APPENDIX E
Assumptions of Distribution

Section I
Introduction

E-1. One of the essential decisions that precedes many statistical calculations is determining the statistical distribution. Whether the data can be classified as normally distributed, lognormally distributed, meeting some other distribution, or meeting no distributional assumption, dictates how subsequent calculations and statistical tests are chosen and conducted. Distributional assumptions are common in statistical analyses, especially assumptions of normality. Data from environmental studies tend to be skewed rather than following a classical bell-shaped curve, or normal distribution. Thus, verifying distributional assumptions is critical to a successful statistical analysis.

E-2. To provide an objective basis for making this decision, statistical tests are available and discussed in this Appendix. Tests can be applied to the untransformed data when testing for normality or to the log-transformed data when testing for lognormality. Normal probability plots should also be constructed and examined as described in Appendix J.

Section II
Probability Distributions

E-3. Introduction. Many statistical tests and models are appropriate only for data that follow a particular distribution. For a continuous variable X (e.g., the concentration of a contaminant), the distribution is modeled by a mathematical function of the form: $P = P(X)$, where $P(X)$ is referred to as the *probability density function* or *probability distribution.* A plot of P versus X generates a curve. The area (integral) under the curve between any two points, X_a and X_b, gives the probability that the random variable X lies between the two points, $P(X_a \le X \le X_b)$, which will be a number between 0 and 1. The total area under the entire curve is always 1. Figure E-1 plots $P(X)$ and shows how $P(5 < X < 6)$ would be found.

E-3.1. A common use of probability density functions is to calculate population percentiles for the distribution. For example, if $X_{0.95}$ is the value such that $P(X \le X_{0.95}) = 0.95$, then $X_{0.95}$ is referred to as the 95th (population) percentile or 0.95 quantile of X. In general, X_p denotes the p100th percentile or p quantile of X. Appendix D covers techniques to estimate the population percentile from sample data.

E-3.2. Two of the most important distributions for tests involving environmental data are the normal and the lognormal probability distributions. When a parametric statistical test is performed on some set of measured values of X ($x_1, x_2, \ldots, x_n$), some specific probability density function, *P*(*X*), is either known or assumed. This section will provide guidance for determining if the distributional assumption of a given statistical test is satisfied; in particular, the assumption of normality, as this assumption is fundamental to virtually all parametric statistical tests.

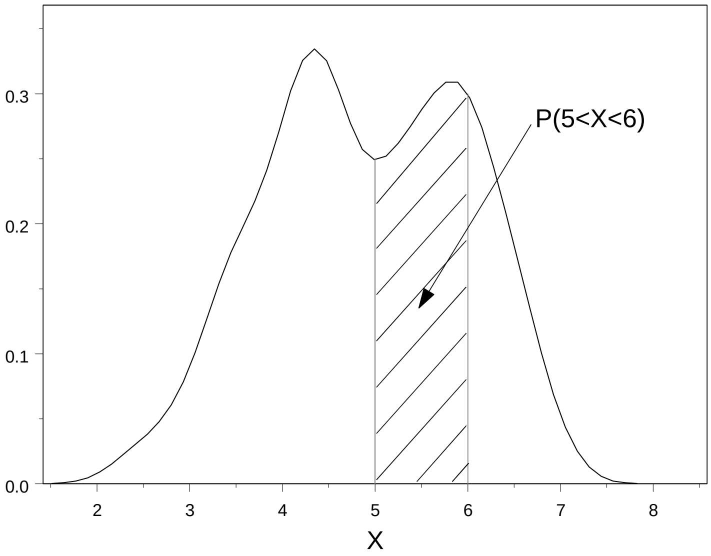

Figure E-1. Probability density function.

E-3.1. *Normal Distribution.* If the variable *X* possesses a *normal* or *Gaussian* distribution (i.e., is said to be *normally distributed*), then the probability density function for *X* is

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right) . \qquad \text{(E-1)}$$

E-3.1.1. A plot of *X* versus *P*(*X*) generates a bell-shaped curve. Two such curves are shown in Figure E-2. The function *P*(*X*) depends on two parameters (constants), the *population mean*, μ, and the *population standard deviation*, σ, where $\sigma > 0$. It is often useful to work with the square of the standard deviation, σ^2, which is referred to as the *population variance*. Note that the normal distribution is symmetrically centered about the mean, μ, and tapers off rapidly at the tails. Because exactly 50% of the distribution falls below the mean, the median (50th percentile) of the normal distribution is equal to the mean. The value of the parameter σ affects the shape of the distribution. In particular, as shown in Figure E-2, as the value of the standard deviation is increased from σ_1 to some value $\sigma_2 > \sigma_1$, the "spread" of the distribution about the mean increases. Because a normal distribution depends upon the parameters, μ and σ, it is often denoted by $N(\mu,\sigma)$.

