APPENDIX O
Measures of Correlation

O-1. Introduction. A correlation coefficient provides a measure of the degree of association between two variables or measurements. For example, the degree of association between pH and the concentration of a dissolved metal in groundwater may be of interest. The primary objective of calculating a correlation coefficient is to determine whether one variable increases or decreases as the second variable increases, or whether the two variables vary independently of one another.

O-1.1. In environmental applications, a correlation coefficient may be used to determine the strength of an association. For example, numerous groundwater sites contaminated with chlorinated solvents also have high dissolved iron concentrations. Is it possible to determine whether the high iron locations are the same as where chlorinated solvent levels are also high? A correlation coefficient for the relationship provides a quantitative measure of the degree of association of these measured parameters.

O-1.2. A high correlation coefficient does not prove cause and effect. When the correlation between two variables is high, the relationship is strong; but one cannot conclude that one variable causes the other variable to increase or decrease without further evidence. Measuring and identifying correlation is often critical for environmental data, which are frequently correlated over time or space, or both.

O-1.3. Classical statistical methods typically assume data are not correlated. If correlations are not identified before data are statistically evaluated, then statistical methods can provide misleading results. There are also statistics that depend upon correlation in the data, such as geostatistics (Appendix R), and there are methods available for "detrending" or "uncorrelating" data under certain circumstances. These cases are beyond the scope of this discussion, and may be best addressed by a statistician.

O-1.4. Several different correlation coefficients for measuring the degree of association between two variables will be discussed. The correlation coefficients share common properties. Each is a dimensionless quantity with values ranging from –1 to 1. A positive correlation coefficient for two variables indicates that one variable tends to increase as the other variable increases. A negative correlation indicates that one variable tends to decrease as the other variable increases. The highest possible degree of correlation occurs when the absolute value of the correlation coefficient equals one. When two variables are truly independent, the behavior of one variable cannot be predicted from the other variable, and the correlation coefficient is zero. The references EPA 600/R-96/084, QA/G-9 and Conover (1980) contain additional details about measures of correlation.

O-2. Correlation Coefficients as Hypothesis Testing.

O-2.1. *Introduction*. Calculated values of a correlation coefficient for a set of actual measurements are rarely identically equal to zero when a true correlation is absent (when the true correlation coefficient $\gamma = 0$). Therefore, a hypothesis test is done to determine the presence or absence of a significant correlation. Hypothesis tests are discussed in additional detail in Appendices L, M, and N.

O-2.1.1. The significance of the correlation is often evaluated using a hypothesis test in the form:

H_0: $\gamma = 0$, H_A: $\gamma \neq 0$.

O-2.1.1.1. The correlation coefficient for a set of measured results (x_i, y_i) is initially calculated. The calculated (sample) correlation coefficient, $\hat{\gamma}$, is viewed as an approximation of the population correlation coefficient, γ, for the X and Y variables.

O-2.1.1.2. The probability, p, of obtaining the calculated value when X and Y are not correlated (when the true correlation coefficient $\gamma = 0$) is then determined. The probability is typically calculated by statistical software.

O-2.1.1.3. If p is sufficiently small (e.g., $p \leq \alpha = 0.05$ or 0.01), then a correlation exists. More accurately, the null hypothesis that the true correlation coefficient is zero is rejected (with a level of confidence of at least $1 - \alpha$).

O-2.1.1.4. When statistical software is unavailable, the largest possible absolute value of a correlation coefficient that can occur when X and Y are not correlated is obtained from a table. The tabular value for the $1 - \alpha$ level of confidence is subsequently compared to the calculated value. If the calculated value is larger than the value obtained from the table, the null hypothesis is rejected, and the correlation coefficient is not equal to zero.

O-2.1.2. Directions and an example for using a correlation coefficient statistical test are in Paragraphs O-2.2 and O-2.3, respectively.

O-2.1.3. Typically, a correlation coefficient is viewed to be significantly different from zero if the p value is less than a specified significance level, usually taken to be between 0.1 and 0.01. The p value is discussed in more detail in Appendices L, M, and N. Various values for the absolute value of the correlation coefficient, $|\gamma|$, qualitatively describe the degree of association below:

Absolute value of correlation coefficient	Degree of relationship
$\lvert\gamma\rvert < 0.50$	Extremely Weak
$0.50 < \lvert\gamma\rvert < 0.75$	Weak
$0.75 < \lvert\gamma\rvert < 0.90$	Moderate
$0.90 < \lvert\gamma\rvert < 0.95$	Moderately Strong
$0.95 < \lvert\gamma\rvert < 1.00$	Strong

O-2.1.4. Four different sample correlation coefficients are discussed below.

