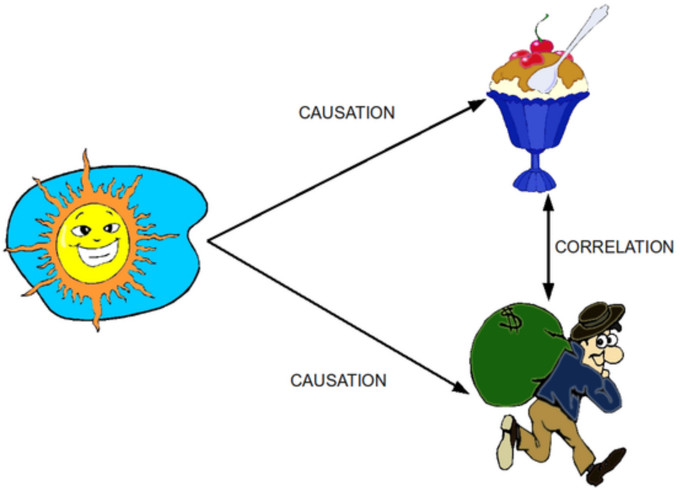
Correlation

EN5423 | Spring 2024

w15_correlation.pdf
(Week 15)

Contents



Intro

Concentrations of atrazine and nitrate in shallow groundwaters are measured in wells over an area of several counties. For each sample, the concentration of atrazine is plotted versus the concentration of nitrate. As atrazine concentrations increase, so do nitrate. How might the strength of this association be measured and summarized? Can nitrate concentrations be used to predict atrazine concentrations?

Streams draining the Sierra Nevada in California usually receive less precipitation in November than in other months. Has the amount of November precipitation gradually changed over the past 70 years?

Streamflow observations at two streamgages appear to respond similarly to precipitation events over the same period of time. If one streamgage has more observations than the other, can the observations from that streamgage be used to fill in portions of the streamflow record missing from the other streamgage?

These examples require a ***measure of the strength of association between two continuous variables***. Correlation coefficients are one class of measures that can be used to determine this association. ***Three correlation coefficients*** are discussed in this chapter. Also discussed is how the ***significance of an association*** can be tested to determine whether the observed pattern differs from what is expected entirely owing to chance. For measurements of correlation between noncontinuous or grouped variables, see chapter 14 on USGS SMWR.pdf.

Whenever a correlation coefficient is calculated, the data ***should first be plotted on a scatterplot***. No single numerical measure can substitute for the visual insight gained from a plot. Many different patterns can produce the same correlation coefficient, and similar strengths of relations can produce differing coefficients depending on the curvature of the relation. Recall that in figure 1 in week 02, eight plots showed the relation between two variables, all with a linear correlation coefficient of 0.70; yet the data were radically different! ***It is important to never compute correlation coefficients without plotting the data first.***

1 Characteristics of Correlation Coefficients

• Correlation coefficients measure the strength of ***association*** between ***two continuous variables***. Of interest is whether one variable generally increases as the second increases, whether it decreases as the second increases, or whether their patterns of variation are totally unrelated.

• Correlation measures observed covariation between two variables; that is, how one varies by the other. It does not provide evidence for ***causal relation between the two variables***. A change in one variable may cause change in the other, for example, precipitation changes cause runoff changes.

• Two variables may also be correlated because they share the same cause, for example, changes to concentrations of two constituents measured at a variety of locations are caused by variations in the quantity or source of the water. In trend analysis, correlation is used to measure the change in one variable respective to time.

• Evidence for ***causation must come from outside the statistical analysis***, through knowledge of the processes involved. Measures of correlation have the characteristic of being dimensionless and scaled to lie between values of -1 and 1. When there is no correlation between two variables, correlation is equal to ***zero***. When one variable increases as the second increases, correlation is ***positive***. When the variables vary together but in opposite directions, correlation is ***negative***.

• When using a two-sided test, the following statements about the null, H_0, and alternative hypotheses, H_A, are equivalent:

H_0: No correlation exists between x and y (correlation = 0), or x and y are independent.

H_A: x and y are correlated (correlation $\neq$ 0), or x and y are dependent.

1.1 Characteristics of Correlation Coefficients

• Data may be correlated in either a ***linear*** or ***nonlinear*** fashion. When y generally increases or decreases as x increases, the two variables are defined as possessing a monotonic correlation.

