximate (and more popular) confidence interval for the population correlation is produced by first computing a confidence interval for Fisher's *Z* transformation of the correlation, and then transforming back to the original correlation scale.

$$(L_Z, U_Z) = \frac{1}{2}\ ln\left(\frac{1+r}{1-r}\right) \pm \frac{z_{1-\alpha/2}}{\sqrt{n-3}}$$

Convert L_Z and U_Z to the lower and upper correlation confidence limits using

$$L_\rho = \frac{e^{2L_Z}-1}{e^{2L_Z}+1} \quad \text{and} \quad U_\rho = \frac{e^{2U_Z}-1}{e^{2U_Z}+1}$$

Hypothesis Tests

The formulation of the hypothesis test depends on whether the correlation is to be tested against 0, or against some other value.

Testing H_0: $\rho = 0$

Commonly the researcher wishes to know if the correlation is different from 0, or in other words, if correlation exists. In this case, a standard *t*-test may be used:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

The *t* distribution with $n-2$ degrees of freedom may be used to generate one- or two-sided test *p*-values from this test statistic. As an aside, this test is equivalent to testing whether the linear regression slope is 0.

Testing H_0: $\rho = \rho_0$, where $\rho_0 \neq 0$

When testing against a correlation other than zero, either the direct distribution of the correlation may be used, or a Normal approximation based on Fisher's Z transformation of the correlation may be used.

Correlation Distribution Test

Although less commonly used due to the complexity in calculation, the probabilities (*p*-values) for this hypothesis test of the correlation may computed directly from the correlation distribution. As in the case of the confidence intervals, these calculations are outside the scope of this documentation, but details may be found in Pearson and Hartley (1984) and Guenther (1977). This test is preferred, even though it is not as commonly described in correlation chapters of texts. An assumption of this formulation is that the two variables are characterized by a bivariate Normal distribution.

Normal Approximation Test

An approximate (and more popular) test statistic for the population correlation is based on Fisher's *Z* transformation of the correlation. This transformation has an approximate standard Normal distribution, and thus the standard Normal probabilities may be used to obtain one- or two-sided probabilities (*p*-values). The test statistic is

$$Z = \frac{\frac{1}{2}\,ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\,ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{\frac{1}{n-3}}}$$

Spearman Rank Correlation Details

The Spearman rank correlation (r_S) is calculated by ranking the observations for each of the two variables and then computing the Pearson correlation on the ranks. When ties are encountered, the average rank is used.

Confidence Intervals

The Fisher's Z transformation (Normal approximation) methods are used to produce confidence intervals. One adjustment is made to the variance of *Z*, according the recommendation of Bonett and Wright (2000). The adjustment is to change the variance from $1 / (n-3)$ to $(1 + r_S^2/2) / (n-3)$. It should be noted that these

approximate formulas are suggested to be used only when the Spearman rank correlation is less than 0.9 and when *n* is 10 or greater.

Hypothesis Tests

The formulation of the hypothesis test depends on whether the correlation is to be tested against 0, or against some other value.

Testing H_0: $\rho = 0$

To test whether the Spearman rank correlation is significantly different from 0, the same method as that of the Pearson correlation is used, namely,

$$t = \frac{r_S}{\sqrt{\frac{1 - {r_S}^2}{n-2}}}$$

The *t* distribution with $n-2$ degrees of freedom may be used to generate one- or two-sided test *p*-values from this test statistic.

Testing H_0: $\rho = \rho_{S0}$, where $\rho_{S0} \neq 0$

When testing against a correlation other than zero, the Normal approximation based on Fisher's Z transformation of the correlation is used, but with the variance recommended by Bonett and Wright (2000). The test statistic is

$$Z = \frac{\frac{1}{2}\, ln\left(\frac{1+r_S}{1-r_S}\right) - \frac{1}{2}\, ln\left(\frac{1+\rho_{S_0}}{1-\rho_{S_0}}\right)}{\sqrt{\frac{1+\frac{{r_S}^2}{2}}{n-3}}}$$

Kendall's Tau Correlation Details

Kendall's Tau correlation is also a correlation based the ranks of the observations, and is thus another nonparametric alternative to the Pearson correlation.