O-2.1.4.1. Pearson's *r*.

O-2.1.4.2. Spearman's rho (ρ).

O-2.1.4.3. Serial correlation coefficient.

O-2.1.4.4. Kendall's tau (τ).

O-2.1.5. Pearson's *r* measures the degree of correlation between two variables for linear relationships. Kendall's τ and Spearman's ρ measure the degree of any monotonic relationship between two variables. Two variables, *X* and *Y*, are monotonically correlated if, overall, *Y* consistently increases or decreases as *X* increases. Note that *X* and *Y* will not be monotonically correlated if, as *X* increases, *Y* increases then decreases (or decreases then increases).

O-2.2. *Directions for a Correlation Coefficient Statistical Test*. Calculate the test statistic:

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

O-2.2.1. Use Table B-23 of Appendix B to find the critical value $t_{1-\alpha/2,\nu}$, which is (1 – α/2)100th percentile of the Student's *t* distribution with degrees of freedom $\nu = n-2$.

O-2.2.1.1. Conclude that the correlation is significantly different from zero if

$$\lvert t \rvert > t_{1-\alpha/2,\nu}.$$

O-2.2.1.2. Otherwise, state that there is insufficient evidence to conclude that the correlation coefficient is different from zero.

O-2.2.2. A one-tailed test can be performed in a similar manner by replacing $\alpha/2$ by α. For example, to test whether a correlation exceeds zero, compare t with $t_{1-\alpha,n-2}$. If $t > t_{1-\alpha,n-2}$ conclude that the correlation is larger than zero. Otherwise conclude that the true correlation may be less than or equal to zero.

O-2.3. *Example of a Test for a Correlation Coefficient*. Consider the following data set for chromium and lead in subsurface soil background (in mg/kg).

Sample	**Chromium (*X*)**	**Lead (*Y*)**
EPC-BG01	4.60	3.50
EPC-BG02	5.29	4.16
EPC-BG03	4.26	4.19
EPC-BG04	5.28	3.91
EPC-BG05	4.53	3.66
EPC-BG06	5.74	4.31
EPC-BG07	5.86	4.19
EPC-BG08	3.84	3.35

O-2.3.1. The objective is to test if the correlation coefficient is different from zero, based on 90% level of confidence.

O-2.3.2. For 90% confidence, $\alpha = 0.10$.

O-2.3.3. The correlation coefficient was calculated in Paragraph O-2.4.2 and equals $r = 0.72$.

O-2.3.4. The test statistic is

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.7229}{\sqrt{\frac{1-(0.7229)^2}{8-2}}} = 2.563$$

with $\nu = 8 - 2 = 6.$

O-2.3.5. The critical value is $t_{1-\alpha/2,n-2} = t_{0.95,6} = 1.943$.

O-2.3.6. Comparing the test statistic to the critical value, $t = 2.563 > 1.943$. With at least 90% confidence, the correlation coefficient is significantly different from zero. However, given the magnitude of r, the linear association between chromium and lead could be qualitatively described as "weak."

O-2.4. *Pearson's* r**.** The Pearson's r is a parametric measure of correlation for *linear* relationship between two variables. A linear association implies that, as one variable increases, so does the other in a uniform manner (i.e., linearly), or as one variable decreases the other increases linearly. A value of +1 implies a perfect positive linear correlation, i.e., that all the data pairs (x_i, y_i) lie on a straight line with a positive slope. A value of –1 implies perfect negative linear correlation. Directions and an example for Pearson's correlation coefficient are presented in Paragraphs O-2.4.1 and O-2.4.2.

