APPENDIX F
Testing for Normality

Section I
Methods for Determining Normality

F-1. Introduction. As previously stated, the assumption of normality is important because it is required for many statistical tests. A normal, or Gaussian, distribution is one of the most common probability distributions used for the analysis of environmental data. A normal distribution is a reasonable model of the behavior of certain random phenomena and can often be used to approximate other probability distributions. In addition, the central limit theorem and other limit theorems state that, as the sample size gets large, some of the sample summary statistics (e.g., the sample mean) behave as if they are a normally distributed variable. As a result, a common assumption associated with parametric tests or statistical models is that the errors associated with data or models follow a normal distribution. Therefore, this Appendix will focus on statistical tests that are used to determine whether normality can be reasonable assumed for a set of measured results.

F-1.1. In general, any distribution assumption should be verified using a combination of graphical plots and statistical tests. Environmental data commonly exhibit frequency distributions that are non-negative and positively skewed (i.e., possess long right tails). Several parametric probability distributions have these properties, including the Weibull, gamma, and lognormal distributions. The methods for testing for normality described in this Appendix can be used to test for lognormality if a logarithmic transformation has been applied to the data.

F-1.2. There are many methods available for verifying the assumption of normality, ranging from simple to complex. They are listed in Table F-1 below. It should be noted that statistical tests for normality do **not** actually demonstrate normality but the **lack** of normality. They rely on the probability a given data set is normal (e.g., statistical software typically reports a "*p* value" for the hypothesis that the population distribution is normal). If the probability is low (e.g. $p < 0.01$), one "rejects the assumption of normality," that is, one concludes, based upon weight of evidence, that the data set is not normal. However, if the assumption of normality is not rejected, then, strictly speaking, the statistical test is inconclusive; the data may or may not be normal. This constitutes an additional reason to visually examine the data set for normality and to decide whether to proceed with a statistical test that requires normality. In practice, if the assumption of normality is not rejected and graphical plots suggest normality, the statistical tests that rely upon normality are typically used.

Table F-1.
Methods Available To Verify the Assumption of Normality

Test	Sample Size, *n*	Recommended Use
Graphical Methods	Any	Highly recommended in conjunction with test methods.
Shapiro-Wilk *W* Test	≤ 50	Highly recommended (D'Agostino's test may be used when sample size is between 50 and 1000).
Filliben's Statistic	≤ 100	Highly recommended.
Coefficient of Variation Test	Any	Only use to quickly discard an assumption of normality and for screening only.
Geary's Test	> 50	Useful when tables for other tests are not available.
Studentized Range Test	≤ 1000	Use for screening purposes only.
Chi-square Test	Large	Useful for grouped data and when the comparison distribution is known.
Lilliefors Kolmogorov-Smirnoff Test	> 50	Useful when tables for other tests are not available.

F-2. Graphical Methods.

F-2.1. Graphical methods present qualitative information about data sets that may not be apparent from statistical tests. Histograms and normal probability plots are some graphical methods that are useful for determining whether data follow a normal curve. The histogram of a normal distribution is bell-shaped. The normal probability plot (Appendix J) of a normal distribution follows a straight line. For non-normally distributed data, there will be large deviations in the tails or middle of a normal probability plot. Extreme deviations from normality are often readily identified from graphical methods. However, in many instances the decision is not straightforward. Using a plot to decide whether a data set is normally distributed involves making a subjective decision; formal test procedures are usually necessary to test the assumption of normality.

F-2.2. In general, both statistical tests and graphical plots should be used to evaluate normality. The assumption of normality should not be rejected on the basis of a statistical test alone. In particular, when a large number of data are available, statistical tests for normality can be sensitive to very small (i.e., negligible) deviations in normality. Therefore, if a very large number of data are available, a statistical test may reject the assumption of normality when the data set, as shown using graphical methods, is essentially normal and the deviation from normality too small to be of practical significance.

