APPENDIX L
Hypothesis Testing—Simple Cases

L-1. Introduction. This Appendix provides an extensive discussion of the statement of hypotheses (null and alternative) and the consequences deriving from that choice. Also, a general introduction of the basic types of hypothesis testing commonly employed in environmental operations is provided. Further reading on the foundations of hypothesis testing can be found in EPA 600/R-96/055, QA/G-4. Additional reading on the one-sample hypothesis tests presented below can be found in EPA 600/R-96/084, QA/G-9.

L-2. Translating Objectives into Statistical Hypotheses. A data user's question, or a decision rule from the DQO process, must be translated into a precise statistical statement to be tested using environmental data. Such a statement is called a hypothesis. It includes a null hypothesis (H_0) and an alternative hypothesis (H_A). The null hypothesis is a baseline condition presumed to be true in the absence of strong evidence to the contrary, and the alternative hypothesis is the opposite condition that bears the burden of proof. In other words, unless it is demonstrated that the alternative hypothesis is correct based upon weight of evidence, the baseline condition is retained.

L-2.1. A hypothesis test consists of the following elements.

L-2.1.1. It has a quantitative population parameter of interest describing the feature of the environment that the data user is investigating, such as a mean, median, or proportion,

L-2.1.2. It has a numerical value to which the parameter of interest will be compared, such as a regulatory or risk-based threshold or a similar parameter from another population (i.e., comparison to a reference site) or time (i.e., comparison to a prior time).

L-2.1.3. It has a relation that specifies precisely how the parameter will be compared to the numerical value, such as "is equal to" or "is greater than."

L-2.2. *If the data user is interested in drawing inferences about only one population*, the null and alternative hypotheses are stated in terms that relate the true value of the parameter to some fixed threshold value. A typical example of this one-sample problem in environmental studies is when the concentration of a contaminant is compared to a fixed regulatory limit or threshold value. For example, a data user may wish to determine whether the true mean concentration (μ) of the herbicide atrazine in groundwater at a hazardous waste site is greater than a fixed threshold value C, determined from a human or ecological risk assessment. If the decision maker wishes to "prove" that the contamination is less than C, it is initially assumed that the true (population) mean concentration is greater than or equal to C. This assumption is known as the null hypothesis and is denoted as H_0. If the data provide compelling evidence that the null hypothesis is false, then the null hypothesis is rejected and it would be concluded that the population mean concentration is less than C. The opposite conclusion is known as the alternative

hypothesis and is denoted as H_A or H_1. For this example, the null and alternative hypotheses can be stated as follows:

$$H_0 : \mu \geq C, \quad H_A : \mu < C \ .$$

L-2.2.1. The null hypothesis (H_0) is the mean is greater than or equal to the threshold value *C*. The alternative hypothesis (H_A) is the opposite condition: the mean is less than the threshold value *C*.

L-2.2.2. If the decision maker wishes to demonstrate that the true mean is greater than the threshold value, the data must provide compelling evidence to reject this presumption, and the hypotheses can be stated as follows:

$$H_0 : \mu \leq C, \quad H_A : \mu > C \ .$$

L-2.2.3. Note that, thus far, two possible null hypotheses, $\mu \leq C$ and $\mu \geq C$, have been discussed. Depending upon the data quality objectives of the project, *it is possible to legitimately assign either alternative to the null hypothesis*. Because of this freedom or ambiguity, the most appropriate assignment must be determined from the project's data quality objectives.

L-2.2.4. Lastly, it should be noted that the null and alternative hypotheses for the examples presented above would be used for a *one-sample, one-tailed* statistical test. Typically, the sample mean of some set of measured concentrations would be statically compared to the threshold, *C*. The test is *one-sample* in nature because one data set (from one population) is used to calculate the test statistic, the sample mean. If, however, the statistical test entailed the use of two different data sets, in which each was potentially drawn from a separate population, it would be described as a *two-sample* test. The test is *one-tailed* in nature when the null hypothesis is an inequality. Although less common for environmental applications, the null and alternative hypotheses for the corresponding *one-sample two-tailed test* are as follows:

$$H_0 : \mu = C, \quad H_A : \mu \neq C \quad \text{(i.e., } \mu > C \text{ or } \mu < C\text{)} \ .$$

L-2.2.5. The null hypothesis is that the population mean is equal to *C* and the alternative hypothesis is that the population mean is either greater than or less than *C*.

L-2.3. *If two populations are being compared*, the null and alternative hypotheses are stated in terms that compare the true parameter value of one population to the corresponding true parameter value of the other population. A common example of this two-sample problem is when a potentially contaminated waste site is compared to a reference area using samples collected from the respective areas. In this situation, the hypotheses often are stated in terms of the

difference between the two parameters; for example, the difference between the mean site concentration and the mean background concentration:

$$H_0 : \mu_{Site} - \mu_{Background} \le 0, \quad H_A : \mu_{Site} - \mu_{Background} > 0 \ .$$

L-2.3.1. The hypothesis above would be used for a *two-sample*, *one-tailed* statistical test. As previously stated, the null and alternative hypotheses must be determined from project data quality objectives. Environmental regulations may specify particular null and alternative hypotheses. For example, the null hypothesis for a RCRA facility groundwater monitoring program is as follows: The concentration in down-gradient groundwater is less than or equal to the background concentration. When the null hypothesis is not specified by regulation, however, this determination should be made by carefully considering the consequences of making decision errors and taking the wrong actions. Selecting the null hypothesis is extremely important to the outcome of the decision process. The same set of sample data from a decision unit can lead to different decisions, depending on which possibility was selected as the null hypothesis.

L-2.3.2. Typically, hypothesis tests are established to prove a desired hypothesis. The condition or alternative that requires proof is selected as the alternative or research hypothesis. The alternative hypothesis is accepted (via burden of proof) when the null hypothesis is rejected (that is, disproved) based upon the weight of the evidence.

L-2.4. EPA 600/R-96/055, QA/G-4 recommends that the null hypothesis be defined as the true condition associated with the "more severe decision error"; that is, the more undesirable outcome if a wrong decision were made. For example, when the mean concentration of a contaminant is compared to a risk-based action level, C, the most severe decision error often consists of concluding $\mu < C$ when $\mu \ge C$ is the true condition. Therefore, as per EPA guidance, the null hypothesis is often $\mu \ge C$. In other words, it would typically be assumed that the site is "dirty" (H_0: $\mu \ge C$) until the weight of evidence demonstrates that the site is "clean" (H_A: $\mu < C$), the hypothesis that one wishes to demonstrate.

L-2.5. Rather than defining the null hypothesis based on the most severe condition, a second approach consists of defining the null hypothesis based on the least probable condition (or, equivalently, the alternative hypothesis based on the most probable condition). According to this approach, if a large amount of existing information suggests that one hypothesis is extremely likely, then this hypothesis would be defined as the alternative hypothesis. The advantage of this approach is that a large number of data may not be necessary to provide overwhelming evidence that the null hypothesis is false. For example, if the waste from an incinerator was previously hazardous and the waste process has not changed, it may be more cost-effective to define the alternative hypothesis as "the waste is hazardous" (H_A: $\mu \ge C$) and the null hypothesis as "the waste is not hazardous" (H_0: $\mu < C$). This approach generally will not result in the same null hypothesis as the approach EPA recommends. The most protective alternative for H_0 will not necessarily be the least probable alternative for H_0 (i.e., the most probable alternative for H_A).