E-3.1.2. The normal distribution is critical because measurement data (e.g., a set of concentration measurements) can often be modeled by it. When it is known or it can be assumed that a set of measurements, $x_1, x_2, \ldots, x_n$, follow a normal distribution, then the sample mean, $\bar{x}$, is a good estimate of the population mean, μ. Also the sample standard deviation, s, is a good estimate for the population standard deviation, σ. (Refer to Appendix D for the definitions of the sample mean and standard deviation.)

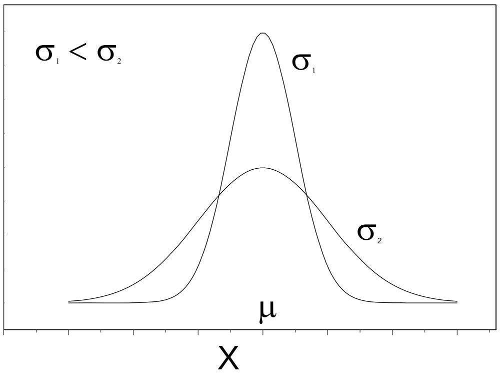

Figure E-2. Normal distribution.

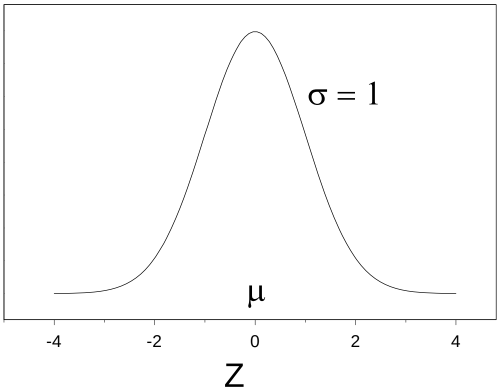

Figure E-3. Standard normal (*Z*) curve.

E-3.1.3. It can be shown, if the random variable X possesses a normal distribution, then the random variable

$$Z = \frac{(X - \mu)}{\sigma} \tag{E-2}$$

has a standard normal distribution, $N(0,1)$. The probability density function of the standard normal distribution is illustrated in Figure E-3. Using the notation from above, we can denote the $p100^{\text{th}}$ percentile (p quantile) of Z as Z_p. The standard normal distribution is important since the percentiles Z_p are commonly listed in statistical tables like Table B-15.

E-3.1.4. For example, if random variable X is $N(3,2)$, we can use Table B-15 to find $X_{0.95}$ as follows. Find the closest value to 0.95 in the interior of Table B-15. In this case 0.9495 and 0.9505 are equally distant. Find $Z_{0.95}$ by the value to the far left of the row found in the last step and the top of the column. Here, it is necessary to interpolate between 1.64 and 1.65 to get $Z_{0.95} = 1.645$. Figure E-4 demonstrates that 95% of the area under the standard normal density curve (the shaded area) lies to the left of 1.645. Returning to the stated problem, solve Equation E-2 for X to get:

$$X_p = \mu + Z_p \sigma \qquad \text{(E-3)}$$

so in this example,

$$X_{0.95} = 3 + 1.645(2) = 6.29 .$$

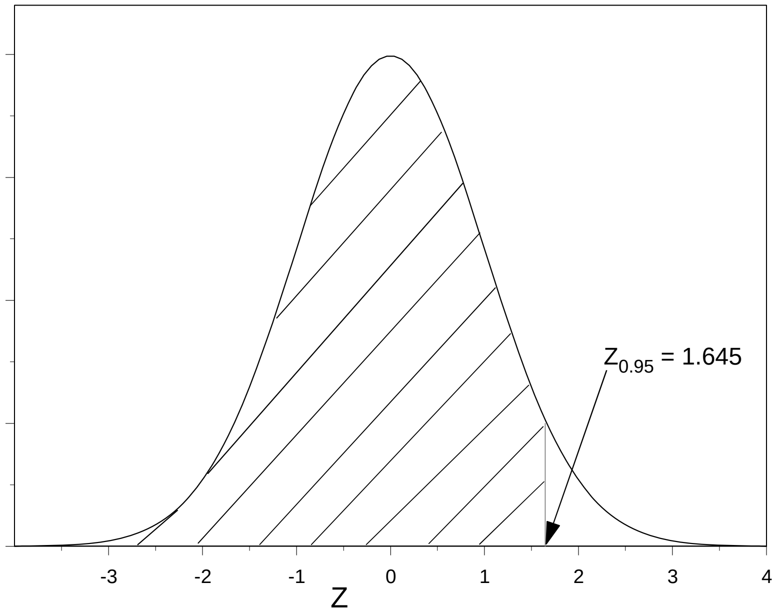

Figure E-4. 95th percentile of the standard normal distribution.

E-3.1.5. Because the standard normal distribution is symmetrical about a mean of zero, $Z_{1-\alpha} = -Z_\alpha$. Thus, the area of the standard normal curve that falls between $Z_{1-\alpha}$ and Z_α is equal to 1 – 2α (e.g., for α = 0.05, 90% of the distribution falls between $Z_{0.05}$ = –1.645 to $Z_{0.95}$ = 1.645). It follows from Equation E-1 that, in terms of the variable *X*, the proportion 1 – 2α (equivalently, 100(1 – 2α)%) of the distribution falls between $X_\alpha = \mu + Z_\alpha \sigma$ and $X_{1-\alpha} = \mu + Z_{1-\alpha} \sigma$. Because $Z_{1-\alpha} = -Z_\alpha$, 100(1 – 2α)% of the distribution falls within $\mu \pm Z_{1-\alpha}\sigma$. Some examples are presented below:

- 90% of the distribution (α = 0.05) falls within the interval $\mu \pm 1.645\sigma$.
- 95% of the distribution (α = 0.025) falls within the interval $\mu \pm 1.960\sigma$.
- 99% of the distribution (α = 0.005) falls within the interval $\mu \pm 2.576\sigma$.

- 99.9% of the distribution (α = 0.0005) falls within the interval $\mu \pm 3.291\sigma$.

E-3.1.6. Thus, approximately 95% of the distribution falls within two standard deviations of the mean ($\mu \pm 2\sigma$) and over 99% (in fact, about 99.7%) of the distribution falls within three standard deviations of the mean ($\mu \pm 3\sigma$). It can similarly be shown that about 68% of the distribution falls within one standard deviation of the mean.

E-3.1.7. Finally, a useful property of the normal distributions is that that any linear combination of normally distributed variables will also be normally distributed. In particular, let

$$Y = \frac{\left(X_1 + X_2 + \cdots + X_n\right)}{n}$$

where each random variable X_i follows the same normal distribution $N(\mu, \sigma)$. It can be shown that the random variable *Y* is distributed as

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

This is extremely useful because the definition of *Y* is very similar to the definition of the sample mean, $\bar{x}$, presented in Appendix G. Thus, if the variable *X* is normally distributed, with mean μ and standard deviation σ, a set of *n* measurements of *X* are taken, the sample mean $\bar{x}$ is calculated for the set of *n* measurements, and this process could be repeated indefinitely. The resulting distribution of values of the sample mean will be normally distributed with mean and standard deviation:

$$\mu_{\bar{x}} = \mu, \quad \sigma_{\bar{x}} = \sigma / \sqrt{n}$$

E-3.1.8. It also follows that

$$Z = \frac{\left(\bar{x} - \mu\right)}{\left(\sigma / \sqrt{n}\right)} \qquad \text{(E-4)}$$

will follow a standard normal distribution. Although σ is not typically known, it can be shown that for sufficiently large *n*,

$$Z = \frac{\left(\bar{x} - \mu\right)}{\left(s / \sqrt{n}\right)} \qquad \text{(E-5)}$$

is closely approximated by a standard normal distribution. Furthermore, if X is normally distributed and n is large, then an approximate $p100\%$ upper bound can be calculated for the population mean from the above equation.

$$\mu \leq \bar{x} + Z_p(s/\sqrt{n}) \tag{E-6}$$

E-3.1.9. The right side of inequality is approximately the *p100% upper one-sided confidence limit (UCL) of the population mean*. For example, if $p = 0.95$, then the right side of the inequality is the 95% UCL of the population mean. For $p = 0.95$, the population mean μ will be less than the UCL an average of 95 out of 100 times. The calculation of a 95% UCL is typically used in environmental risk assessments.