Kendall's Tau correlation coefficient is calculated from a sample of *n* data pairs (X, Y) by first creating a variable U as the ranks of X and a variable V as the ranks of Y (ties replaced with average ranks). Kendall's Tau is then calculated from U and V using

$$\tau = \frac{2(n_C - n_D)}{\sqrt{n(n-1) - T_X}\sqrt{n(n-1) - T_Y}}$$

$$T_X = \sum_{i=1}^{S_X} \left(t_{(X)i}^2 - t_{(X)i}\right)$$

$$T_Y = \sum_{i=1}^{S_Y} \left(t_{(Y)i}^2 - t_{(Y)i}\right)$$

The parameter n_C is the total number of concordant pairs and n_D is the total number of discordant pairs. A pair of points (U_i, Y_i) and (U_j, V_j) is said to be **concordant** when either of the following statements is true: 1. ($U_i < U_j$) and ($V_i < V_j$); 2. ($U_i > U_j$) and ($V_i > V_j$). Similarly, a pair of points is said to be **discordant** when either of the following statements is true: 1. ($U_i < U_j$) and ($V_i > V_j$); 2. ($U_i > U_j$) and ($V_i < V_j$). Pairs in which ($U_i = U_j$) or ($V_i = V_j$) are not classified as concordant or discordant and are ignored.

The value of $t_{(X)i}$ is the number of ties in the i[th] set of ties of the X variable. There are S_X sets of ties in the X variable. The value of $t_{(Y)i}$ is the number of ties in the i[th] set of ties of the Y variable. There are S_Y sets of ties in the Y variable.

Confidence Intervals

The Fisher's Z transformation (Normal approximation) methods are used to produce confidence intervals. One adjustment is made to the variance of *Z*, according the recommendation of Fieller, Hartley, and Pearson (1957). The adjustment is to change the variance from 1 / (*n* – 3) to 0.437 / (*n* – 4). It should be noted that these approximate formulas are suggested to be used only when the Kendall's Tau correlation is less than 0.8.

Hypothesis Tests

The formulation of the hypothesis test depends on whether the correlation is to be tested against 0, or against some other value.

Testing H_0: $\tau = 0$

The statistical significance of Kendall's tau is tested using

$$Z = \frac{S + \delta}{\sigma_S}$$

where

$$S = n_C - n_D$$

$$\sigma_S^2 = \frac{(n^2 - n)(2n + 5) - T_X'' - T_Y''}{18} + \frac{T_X' T_Y'}{9(n^2 - n)(n - 2)} + \frac{T_X T_Y}{2(n^2 - n)}$$

$$T_X' = \sum_{i=1}^{S_X} \left(t_{(X)i}^2 - t_{(X)i}\right)\left(t_{(X)i} - 2\right)$$

$$T_X'' = \sum_{i=1}^{S_X} \left(t_{(X)i}^2 - t_{(X)i}\right)\left(2t_{(X)i} + 5\right)$$

$$T_Y' = \sum_{i=1}^{S_X} \left(t_{(Y)i}^2 - t_{(Y)i}\right)\left(t_{(Y)i} - 2\right)$$

$$T_Y'' = \sum_{i=1}^{S_X} \left(t_{(Y)i}^2 - t_{(Y)i}\right)\left(2t_{(Y)i} + 5\right)$$

$$\delta = \begin{cases} -1 \text{ if } S > 0 \\ 1 \text{ if } S < 0 \end{cases}$$

The distribution of *Z* is approximately normal.

Testing H_0: $\tau = \tau_0$, where $\tau_0 \neq 0$

When testing against a correlation other than zero, the Normal approximation based on Fisher's Z transformation of the correlation is used, but with the variance recommended by Fieller, Hartley, and Pearson (1957). The test statistic is

$$Z = \frac{\frac{1}{2}\,ln\left(\frac{1+\tau}{1-\tau}\right) - \frac{1}{2}\,ln\left(\frac{1+\tau_0}{1-\tau_0}\right)}{\sqrt{\frac{0.437}{n-4}}}$$

Data Structure

Correlation data are entered as two columns on the spreadsheet. A frequency column may also be used to supply counts for each row, but a frequency column is not required. An example dataset, consisting of mold spore growth and moisture levels (humidity) for 20 locations, is presented below. The data are contained in the Spore Growth dataset.

Spore Growth dataset

Growth	Moisture
1.8	33
3.3	57
5.4	81
6.2	80
4.5	63
1.8	38
2.2	45
4.5	63
.	.
.	.
.	.

Missing Values

Rows with missing values in either of the analyzed columns are ignored.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Dependent Variable

Y Axis Variable(s)

Specify one or more numeric Y axis variables (columns). If more than one column is specified, a separate analysis is displayed for each column. You may type the column names or numbers directly, or you may use the column selection tool by clicking the column selection button to the right.

X Axis Variable

Specify the numeric X axis variable (column). You may type the column name or number directly, or you may use the column selection tool by clicking the column selection button to the right.