• This correlation may be nonlinear; for example, when plotted they have exponential patterns, linear patterns, or patterns similar to power functions when both variables are nonnegative.

• A special case of monotonic correlation is linear correlation, where a plot of y versus x has a linear pattern. ***Three measures of correlation*** are in common use—product moment or **Pearson's r**, **Spearman's rho (ρ)**, and **Kendall's tau (τ)**.

• The more commonly used Pearson's r is a measure of linear correlation, whereas Spearman's ρ and Kendall's τ measure ***monotonic correlation***. The last two correlation coefficients are based on ranks and measure monotonic relations such as that in **Figure 1**. These two metrics are also resistant to the effects of outliers because they are rank-based.

• Pearson's r is only appropriate when plots of x and y indicate a linear relation between the two variables, such as shown in **Figure 2**.

• ***None of the measures*** are appropriate to assess nonmonotonic relations where the pattern doubles back on itself, like that in **Figure 3**. A monotonic, but not linear, association between two variables is illustrated in **Figure 1**.

• If the Pearson's r correlation coefficient were calculated to measure the strength of the association between the two variables, the nonlinearity would result in a low value that would not reflect the strong association between the two variables that is apparent in the figure.

• This illustrates ***the importance of plotting the data before deciding which correlation measure would appropriately represen***t the relation between data.

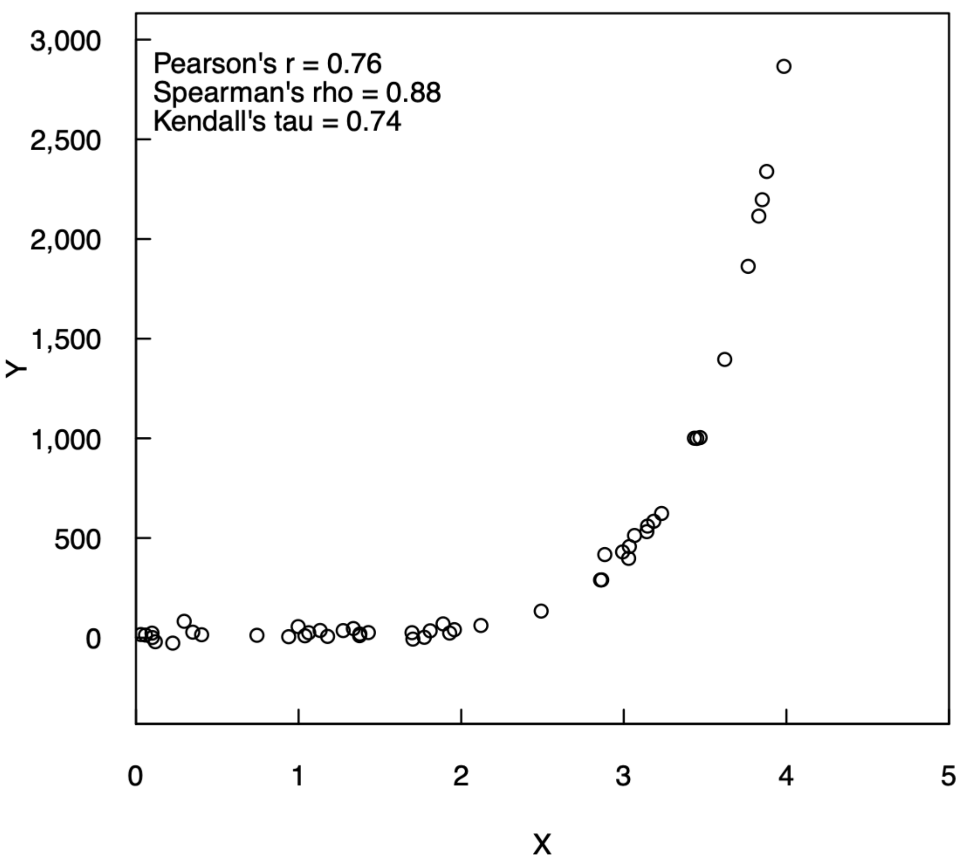


Figure 1. Plot showing monotonic, but nonlinear, correlation between *x* and *y*.