Table L-1.
Commonly Used Statements of Statistical Hypotheses

Type of Decision	Null Hypothesis	Alternative Hypothesis
Compare environmental conditions to a fixed threshold value, such as a regulatory standard or acceptable risk level; presume that the true condition is less than the threshold value.	H_0: $\Theta \leq C$	H_A: $\Theta > C$
Compare environmental conditions to a fixed threshold value; presume that the true condition is greater than the threshold value.	H_0: $\Theta \geq C$	H_A: $\Theta < C$
Compare environmental conditions to a fixed threshold value; presume that the true condition is equal to the threshold value and the data user is concerned whenever conditions vary significantly from this value.	H_0: $\Theta = C$	H_A: $\Theta \neq C$
Compare environmental conditions associated with two different populations to a fixed threshold value (δ_0), such as a regulatory standard or acceptable risk level; presume that the true condition is less than the threshold value. If it is presumed that conditions associated with the two populations are the same, the threshold value is 0.	H_0: $\Theta_1 - \Theta_2 \leq \delta_0$ If $\delta_0 = 0$, H_0: $\Theta_1 - \Theta_2 \leq 0$ H_0: $\Theta_1 \leq \Theta_2$	H_A: $\Theta_1 - \Theta_2 > \delta_0$ If $\delta_0 = 0$, H_A: $\Theta_1 - \Theta_2 > 0$ H_A: $\Theta_1 > \Theta_2$
Compare environmental conditions associated with two different populations to a fixed threshold value (δ_0), such as a regulatory standard or acceptable risk level; presume that the true condition is greater than the threshold value. If it is presumed that conditions associated with the two populations are the same, the threshold value is 0.	H_0: $\Theta_1 - \Theta_2 \geq \delta_0$ If $\delta_0 = 0$, H_0: $\Theta_1 - \Theta_2 \geq 0$ H_0: $\Theta_1 \geq \Theta_2$	H_A: $\Theta_1 - \Theta_2 < \delta_0$ If $\delta_0 = 0$, H_A: $\Theta_1 - \Theta_2 < 0$ H_A: $\Theta_1 < \Theta_2$
Compare environmental conditions associated with two different populations to a fixed threshold value (δ_0), such as a regulatory standard or acceptable risk level; presume that the true condition is equal to the threshold value. If it is presumed that conditions associated with the two populations are the same, the threshold value is 0.	H_0: $\Theta_1 - \Theta_2 = \delta_0$ If $\delta_0 = 0$, H_0: $\Theta_1 - \Theta_2 = 0$ H_0: $\Theta_1 = \Theta_2$	H_A: $\Theta_1 - \Theta_2 \neq \delta_0$ If $\delta_0 = 0$, H_A: $\Theta_1 - \Theta_2 \neq 0$ H_A: $\Theta_1 \neq \Theta_2$

L-2.6. Table L-1 summarizes common environmental decision rules and the corresponding hypotheses. The population parameter of interest (e.g., μ) in this table is denoted by the symbol Θ and the difference between two population parameters is denoted as $\Theta_1 - \Theta_2$, where Θ_1 represents the parameter of the first population (such as a constituent from a hazardous waste site) and Θ_2 represents the parameter of the second population (such as a constituent from background). The use of Θ is intended to avoid using the terms "population mean" or "population median" repeatedly because the structure of the hypothesis test remains the same regardless of the population parameter. The fixed threshold value is denoted as C, and the difference between two parameters is denoted as δ_0 (often the null hypothesis is defined such that $\delta_0 = 0$).

L-2.7. As previously discussed, hypothesis tests may be one-tailed or two-tailed, depending on the specified null and alternative hypotheses. The first, second, fourth, and fifth rows of Table L-1 are examples of one-tailed hypothesis tests. The third and sixth rows are examples of two-tailed tests. Most hypotheses connected with environmental monitoring are one-tailed because high pollutant levels can cause harm to humans or ecosystems, whereas lowered concentrations are of little, if any, concern.

L-3. Decision Errors Associated with Hypothesis Tests. Table L-2 presents all of the possible scenarios that can result from a statistical hypothesis test. Two correct decisions and two incorrect decisions are possible. The probability of each event is presented in parenthesis.

Table L-2.
Conclusions Associated with Any Statistical Hypothesis Test

		True Hypothesis (Actual site conditions)	
		H_0 True	**H_a True**
Decision (Conclusion from sample data)	**Do Not Reject H_0**	Correct decision Confidence Level = $(1 - \alpha)100\%$	Incorrect decision False Acceptance of H_0 Type II error tolerance = β
	Reject H_0	Incorrect decision False Rejection of H_0 Type I error tolerance = α	Correct decision Power of test = $(1 - \beta)100\%$

L-3.1. The two incorrect answers for a hypothesis test are the following.

L-3.1.1. *False rejection of* H_0*, or Type I error*. The null hypothesis is rejected when the null hypothesis is true. The probability for a Type I error is defined as the level of significance. The maximum allowable probability for a Type I error is typically denoted by the symbol α. The level of confidence is defined as one minus the level of significance. Thus, the minimum level of confidence for a correct decision is $1 - \alpha$.

L-3.1.2. *False acceptance or Type II error.* The null hypothesis is accepted (more accurately, not rejected) when the null hypothesis is false. The maximum allowable probability for a Type II error is denoted by the symbol β. The power of the test is defined as one minus the Type II error probability. Therefore, the minimum power is $1 - \beta$.

L-3.2. A false rejection decision error occurs when it is concluded, from the observed data, that the null hypothesis is false when it is actually true. (This is sometimes called a "false positive.") A false acceptance decision error occurs when it is concluded that the null hypothesis is true when it is really false. (This is sometimes called a "false negative.") For example, suppose the null hypothesis states that the true value of the parameter of interest exceeds the action level. If the null hypothesis is actually correct and the sample data, by chance, contained an abnormally

large proportion of low values, it would be concluded that the true value did not exceed the action level; therefore, a false rejection decision error would occur.

L-3.3. Three different equivalent approaches can be used to perform hypothesis tests: "The confidence interval," "*p*-value," and "critical value" approaches. Table L-3 illustrates the use of each of these three approaches for hypothesis testing.

Table L-3.
Relationship Between Hypothesis Tests and Confidence Intervals

Hypotheses	*p*-Value Approach Reject H_0 when	Critical Value Approach Reject H_0 when	Confidence Interval Approach Reject H_0 when
$H_0 : \Theta = C$ $H_A : \Theta \neq C$	$p < \alpha$	Test statistic less than or greater than critical values. Example: $t < t_{\alpha/2,n-1}$ or $t > t_{1-\alpha/2,n-1}$	Two-sided $1-\alpha$ confidence interval for Θ does not contain C
$H_0 : \Theta \geq C$ $H_A : \Theta < C$	$p < \alpha$	Test statistic less than "critical value. Example: $t < t_{\alpha,n-1}$.	One-sided $1-\alpha$ upper confidence interval limit for Θ is less than C: UCL < C
$H_0 : \Theta \leq C$ $H_A : \Theta > C$	$p < \alpha$	Test statistic exceeds "critical value." Example: $t > t_{1-\alpha,n-1}$.	One-sided $1-\alpha$ lower confidence limit for Θ is greater than C: LCL > C

L-3.4. Table L-3 lists the possible null hypotheses for a one-sample statistical test. The objective is to determine if some population parameter of interest, Θ (the value of which is typically known) equals, is less than, or is greater than some fixed threshold value *C*. For the critical value approach for hypothesis testing, the decision to reject the null hypothesis is essentially determined by calculating some sample test statistic and comparing the value of the sample test statistic to a threshold or "critical value" for the sample statistic. If the sample statistic is greater than or less than the "critical value" (depending upon the null hypothesis selected), the null hypothesis is rejected.

L-3.5. Confidence intervals are directly related to hypothesis tests. Whenever a hypothesis test can be used to evaluate a parameter of interest (such as the mean, variance, median, etc.), a confidence interval also can be estimated and used to evaluate the same parameter. An equivalent approach consists of the following: Use the sample data to derive an estimate of the population parameter $\hat{\Theta}$, construct a confidence interval for Θ using the estimate $\hat{\Theta}$, and determine whether *C* falls within the confidence interval for Θ. If *C* does not fall within the confidence interval for Θ, then the null hypothesis is rejected. This is referred to as the "confidence interval approach for hypothesis testing."

L-3.6. A third approach for hypothesis testing is referred to as the "*p*-value approach for hypothesis testing." The "*p*-value" is the probability of obtaining the calculated sample statistic if the null hypothesis is true. If the *p*-value is sufficiently small, that is, if $p < \alpha$, where α is the Type I error tolerance, then the null hypothesis is rejected. All three approaches are illustrated below. This document predominately uses the critical value approach for hypothesis tests.

L-4. Illustration of Hypothesis Testing. To illustrate hypothesis testing, a one-population test to threshold value *C* is considered, with the following null and alternative hypotheses:

$$H_0 : \mu \geq C, \quad H_A : \mu < C \ .$$

Assume that the variable *X* is normally distributed with an *unknown* population mean μ but a *known* standard deviation σ. A single sample measurement *x* is compared to the threshold value, *C*, to determine whether or not to reject the null hypothesis, H_0: $\mu \geq C$. Because the standard deviation of the population (σ) typically would not be known for environmental applications, the example is not realistic, but serves only to illustrate the concept of hypothesis testing. Figure L-1 illustrates the decision errors for hypothesis testing.