E-3.1.10. Lastly, it should be noted that the UCL is useful because of the central limit theorem. According to the central limit theorem, Equation E-6 is approximately valid for n sufficiently large regardless of whether or not the measurement variable X is normally distributed. The central limits theorem is discussed below.

E-4. Central Limit Theorem. The central limit theorem states:

> If a variable X possesses ANY probability distribution with mean (μ) and finite standard deviation (σ), then the sample mean ($\bar{x}$) will be approximately normally distributed with mean (μ) and standard deviation ($\sigma/\sqrt{n}$)) if n is sufficiently large.

E-4.1. In other words, if a set of n data points is collected and the sample mean is calculated, and this process is repeated many times and all the resulting values of sample mean are plotted (on a histogram), then the resulting distribution will be approximately normal if n is large (i.e., $n > 50$). As the size of the sample increases, the mean of that sample acts increasingly as if it came from a normal distribution regardless of the true distribution of the individual values. As a consequence, statistical tests that require normality may be performed using the sample mean. Thus, large sample sizes are desirable within the limits imposed by available resources.

E-4.2. The central limit theorem is important for environmental applications, because the mean of a random sample of observations or measurements is frequently of interest (for example, to calculate an exposure point concentration for a risk assessment). Furthermore, no actual environmental data set is completely normal. The assumption of normality for any data set will always be an approximation. In many cases, the normality based statistical tests are not overly affected by a small or even moderate deviation from normality as the tests are robust (sturdy) and perform tolerably well, unless gross non-normality is present. The central limit theorem ensures that tests become increasingly tolerant of deviations from normality as the number of individual samples constituting the sample mean increases.

E-5. **Student's *t* Distribution**. The Student's *t* distribution is a continuous probability distribution that is similar in shape to the standard normal distribution. Like the standard normal distribution, the *t* distribution is a bell-shaped curve that is symmetrical about a mean of zero. However, the *t* distribution is somewhat flatter in the center and possesses fatter tails than the standard normal distribution. Furthermore, the shape of the *t* distribution is dependent upon the "degrees of freedom," ν (the Greek letter nu). Each value of ν (ν = 1, 2, 3 …) gives rise to a different *t* distribution curve. The degree of "fatness" in the tails of a *t* distribution depends upon the value of ν. As ν increases, the *t* distribution approaches a normal distribution. These properties are illustrated in Figure E-5. For most practical applications, the *t* distribution may be approximated using a standard normal distribution when $\nu > 30$. The mathematical function that defines the probability distribution is more complex than that for the normal distribution and is not presented.

E-5.1. The standard normal curve is used when the mean (μ) and standard deviation (σ) of a normally distributed population of interest are known. When only an estimate of the standard deviation (s) is available from a sample, the *t* distribution applies. More precisely, if the variable X possesses a normal distribution, then the variable:

$$t_{\nu} = \frac{\bar{x} - \mu}{s / \sqrt{n}} \tag{E-7}$$

possesses a *t* distribution with $\nu = n - 1$ degrees of freedom. The p100% percentiles (p quantiles) of the *t* distribution are denoted as $t_{p,v}$. This value can be found using Table B-23. Find the row matching the degrees of freedom, ν, on the left side of the table. Find the column containing the value p along the top of the table. The value of $t_{p,v}$ is found at the intersection of this row and column. For example, $t_{0.95,10} = 1.812$.

E-5.2. Note that the equation that defines $t_{p,v}$ provides the basis for calculating an upper bound for the mean (μ) when μ is unknown but the sample mean is normally distributed. It can be shown that

$$\mu \leq \bar{x} + t_{p,\nu}(s / \sqrt{n}) \tag{E-8}$$

where the sample mean ($\bar{x}$) and the sample standard deviation (s) are calculated for some set of n data points and the value $t_{p,v}$ is obtained from Table B-23. Roughly speaking, the probability that the population mean will be less than or equal to the right side of the above inequality is p100%. The right side of the above inequality is referred to as the upper one-sided p100% confidence limit of the population mean or simply as the 95% UCL of the population mean.