F-3. Shapiro-Wilk Test for Normality.

F-3.1. *General*. One of the most powerful and most commonly employed tests for normality is the *W* test by Shapiro and Wilk, also called the Shapiro-Wilk test. The Shapiro-Wilk test is an effective method for testing whether a data set has been drawn from an underlying normal distribution. It can also evaluate lognormality if the test is conducted on logarithms of the data. This test is similar to computing a correlation between the quantiles of the standard normal distribution and the ordered values of a data set. If the normal probability plot is approximately linear (the data follow a normal curve), the test statistic will be relatively high. If the normal probability plot has curvature that is evidence of non-normality in the tails of a distribution, the test statistic will be relatively low. The Shapiro-Wilk test is recommended in several EPA guidance documents and in many statistical texts. It is designed so that the burden of proof rests on showing evidence that the data are not normally distributed. (In terms of hypothesis testing, the Shapiro-Wilk test is based on H_0 that the data are normally distributed. Hypothesis testing is addressed in detail in Appendices L, M, and N.)

F-3.1.1. The Shapiro-Wilk test is good for evaluating whether a sample set of data has been drawn from a normal or lognormal distribution. However, this test will not have very much power to reject the null hypothesis of normality or lognormality if the sample size is very small (i.e., the test would fail to detect non-normal behavior when the sample size is small). The method for calculating the *W* statistic is presented below in Paragraph F-3.2.

F-3.1.2. As this test is laborious to compute by hand, statistical software packages such as *SAS, WQ Stat, Statistica,* and the *Data Quality Assessment Toolbox (QA/G-9D)* are recommended. An example calculation is presented below in Paragraph F-3.3.

F-3.1.3. D'Agostino's test is an extension of the Shapiro-Wilk test. It is based on an estimate of the standard deviation obtained using the ranks of the data. This estimate is compared to the usual estimate of the standard deviation, which is appropriate for the normal distribution. The D'Agostino's test is recommended for sample sizes between 50 and 1000.

F-3.1.4. Another test related to the *W* test is Filliben's statistic, also called the probability plot correlation coefficient. This statistic measures the linearity of the points on the normal probability plot. Similar to the Shapiro-Wilk test, if the normal probability plot is approximately linear (the data follow a normal curve), the correlation coefficient will be relatively high. If the normal probability plot contains significant curves (the data do not follow a normal curve), the correlation coefficient will be relatively low. Filliben's statistic is recommended for sample sizes less than or equal to 100. Although easier to compute than the Shapiro-Wilk test, Filliben's statistic is still difficult to compute by hand. It is available in the *Data Quality Assessment Toolbox (QA/G-9D)* and various software packages.

F-3.2. *Directions for the Shapiro-Wilk* W *Test*. Order the data points, $x_{(1)}$, $x_{(2)}$, ..., $x_{(n)}$, where $x_{(1)}$ is the smallest value and $x_{(n)}$ is the largest value of the n observations.

F-3.2.1. Estimate the sample standard deviation, s. Compute the Shapiro-Wilk test statistic:

$$W = \left[\frac{b}{s\sqrt{n-1}} \right]^2$$

where

$$b = \sum_{i=1}^{k} b_i = \sum_{i=1}^{k} a_{n-i+1} [x_{(n-i+1)} - x_i] .$$

F-3.2.2. The coefficients a can be found for any sample size between 3 and 50 in Table B-19 of Appendix B. The value k is the greatest integer less than or equal to $n/2$.

F-3.2.2.1. Reject normality if the calculated statistic $W < W_\alpha$, where the critical values W_α are listed in Table B-20 of Appendix B.

F-3.2.2.2. If $W \geq W_\alpha$, do not reject the assumption of normality. Typically, one assumes the data are approximately normal for further statistical analysis.

F-3.3. *Example of Shapiro-Wilk* W *Test*. Consider using the Shapiro-Wilk to test the sub-surface soil background chromium results for normality. The results (in mg/kg) are as follows: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, and 3.84.