L-4.1. *Type I Error Tolerance and the Rejection of the Null Hypothesis*. If the null hypothesis is true with $\mu = C$, a distribution of measured values of *X* would be obtained, as shown by the blue normal curve centered about $\mu = C$. The probability that a measurement, *x,* would be less than the critical value, X_α, is equal to α (refer to the region shaded in blue). The value of X_α depends upon the α value selected. The value of α is determined from the project's data quality objectives but is usually some acceptably small positive number (e.g., $\alpha = 0.01$ or 0.05). As the probability a measurement, *x*, will be less than X_α is acceptably small when $\mu = C$, the null hypothesis (H_0: $\mu \geq C$) is rejected when a measurement of $x < X_\alpha$ is obtained. (The null hypothesis is retained when $x > X_\alpha$.) The value α represents the tolerance for Type I error; that is, the maximum acceptable probability for rejecting H_0 when H_0 is actually true. When H_0 is $\mu \geq C$, the Type I error can be roughly described as the probability of concluding that a "dirty" site is "clean."

L-4.1.1. When *X* is normal with known standard deviation, σ, it is convenient to "standardize" the variable *X* using the linear transformation:

$$Z = \frac{X - \mu}{\sigma} \ .$$

L-4.1.2. The variable *Z* is a standard normal variable. If $x < X_\alpha$, it follows that

$$z = \frac{x - \mu}{\sigma} < Z_\alpha = \frac{X_\alpha - \mu}{\sigma} \ .$$

L-4.1.3. The quantity Z_α is the $\alpha 100^{th}$ percentile of the standard normal distribution. Thus, if the null hypothesis $\mu = C$ is true and $x < X_\alpha$, then

$$z = \frac{x - C}{\sigma} < Z_\alpha \ .$$

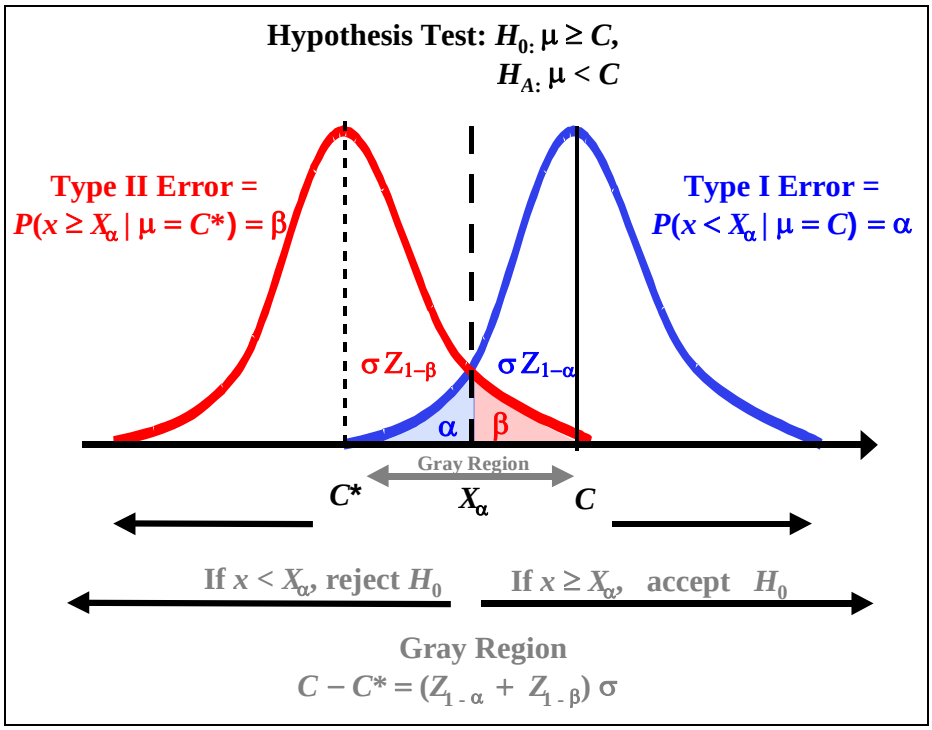

Figure L-1. Decision errors associated with a hypothesis test.

L-4.1.4. Because H_0 is rejected when $x < X_\alpha$, it may be also be rejected when the test statistic $z < Z_\alpha$. In this context, the percentile Z_α is called the "critical value." If the sample statistic z is less than the "critical value" Z_α, it is often stated that the null hypothesis is rejected at the "α100% level of significance" or, equivalently, at the "$(1 - \alpha)$100% level of confidence. " This is a convenient approach as the sample test statistic z can be calculated and compared to a desired percentile of the standard normal distribution (Z_α), which is readily available from a statistical table. The comparison of a sample statistic such as z to some percentile Z_α to determine whether or not to reject H_0 is referred to as the "critical value approach."

L-4.1.5. Statistical software provides an alternative to the critical value approach (for determining whether H_0 should be rejected), referred to as the "*p* value approach." For this particular example, given that a measure *x* from a normal distribution with known standard deviation (σ) is taken, the software also initially assumes that the null hypothesis is true (i.e., sets $\mu = C$), and calculates *z*. The calculated value is assumed to be equal to some percentile, $Z_{p,}$ of the standard normal distribution. Rather than reporting the statistic *z* and comparing it to the percentile $Z_{\alpha,}$ the software outputs the fraction of the normal probability distribution, *p*, that falls below the calculated value of *z* when $\mu = C$. This value is referred to as the "*p* value." The *p* value is the probability of obtaining a measured result of *x* (or a result different than the null hypothesis) when the null hypothesis is true ($\mu = C$). If *p* is sufficiently small relative to α (i.e., $p < \alpha$), the null hypothesis is rejected.

L-4.1.6. The third approach is referred to as the "confidence interval approach for hypothesis testing." It entails calculating a confidence interval for the population mean μ. In this situation, the best estimate of μ is the single measurement *x*. Because rejecting the null hypothesis requires

$$\frac{x - C}{\sigma} < Z_{\alpha}$$

and $Z_{\alpha} = -Z_{1-\alpha,}$ it follows that the null hypothesis would be rejected if:

$$\text{UCL} = x + Z_{1-\alpha}\,\sigma < C \ .$$

L-4.1.7. The left side of the inequality is the one-sided $(1 - \alpha)100\%$ upper confidence limit for μ for a normal distribution with known standard deviation σ. Therefore, the null hypothesis is rejected if the UCL for μ is less than *C*. More information on confidence limits is contained in Appendix N.

L-4.1.8. The strategies discussed above are generally applicable for hypothesis tests, but the critical value approach is predominately used in this document.

L-4.2. *Type II Error and Power*. The discussion above focused on the criteria for rejecting the null hypothesis. The alternative hypothesis is discussed here. When the alternative hypothesis is true with $\mu = C^* < C$ (when the mean [μ] is equal to some value $C^* < C$), a normal distribution of measurements centered about $\mu = C^*$ will be obtained (refer to the red normal curve). When $\mu = C^*$, the probability $x > X_{\alpha}$ equals β (refer to the red shaded region). Because the null hypothesis is retained when $x > X_{\alpha}$, β is equal to the probability of retaining the null hypothesis (H_0: $\mu \geq C$) when the null hypothesis is false (i.e., when $\mu = C^* < C$). The value of β determined from project data quality objectives represents the maximum tolerance for Type II error; that is, the maximum tolerable probability for erroneously retaining the null hypotheses. In terms of an

environmental investigation, the Type II error can be roughly described as the probability of concluding that a clean site is dirty. The power of the hypothesis test is defined as $1 - \beta$ and is equal to the probability of accepting the alternative hypothesis ($\mu = C^* < C$) when the alternative hypothesis is true (the probability of concluding that a clean site is clean).

L-4.2.1. Note that, to calculate the Type II error or the power of a test, the Type I error must first be specified. Also, note that, in this example, the Type II error tolerance and power is for some *pre-specified* value $C^* < C$. Paragraph L-5.2 illustrates how to calculate the power once α and C^* are specified for a normally distributed variable X with a known population standard deviation.