O-2.4.1. *Directions for Pearson's Correlation Coefficient*. Let $x_1, x_2,..., x_n$ represent one variable (X) of the n data points and let $y_1, y_2,..., y_n$ represent the corresponding values of a second variable (Y). The Pearson correlation coefficient, r, for the sample of (x_i, y_i) pairs is computed by:

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

O-2.4.2. *Example of Pearson's Correlation Coefficient*. Consider the following data set for $n = 8$ chromium and lead in subsurface soil background (in mg/kg):

Sample	**Chromium(X)**	**Lead(Y)**
EPC-BG01	4.60	3.50
EPC-BG02	5.29	4.16
EPC-BG03	4.26	4.19
EPC-BG04	5.28	3.91
EPC-BG05	4.53	3.66
EPC-BG06	5.74	4.31
EPC-BG07	5.86	4.19
EPC-BG08	3.84	3.35

O-2.4.2.1. For chromium,

$$\sum_{i=1}^{8} x_i = 4.60 + 5.29 + 4.26 + 5.28 + 4.53 + 5.74 + 5.86 + 3.84 = 39.4$$

$$\sum_{i=1}^{8} x_i^2 = 4.60^2 + 5.29^2 + 4.26^2 + 5.28^2 + 4.53^2 + 5.74^2 + 5.86^2 + 3.84^2 = 197.7 \ .$$

So, $\bar{x} = 39.4/8 = 4.925$ and

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}} = \sqrt{\frac{197.7 - (8 \times 4.925^2)}{7}} = 0.7226 \ .$$

O-2.4.2.2. For lead,

$$\sum_{i=1}^{8} y_i = 3.50 + 4.16 + 4.19 + 3.91 + 3.66 + 4.31 + 4.19 + 3.35 = 31.27$$

$$\sum_{i=1}^{8} y_i^2 = 3.50^2 + 4.16^2 + 4.19^2 + 3.91^2 + 3.66^2 + 4.31^2 + 4.19^2 + 3.35^2 = 123.2 \, .$$

So, $\bar{y} = 31.27/8 = 3.909$ and

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1}} = \sqrt{\frac{123.2 - (8 \times 3.909^2)}{7}} = 0.3632 \, .$$

O-2.4.2.3. The "cross term" dependent upon the product of chromium and lead is:

$$\sum_{i}^{n} x_i y_i = (4.60 \times 3.50) + (5.29 \times 4.16) + (4.26 \times 4.19) + (5.28 \times 3.91) +$$
$$(4.53 \times 3.66) + (5.74 \times 4.31) + (5.86 \times 4.19) + (3.84 \times 3.35) = 155.3.$$

So,

$$r = \frac{155.3 - (8 \times 4.925 \times 3.909)}{7 \times 0.7226 \times 0.3729} = 0.72 \, .$$

Paragraphs O-2.4.3 and O-2.4.4 will demonstrate how to test whether the sample correlation coefficient indicates that the population correlation coefficient differs from zero.

O-2.4.3. *Discussion*. Although two independent variables will produce a correlation coefficient of zero, it should be noted that a calculated correlation coefficient that is equal to or near zero does not demonstrate the absence of a significant relationship between the two variables. For example, because Pearsons' *r* does not detect non-linear relationships, a strong non-linear relationship could result in a value of *r* equal to zero.

O-2.4.3.1. The data from the previous example are illustrated in Figure O-1. Correlation coefficients should be used with scatter plots to determine whether a low value of Pearson's *r* is due to a non-linear relationship or a lack of association.

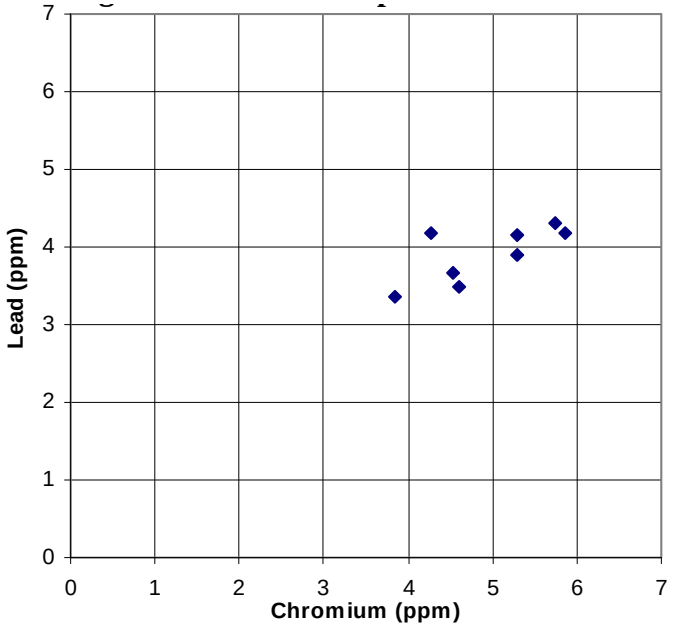

Figure O-1. Scatter plot for chromium and lead.