F-3.3.1. Hypothesis test for Shapiro-Wilk W test:

H_0: The data are normally distributed.

H_A: The data are not normally distributed.

F-3.3.2. Estimate the sample standard deviation, $s = \sqrt{0.5255} = 0.7249$.

F-3.3.3. Compute Shapiro-Wilk test statistic W, where $n = 8$, $k = 8/2 = 4$ and $b = 1.859$:

$$W = \left[\frac{b}{s\sqrt{n-1}} \right]^2 = \left[\frac{1.859}{0.7249\sqrt{8-1}} \right]^2 = 0.9395 .$$

Using an α level of 0.05 and n = 8, we find the critical value, W_α, from Table B-20 to be 0.818. As $W > 0.818$, there is insufficient evidence to reject the assumption of normality.

$x_{(i)}$	$x_{(n-i+1)}$	$x_{(i)} - x_{(n-i+1)}$	$a_{(n-i+1)}$	b_i
3.84	5.86	2.02	0.6052	1.2200
4.26	5.74	1.48	0.3164	0.4683
4.53	5.29	0.76	0.1743	0.1325
4.60	5.28	0.68	0.0561	0.0381
5.28	4.60	–0.68		
5.29	4.53	–0.76		
5.74	4.26	–1.48		
5.86	3.84	–2.02		

F-4. Coefficient of Variation. The coefficient of variation (CV) may be used to quickly determine whether or not data follow a normal curve by comparing the sample CV to 1. However, the CV evaluation is not reliable. The use of the CV is valid only for some environmental applications if the data represent a non-negative characteristic, such as contaminant concentrations. If the CV is much greater than 1, the data should not be modeled with a normal curve. However, this method *should not be used to conclude the opposite;* do not conclude that the data can be modeled with a normal curve if the CV is less than 1. Furthermore, the sample CV $(s/\bar{x})$ can be greater than 1 when the population CV (σ/μ) is between 0.5 and 1. This is because of the sample CV being a random variable and estimating the true CV with some degree of error (EPA 68-W0-0025). This test is to be used only in conjunction with other statistical tests or when graphical representations of the data indicate extreme departures from normality. Details for estimating the CV are presented in Appendix D.

F-5. Range Tests

F-5.1. *General*. Range tests for normality have been developed based on the knowledge that virtually 100% of the area of a normal curve lies within plus and minus 5 standard deviations from the mean. Two such tests, which are both simple to apply, are the Studentized range test and Geary's test. Both of these tests use a ratio of an estimate of the sample range to the sample standard deviation. Very large and very small values of the ratio then imply that the data are not well modeled by a normal curve. These range tests are not as reliable as the previously discussed tests, and are recommended only if computer procedures or look-up tables for the other tests are not available. However, both range tests are relatively simple to use, so they are presented here.

F-5.1.1. The Studentized range test compares the range of the sample to the sample standard deviation. Tables of critical values for sample sizes up to 1000 (Table B-21 of Appendix B) are available for determining whether the absolute value of this ratio is significantly large.

F-5.1.2. Directions to conduct the Studentized range test and an example of this test follow in Paragraph F-5.2.

F-5.1.3. The Studentized range test does not perform well if the data are asymmetric and if the tails of the data are heavier than the normal distribution. In addition, this test may be sensitive to extreme values. Unfortunately, lognormally distributed data, which are common in environmental applications, have these characteristics. If the data appear to be lognormally distributed, then this test should not be used. In most cases, the Studentized range test performs as well as the Shapiro-Wilk test and is easier to apply.

F-5.1.4. Alternatively, Geary's test uses the ratio of the mean deviation of the sample to the sample standard deviation. This ratio is then adjusted to approximate a standard normal distribution.