L-4.2.2. When the mean (μ) is equal to some value greater than C (when it falls somewhere to the right of C), the probability that the null hypothesis will be rejected is acceptably small, less than α. The probability that the null hypothesis will be retained will be greater than $1 - \alpha$. In terms of an environmental study, when $\mu > C$, the probability that a dirty site will be identified as dirty will be acceptably high. Similarly, when the mean (μ) is equal to some value less than C^*, the probability of retaining the null hypothesis (H_0: $\mu \geq C$) will be less than β. The probability of correctly rejecting the null hypothesis (and accepting H_A: $\mu < C$) will be greater than $1 - \beta$. When $\mu < C^*$, the probability that a clean site will be identified as clean will be acceptably high. However, when μ lies between C and C^*, the probability of making a correct decision will be low (the Type II error will be higher than β). This range of values, $C - C^*$, is called the "gray region" or the "minimum detectable difference." Because reliable decisions cannot be made for differences smaller than $C - C^*$, the difference $C - C^*$ may be viewed as the "resolution" of the statistical design.

L-4.2.3. Statistical tests cannot control both types of error simultaneously. Generally, a hypothesis test is set up in a manner that committing false rejection (Type I) is considered the more serious error and is controlled by the test, and committing false acceptance (Type II) is considered not as serious an error and is not controlled by the test. The data user specifies the probability limit, α, by the data user's tolerance for committing false rejection (Type I). Determining how large a risk the project team is willing to tolerate for Type I errors must be done *before the fact*, especially when the consequences of making such an error are very serious (Milton and Arnold, 1990). If the null hypothesis is not rejected after the test is performed, then the Type II error or the power (one minus the Type II error) is calculated. If the Type II error is not sufficiently small (or equivalently, the power is not sufficiently large), additional sampling would be considered. In general, increasing the sample size simultaneously reduces both Type I and Type II errors.

L-4.2.4. If the sample mean, $\bar{x}$, for a set of n measurements, rather than a single measurement, were compared to the threshold, C, to determine whether or not to reject the null hypothesis (H_0: $\mu \geq C$), then the minimum detectable difference would be given by:

$$C - C^* = \left(\sigma / \sqrt{n}\right)\left(Z_{1-\alpha} + Z_{1-\beta}\right) .$$

L-4.2.5. The number of random samples that must be collected can be solved from the above equation:

$$n = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2 \sigma^2}{\left(C - C^*\right)^2} .$$

L-4.2.6. Hence, the number of samples is dependent upon α, β, σ, and $C - C^*$. The number of samples increases as the tolerance of Type I and Type II error, α and β, decreases (as $Z_{1-\alpha}$ and $Z_{1-\beta}$ increase as α and β decrease). The number of samples also increases as the variance (σ^2) increases and $C - C^*$ decreases. This is reasonable because the variance is a measure of the variability of the underlying environmental population and $C - C^*$ is a measure of the resolution of the statistical design. The number of samples increases as variability or heterogeneity of the underlying populations increases. As the probability of making a correct decision when the true mean lies in the gray area is low, the quantity $C - C^*$ essentially represents the smallest difference between the mean contaminant concentration and the threshold level that can be tolerated or that is deemed to be important for the overall statistical design. The sample size increases when smaller differences become significant for the statistical design.

L-5. Statistical Power Associated with Hypothesis Tests. As previously stated, the power of a statistical hypothesis test is defined as the likelihood that the null hypothesis is correctly rejected at a fixed level of significance, α, when the alternative hypothesis is truly correct. Power is related to Type II errors, or false rejection. The power of a statistical test is $1 - \beta$ where β is the probability of a false acceptance or Type II error. Therefore, as the power of a statistical test increases, the probability of a false acceptance decreases.

L-5.1. *Introduction*. To calculate the power of a statistical test, first determine the event that the test rejects the null hypothesis, H_0, in a form that does not contain any unknown parameters. There must be a predetermined level of significance, α, so there is a set criterion for rejecting the null hypothesis. The power is the calculated probability for rejecting the null hypothesis when the alternative hypothesis is assumed to be true. Unfortunately, the specific algorithm for calculating power is highly dependent upon the nature of the statistical test and power calculations are often complex. Paragraph L-5.2 presents directions for calculating the power for a hypothesis test of the form:

$$H_0 : \mu \le C, \quad H_A : \mu > C .$$

(Refer to Figure L-1.) The variable of interest is assumed to be normally distributed and the population standard deviation is known. The assumption that the population standard deviation

(σ) is known severely limits the utility of the approach. However, it constitutes, perhaps, the simplest method to estimate power. In practice, an estimate of σ could be used to estimate the power if the uncertainty associated with the estimate was sufficiently small.

L-5.2. *Example for Calculating the Power of a One-Tailed Test* (from Mason et al., 1989). This procedure is strictly applicable only when the variable *X* is *normally* distributed with a *known* standard deviation. The procedure could potentially be used (to estimate the power) when the (population) standard deviation is not known and the sample is mean is calculated from a large number of samples (e.g., $n > 100$).

L-5.2.1. Suppose

$$H_0 : \mu \le 10, \quad H_A : \mu > 10 \ .$$

Assume a known standard deviation of $\sigma = 2$ for a normally distributed population. Let the Type I error tolerance for rejecting the null hypothesis $\alpha = 0.05$ and the sample size $n = 25$. Note that the threshold value $C = 10$. Let $C^* = 11$ in this example. Thus the "resolution" for the test, $C^* - C = 1$. Under the null hypothesis, the largest mean $\mu_0 = 10$. It follows that the power of the test is as follows:

$$1-\beta = P(\bar{x} > 10 \,|\, \mu = 11) = P\left\{\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} > Z_{1-\alpha}\right\} = P\left\{\frac{\bar{x}-10}{2/\sqrt{25}} > 1.645\right\} = P\left\{\frac{\bar{x}-11}{2/\sqrt{25}} > 1.645 - \frac{11-10}{2/\sqrt{25}}\right\}$$
$$= P(Z > -0.855) = 1 - P(Z \le -0.855) = 0.804 \ .$$

$Z_{1-\alpha}$ is the $(1-\alpha)100^{\text{th}}$ percentile of the standard normal distribution, which is provided in Table B-15 of Appendix B.

L-5.2.2. More generally, when comparing the sample mean (of a normally distributed variable with standard deviation σ) to some decision limit μ_0 using the null hypothesis, $H_0 : \mu \le \mu_0 = C$, the power at $\mu = \mu_1 = C^*$ is as follows:

$$1-\beta = P(\bar{x} > \mu_0 \,|\, \mu = \mu_1) = 1 - P\left\{Z \le \left(Z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)\right\} \ .$$

L-5.2.3. For this particular example, the experiment has a probability of 0.804 of correctly rejecting the null hypothesis when the true population mean is $\mu = 11$. If this power is not ac-

ceptably large, the sample size must be increased to maintain the same significance level. For example, a sample size $n = 50$ would produce the following power:

$$1-\beta = 1-P\left\{Z \le \left(Z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)\right\} = 1-P\left\{Z \le \left(1.645 - \frac{11-10}{2/\sqrt{50}}\right)\right\}$$
$$= 1-P(Z \le -1.891) = 0.971 .$$

L-6. Tests for the Mean.

L-6.1. *One-Sample* t-*test (Simple Random, Systematic Random, or Composite Sampling)*. Given a random sample of size *n* (or a composite sample of size *n*, each composite consisting of *k* aliquots), the one-sample *t*-test is parametric test that can be used to test hypotheses involving the mean (μ) of the population from which the sample was selected. The *t*-test is used when the population standard deviation is unknown but normality can be assumed.

L-6.1.1. *Introduction*.

L-6.1.1.1. The primary assumptions required for validity of the one-sample *t-test* are that the sample is random (data values are independent) and that the sample mean ($\bar{x}$) has an approximately normal distribution. Note that, according to the Central Limit Theorem, the sample mean will be approximately normally distributed for a large *n*. Unfortunately, the value of *n* that is sufficiently large enough to normalize the sample mean is seldom known. For environmental data, normality is not typically assumed for the sample mean unless *n* is very large (e.g., *n* > 100). Small sample sizes are common for environmental studies. As the sample mean is normal if *X* is normal, in practice, a data set consisting of *n* values of *X* is tested for normality and the *t*-test is used if the assumption of normality is not rejected.