O-2.4.3.2. Pearson's *r* can be sensitive to the presence of one or two extreme values, especially when sample sizes are small. Such values may result in a high correlation, suggesting a strong linear trend, when only a moderate or weak trend is present. This may happen, for instance, if a single (*x*, *y*) pair has very high values for both measurements while the remaining data values are uncorrelated. For example, Figure O-2 plots an example where a very large outlier exists. Including the outlier leads to a sample correlation coefficient of 0.96. Without this value, the sample correlation coefficient falls to –0.10. Extreme values may also lead to low sample correlation coefficients, thus tending to mask a strong linear trend. This may happen if all the (*x, y*) pairs except one (or two) tend to cluster tightly about a straight line, and the exceptional point has a very large *X* value paired with a moderate or small *Y* value (or vice versa). Because of the influences of extreme values, it is wise to use a scatter plot in conjunction with a Pearson correlation coefficient.

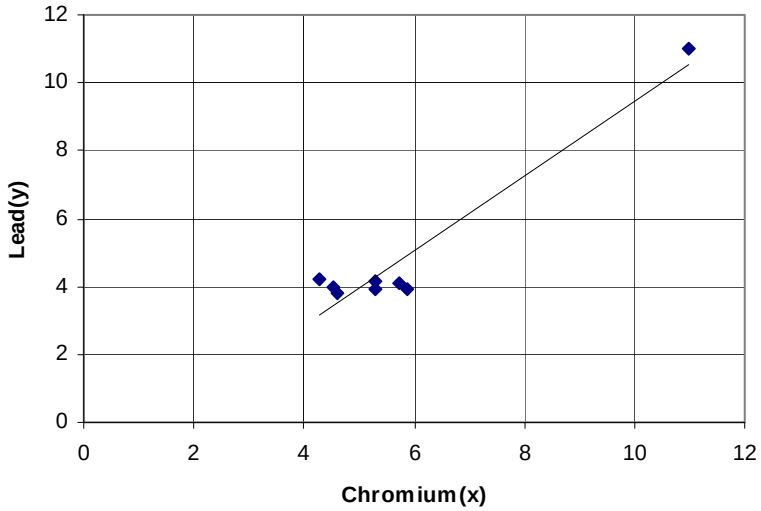

Figure O-2. Scatter plot with outlier.

O-2.4.3.3. An important property of Pearson's *r* is that it is unaffected by changes in location of the data (adding or subtracting a constant from all of the *X* or *Y* measurements) and by changes in scale of the data (multiplying the *X* or *Y* values by a positive constant). Linear transformations on the data pairs do not affect the correlation coefficient of the measurements. For example, if one variable in the pair was temperature in degrees Celsius, then the correlation would not change if Celsius is converted to Fahrenheit.

O-2.4.3.4. However, Pearson's *r* is not invariant to non-linear transformations. If non-linear transformations of the measurements are made, then the Pearson correlation coefficient between the transformed values will differ from the Pearson correlation coefficient of the original measurements. For example, if *X* and *Y* represent PCB and dioxin concentrations in soil, respectively, and $U = \mathrm{Log}(X)$ and $V = \mathrm{Log}(Y)$, then the Pearson correlation coefficients between *X* and *Y* and between *U* and *V* will be different because the logarithmic transformation is a nonlinear transformation.

O-2.4.3.5. It should be further noted that statistical tests that use *r* to estimate the population correlation coefficient rely on the assumption that the true relationship between the variables *X* and *Y* follows a bivariate normal distribution. If either variable *X* or *Y* is not normal, then together *X* and *Y* are not likely to follow a bivariate normal distribution. For more details see Snedecor and Cochran (1982).

O-2.5. *Spearman's rho*. Spearman's rank correlation coefficient measures monotonic correlation for ordinal data (data that can be ranked) and is nonparametric (i.e., can be used when the data are not normally distributed).

O-2.5.1. *Introduction*. Data may be either linearly or non-linearly correlated. When one variable tends to consistently increase or decrease as another variable increases, the two variables possess a monotonic correlation. Unlike Pearson's *r*, Spearman's rho, ρ, may be used to measure the strength of both linear and nonlinear relationships.