Frequency Variable

Specify an optional frequency (count) variable. This variable contains integers that represent the number of observations (frequency) associated with each observation. If left blank, each observation has a frequency of one. This variable lets you modify that frequency. This is especially useful when your data are already tabulated and you want to enter the counts.

Reports Tab

The following options control which reports and plots are displayed.

Summaries

Run Summary

Check this box to obtain a table of summary information, including the names of the X and Y variables, frequency variable information, and processed rows information.

Column Summary Statistics

Check this box to obtain a report with the count, mean, standard deviation, minimum, and maximum for each of the X and Y axis variables.

Confidence Intervals

Confidence Level

This confidence level is used for all reported confidence intervals of the correlation. Typical confidence levels are 90%, 95%, and 99%, with 95% being the most common.

Limits

Specify whether a two-sided or one-sided confidence interval of the correlation is to be reported.

Two-Sided

For this selection, the lower and upper limits of ρ are reported, giving a confidence interval of the form (Lower Limit, Upper Limit).

One-Sided Upper

For this selection, only an upper limit of ρ is reported, giving a confidence interval of the form (-1, Upper Limit).

One-Sided Lower

For this selection, only a lower limit of ρ is reported, giving a confidence interval of the form (Lower Limit, 1).

Confidence Intervals – Pearson Correlation

Pearson Correlation Confidence Intervals

Check this box to obtain the correlation distribution confidence interval and the Normal approximation confidence interval of the Pearson correlation.

Confidence Intervals – Nonparametric (Rank Correlation)

Spearman Rank Correlation Confidence Interval

Check this box to obtain the Spearman rank correlation confidence interval. This confidence interval is based on Fisher's Z transformation.

Kendall's Tau Correlation Confidence Interval

Check this box to obtain the Kendall's Tau correlation confidence interval. This confidence interval is based on Fisher's Z transformation.

Tests

ANOVA

Indicate whether to display this report.

Alpha

Alpha is the significance level used in the hypothesis tests. A value of 0.05 is most commonly used, but 0.1, 0.025, 0.01, and other values are sometimes used. Typical values range from 0.001 to 0.20.

Alternative Hypothesis Value

This is selection is used to specify the hypothesized value of the population correlation under the null hypothesis. This is the correlation to which the sample correlation will be compared. Specify whether the correlation will be tested against a value of 0, or a specified custom value. The formulas for testing against 0 are different than those used for testing against a non-zero value. See the documentation for details."

Alternative Hypothesis Custom Value

This is the custom hypothesized value of the population correlation under the null hypothesis. This is the correlation to which the sample correlation will be compared.

Alternative Hypothesis Direction

Specify the direction of the alternative hypothesis. It is typically recommended that the direction of the alternative hypothesis be determined before examining the value of the sample correlation.

Tests – Pearson Correlation

Pearson Correlation Test

Check this box to obtain a statistical test of the Pearson correlation. The formulas for testing against 0 are different than those used for testing against a non-zero value. See the documentation for details.

Tests – Nonparametric (Rank Correlation)

Spearman Rank Correlation Test

Check this box to obtain a statistical test of the Spearman rank correlation. The formulas for testing against 0 are different than those used for testing against a non-zero value. See the documentation for details."

Kendall's Tau Correlation Test

Check this box to obtain a statistical test of Kendall's Tau correlation. The formulas for testing against 0 are different than those used for testing against a non-zero value. See the documentation for details.

Report Options Tab

These options specify the number of decimal places shown when the indicated value is displayed in a report. The number of decimal places shown in plots is controlled by the Tick Labels buttons on the Axis Setup window.

Report Options

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Report Options – Decimal Places

Summary Statistics ... P-Value

Specify the number of digits after the decimal point to display on the output of values of this type. Note that this option in no way influences the accuracy with which the calculations are done.

Enter 'All' to display all digits available.

Plots Tab

These options specify which plots are produced as well as the plot format.

Select Plots

Y vs X Plot

Check this box to obtain a scatter plot of Y versus X.

Plot Options

Y vs X Plot Size

This option controls the size of the Y vs X plot.

- **Small**
  Each plot is about 2.5 inches wide.
- **Large**
  Each plot is about 5.5 inches wide.

Example 1 – Pearson Correlation Analysis

This section presents an example of analyzing the Pearson correlation of the Spore Growth dataset. In this example, the Y axis variable will be Growth (Spore Growth), while the X axis variable will be Moisture (Humidity). A confidence interval and a test against a null correlation value of 0 will be obtained.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Correlation window.