L-6.1.1.2. Because the sample mean and standard deviation are very sensitive to outliers, the *t*-test should be preceded by a test for outliers (Appendix E). The *t*-test is also adversely affected by censored results. Directions for a one-sample *t*-test are presented in Paragraph L-6.1.2, followed by an example in Paragraph L-6.1.3.

L-6.1.2. *Directions for a One-Sample* t-*test*. The steps for a one-sample *t*-test are presented for Case 1: $H_0 : \mu \le C , H_A : \mu > C$; and Case 2: $H_0 : \mu \ge C , H_A : \mu < C$. The steps for Case 2 are given in braces {}. Let $x_1, x_2, \ldots, x_n$ represent the *n* data points from a normal distribution. These could be either *n* individual samples or *n* composite samples consisting of *k* aliquots each.

L-6.1.2.1. Verify that the data come from a normal distribution using tests presented in Appendices F and J, such as the Shapiro-Wilk test (Paragraph F-3.2) and a normal probability plot (Paragraph J-5.5).

L-6.1.2.2. Calculate the sample mean, $\bar{x}$, and the standard deviation, *s* (Appendix D).

L-6.1.2.3. Use Table B-23 of Appendix B to find the critical value, $t_{1-\alpha,\nu}$, such that 100(1 – α)% of the *t* distribution with $\nu = n-1$ degrees of freedom is below this value. For example, if $\alpha = 0.05$ and $n = 16$, then $n - 1 = 15$ and $t_{0.95,15} = 1.753$.

L-6.1.2.4. Calculate the test statistic *t* for the data set:

$$t = \frac{\bar{x} - C}{s/\sqrt{n}}.$$

L-6.1.2.5. Compare the calculated test statistic *t* with the critical value $t_{1-\alpha,\nu}$ (from Table B-23):

L-6.1.2.5.1. If $t > t_{1-\alpha,\nu}$ $\{t < -t_{1-\alpha,\nu}\}$, H_0 may be rejected. Go to L-6.1.2.7.

L-6.1.2.5.2. If $t \leq t_{1-\alpha,\nu}$ $\{t \geq -t_{1-\alpha,\nu}\}$, there is not enough evidence to reject H_0 and the false acceptance error rate should be verified. Go to L-6.1.2.6.

L-6.1.2.6. If H_0 is not rejected, calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates. The power of the test can be estimated using Paragraph L-5.2, assuming the true values for the mean and standard deviation are those obtained in the sample. A power curve of the test can be generated using software packages such as the Decision Error Feasibility Trial (DEFT) software (EPA QA/G-4D).

L-6.1.2.6.1. If only one false acceptance error rate (β) has been specified (at μ_1), it is possible to approximately calculate the sample size that achieves the DQOs, assuming the true mean and standard deviation are equal to the values estimated from the sample, instead of calculating the power of the test. A derivation of the following formula is provided in Appendix A of EPA QA/G-4D.

L-6.1.2.6.2. Calculate:

$$m = \frac{s^2 (Z_{1-\alpha} + Z_{1-\beta})^2}{(\mu_1 - C)^2} + (0.5) Z_{1-\alpha}^2$$

where Z_p is the $p100^{th}$ percentile of the standard normal distribution (Table B-15, Appendix B).

L-6.1.2.6.3. Round *m* up to the next integer. If $m \leq n$, the false acceptance error rate has

been satisfied. If $m > n$, the false acceptance error rate has not been satisfied.

L-6.1.2.7. Summary of results for one-sample *t*-test:

L-6.1.2.7.1. H_0 is rejected. One concludes $H_A : \mu > C$ $\{H_A : \mu < C\}$.

L-6.1.2.7.2. H_0 is not rejected and the false acceptance error rate is satisfied. One concludes $H_A : \mu \le C$ $\{H_A : \mu \ge C\}$; or

L-6.1.2.7.3. H_0 is not rejected but the false acceptance error rate is not satisfied. The null hypothesis must be retained but the conclusions are uncertain since the sample size is too small.

L-6.1.2.8. Report the results of the test, sample size, sample mean, standard deviation, and *t* and $t_{1-\alpha,\nu}$. Note that the calculations for the *t*-test are the same for both simple random or composite random sampling. The use of compositing usually results in a smaller value of *s* than simple random sampling.

L-6.1.3. *Example of One-Sample* t-*Test for Simple and Systematic Random Samples with or without Compositing*. Suppose total chromium in subsurface soil (below 5 feet from ground surface) at Site A is to be compared to a regulatory threshold of *C* = 2.0 mg/kg using the following test with 95% level of confidence:

$H_0 : \mu \ge 2$, $H_A : \mu < 2$.

L-6.1.3.1. Table L-4 presents the data. All chromium concentrations were detected, so no proxy concentrations are needed to evaluate the data.

L-6.1.3.2. Verify that the data follow a normal distribution. The Shapiro-Wilk test for normality shows evidence that the data follow a normal distribution because the test's *p* value was 0.8489 and is > 0.05.

L-6.1.3.3. Calculate the mean and standard deviation: $\bar{x} = 4.619$ and $s = 0.8980$.

L-6.1.3.4. Because we want a 95% level of confidence, $\alpha = 0.05$. Also, because $n = 36$, $\nu = n - 1 = 36 - 1 = 35$.

L-6.1.3.5. Using Table B-23 of Appendix B and linear interpolation, the critical value is 1.6905.

$t_{1-\alpha,\nu} = t_{0.95,35} = (1.697 + 1.684)/2 = 1.6905$.

L-6.1.3.6. The test statistic is

$$t = \frac{\bar{x} - C}{s/\sqrt{n}} = \frac{4.619 - 2.0}{0.8980/\sqrt{36}} = 17.50 .$$

L-6.1.3.7. Comparing the calculated test statistic, t, with the critical value, $t_{1-\alpha,\nu}$, we see that $t \geq -t_{1-\alpha,df}$ $(17.5 \geq -1.6905)$ and so we cannot reject H_0 and we must check that the false acceptance rate has been achieved.

Table L-4.
Example L-6.1.3 Data

Site A sample location	Top depth of sample	Bottom depth of sample	Chromium (total) concentration (mg/kg)		Site A sample location	Top depth of sample	Bottom depth of sample	Chromium (total) concentration (mg/kg)
EPC-SB01	9	10	2.95		EPC-SB07	9	10	5.1
EPC-SB01	14	15	5.17		EPC-SB07	14	15	4.94
EPC-SB01	19	20	4.8		EPC-SB07	19	20	4.76
EPC-SB02	9	10	4.53		EPC-SB08	9	10	4.62
EPC-SB02	14	15	4.01		EPC-SB08	14	15	4.72
EPC-SB02	19	20	5.91		EPC-SB08	19	20	4.73
EPC-SB03	9	10	3.96		EPC-SB09	9	10	3.21
EPC-SB03	14	15	4.81		EPC-SB09	14	15	4.14
EPC-SB03	19	20	5.27		EPC-SB09	19	20	4.85
EPC-SB04	9	10	5.99		EPC-SB10	9	10	4.25
EPC-SB04	14	15	4.6		EPC-SB10	14	15	5.09
EPC-SB04	19	20	5.51		EPC-SB10	19	20	3.68
EPC-SB05	9	10	4.72		EPC-SB11	9	10	5.12
EPC-SB05	14	15	3.56		EPC-SB11	14	15	6.6
EPC-SB05	19	20	4.22		EPC-SB11	19	20	6.19
EPC-SB06	9	10	3.91		EPC-SB12	9	10	3.15
EPC-SB06	14	15	5.81		EPC-SB12	14	15	4.11
EPC-SB06	19	20	4.48		EPC-SB12	19	20	2.8

L-6.1.3.8. Suppose the false acceptance rate is $\beta = 0.20$.

L-6.1.3.9. The power of this test is verified by assuming that the true values for the mean and standard deviation are those obtained in the sample. A power curve of the test was generated using DEFT software, as shown in the figure below. The probability of accepting the null hypothesis is plotted for a range of assumed true mean concentrations. For the regulatory threshold concentration of 2.0, a 95% (i.e., α = 0.05) chance of accepting the null hypothesis is requested. A 20% (β) probability of accepting the null hypothesis when the true concentration is $\mu_1 = 1.0$ is also requested (80% power). A sample size of seven is suggested for this request. For the sample

mean, this plot shows the probability of deciding that the true mean is higher than the regulatory threshold is nearly 100%, which means the test has strong power.