F-5.1.5. Directions for calculating Geary's test are presented below in Paragraph F-5.3

F-5.1.6. This test does not perform as well as the Shapiro-Wilk test or the Studentized range test. However, because Geary's test statistic is based on the normal distribution, critical values for all possible sample sizes are available. An example application of Geary's test follows in Paragraph 5-4.

F-5.2. *Directions and an Example of Studentized Range Test.*

F-5.2.1. *Directions.*

- Calculate sample range (R) and sample standard deviation (s).
- Calculate the ratio R/s.
- Compare to the critical values for R/s given in Table B-21 (labeled a and b).

If the calculated value of R/s falls outside the two critical values, then the data do not follow a normal curve.

F-5.2.2. *Example*. Consider using the Studentized range test to determine if the subsurface soil background chromium results can be modeled using a normal curve. The results are (in mg/kg) as follows: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, and 3.84.

Sample range $R = 5.86 - 3.84 = 2.02$

Sample standard deviation $s = \sqrt{0.5255} = 0.7249.$

$R / s = 2.02 / 0.7249 = 2.787$.

The critical values for *R/s* in Table B-21 for *n* = 8 and α = 0.05 are 2.50 and 3.399. As 2.787 falls between these values, the assumption of normality is not rejected.

F-5.3. *Directions for Calculating Geary's Test*. Calculate the sample mean $\bar{x}$, the sample sum of squares (SSS), and the sum of absolute deviations (SAD):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{SSS} = (n-1)s^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

$$\text{SAD} = \sum_{i-1}^{n} \left| x_i - \bar{x} \right| .$$

F-5.3.1. Calculate Geary's test statistic

$$a' = \frac{\text{SAD}}{\sqrt{n(\text{SSS})}} .$$

F-5.3.2 Test *a* for significance by computing

$$z = \frac{a' - 0.7979}{0.2123/\sqrt{n}} .$$

Here, 0.7979 and 0.2123 are constants used to achieve normality.

F-5.3.3. Use Table B-15 of Appendix B to find the critical value $Z_{1-\alpha}$ such that 100(1 – α)% of the normal distribution is below $Z_{1-\alpha}$. For example, if α = 0.05, then $Z_{1-\alpha}$ = 1.645. The statistic a' is sufficiently small or large to conclude the data are not normally distributed if $|z| > Z_{1-\alpha}$.

F-5.4. *Example of Geary's Test*. Consider using Geary's test to see if the subsurface soil background chromium results can be modeled using a normal curve. The results are (in mg/kg) as follows: 4.60, 5.29, 4.26, 5.28, 4.53, 5.74, 5.86, and 3.84.

F-5.4.1. Calculate the sample mean $\bar{x}$:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{8}(4.60 + 5.29 + 4.26 + 5.28 + 4.53 + 5.74 + 5.86 + 3.84) = 4.925 \ .$$

F-5.4.2. Calculate the SSS:

$$\text{SSS} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

$$\sum_{i=1}^{n} x_i^2 = (4.60)^2 + (5.29)^2 + (4.26)^2 + (5.28)^2 + (4.53)^2 + (5.74)^2 + (5.86)^2 + (3.84)^2 = 197.73$$

$$\frac{(\sum_{i=1}^{n} x_i)^2}{n} = \frac{(39.4)^2}{8} = 194.05 \ .$$

So, $\text{SSS} = 197.73 - 194.05 = 3.68$.

F-5.4.3. Calculate the sum of absolute deviations (SAD):

$$\begin{aligned}\text{SAD} = \sum_{i=1}^{n} |x_i - \bar{x}| = {} & |4.60 - 4.925| + |5.29 - 4.925| + |4.26 - 4.925| \\ & + |5.28 - 4.925| + |4.53 - 4.925| + |5.74 + 4.925| \\ & + |5.86 - 4.925| + |3.84 - 4.925| = 4.94 \ .\end{aligned}$$

F-5.4.4. Calculate Geary's test statistic:

$$a' = \frac{\text{SAD}}{\sqrt{n(\text{SSS})}} = \frac{4.94}{\sqrt{8(3.68)}} = 0.910 \ .$$

F-5.4.5. Test a' for significance by computing:

$$z = \frac{a' - 0.7979}{0.2123/\sqrt{n}} = \frac{0.910 - 0.7979}{0.2123/\sqrt{8}} = 1.49 \ .$$

Here, 0.7979 and 0.2123 are constants used to achieve normality.