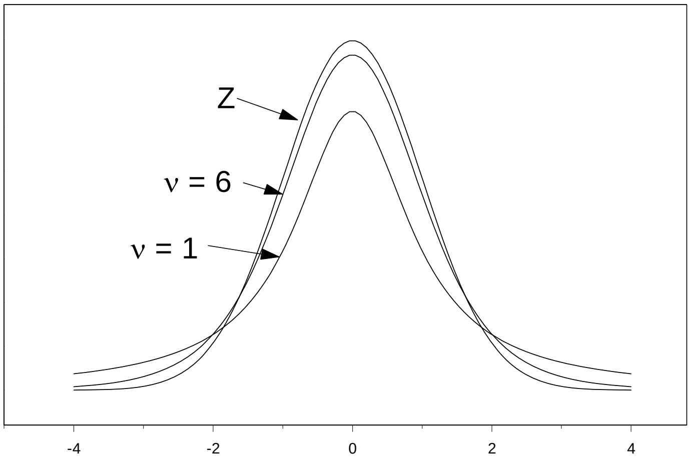

Figure E-5. Comparison of *t*-distribution to standard normal.

E-6. Lognormal Distribution. It is not uncommon for environmental data to follow a lognormal distribution. Data collected from contaminated sites often possess a skewed probability distribution that is easily modeled by a lognormal distribution (EPA 600/R-97/006). This occurs because contaminant concentrations are constrained to be non-zero values, with very high values near a source and declining contaminant concentrations away from source areas.

E-6.1. The lognormal distribution is a continuous, non-symmetrical, positively skewed probability distribution that is bounded to the left by zero. However, like the normal distribution, the lognormal distribution is completely characterized by two parameters that represent the population mean and standard deviation of the log-transformed distribution. Several lognormal distributions are shown in Figure E-6.

E-6.2. There is a simple relationship between the normal and lognormal distributions. If X is lognormally distributed, then $Y = Ln(X)$ is normally distributed. Though the probability distribution is a non-symmetrical, positively skewed curve (where the median of the distribution is less than the mean), the probability distribution for $Y = Ln(X)$ is the symmetrical, bell-shaped normal curve. It is a common practice to transform data using the natural log function to achieve approximate normality prior to conducting statistical tests. Just as the notation $N(\mu, \sigma)$ was used to denote a normal distribution, a lognormal distribution will be denoted by $\Lambda(\mu, \sigma^2)$, where μ and σ^2, denote the population mean and variance, respectively, of the normally distributed variable $Y = Ln(X)$ (*rather than the lognormally distributed variable* X). For brevity, the following notation will be used to indicate that X possesses a log normal distribution: $X \sim \Lambda(\mu, \sigma^2)$, or, equivalently, $Ln(X) \sim N(\mu, \sigma)$.

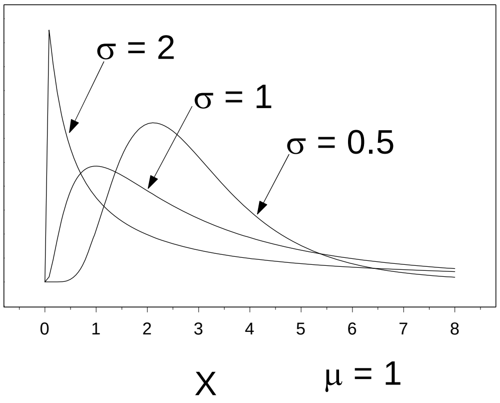

Figure E-6. Lognormal distributions.

E-6.3. Because any linear combination of normally independent distributed variables will be also be normally distributed, owing to the relationship $Y = Ln(X)$, the product a set of independent lognormally distributed variables will also be lognormally distributed. For example, if $X_1 \sim \Lambda(\mu_1, \sigma_1^2)$, and $X_2 \sim \Lambda(\mu_2, \sigma_2^2)$, then

$$X_1 X_2 \sim \Lambda(\mu_{1+}\mu_2, \sigma_1^2 + \sigma_2^2)$$

$$X_1/X_2 \sim \Lambda(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2).$$

Also, if $X \sim \Lambda(\mu, \sigma^2)$, then

$$cX^b \sim \Lambda(a\mu + b, b^2\sigma^2)$$

where c and b are constants, where $c = \exp(a) > 0$ and $b \neq 0$.