O-2.5.1.1. It is calculated by first replacing each value *x*, by its rank R(*x*) (1 for the smallest *x* value, 2 for the second smallest, etc.) and each value *y* by its rank R(*y*). These pairs of ranks are then treated as the (*x*, *y*) data and Spearman's rank correlation is calculated using the same formula as for Pearson's correlation.

O-2.5.1.2. Directions and an example for calculating a Spearman's rank correlation coefficient are contained in the Paragraphs O-2.5.2 and O-2.5.3.

O-2.5.1.3. Because meaningful (monotonically increasing) transformations of the data will not alter the ranks of the respective variables (the ranks for Log(*x*) will be the same as the ranks for *x*), Spearman's correlation will not be altered by non-linear increasing transformations of *x* and *y*. For instance, the Spearman correlation between PCB and dioxin concentrations (*x* and *y*) in soil will be the same as the correlation between their logarithms, Log(*x*) and Log(*y*). Because Spearman's ρ is a nonparametric measure of correlation, it is invariant for monotonic increasing transformations and is less sensitive to extreme values than Pearson's correlation. However, Pearson's *r* has *higher statistical power* than Spearman's ρ.

O-2.5.2. *Directions for the Spearman's Rank Correlation Coefficient*. Let

$$R(x_1), R(x_2), \ldots, R(x_n)$$

represent a set of ranks of the *n* data points for the variable *X* and let

$$R(y_1), R(y_2), \ldots, R(y_n)$$

represent a set of ranks of a second variable *Y* of the *n* data points. The Spearman sample correlation coefficient, ρ, for *X* and *Y* is computed by:

$$\rho = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{R(x_i)-\overline{R}(x)}{s_{R(x)}}\right)\left(\frac{R(y_i)-\overline{R}(y)}{s_{R(y)}}\right) = \frac{\sum_{i=1}^{n} R(x_i)R(y_i) - n\,\overline{R}(x)\overline{R}(y)}{(n-1)\,s_{R(x)}\,s_{R(y)}} \quad .$$

O-2.5.3. *Example of Spearman's Correlation Coefficient*. Consider the following data set for chromium and lead in subsurface soil background (in mg/kg):

Sample	Chromium(*X*)	Lead(*Y*)
EPC-BG01	4.60	3.50
EPC-BG02	5.29	4.16
EPC-BG03	4.26	4.19
EPC-BG04	5.28	3.91
EPC-BG05	4.53	3.66
EPC-BG06	5.74	4.31
EPC-BG07	5.86	4.19
EPC-BG08	3.84	3.35

O-2.5.3.1. First the data must be ranked:

Sample	Chromium	Rank (*X*)	Lead	Rank (*Y*)
EPC-BG01	4.60	4	3.50	2
EPC-BG02	5.29	6	4.16	5
EPC-BG03	4.26	2	4.19	6.5
EPC-BG04	5.28	5	3.91	4
EPC-BG05	4.53	3	3.66	3
EPC-BG06	5.74	7	4.31	8
EPC-BG07	5.86	8	4.19	6.5
EPC-BG08	3.84	1	3.35	1

O-2.5.3.2. Notice that two of the lead values are equal, so their rank is assigned to be the average of ranks 6 and 7.

O-2.5.3.3. For chromium, $\overline{R}(x) = 4.5$, and $s_{R(x)} = 2.45$.

O-2.5.3.4. For lead, $\overline{R}(y) = 4.5$, and $s_{R(y)} = 2.43$.

O-2.5.3.5. The sum of the cross-products for chromium and lead ranks is:

$$\sum_{i=1}^{8} R(x_i) R(y_i) = (1\times 1) + (2\times 6.5) + (3\times 3) + (4\times 2) + (5\times 4) + (6\times 5) + (7\times 8) + (8\times 6.5)$$
$$= 189 .$$

O-2.5.3.6. The correlation coefficient is

$$\rho = \frac{189 - (8\times 4.5\times 4.5)}{7\times 2.45\times 2.43} = 0.647 .$$