1 Open the Spore Growth dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Spore Growth.NCSS**.
- Click **Open**.

2 Open the Correlation window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Correlation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Correlation window, select the **Variables tab**.
- Set the **Y Axis Variable** box to **Growth**.
- Set the **X Axis Variable** box to **Moisture**.

4 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Run Summary Section

Parameter	Value	Parameter	Value
Y Axis Variable	Growth	Rows Processed	20
X Axis Variable	Moisture	Rows Used in Estimation	20
Frequency Variable	None	Rows with X Missing	0
Sum of Frequencies	20	Rows with Freq Missing	0

This report summarizes the variables, frequencies, and processed rows.

Column Summary Section

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
Growth	20	3.75	1.62	1.10	7.40
Moisture	20	58.40	14.82	33.00	81.00

This report presents the count, mean, standard deviation, minimum, and maximum of the two variables.

Pearson Correlation Confidence Interval Section

Pearson Correlation Confidence Interval Section
Two-Sided Confidence Interval of ρ

Pearson Correlation	Count	R Distribution 95% Confidence Limits Lower	R Distribution 95% Confidence Limits Upper	Normal Approximation 95% Confidence Limits Lower	Normal Approximation 95% Confidence Limits Upper
0.8763	20	0.6991	0.9470	0.7085	0.9503

The estimated Pearson correlation is 0.8763. This report presents two sets of confidence limits: one based on the exact correlation distribution, and the other based on a Normal approximation using Fisher's Z-transformation. The first is the more accurate of the two, while the second is more commonly used due to its ease in calculation. Calculation details are given in the Pearson Correlation Details section earlier in this chapter of the documentation.

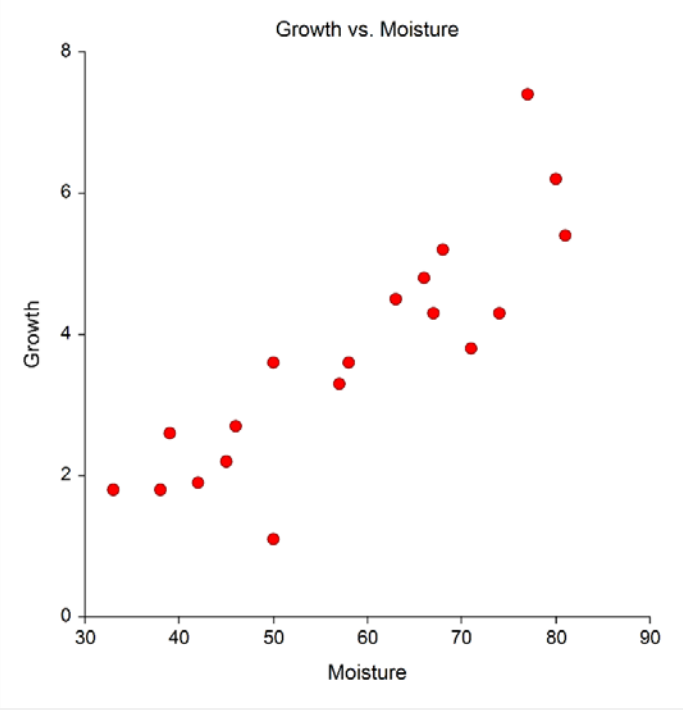
Pearson Correlation Test Section

Pearson Correlation Test Section
H0: ρ = 0

Alternative Hypothesis	Pearson Correlation	Count	df	T-Value	P-Value	Reject H0 at α = 0.05?
ρ ≠ 0	0.8763	20	18	7.7167	0.0000	Yes

This report gives the two-sided statistical test to determine whether the correlation is different from 0. The very small P-value indicates strong evidence that the correlation is not 0.

Scatter Plot Section



The plot shows the visual representation of the X and Y axis data. This plot may be useful for finding outliers and nonlinearities.

Example 2 – Testing Against a Nonzero Null Hypothesis Value

This section presents an example of testing whether the correlation between spore growth and humidity is greater than 0.5. In this example, the Y axis variable will be Growth (Spore Growth), while the X axis variable will be Moisture (Humidity). The dataset used is the Spore Growth dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 2** by clicking on Open Example Template from the File menu of the Correlation window.

1 Open the Spore Growth dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Spore Growth.NCSS**.
- Click **Open**.

2 Open the Correlation window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Correlation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Correlation window, select the **Variables tab**.
- Set the **Y Axis Variable** box to **Growth**.
- Set the **X Axis Variable** box to **Moisture**.