E-6.4. The lognormal distribution $\Lambda(\mu, \sigma^2)$ is mathematically described by:

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{X} \exp\left(- \frac{\left(Ln(X) - \mu\right)^2}{2\sigma^2} \right) \quad . \tag{E-9}$$

The population mean, μ_X, and standard deviation, σ_X, of the lognormally distributed variable X are calculated as:

$$\mu_X = \exp\left(\mu + \frac{\sigma^2}{2} \right) \tag{E-10}$$

$$\sigma_X^2 = \exp\left(2\mu + \sigma^2\right)\left[\exp\left(\sigma^2\right) - 1\right] \ = \mu_X^2\,[\exp(\sigma^2) - 1]\,. \qquad \text{(E-11)}$$

It follows that the (population) coefficient of variation of *X* is

$$CV = \mu_X \,/\, \sigma_X = \left[\exp(\sigma^2) - 1\right]^{1/2}.$$

The *p*100% population percentile (*p* quantile), X_p, can be found from the corresponding *p*100% percentile of the standard normal distribution, Z_p, as follows:

$$X_P = \exp\left(\mu + Z_p\sigma\right). \qquad \text{(E-12)}$$

E-7. Binomial Distribution. The binomial distribution is useful in describing the number of successful outcomes, K, from a set number of observations, n. The distribution is considered binomial if the following conditions are satisfied (Moore, 1999):

- The number of observations, *n*, is fixed.

- The n observations are all independent; that is, each observation has no effect on any other.

- Each observation falls into one of two mutually exclusive categories: Each observation is either a “success” or a “failure.”

- The probability each observation is a “success” is *p*. (The probability each observation is a “failure” is 1− *p*).

E-7.1. A common example that gives rise to a binomial distribution would be counting the number of heads (successes) obtained from flipping a coin a set number of times. As the number of successful outcomes, K, is a discrete rather than continuous random variable, then the value of the variable K can equal any integer value from 0 to n. The binomial probability distribution is described mathematically by:

$$P\left(K = k\right) = \frac{n!}{k!\,\left(n-k\right)!}\,p^k\,(1-p)^{n-k} \ . \qquad \text{(E-13)}$$

The population mean, μ, and standard deviation, σ, are given by:

$$\mu = np \qquad \text{(E-14)}$$

$$\sigma = \sqrt{np\left(1-p\right)} \ . \qquad \text{(E-15)}$$

E-7.2. Table B-1 gives probabilities for the binomial in terms of cumulative probability distribution. That is, the table reports:

$$P(K \leq k) = \sum_{i=1}^{k} \frac{n!}{i!(n-i)!} p^{i} (1-p)^{n-i} \quad . \qquad \text{(E-16)}$$

For example, for $n = 4$ and $p = 0.5$, $P(K \leq 2) = 0.6875$.

E-7.3. The binomial distribution under certain conditions can be related to the normal distribution (and the Poisson distribution, as seen in Paragraph E-8). In particular, as n becomes large, the binomial distribution gets close to a normal distribution with mean, np, and standard deviation, $\sqrt{np(1-p)}$. As a rule, this approximation should be used only when both np and $n(1-p)$ are larger than 10 (Moore, 1999).

E-8. Poisson Distribution. The Poisson distribution is useful in describing the number of occurrences of an event over a fixed interval of time. A distribution is considered a Poisson distribution if the following conditions are satisfied:

- The event is a rare occurrence.

- The occurrence of two or more events in a small interval of time is zero.

- A large number of independent observations are made.

- The average number of occurrences, λ, over some fixed interval of time is constant (Mason et al., 1989).

E-8.1. The Poisson distribution is typically used to describe or predict rare events. Data from a Poisson distribution must be independent and must be composed of only two responses, such as detected or not detected. Poisson distributions are common when counting the number of detected or not detected occurrences with environmental data that contain only a small percentage of detected concentrations. The probability for one of the two mutually exclusive outcomes must be small. Therefore, the Poisson distribution can be used for highly censored environmental data because the detection of an analyte in a sample would constitute a rare event. This often occurs for background data when organics are being analyzed (most of the results are reported as not detected).

E-8.2. The Poisson distribution can be used with background data to calculate upper limits for the number of detections for each organic analyte. The limits would subsequently be compared to the study area data to determine if detections for a given organic analyte are being obtained more frequently for the study area than for the background area.