O-2.6. *Serial Correlation Coefficient*. The serial correlation coefficient is a measure of the extent to which successive observations (either in time or space) are related. The primary difference between the serial correlation coefficient and other measures of correlation is the manner in which the correlation coefficient is used and the manner in which one of the variables is scaled. For example, the serial correlation coefficient is frequently used to determine the behavior of some variable of interest X with respect to time (t). Frequently, the variable X is measured at equally spaced time intervals, so that the data points are of the form (x_1, t_1), (x_2, t_2),…., (x_n, t_n). The serial correlation coefficient may be a parametric or non-parametric measure of correlation, depending upon how it is calculated. For example, if variable X is being evaluated with respect to time t, Spearman's ρ is essentially being calculated if the values of X are replaced with the corresponding ranks. Directions and examples for calculating a serial correlation coefficient are presented in the following two Paragraphs.

O-2.6.1. *Directions to Calculate the Serial Correlation Coefficient*.

O-2.6.1.1. For a sequence of data points taken serially in time, or "one-by-one in a row," the serial correlation coefficient can be calculated by replacing the sequencing variable by the numbers 1 through n and calculating Pearson's correlation coefficient with x being the actual data values, and y being the numbers 1 through n. For example, for a sequence of samples collected every 10 feet along a straight transit line at a waste site, the distances on the transit line of the data points are replaced by the numbers 1 through n, for samples taken at 10-foot intervals (first 10-foot sample point = 1, the 20-foot sample point = 2, the 30-foot sample point = 3, etc.)

O-2.6.1.2. To calculate the serial correlation coefficient, let $x_1, x_2,..., x_n$ represent the data values collected in sequence over equally spaced periods. Label the periods 1, 2..., n to match the data values. Use the directions above to calculate the Pearson's Correlation Coefficient between the data, x, and the time-periods, y.

O-2.6.2. *Estimating the Serial Correlation Coefficient*. Consider benzene results taken from quarterly groundwater samples at well MW01 in Site A from 1998–2000. Benzene has been detected during all of these sampling events, so no proxy concentrations were derived. Also, notice how the numbers 1 through 10 replace the actual sample dates.

Time	Jan-98	Apr-98	Jul-98	Oct-98	Apr-99	Jul-99	Oct-99	Apr-00	Jul-00	Oct-00
Time Period Number	1	2	3	4	5	6	7	8	9	10
Concentration (µg/L)	12.2	3.79	3.42	5.47	0.81	1.84	7.56	4.3	2.68	6.17

O-2.6.2.1. For the concentration (X),

$\sum_{i=1}^{10} x_i = 48.24$, $\sum_{i}^{10} x_i^2 = 329.8$, $\bar{x} = 4.824$, and $s_x = 3.284$.

O-2.6.2.2. For the time period (*Y*),

$\sum_{i=1}^{10} y_i = 55$, $\sum_{i}^{10} y_i^2 = 385$, $\bar{y} = 5.5$, and $s_y = 3.028$.

The cross term is:

$\sum_{i=1}^{10} x_i y_i = 240.2$.

O-2.6.2.3. Using Paragraph O-2.4.1, we see that the Pearson correlation coefficient, *r*, between the concentration (*X*) and the time period (*Y*) gives a serial correlation coefficient of:

$$r = \frac{240.2 - (10 \times 4.824 \times 5.5)}{9 \times 3.284 \times 3.028} = -0.2813.$$

O-2.7. *Kendall's Coefficient of Rank Correlation*. In instances where data do not follow a normal or other known distribution, it is still possible to test for the significance of association between two variables. Kendall's coefficient of rank correlation, also referred to as *Kendall's* τ (the Greek letter tau), is a measure of correlation that may be used for variables that are at least ordinal in nature (i.e., variables with values that can be ranked). It is frequently encountered in ecological applications such as counting of fish species in a stream in different seasons.

O-2.7.1. *Introduction*. Kendall's τ does not assume any particular data distribution and accommodates censored values. Non-detected results should be assigned a value smaller than the lowest measured value. As the test depends only upon signs of the differences between data points (or the ranks), information about magnitudes of these differences is not used; as a result, the test possesses less power than its parametric counterpart, Pearson's *r* (i.e., a larger number of data points are required to identify a correlation using Kendall's τ). However, Kendall's τ is advantageous because assumptions about the underlying data distribution are not required, and it is less sensitive to outliers and censored values than a parametric test.