L-6.1.3.10. The sample size needed to achieve the false rejection rate of 0.20 when $\mu_1 = 1$ is:

$$m = \frac{s^2 (Z_{1-\alpha} + Z_{1-\beta})^2}{(\mu_1 - C)^2} + (0.5) Z_{1-\alpha}^2 = \frac{0.8980^2 (1.645 + 0.8417)^2}{(1-2)^2} + (0.5) 1.645^2 = 6.34 .$$

Rounding up to the next integer, $m = 7$ (the reported value for "Sample Size" in Figure L-2).

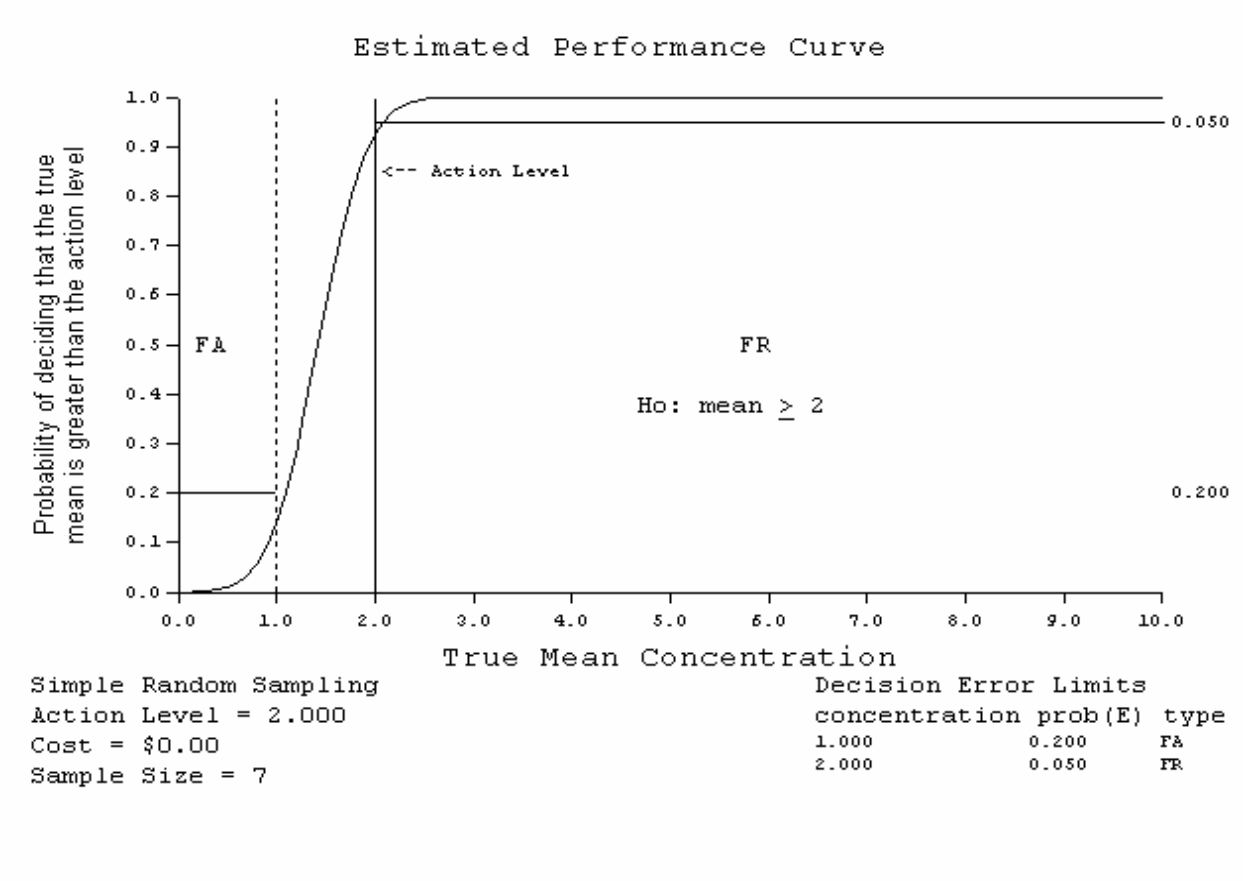

Figure L-2. Power curve for the one-sample *t*-test for simple random sampling.

L-6.1.3.11. Because more than seven samples have been collected (in fact, 36 samples have been collected), the false acceptance error rate has been satisfied. Therefore, we have evidence to suggest the true mean for chromium in Site A subsurface soil is greater than the regulatory threshold of 2.0 mg/kg on average.

L-6.2. *One Sample* t-*Test for the Mean (Stratified Random Sampling)*. Directions for a one-sample *t*-test for a stratified random sample followed by an example are presented in Paragraphs L-6.2.1 and L-6.2.2, respectively.

L-6.2.1. *Directions for a One-Sample* t-*Test for a Stratified Random Sample*. The steps for a one-sample *t*-test are presented for: Case 1: $H_0 : \mu \le C , H_A : \mu > C$; and Case 2: $H_0 : \mu \ge C , H_A : \mu < C$. The steps for Case 2 are given in braces {}.

L-6.2.1.1. Let h = 1, 2, 3,…L represent the L strata and n_h represent the sample size of stratum h. The i^{th} sample from stratum h is presented by $x_{h,i}$.

L-6.2.1.2. Verify that the data come from a normal distribution using tests presented in Appendices F and J, such as the Shapiro-Wilk test (Paragraph F-3.2) and a normal probability plot (Paragraph J-5.5).

L-6.2.1.3. Calculate the stratum weights w_h using the proportion of the volume in stratum h,

$$w_h = \frac{v_h}{\sum_{h=1}^{L} v_h}$$

where v_h is the surface area (or volume) of stratum h divided by the total surface area (or volume) over all strata.

L-6.2.1.4. For each stratum, calculate the sample stratum mean

$$\bar{x}_h = \frac{\sum_{i=1}^{n_h} x_{h,i}}{n_h}$$

and the sample stratum standard error

$$s_h^2 = \sum_{i-1}^{n_h} \frac{(x_{h,i} - \bar{x}_h)^2}{n_h - 1}.$$

L-6.2.1.5. Calculate overall mean and variance:

$$\bar{x}_{ST} = \sum_{h-1}^{L} w_h \bar{x}_h \ , \quad s_{ST}^2 = \sum_{h-1}^{L} w_h^2 \frac{s_h^2}{n_h} \ .$$

L-6.2.1.6. Calculate the degrees of freedom

$$v = \frac{\left(s_{ST}^2\right)^2}{\sum_{h=1}^{L} \frac{w_h^4 s_h^4}{n_h^2 (n_h - 1)}} \ .$$

L-6.2.1.7. Use Table B-23 of Appendix B to find the critical value, $t_{1-\alpha,\nu}$, so that (1 – α)100% of the *t*-distribution with the above degrees of freedom (rounded to the next highest integer) is below $t_{1-\alpha,\nu}$.

L-6.2.1.8. Calculate the sample value (statistic):

$$t = \frac{\bar{x}_{ST} - C}{\sqrt{s_{ST}^2}} \ .$$

L-6.2.1.9. Compare the calculated test statistic, *t*, to the critical value $t_{1-\alpha,\nu}$,. If $t > t_{1-\alpha,\nu}$ {$t < -t_{1-\alpha,\nu}$,} H_0 may be rejected. If $t \le t_{1-\alpha,\nu}$ {$\ge -t_{1-\alpha,\nu}$}, there is not enough evidence to reject H_0 and the false acceptance error rate should be verified.

L-6.2.1.10. If H_0 was not rejected, calculate either the power of the test or the sample size necessary to achieve the false rejection and false acceptance error rates. The results of the test could be:

L-6.2.1.10.1. H_0 was rejected so it seems that the true mean is less than *C* {greater than *C*}.

L-6.2.1.10.2. H_0 was not rejected and the false acceptance error rate was satisfied and it appears that the true mean is greater than *C* {less than *C*}; or,

L-6.2.1.10.3. H_0 was not rejected and the false acceptance error rate was not satisfied and it appears that the true mean is greater than *C* {less than *C*} but conclusions are uncertain since the sample size was too small.