E-8.3. The Poisson distribution may be used for highly censored environmental data in one of two ways. In the first approach, X denotes the number of times an analyte is detected. If the variable X follows a Poisson distribution, then the probability density function is described mathematically by:

$$P(X = x) = \frac{\mu^{x} e^{-\mu}}{x!} \tag{E-17}$$

where μ denotes the mean of the Poisson distribution (such as the average number of times the analyte is detected). For example, if n analyses are performed (n background wells are analyzed for an analyte) and the analyte is detected k times, then the average number of detections, μ, is approximately:

$$\mu \approx \bar{x} = \frac{k}{n}.$$

Data following a Poisson distribution have an equal mean and variance (i.e., $\mu = \sigma^2$).

E-8.4. When n is large and p is small, the binomial distribution and the Poisson distribution give similar results. If follows from Equation E-14 that the probability of detecting the given analyte k out of n times can be calculated using the binomial distribution using the relationship:

$$p = \frac{\mu}{n} \approx \frac{k}{n^2} \ .$$

Therefore,

$$\left\{ P(X = x) = \frac{\mu^{x} e^{-\mu}}{x!} \right\} \approx \left\{ P(K = k) = \frac{n!}{k!(n-k)!} (\mu / n)^{k} (1 - (\mu / n))^{n-k} \right\}.$$

E-8.5 For example, if $k = 6$ and $n = 100$, then

$$\mu \approx \bar{x} = \frac{6}{100} = 0.06$$

and

$$p = \frac{\mu}{n} \approx \frac{k}{n^2} = \frac{6}{100^2} = 0.0006.$$

Using the Poisson distribution, we find that the probability of one detection is

$$P(X = 1) = \frac{0.06^1 e^{-0.06}}{1!} = 0.056506 \ .$$

Using the binomial distribution, we find that the probability is:

$$P(K = 1) = \frac{100!}{1!\,(100 - 1)!}(0.0006)^1 (1 - 0.0006)^{100-1} = 0.056539 \ .$$

As previously stated, these probabilities are very similar as p is small and n is large.

E-8.6. In a second approach, X may denote the concentration per sample rather than the number of detections. In this context, sometimes referred to as the “molecular approach,” n samples are analyzed, the analyte is detected in the i^{th} sample at a concentration of x_i, and units for the n measurements are selected such that $x_i > 1$. For example, $x_i = 2\,\mu\text{g}/\text{L} = 2\,\text{ppb}$. In this example, the i^{th} sample is detected at two units or occurrences per billion units of sample examined. (The Poisson distribution is appropriate since the ratio of analyte to sample is small.) The mean concentration per sample (mean number of units per billion units of sample examined) will be:

$$\mu \approx \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}. \tag{E-18}$$

Using this approach, we can readily calculate the probability that the analyte will be detected at a concentration X when X is a whole number.

E-8.7. Note the difference between the two approaches. For the first approach, the mean number of detections for a set of n samples is being calculated. A detection, regardless of the magnitude of the reported concentration greater than the detection limit, consists of a unit count for the calculation of the mean. In the second approach, the mean concentration or number of counts per sample is being calculated; thus, the magnitude of detected concentrations for an individual sample influences the estimation of the mean.

E-8.8. A useful property of the Poisson distribution is that, if the independent variables X_1, $X_2 \ldots X_n$ possess Poisson distributions with means μ_1, $\mu_2 \ldots \mu_n$, respectively, then the sum of the variables

$$Y = \sum_{i=1}^{n} X_i$$

has a Poisson distribution with mean

$$\mu_Y = \sum_{i=1}^{n} \mu_i \quad .$$

Therefore, if all of the means $\mu_i = \mu$, it follows that $\mu_Y = n\mu$ and $\mu = \frac{\mu_Y}{n} \approx \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$.

E-8.9. As the parameter, μ, becomes very large, the Poisson distribution can also be approximated by a normal distribution. In this case the mean and variance of the normal distribution equal to μ.

E-9. Nonparametric (Distribution Free). Nonparametric statistical methods are used when it is inappropriate to assume some underlying distribution for a data set (when a data set does not conform to some desired theoretical probability distribution). Sometimes it is difficult to verify or satisfy the assumptions that are associated with parametric distributions, such as normal and lognormal distributions for environmental data sets. Using parametric statistical tests when the appropriate assumptions have not been met can result in inaccurate conclusions. In this situation, nonparametric (distribution free) statistical procedures would be appropriate and recommended (Gilbert, 1987; Hahn and Meeker, 1991).