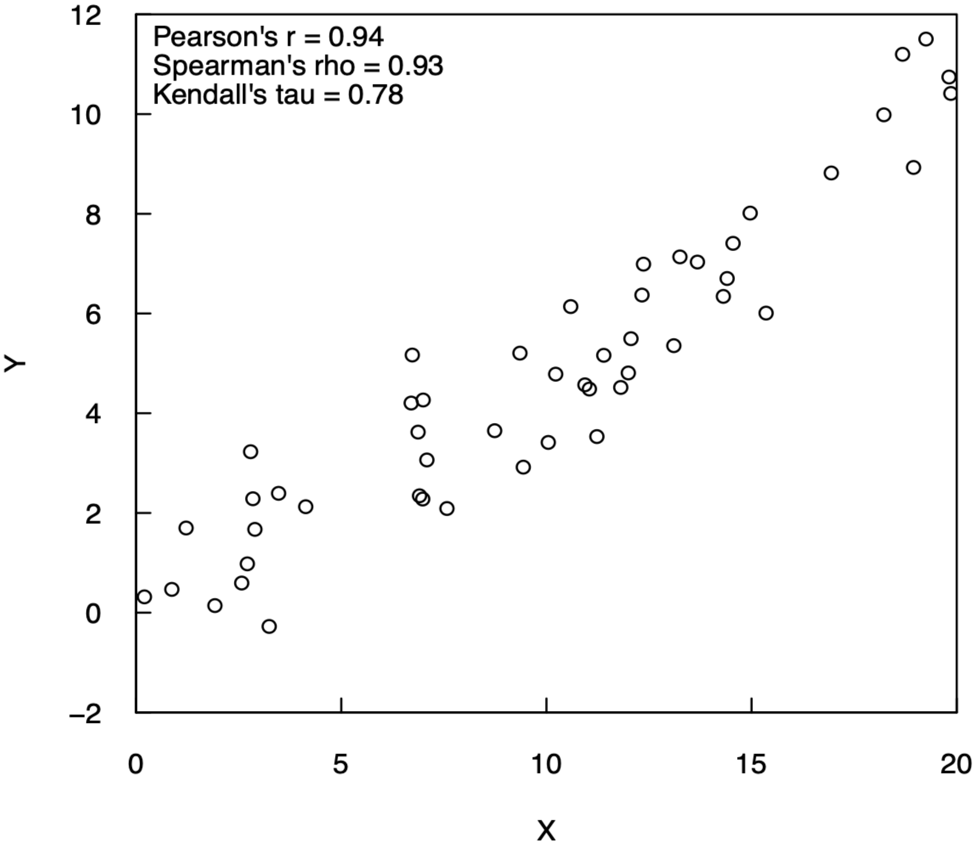


Figure 2. Plot showing monotonic linear correlation between *x* and *y*.