L-6.2.1.10.4. If H_0 is not rejected, determine whether the power is adequate. Statistical software such as DEFT can be used for this purpose. DEFT uses the following approximation to calculate the number of samples required for each stratum to achieve a power of 1 – β at some desired value μ_1:

$$n'_h = \left[\sum_{h=1}^{L} w_h s_h\right] \times \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{\left(C - \mu_1\right)^2} \times w_h s_h \text{, } h = 1, \ldots, L \text{ .}$$

The value n'_h is rounded up to a whole number. The power is adequate if the calculated sample size is less than or equal to the actual sample size for each stratum: $n'_h \le n_h$ for $h = 1,\ldots, L$.

L-6.2.2. *Example of a One-Sample* t-*Test for a Stratified Random Sample*. Suppose the total chromium in subsurface soil data used in the previous example (Paragraph L-6.2.1) came from a stratified sampling effort. Two strata were sampled, stratum A and stratum B, where stratum B makes up one-third of the area to be investigated. The objective is to compare the chromium concentration at Site A to a regulatory threshold of 2.0 mg/kg, based on a 95% level of confidence.

$H_0 : \mu \ge 2$, $H_A : \mu < 2$.

L-6.2.2.1. Table L-5 presents the data. All chromium concentrations were detected so no proxy concentrations are needed to evaluate the data.

$L = 2 \qquad n_A = 24 \qquad n_B = 12 \qquad w_A = 0.75 \qquad w_B = 0.25$

L-6.2.2.2. Verify that the data follow a normal distribution for each stratum. The Shapiro-Wilk test was performed for each stratum and results indicated that the data for each follow a normal distribution because the tests' *p* values were greater than 0.05.

L-6.2.2.3. The mean and standard deviation of the data were calculated per stratum; $\alpha = 0.05$ because we want a 95% level of confidence:

$\bar{x}_A = 4.674, \quad s_A = 1.027, \quad n_A = 24$
$\bar{x}_B = 4.508, \quad s_B = 0.5827, \quad n_B = 12$

Table L-5.
Data for Example L-6.2.2

Stratum	Site A sample location	Top depth of sample	Bottom depth of sample	Chromium (total) concentration (mg/kg)	Stratum	Site A sample location	Top depth of sample	Bottom depth of sample	Chromium (total) concentration (mg/kg)
A	EPC-SB01	9	10	2.95	B	EPC-SB07	9	10	5.1
A	EPC-SB01	14	15	5.17	B	EPC-SB07	14	15	4.94
A	EPC-SB01	19	20	4.8	B	EPC-SB07	19	20	4.76
A	EPC-SB02	9	10	4.53	B	EPC-SB08	9	10	4.62
A	EPC-SB02	14	15	4.01	B	EPC-SB08	14	15	4.72

tum	Site A sample location	Top depth of sample	Bottom depth of sample	Chromium (total) concentration (mg/kg)	tum	Site A sample location	Top depth of sample	Bottom depth of sample	Chromium (total) concentration (mg/kg)
A	EPC-SB02	19	20	5.91	B	EPC-SB08	19	20	4.73
A	EPC-SB03	9	10	3.96	B	EPC-SB09	9	10	3.21
A	EPC-SB03	14	15	4.81	B	EPC-SB09	14	15	4.14
A	EPC-SB03	19	20	5.27	B	EPC-SB09	19	20	4.85
A	EPC-SB04	9	10	5.99	B	EPC-SB10	9	10	4.25
A	EPC-SB04	14	15	4.6	B	EPC-SB10	14	15	5.09
A	EPC-SB04	19	20	5.51	B	EPC-SB10	19	20	3.68
A	EPC-SB05	9	10	4.72	A	EPC-SB11	9	10	5.12
A	EPC-SB05	14	15	3.56	A	EPC-SB11	14	15	6.6
A	EPC-SB05	19	20	4.22	A	EPC-SB11	19	20	6.19
A	EPC-SB06	9	10	3.91	A	EPC-SB12	9	10	3.15
A	EPC-SB06	14	15	5.81	A	EPC-SB12	14	15	4.11
A	EPC-SB06	19	20	4.48	A	EPC-SB12	19	20	2.8

L-6.2.2.4. The overall mean and variance are:

$$\bar{x} = (0.75 \times 4.674) + (0.25 \times 4.508) = 4.633$$

$$s^2 = \left(0.75^2 \times \frac{1.027^2}{24}\right) + \left(0.25^2 \times \frac{0.5827^2}{12}\right) = 0.02472 + 0.001768 = 0.2649 .$$

L-6.2.2.5. The degrees of freedom are (rounded to the next highest integer):

$$\nu = \frac{(0.02649)^2}{\frac{0.75^4 \times 1.027^4}{24^2(24-1)} + \frac{0.25^4 \times 0.5827^4}{12^2(12-1)}} = 26.13 \approx 27 .$$

L-6.2.2.6. Table B-23 of Appendix B gives the critical value $t_{1-\alpha,\nu} = 1.703$.

L-6.2.2.7. The test statistic is

$$t = \frac{\bar{x} - C}{s} = \frac{4.633 - 2.0}{\sqrt{0.02649}} .$$

L-6.2.2.8. Compare the calculated test statistic t with the critical value $t_{1-\alpha,\nu}$. Because $t \geq -t_{1-\alpha,\nu}$ ($16.18 \not\leq -1.703$), we cannot reject H_0 and must check that the false acceptance rate has been achieved.

L-6.2.2.9. As in Paragraph L-6.1.3.9, a 20% (β) probability of accepting the null hypothesis when the true concentration is 1.0 is also requested (80% power). A power curve of the test was generated using DEFT software in Figure L-3 (by entering the sample standard deviation s_i and the weight w_i for each stratum). The required sample size for stratum A is equal to 5 and that for stratum B is equal to 2 (a total sample size of 7). The required power is achieved as actual the sample sizes for strata A and B are 24 and 12, respectively (a total of 36 samples).

L-6.3. *The Chen Test*. Environmental data such as concentration measurements are often confined to positive values and appear to follow a distribution with most of the data values relatively small or near zero, but with a few relatively large values. Underlying such data is a distribution that is not symmetrical (like a normal distribution) but is skewed to the right (like a lognormal distribution). Given a random sample of size n from a right-skewed distribution, the Chen test can be used to compare the mean (μ) of the distribution with a threshold level or regulatory value. This test assumes that the data arise from a right-skewed distribution and a random sample has been employed. Chen's test is a generalization of the *t*-test, with slightly more complicated calculations involving the sample mean, standard deviation, and skewness. Directions for conducting the Chen test are presented in Paragraph L-6.3.1, followed by an example in Paragraph L-6.3.2.

L-6.3.1. *Directions for Conducting the Chen Test*. Let $x_1, x_2, \ldots, x_n$ represent the n data points. Let C denote the threshold level of interest. The null hypothesis is $H_0 : \mu \le C$ and the alternative is $H_A : \mu > C$; the level of significance is α.

L-6.3.1.1. If, at most, 15% of the data points are below the detection limit and C is much larger than the DL, then replace values (< DL) with a proxy value (Appendix C).

L-6.3.1.2. Visually check the assumption of right-skewness by inspecting a histogram or frequency plot for the data.

L-6.3.1.3. Calculate the sample mean, $\bar{x}$, and the standard deviation, s (Appendix D).