O-2.7.1.1. Kendall's τ is also invariant with respect to monotonic transformations of the variables. For example, the calculated value of τ will be identical to the calculated value for log-transformed variables. See the discussion at the end of Paragraph O-2.5 for more details. It should also be noted that for the same data, the value for Kendall's τ is generally lower than for

Spearman's *r* (Conover, 1980). However, statistical tests for $\gamma = 0$ are generally in agreement between the two.

O-2.7.1.2. Kendall's τ for small sample sizes is appropriate for data with fewer than 40 samples (Gilbert, 1987); the EPA suggests using this method with data sets fewer than 10 samples. Tied observations (when two or more measurements are equal) degrade the statistical power and should be avoided, if possible, by recording the data to sufficient accuracy. If the number of samples becomes too large, the calculations become cumbersome to do by hand. Directions for calculating Kendall's τ for a small sample size (less than 10 samples) are presented in Paragraph O-2.7.2 and an example is presented in Paragraph O-2.7.3. Extensions of Kendall's τ for larger sample sizes are explained with the Mann-Kendall test for trends in Appendix P. In that Appendix, the time variable corresponds to the *X* variable here, and the *X* variable in Appendix P corresponds to the *Y* variable here.

O-2.7.2. *Directions for Kendall's Coefficient of Rank Correlation*. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ represent pairs of measurements of variables *X* and *Y*. Order the pairs from least to greatest *by the x value* $(x_{(1)}, y_{x_{(1)}}), (x_{(2)}, y_{x_{(2)}}), \ldots, (x_{(n)}, y_{x_{(n)}})$. Here the notation $y_{x_{(i)}}$ indicates the *Y* measurement that corresponds to the i^{th} *X* measurement ordered from least to greatest. The test statistic *S* is then calculated:

$$S = S^+ - S^-$$

where S^+ is the number of positive ("concordant") pairs: $(y_{x_{(i)}}, y_{x_{(j)}})$ with $i < j$ and $y_{x_{(i)}} < y_{x_{(j)}}$. Likewise, S^- is the number of negative ("discordant") pairs: $(y_{x_{(i)}}, y_{x_{(j)}})$ with $i < j$ and

$$y_{x_{(i)}} > y_{x_{(j)}}.$$

It can be shown that there are a total of $n(n-1)/2$ possible pairwise comparisons for a set of *n* pairs $(y_{x_{(i)}}, y_{x_{(j)}})$. The sample statistic Kendal1's τ, is:

$$\tau = \frac{S}{n(n-1)/2} \quad .$$

Note that differences of zero are not included in the test statistic (and should be avoided, if possible, by recording data to sufficient accuracy). However, an adjustment for ties may be made by calculating Kendall's "tau b," τ_b

$$\tau_b = \frac{S}{\sqrt{\left(\frac{n(n-1)}{2} - n'_X\right)\left(\frac{n(n-1)}{2} - n'_Y\right)}} .$$

The quantities n'_X and n'_Y denote the number of ties for the *X* variable and *Y* variable, respectively. In particular, if there are *n* pairs of values (x_i, y_j), so that the measured values of *X* are $x_1, x_2, \ldots x_n$, then n'_X is the number of pairs (x_i, x_j), where *i* > *j,* for which $(x_i - x_j) = 0$ or for which this difference cannot be determined to be either positive or negative because of data censoring. For example, assume that there are multiple censoring limits for non-detects (e.g., < 3 and < 5), and *X* is the set of *n* = 5 values {< 1, < 3, < 5, 2, 10} with the corresponding *Y* values {2, 4, 5, 7, 9}, so that, for example, the first pair of results (x_1, y_1) is $(<1, 2)$. There are five tied pairs for the measured values of *X*: (< 1, < 3), (< 1, < 5), (< 3, < 5), (< 3, 2), and (< 5, < 2). Therefore, $n'_X = 5$. As there are no tied values for Y, $n'_Y = 0$. Note that when $n'_X = n'_Y = 0, \tau_b = \tau$. Tied values tend produce larger values for τ_b relative to the corresponding values for τ.

O-2.7.2.1. Table O-1 presents the resulting matrix of differences when applying the steps above. Fill in the blank spaces with a 1 if the value at the top of the column exceeds the value at the left of the row. Fill in 0 if they are equal, and fill in –1 otherwise. Then sum the values across rows and add up the sums to get *S*.