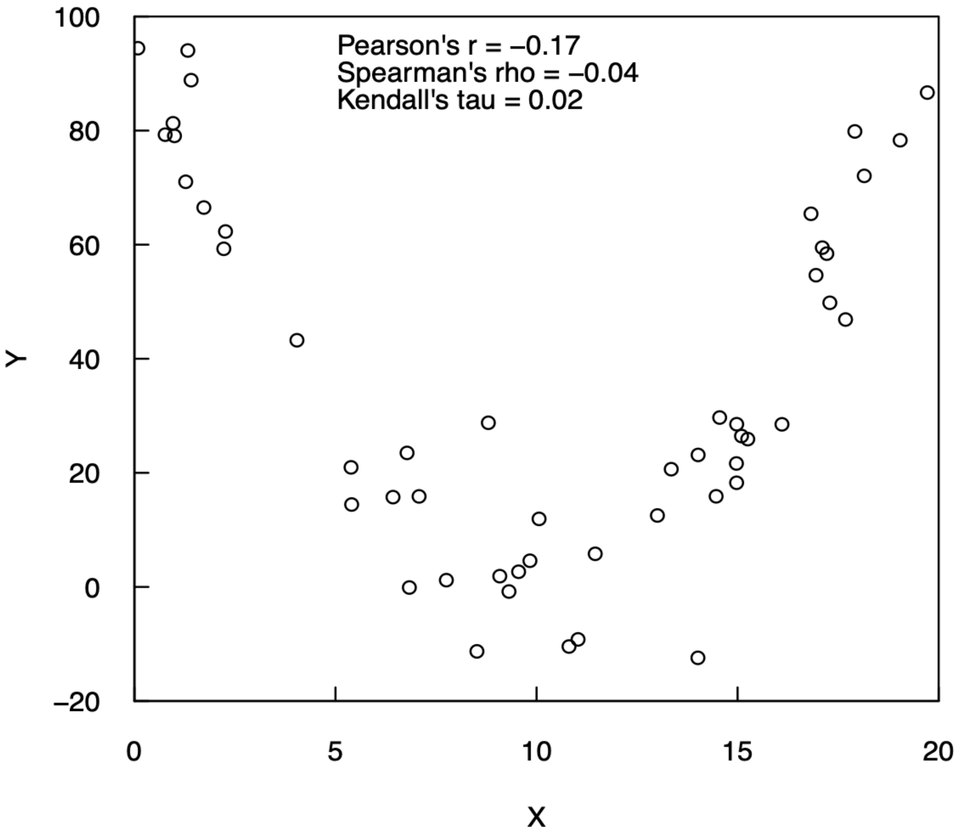


Figure 3. Plot showing nonmonotonic relation between *x* and *y*.

2 Characteristics of Correlation Coefficients

• Pearson's *r* is the most commonly used measure of correlation and sometimes called the ***linear correlation coefficient*** because *r* measures the linear association between two variables.

• If the data lie exactly along a straight line with a positive slope then $r = 1$; if the straight line has a negative slope then $r = -1$.

• When considering the use of Pearson's *r*, this assumption of linearity makes inspection of a plot even more important for *r* than for other correlation metrics, because a small value of Pearson's *r* may be the result of curvature or outliers.

• As in **Figure 1**, *x* and *y* may be strongly related in a nonlinear fashion and the value of the Pearson's *r* measure may not be statistically significant.

2.1 Characteristics of Correlation Coefficients

Pearson's *r* is computed from equation 1:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) \qquad \textbf{Eq. (1)}$$

where

n is the number of observations;

$\bar{x}$ and $\bar{y}$ are the means of *x* and *y*, respectively;

s_x and s_y are the standard deviations of *x* and *y*, respectively; and

r is a dimensionless value that is not affected by scale changes in the *x* and *y* observations, for example, converting streamflows in cubic feet per second into cubic meters per second.

This dimensionless property results from dividing by s_x and s_y, the sample standard deviations of the *x* and *y* variables, respectively (eq. 1).

2.2 Hypothesis Tests

• Pearson's r is not resistant to outliers because it is computed by using ***nonresistant measures—means and standard deviations.***

• Pearson's *r* also assumes that the variability in *y* cannot increase (or decrease) with increasing *x*. In linear regression (later weeks), variables with the property of having constant variability in *y* with increasing *x* are said to be ***homoscedastic***.

• Skewed variables often demonstrate outliers and increasing variance; thus *r* is often not useful for describing the correlation between skewed hydrologic variables.

• Transforming the data to reduce skewness and linearize the relation between *x* and *y* in order to compute Pearson's *r* is a ***common practice that is often used in hydrologic data analysis*** and is explored further in later section.

• If these assumptions are met, the statistical significance of *r* can be tested under the null hypothesis that *r* is not significantly different from zero (that is, there is no correlation) or, in terms of the null and alternate hypothesis, $H_0: r = 0$ or $H_A: r \neq 0$.

• The test statistic t_r is computed by equation 2 and compared to a table of the *t*-distribution with *n* - 2 degrees of freedom.

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad \textbf{Eq. (2)}$$

Example 1: Pearson's *r*

Sample	**x**	**y**
1	2	1.22
2	24	2.20
3	99	4.80
4	197	1.28
5	377	1.97
6	544	1.46
7	632	2.64