F-5.4.6. Using Table B-15 of Appendix B to find the critical value $Z_{1-\alpha}$, where α = 0.05, then $Z_{1-\alpha}$ = 1.645. Since $1.49 \not> 1.645$, there is not enough information to conclude that the data do not follow a normal distribution.

F-6. Goodness-of-Fit Tests. Goodness-of-fit tests are not practical to do manually. Because these are included in most statistical software packages, detailed instructions for doing them are not included. Following is a brief overview of these tests with recommendations for their use.

F-6.1. Goodness-of-fit tests are used to determine whether data conform to some theoretical probability distribution. However, unlike the tests previously discussed, these tests can be used to see if a data set fits any specified probability distribution, not just the normal distribution. In contrast, the Shapiro-Wilk test can be used only to determine whether a data set is normally distributed.

F-6.2. There are many different goodness-of-fit tests. One classic test is the chi-square test, which partitions the data into groups, comparing these to the expected groups from a known distribution. There are no fixed methods for selecting these groups, and this test requires a large sample size because at least five observations per group are required to implement it. In addition, the chi-square test does not have the power of the Shapiro-Wilk test or some of the other tests mentioned. For these reasons, the chi-square test is not recommended.

F-6.3. Another way of using a goodness-of-fit test is based on the empirical distribution function. Empirical distribution functions estimate the true cumulative distribution functions underlying a set of data. An empirical distribution is generated from the data set and compared to the theoretical cumulative distribution. If the empirical distribution function is not close to the given cumulative distribution function, then there is evidence that the data do not come from that function.

F-6.4. Various methods have been used to measure the discrepancy between the sample empirical distribution function and the theoretical cumulative distribution function. These measures are referred to as empirical distribution function statistics. The best known of these is the Kolmogorov-Smirnov (K-S) statistic. The K-S approach is appropriate if the sample size exceeds 50 and if $F(x)$ represents a specific distribution with known parameters (e.g., a normal distribution with μ = 100 and σ^2 = 30). A modification to the test, called the Lilliefors K-S test, is appropriate when $n > 50$ for testing that the data are normally distributed and when the $F(x)$ is based on an estimated mean and variance.

F-6.5. Unlike the K-S type statistics, most empirical distribution function statistics are based on integrated or average values between the empirical and cumulative distribution functions. The two most powerful are the Cramer-von Mises and Anderson-Darling statistics. Extensive simulations show that the Anderson-Darling empirical distribution function statistic is as effective as any, including the Shapiro-Wilk statistic, when testing for normality. However, the Shapiro-Wilk test is applicable only to a normal distribution, while the Anderson-Darling method is more general. Because it is unlikely that the user of this manual will ever need to use these tests, they will not be described further. When using a computer software package, a *p* value is typically given. If the *p* value is low (i.e., typically less than 0.01 to 0.1), then the assumption of normality is rejected.

Section II
Data Transformations

F-7. Introduction. Any mathematical function $f(x)$ that is applied to every point in a data set, x, is called a transformation (e.g., $Ln(x)$ is calculated for every data value x). For the transformation

$$y = f(x)$$

the values of x are the original data values and the corresponding values $y = f(x)$ are the transformed data values. An inverse transformation is a function, $f^{-1}(x)$, which, when applied to all of the transformed data values, results in the original data values:

$$f^{-1}(y) = f^{-1}[f(x)] = x \quad .$$

F-7.1. For example, if $y = Ln(x)$, then $f^{-1}(y) = exp(y)$ because $exp[Ln(x)] = x$.