4 Specify the reports.

- On the Correlation window, select the **Reports tab**.
- Uncheck the Run Summary and Column Summary Statistics checkboxes.
- Uncheck the Pearson Correlation Confidence Intervals checkbox.
- Set **H0: ρ =** to **Custom**.
- Set **Custom** to **0.5**.
- Set **Ha** to **ρ > Custom Value (one-sided)**.

5 Specify the plot.

- On the Correlation window, select the **Plots tab**.
- Uncheck the Y vs X plot checkbox.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Pearson Correlation Test Section

Pearson Correlation Test Section
H0: ρ = 0.5

			Correlation		------- Normal Approximation -------		
Alternative Hypothesis	Pearson Correlation	Count	Distribution P-Value	Reject H0 at α = 0.05?	Z-Value	P-Value	Reject H0 at α = 0.05?
ρ > 0.5	0.8763	20	0.0007	Yes	3.3407	0.0004	Yes

This report gives two one-sided statistical tests to determine whether the correlation is statistically greater than 0.5. Both statistical tests exhibit strong evidence that the true correlation is greater than 0.5.

Example 3 – Rank Correlation Coefficients

This section presents an example of analyzing the correlation of spore growth and humidity based on the Spearman rank correlation and Kendall's Tau. In this example, the Y axis variable will be Growth (Spore Growth), while the X axis variable will be Moisture (Humidity). The dataset used is the Spore Growth dataset.

You may follow along here by making the appropriate entries or load the completed template **Example 3** by clicking on Open Example Template from the File menu of the Correlation window.

1 Open the Spore Growth dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Spore Growth.NCSS**.
- Click **Open**.

2 Open the Correlation window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Correlation** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Correlation window, select the **Variables tab**.
- Set the **Y Axis Variable** box to **Growth**.
- Set the **X Axis Variable** box to **Moisture**.

4 Specify the reports.

- On the Correlation window, select the **Reports tab**.
- Uncheck the Run Summary and Column Summary Statistics checkboxes.
- Uncheck the Pearson Correlation Confidence Intervals checkbox.
- Check the **Spearman Rank Correlation Confidence Interval** checkbox.
- Check the **Kendall's Tau Correlation Confidence Interval** checkbox.
- Uncheck the Pearson Correlation Test checkbox.
- Check the **Spearman Rank Correlation Test** checkbox.
- Check the **Kendall's Tau Correlation Test** checkbox.

5 Specify the plot.

- On the Correlation window, select the **Plots tab**.
- Uncheck the Y vs X plot checkbox.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Spearman Rank Correlation Confidence Interval Section

Spearman Rank Correlation Confidence Interval Section
Two-Sided Confidence Interval

Spearman Correlation	Count	Normal Approximation 95% Confidence Limits Lower	Upper
0.8896	20	0.6954	0.9627

The estimated Spearman rank correlation is 0.8896. Calculation details are given in the Spearman Rank Correlation Details section earlier in this chapter of the documentation.

Kendall's Tau Correlation Confidence Interval Section

Kendall's Tau Correlation Confidence Interval Section
Two-Sided Confidence Interval

Kendall's Tau Correlation	Count	Normal Approximation 95% Confidence Limits Lower	Upper
0.7326	20	0.5445	0.8506

The estimated Kendall's Tau correlation is 0.7326. Calculation details are given in the Kendall's Tau Correlation Details section earlier in this chapter of the documentation.

Spearman Rank Correlation Test Section

Spearman Rank Correlation Test Section
H0: ρ = 0

Alternative Hypothesis	Spearman Correlation	Count	df	T-Value	P-Value	Reject H0 at α = 0.05?
ρ ≠ 0	0.8896	20	18	8.2635	0.0000	Yes

This report gives the two-sided statistical test to determine whether the Spearman rank correlation is different from 0. The very small P-value indicates strong evidence that the Spearman correlation is not 0. Calculation details are given in the Spearman Rank Correlation Details section earlier in this chapter of the documentation.

Kendall's Tau Correlation Test Section

Kendall's Tau Correlation Test Section
H0: ρ = 0

			Normal Approximation		
Alternative Hypothesis	Kendall's Tau Correlation	Count	Z-Value	P-Value	Reject H0 at α = 0.05?
ρ ≠ 0	0.7326	20	4.4263	0.0000	Yes

This report gives the two-sided statistical test to determine whether the Kendall's Tau correlation is different from 0. The very small P-value indicates strong evidence that the Kendall's Tau correlation is not 0. Calculation details are given in the Kendall's Tau Correlation Details section earlier in this chapter of the documentation.