8	3452	2.34
9	6587	4.84
10	8271	2.96

- Compute the means and standard deviations of x and y:

np.mean(x) = 2018.5
np.std(x) = 3052.459
np.mean(y) = 2.571
np.std(y) = 113145

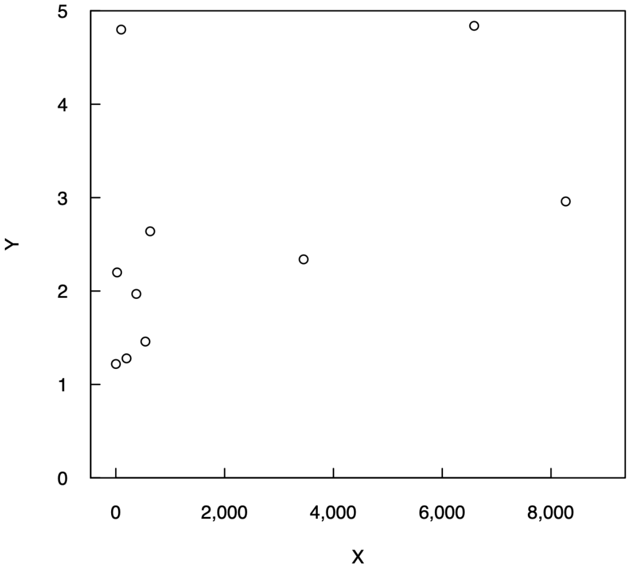


Figure 5. Plot of example 8.1 data showing one outlier present (the third value in the dataset (99, 4.80)).

These values can be used with equation 1 to compute Pearson's r,

$$r = \frac{1}{9}\sum_{i=1}^{n}\left(\frac{x_i - 2018.5}{3052.459}\right)\left(\frac{y_i - 2.571}{1.3145}\right) = 0.457$$

- To test for whether r is significantly different from zero, and therefore y is linearly dependent on x,

$$t_r = \frac{0.4578309\sqrt{(10-2)}}{\sqrt{1-(0.4578309)^2}} = 1.456563$$

with a p-value of 0.18 from a table of the t-distribution. If our alpha was selected as 0.1, H_0: $r = 0$ is not rejected because the p-value is > 0.1, and, therefore we should conclude that y is not linearly dependent (or related) to x.

3 Spearman's Rho (ρ)

• Spearman's ρ is a nonparametric, rank-based correlation coefficient that depends only on the ranks of the data and not the observations themselves. Therefore, ρ is ***resistant to outliers*** and can be implemented even in cases where some of the data are censored, such as concentrations known only as less than an analytical detection limit.

• These properties are important features for applications to water resources. With ρ, differences between data ranked further apart are given more weight, ***similar to the signed-rank test discussed in previous week***; ρ is perhaps easiest to understand then as the linear correlation coefficient computed on the ranks of the data rather than the data themselves.

3.1 Computation of Spearman's ρ

To compute ρ, the data for the two variables are separately ranked from smallest to largest. Ties in *x* or *y* are initially assigned a unique rank. The average rank is then computed from these unique ranks and assigned to each of the tied observations, replacing the unique ranks. Using the ranks of *x* and ranks of *y*, ρ can be computed from the equation:

$$\rho = \frac{\sum_{i=1}^{n} (Rx_i Ry_i) - n\left(\frac{n+1}{2}\right)^2}{n(n^2-1)/12} \qquad \textbf{Eq. (3)}$$

where R_{x_i} is the rank of x_i, R_{y_i} is the rank of y_i, and (n + 1) / 2 is the mean rank of both *x* and *y*. This equation can be derived from substituting R_{x_i} and R_{y_i} for x_i and y_i in the equation for Pearson's *r* (eq. 1) and simplifying.

• If there is a positive correlation, the higher ranks of *x* will be paired with the higher ranks of *y*, and their product will be large.

• For a negative correlation, the higher ranks of *x* will be paired with lower ranks of *y*, and their product will be small. When there is no correlation there will be nothing other than a random pattern in the association between *x* and *y* ranks, and their product will be similar to the product of their average rank, the second term in the numerator of equation 3. Thus, ρ will be close to zero.

3.2 Hypothesis Tests for Spearman's ρ

To compute the test statistic, *S*, for the significance of the ρ value, the rank transform method is used. The values for each variable are ranked separately and the Pearson's *r* correlation is computed from the ranks. The test statistic, *S*, is then given by equation 4

$$S = \sum_{i=1}^{n} (Rx_i - Ry_i)^2 \qquad \textbf{Eq. (4)}$$

where Rx_i is the rank of x_i, and Ry_i is the rank of y_i. The statistical significance of ρ can be tested under the null hypothesis that ρ is not significantly different from zero (that is, there is no correlation) or, in terms of the null and alternate hypothesis, $H_0: \rho = 0$ or $H_A: \rho \neq 0$.