F-7.2. Data transformations are frequently done to obtain normally distributed data sets. By transforming the data, assumptions that are not satisfied in the original data can be satisfied by the transformed data. For example, a right-skewed distribution can often be transformed to be approximately Gaussian (normal) by using a logarithmic transformation or square root transformation. After a data set is transformed, graphical methods and statistical tests verify that the transformed data set is normal. If a transformed data set is normal, then statistical tests that rely on normality are performed using the transformed data. However, finding a transformation that results in a normal data set may be difficult. The selection of a suitable transformation will be dependent upon the nature of the data set and is beyond the scope of this document. Some commonly used transformations will be discussed but only lognormal transformation will be discussed in any detail.

F-7.3. A potential disadvantage of any transformation arises when it is necessary to interpret the results of the statistical evaluation in terms of the untransformed data. For example, in general, if the mean of the transformed data set is calculated, then this quantity will not corre-

spond to the mean of the untransformed data set when an inverse transformation is performed. For example, as previously stated, if $Y = Ln(X)$ is normally distributed with a population mean (μ_Y) and population variance, σ_Y^2, then the mean (μ_Y) corresponds to the population median of X rather than to the population mean of X, μ_X. (Because $X_{p\,=\,0.5}$ is the mean of X and $Z_{0.5} = 0$ in Equation E-12, the median of X is equal to $\exp(\mu)$.)

F-7.4. If a transformation is performed, inverse transformations to the original data set should be avoided. Decisions should be based upon the statistical analyses of only the transformed data. For example, assume that two different data sets are approximately normally distributed with similar variance after transformation. The objective is to determine whether the data sets are significantly different from one another (even though both data sets possess similar variances). The mean of the first transformed data set would be statistically compared to the mean of the second transformed data set. It would be inappropriate to perform inverse transformation for the two means (to express them in the original measurement units) prior to performing the comparison.

F-7.5. While transformations are useful for dealing with data that do not satisfy statistical assumptions, they can also be used for other purposes. Transformations are useful for consolidating data that may be spread out or that have several extreme values. In addition, transformations can be used to derive a linear relationship between two variables, so that linear regression analysis can be applied. Transformations may also make the analysis of data easier by changing the scale into one that is more familiar or easier to analyze.

F-8. Logarithmic. A logarithmic transformation may be useful when the original measurement data follow a lognormal distribution. Data may be lognormally distributed when the variance is proportional to the square of the mean (refer to Equation E-11) or, equivalently, when the coefficient of variation (ratio of standard deviation to mean) is constant over all possible data values:

$$CV = \sigma_X / \mu_X = \text{constant}.$$

F-8.1. For example, if the variance of data collected around 50 ppm is approximately 250, but the variance of data collected around 100 ppm is approximately 1000, then a logarithmic transformation may be useful.

F-8.2. The logarithmic base (either natural or base 10) needs to be consistent throughout the analysis. However, it does not matter whether a natural (Ln) or base 10 (Log) transformation is used because the two transformations are related by a constant:

$$Ln(X) = 2.303\, Log(X).$$

F-8.3. The $Log(x)$ or $Ln(x)$ cannot be transformed when $x = 0$. This is usually not a problem for environmental applications because non-detects are not typically reported as zero but to

some positive reporting (censoring) limit. If some of the original values are zero, it is customary to add a small quantity (ε) to make the data value non-zero, as the logarithm of zero does not exist. However, this introduces some error for the statistical evaluation. The size of ε depends on the magnitude of the non-zero data. It is recommended that the statistical evaluation be performed using several values of ε to determine if it is sensitive to the choice of ε. An initial value of one-tenth of the smallest non-zero value is recommended.

F-9. Square Root.