Table O-1.
Resulting Matrix of Differences

***Y* Measurements**	$y_{x_{(2)}}$	$y_{x_{(3)}}$	. . .	$y_{x_{(n)}}$	**Sum of Row**
$y_{x_{(1)}}$					
$y_{x_{(2)}}$					
. . .					
$y_{x_{(n-1)}}$					
					S

O-2.7.2.2. Use Table B-10 of Appendix B to determine the probability (*p*) using the sample size (*n*) and the absolute value of the statistic *S* if $n \le 10$.

O-2.7.2.3. For testing H_0: $\gamma = 0$ against H_A: $\gamma \neq 0$ at significance level α, reject H_0 if $p < \alpha/2$.

O-2.7.3. *Example of Kendall's Rank Correlation Coefficient*. Consider the same data set presented in Paragraphs O-2.4.2 and O-2.5.2 for chromium and lead in subsurface soil back-

ground (in mg/kg). Although these data are for continuous variables, it is possible to determine the rank correlation between chromium and lead using Kendall's τ.

O-2.7.3.1. First the data must be ordered by the chromium measurements as shown below.

Sample	Chromium	Lead
EPC-BG08	3.84	3.35
EPC-BG03	4.26	4.19
EPC-BG05	4.53	3.66
EPC-BG01	4.60	3.50
EPC-BG04	5.28	3.91
EPC-BG02	5.29	4.16
EPC-BG06	5.74	4.31
EPC-BG07	5.86	4.19

O-2.7.3.2. Then, create Table O-2 for the lead measurements as described in Paragraph O-2.7.2.

O-2.7.3.3. From Table O-2, $S = 15$. There are $n = 8$ pairs of lead and chromium measurements. Therefore, Kendall's tau is:

$$\tau = \frac{S}{n(n-1)/2} = \frac{15}{8(8-1)/2} = 0.536.$$

As there is one tie for the lead measurements (two measurements equal 4.19)

$$\tau_b = \frac{S}{\sqrt{\left(\frac{n(n-1)}{2} - n'_X\right)\left(\frac{n(n-1)}{2} - n'_Y\right)}} = \frac{15}{\sqrt{\left(\frac{8(8-1)}{2} - 0\right)\left(\frac{8(8-1)}{2} - 1\right)}} = 0.546.$$

O-2.7.3.4. To test whether the population correlation coefficient differs from 0 with 90% confidence ($\alpha = 0.05$), look up the value of p corresponding to $S = 15$ for $n = 8$ in Table B-10. Owing to the tied value for lead, $S = 15$ does not appear in the table. Ideally, the data should have been recorded with more accuracy to break the tie. In this case, the value for $S = 14$ will be used to give $p = 0.054 > \alpha/2 = 0.05$. We conclude that the population correlation coefficient does not differ significantly from zero with 90% confidence although further study may be needed.

O-2.8. *Covariance*. A statistic related to the correlation coefficient is covariance. Covariance is a measure of the linear association between two random variables, X and Y. If covariance is positive, large values of X tend to be associated with large values of Y and vice versa. If co-

variance is negative, large values of *X* tend to be associated with small values of *Y* and vice versa. The sample covariance is calculated as

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)}.$$

Table O-2.
Resulting Matrix of Differences

Lead Measurements	4.19	3.66	3.50	3.91	4.16	4.31	4.19	Sum of Row
$y_{x_{(1)}} = 3.35$	1	1	1	1	1	1	1	7
$y_{x_{(2)}} = 4.19$		–1	–1	–1	–1	1	0	–3
$y_{x_{(3)}} = 3.66$			–1	1	1	1	1	3
$y_{x_{(4)}} = 3.50$				1	1	1	1	4
$y_{x_{(5)}} = 3.91$					1	1	1	3
$y_{x_{(6)}} = 4.16$						1	1	2
$y_{x_{(7)}} = 4.31$							–1	–1
								S = *15*

O-2.8.1. Pearson's correlation coefficient is derived from the covariance by dividing covariance by the sample standard deviations of *X* and *Y*.

O-2.8.2. Covariance is rarely used because the magnitude of its value is difficult to interpret. In particular, changes in scale cause changes to the covariance; that is, covariance is not invariant to changes in scale. For example, if *X* is multiplied by 100, its covariance with *Y* will also go up by a factor of 100, while its correlation with *Y* will remain the same.