- For large sample sizes (n > 20), S follows a t-distribution with n - 2 degrees of freedom (the same distribution as the Pearson r test statistic). However, for small sample sizes (n < 20), the rank-transformed test statistic does not fit the distribution of the Pearson's r test statistic well.

- There has been some work to define the exact probabilities associated with ρ values for small sample sizes (see Franklin [1988]) and Maciak [2009] as examples).

4 Kendall's Tau (τ)

- Kendall's τ (Kendall, 1938, 1975), much like Spearman's ρ, measures the strength of the monotonic relation between x and y and is a rank-based procedure.

- Just as with ρ, τ is resistant to the effect of outliers and, because τ also depends only on the ranks of the data and not the observations themselves, it can be implemented even in cases where some of the data are categorical, such as censored observations (for example, observations stated as less than a reporting limit for concentrations or less than a perception threshold for floods).

- Despite these similar properties, ρ and τ use different scales to measure the same correlation, much like the Celsius and Fahrenheit measures of temperature. Though τ is generally lower than ρ in magnitude, their p-values for significance should be quite similar when computed on the same data.

- In general, τ will be lower than values of r for linear associations for any given linearly related data (see fig. 2). Strong linear correlations of r = 0.9 (or above) typically correspond to τ values of about 0.7 (or above). These lower values do not mean that τ is less sensitive than r, but simply that a different scale of correlation is being used.

- As it is a rank correlation method, τ is unaffected by monotonic power transformations of one or both variables. For example, τ for the correlation of log(y) versus log(x) will be identical to that of y versus log(x), and of y versus x.

4.1 Computation of Kendall's τ

- Kendall's τ examines every possible pair of data points, (x_i, y_i) and (x_j, y_j), to determine if the pairs have the same relation to one another—that is, if x_i is greater than y_i and x_j is greater than y_j, or if x_i is less than y_i and x_j is less than y_j.

- Each pair is assessed in this way, keeping track of the number of pairs that have the same relation to one another versus the number of pairs that do not.

- Kendall's τ is most easily computed by ordering all data pairs by increasing x. If a positive correlation exists, the y observations will increase more often than decrease as x increases. For a negative correlation, the y observations will decrease more often than increase as x increases.

- If no correlation exists, the *y* observations will increase and decrease about the same number of times. Kendall's τ is related to the sign test in that positive differences between data pairs are assigned +1 without regard to the magnitude of those differences and negative differences are assigned −1.

- The calculation of τ begins with the calculation of Kendall's *S*, the test statistic. Kendall's *S* measures the monotonic dependence of *y* on *x* and the formula is

$$S = P - M \qquad \textbf{Eq. (5)}$$

The *S* statistic ***is simply the number of concordant pairs*** (denoted as *P*) minus the number of discordant pairs (denoted as *M*). We can think of this conveniently as "*P* for plus" when the slope between the two points is a positive value. We can think of "*M* for minus" when the slope between the two points is a minus value.

- A ***concordant pair*** is a pair of observations where the difference between the *y* observations is of the same sign as the difference between the *x* observations. A ***discordant pair*** is a pair of observations where the difference in the *y* observations and the difference in the *x* observations is of the opposite sign.

- The computation can be simplified by rearranging the data pairs, placing the *n* observations in order based on the *x* observations with x_1 being the smallest *x* to x_n being the largest *x*.

- After this rearrangement, we consider all pairwise comparisons of the *y* observations, where the pairs are sorted by their *x* rank. If we compare (x_i, y_i) to (x_j, y_j) where $i < j$, then a concordant pair is the case where $y_i < y_j$ and a discordant pair is the case where $y_i > y_j$.

- Note that there are $n \cdot \frac{(n-1)}{2}$ possible comparisons to be made among the *n* data pairs. If all *y* observations increased along with the *x* observations, $S = n \cdot \frac{(n-1)}{2}$. In this situation, the Kendall's τ correlation coefficient should equal +1.