F-9.1. An overview rather than a detailed discussion of the square root transformation is presented here. The square root transformation may be used when the data values are small whole numbers, such as bacteriological counts, or the occurrence of rare events, such as violations of a standard over the course of a year. The underlying assumption is that the original data follow a Poisson-like distribution, in which case the mean and variance of the data are equal. According to EPA's SW-846 methodology, if the mean and variance of a data set are equal, indicating data from a Poisson distribution, then the data can be transformed using a square root transformation so the data can achieve normality.

F-9.2. The square root transformation overcorrects when very small values and zeros appear in the original data. In these cases, $\sqrt{X+1}$ is often used as a transformation. The square root transformation may also be useful when developing control charts for intrawell comparisons when the assumption of normality is a concern. For further discussion on control charts, see Appendix K.

F-10. Inverse Sine (Arcsine). An overview rather than a detailed discussion of the inverse sine transformation is presented here. This transformation may be used for binomial proportions based on count data to achieve stability in variance. The resulting transformed data are expressed in radians (angular degrees). According to EPA's SW-846 methodology, if the mean is less than the variance of a data set, indicating data from a negative binomial distribution, then data can be transformed using an arcsine transformation to achieve normality. Special tables must be used to transform the proportions into degrees.

F-11. Box-Cox Transformations. An overview rather than a detailed discussion of the Box-Cox transformation is presented here. The Box-Cox transformation is a complex but useful transformation that takes the original data and raises each data observation to the power γ. Box-Cox is typically used in regression modeling (a statistical methodology used to identify the best-fitting equation for a set of data) and would be done using statistical software. Box-Cox is also performed when a data set is not normal, but it is desirable to produce normally distributed transformed data. A logarithmic transformation is a special case of the Box-Cox transformation. The Box-Cox family of transformations is defined as follows:

$$X^{\gamma} = \begin{cases} X^{\gamma}, & \gamma \neq 0 \\ Log(X), & \gamma = 0 \end{cases}$$

where γ is a parameter that defines the transformation (Hahn and Meeker, 1991).

F-11.1. Note both the logarithmic transformation and the square root transformation are simply Box-Cox transformations with $\gamma = 0$ and $\gamma = 0.5$, respectively. The parameter γ is generally unknown. The objective is to find a value of γ such that the transformed data are normally distributed and the variance is as constant as possible over all possible concentration values. In general, transformations with $\gamma < 1$ are applied to normalize positively skewed data, and transformations with $\gamma > 1$ are used to normalize negatively skewed data. The value of γ required to normalize the data decreases (from 1) as the degree of positive skew increases. For example, a transformation with $\gamma = 0.5$ might be applied for a distribution with a slight positive skew, and a value of $\gamma = 0$ (a log transform) might be applied for a more positively skewed distribution. From Hahn and Meeker (1991): "One may try different values of γ (i.e., $\gamma = 1$, 0.5, 0.33, 0, and –1, corresponding to no transformation, square root, cube root, log, and reciprocal transformations, respectively) to try to find a value (or range of values) that gives a probability plot that is nearly linear. In some cases physical considerations or experience may suggest such a value."

F-11.2. Analytical methods are also available, such as the maximum likelihood technique, to find the optimal γ. A statistical software package would be used to find the value of γ for the best transformation; that is, the value of γ that produces the most normal data set once the transformation is applied. For example, if γ is nearly equal to zero (e.g., $\gamma = 0.03$), then a logarithmic transformation ($\gamma = 0$) would typically be selected and would produce a data set that is the most normally distributed relative to other Box-Cox transformations (such as a square root transformation). Statistical tests that require normality would subsequently be performed using the transformed data. However, as is true of any transformation, one of the disadvantages of Box-Cox is the difficulty in interpreting the transformed data in terms of the original measurement units.

Section III
Recommendations

F-12. General. Analysts can perform tests for normality with samples as small as three; however, the tests lack statistical power owing to the small sample size. For small sample sizes, it is recommended that a normal distribution not be assumed for the data and that a nonparametric statistical test, one that does not assume a distributional form of the data, be selected instead. Ideally, an adequate sample size to provide the necessary power for statistical tests will have been selected prior to data collection.

F-12.1. This document recommends using the Shapiro-Wilk *W* test wherever practical, along with a normal quantile plot and box-plot. The Shapiro-Wilk *W* test is one of most powerful tests for normality, and it is recommended in several EPA guidance documents as the preferred test when the sample size is less than 50. The Anderson-Darling statistic is also recommended (e.g., when available via statistical software). A normal quantile plot is helpful, no matter the sample size, to verify results from any test of normality. In practice, with the use of computers it may be possible to perform more than one fitness test, and determine which fit has the highest *p* value.

F-12.2. In general, with large sample sizes, both D'Agostino's test and the Shapiro-Wilk test will be overly sensitive to small deviations from lognormality or normality and will result in an unknown distribution assignment more often than is appropriate. In these cases, close examination of probability plots and the application of professional judgment in determining the appropriate distributional assumptions will be particularly important.

F-12.3. If the Shapiro-Wilk *W* test is not feasible, then using either Filliben's statistic or the Studentized range test is reasonable. Filliben's statistic performs similarly to the Shapiro-Wilk test. The Studentized range is a simple test to use; however, it is not applicable for non-symmetrical data with large tails. If the data are not highly skewed and the tails are not significantly large (compared to a normal distribution), the Studentized range provides a simple and powerful test that can be calculated by hand. If critical values for these tests (for the specific sample size) are not available, then implementing either Geary's test or the Lilliefors Kolmogorov-Smirnoff test is reasonable. Geary's test is easy to apply and uses standard normal tables similar to Table B-15 of Appendix B, and is widely available in standard textbooks. Lilliefors Kolmogorov-Smirnoff is more statistically powerful but is also more difficult to apply and uses specialized tables not readily available.

F-12.4. Statistical professional judgment based on normal probability plots and results of the statistical tests should be considered when identifying a data value's distribution. If the statistician's professional judgment suggests a different distributional assumption than that determined by the statistical test or tests, the alternative distribution may be assumed as long as the statistician provides a defensible rationale for this decision.

F-12.5. It should be stressed the Shapiro-Wilk *W* test is a good test to use to evaluate whether a set of data has been drawn from a normal or lognormal distribution. However, this test will not have very much power to reject the null hypothesis of normality or lognormality if the sample size is small.

F-12.6. In conclusion, results from tests regarding the assumption of normality should always be reviewed graphically.

F-13. Data Fitting Multiple Distributions. When data are found to fit more than one distribution, there are a few things to consider in making a decision about which distribution would be most appropriate. One thing to consider is the *p* value. After running a test of distributional assumptions (Shapiro-Wilk, chi-square, Kolmogorov-Smirnoff, etc.), it would be appropriate to use the distribution that had the higher *p* value. Consideration should be given to the sample size of the data; data containing just a few samples may not provide enough information about the true distribution.

F-13.1. Another thing to question is the purpose of identifying the data's distribution. If it is to verify a distributional assumption for a statistical test and the data fit multiple distributions, it may be appropriate to perform the test using several statistical methods and evaluate results from each to see what can be learned. If a distributional assumption is needed to estimate a confidence interval or upper confidence limit, then it may be appropriate to identify which distribution would provide the more conservative estimate.

F-13.2. It is often difficult to interpret the results of statistical tests conducted on transformed data in terms of the original units to make these types of comparisons. If transformation produces only a slightly larger *p* value, it seems advisable not to perform the transformation. For example, if data follow a normal and lognormal distribution, a lognormal UCL can be quite larger than the normal UCL estimate owing to the inherent nature of a lognormal distribution. If the UCL should be used to evaluate risk at a site, a lognormal UCL would provide the more conservative estimate of risk.