- When all *y* observations decrease with increasing *x*, $S = -n \cdot \frac{(n-1)}{2}$ and Kendall's τ should equal −1. Therefore, dividing S by $n \cdot \frac{(n-1)}{2}$ will give a value always falling between −1 and +1. This is the definition of Kendall's τ, which measures the strength of the monotonic association between two variables.

$$\tau = \frac{S}{n(n-1)/2} \qquad \textbf{Eq. (6)}$$

4.2 Hypothesis Tests for Kendall's τ

- To test for the significance of τ, *S* is compared to what would be expected when the null hypothesis is true for a given n.

For a two-sided test, $H_0: \tau = 0$, or $H_A: \tau \neq 0$. If τ is further from 0 than expected, H_0 is rejected.

• When $n \leq 10$, the table of exact p-values using the S and n values should be used; this is because the distribution of S for a given n at small sample sizes is not easily approximated.

• Such a table can be found in Hollander and Wolfe (1999). When $n > 10$, a large-sample approximation can be used because the test statistic Z_s (a rescaled version of S) closely approximates a normal distribution with mean, μ_s, equal to zero and variance, σ_S. For the large-sample approximation,

$$Z_s = \begin{cases} \dfrac{S-1}{\sigma_S} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \dfrac{S+1}{\sigma_S} & \text{if } S < 0 \end{cases} \qquad \textbf{Eq. (7)}$$

where $\sigma_s = \sqrt{(n/18) \cdot (n-1) \cdot (2n+5)}$, n is the number of samples and S is defined in equation 5.

• Here the null hypothesis is rejected at significance level α if $| \ Z_s \ |> Z_{crit}$, where Z_{crit} is the value of the standard normal distribution with a probability of exceedance of $\alpha/2$ (for a two-sided test).

• Recall that α is selected by the user and is the probability at which they think the null hypothesis can be rejected. Just as when computing the test statistic for the rank-sum test, a continuity correction must be applied.

• This is reflected in equation 7 by the -1 or +1

• The Kendall package is most useful in the case where some of the *x* or *y* observations are tied, which requires additional modification of the test statistic.

4.3 Correction for Tied Data when Performing Hypothesis Testing Using Kendall's τ

• When tied observations of either *x* or *y* are present in the data, they will produce a 0 rather than + or − when counting the number of *P*'s and *M*'s.

• If the ties are not accounted for when determining the significance of τ, the variability of S (represented by σ_S) will be an overestimate of the actual σ_S and an underestimate of the test statistic. Therefore, an adjustment is needed for σ_S in the equation for the test statistic Z_s to account for the presence of ties in the data. Details of the adjustment to σ_S can be found in Kendall (1975). This adjustment is only applicable to the large-sample approximation.

Note:

When to Use Kendall's τ:

1. **Smaller Sample Sizes**:
   - Kendall's τ is more accurate for smaller sample sizes because it provides an exact p-value for small samples. Spearman's ρ approximates the p-value using a normal distribution which may not be as accurate for small datasets.
2. **Data with Ties and Censored Data**:
   - Kendall's τ handles ties and censored data (e.g., data below a detection limit) better than Spearman's ρ. Kendall's τ test can be applied even when the data contains tied ranks, making it more robust in such scenarios.
3. **Sensitivity to Concordant and Discordant Pairs**:
   - Kendall's τ is based on the number of concordant and discordant pairs. It directly measures the probability of observing concordance or discordance in the ranks, which can be more interpretable in terms of probability.
4. **When Monotonicity is Important**:
   - Kendall's τ specifically measures the strength of monotonic relationships (i.e., relationships that are consistently increasing or decreasing). If you are interested in whether one variable consistently increases as another increases (or consistently decreases), Kendall's τ is more appropriate.

When to Use Spearman's ρ:

1. **Larger Sample Sizes**:
   - Spearman's ρ is computationally simpler and works well for larger sample sizes. It approximates the p-value using the t-distribution, which becomes more accurate with larger datasets.
2. **Linearity and Rank-based Approximations**:
   - Spearman's ρ is better suited for detecting linear relationships on the ranks of the data. If the relationship between the variables is linear (after rank transformation), Spearman's ρ might be more powerful.
3. **General Rank Correlation**:
   - Spearman's ρ is often used as a general measure of rank correlation. It is widely known and used, making it a default choice in many applications where rank-based correlation is needed.