L-6.3.1.4. Calculate the sample skewness

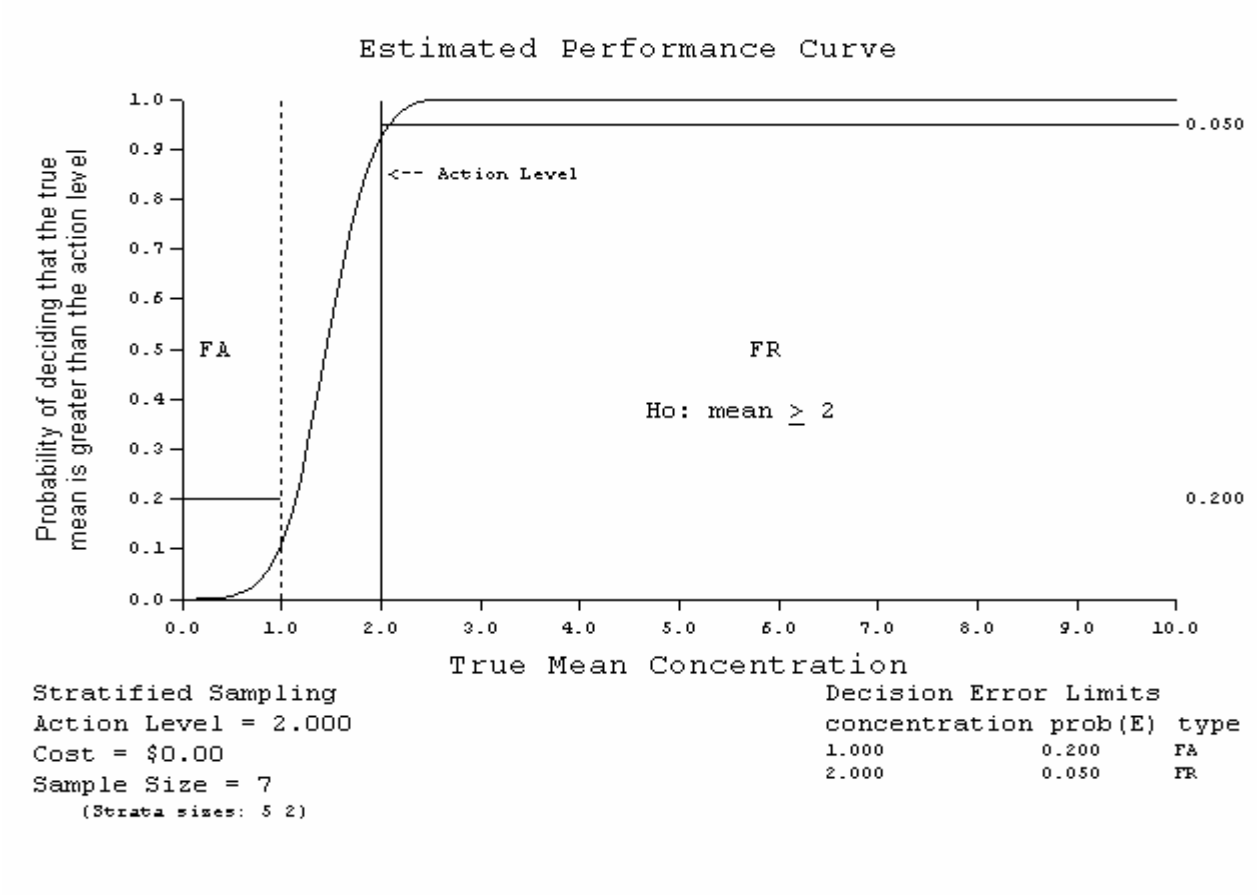

Figure L-3. Power curve for the one-sample *t*-test for stratified sampling.

$$b = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)(n-2)s^3}$$

the quantity

$$a = \frac{b}{6\sqrt{n}}$$

the statistic

$$t = \frac{(\bar{x} - C)}{s/\sqrt{n}}$$

and then compute:

$$z = t + a(1 + 2t^3) + 4a^2(t + 2t^3).$$

The skewness, *b*, should be greater than 1 to confirm that the data are skewed to the right.

L-6.3.1.5. Use Table B-15 in Appendix B to find the critical value, $Z_{1-\alpha}$, such that $(1-\alpha)100\%$ of the standard normal distribution is below $Z_{1-\alpha}$, which is also the $p100^{th}$ percentile of the standard normal distribution. For example, if $\alpha = 0.05$ then $Z_{1-\alpha} = 1.645$.

L-6.3.1.6. Compare *z* with $Z_{1-\alpha}$:

L-6.3.1.6.1. If $z > Z_{1-\alpha}$, H_0 may be rejected and it appears that the true mean is greater than C.

L-6.3.1.6.2. If $z \le Z_{1-\alpha}$, there is not enough evidence to reject H_0 so it appears that the true mean is less than *C*.

L-6.3.2. *Example of the Chen Test*. Suppose surface soil samples (from 0 to 5 feet below ground surface) have been collected at Site B to evaluate arsenic concentrations on site against a regulatory threshold value of 5 mg/kg using a 90% level of confidence ($\alpha = 0.10$) and the following hypothesis test:

$$H_0 : \mu \le 5, \quad H_A : \mu > 5$$

Table L-6 presents the analytical results from samples collected at the site. All arsenic concentrations were detected so no proxy concentrations are needed to evaluate the data.

L-6.4. *The Wilcoxon Signed Rank (One-Sample) Test*. Given a random sample of size *n* (or composite sample size *n*, each composite consisting of *k* aliquots), the Wilcoxon signed rank test is a nonparametric test can be used to test hypotheses regarding the mean or median of the population from which the sample was selected. The mean is used as the parameter of interest in this Appendix, although the median could be used equivalently. The Wilcoxon signed rank test assumes that the data constitute a random sample from a symmetrical, continuous population. (Symmetrical means the underlying population frequency curve is symmetrical about its mean or median.) If the data are not symmetrical, it may be possible to transform them (using a transformation such as a log or square root transformation) so that this assumption is satisfied.

Table L-6.
Analytical Results From Samples Collected at the Site in Example L-6.3.2

Site B sample location	Top depth of sample	Bottom depth of sample	Arsenic Concentration (mg/kg), x_i	$(x_i - \overline{x})^3$
EPC-BG01	1	2	4.84	–0.0024604
EPC-BG01	4	5	4.15	–0.5615156
EPC-BG02	1	2	4.53	–0.0881211
EPC-BG02	4	5	4.72	–0.0165814
EPC-BG03	1	2	4.76	–0.0099384
EPC-BG03	4	5	4.93	-9.112×10^{-5}
EPC-BG04	1	2	4.34	–0.2560479
EPC-BG04	4	5	4.51	–0.1005446
EPC-BG05	1	2	5.01	4.288×10^{-5}
EPC-BG05	4	5	3.83	–1.5011236
EPC-BG06	1	2	4.8	–0.0053594
EPC-BG06	4	5	4.07	–0.7412176
EPC-BG07	0.5	1	7.43	14.796346
EPC-BG07	2	2.5	4.6	–0.0527344
EPC-BG08	1	2	8.12	31.107274
EPC-BG08	4	5	4.96	-3.375×10^{-6}

L-6.4.1. *Introduction*. The Wilcoxon signed rank test is more robust to outliers. The *t*-test is not robust to outliers because the sample mean and standard deviation are strongly influenced by outliers. Although it is less powerful than the *t*-test when the data are normally distributed, it is usually more powerful when the data are not normally distributed. The Wilcoxon signed rank test is more likely than the *t*-test to identify differences for positively skewed distributions. In addition, compared to tests based on ranks, the *t*-test has difficulty accommodating censored values (values below the detection limit).

L-6.4.1.1. Directions for the Wilcoxon signed rank test for a simple random sample and a systematical simple random sample are given below in Paragraph L-6.4.2; Paragraph L-6.4.3 is an example for sample sizes smaller than 20.

L-6.4.1.2. For sample sizes greater than 20, the large sample approximation to the Wilcoxon signed rank test should be used. Directions for this test are given in Paragraph L-6.4.4 followed by an example in Paragraph L-6.4.5.

L-6.4.1.3. Paragraph L-6.4.6 presents sample size calculations for the Wilcoxon signed rank test to achieve a certain power when the sample size is large. An example follows in Paragraph L-6.4.7.

L-6.4.2. *Directions for the Wilcoxon Signed Rank Test for a Simple Random Sample and a Systematic Simple Random Sample*. The following describes the steps for applying the Wilcoxon signed rank test for a sample size (*n*) less than 20 for: Case 1 ($H_0 : \mu \le C, H_A : \mu > C$); and Case 2 ($H_0 : \mu \ge C, H_A : \mu < C$). Modifications for Case 2 are given in braces {}.

L-6.4.2.1. Let $x_1, x_2, \ldots, x_n$ represent the *n* observations.

L-6.4.2.2. If possible, assign values to any measurements below the detection limit with procedures described in Appendix H.

L-6.4.2.3. Subtract *C* from each observation x_i to obtain the difference $d_i = x_i - C$. If any of the differences are zero, delete them and correspondingly reduce the sample size (*n*).

L-6.4.2.4. Assign ranks from 1 to *n* based on ordering the absolute differences $|d_i|$ (i.e., the magnitude of differences ignoring the sign) from smallest to largest. The rank 1 is assigned to the smallest value, the rank 2 to the second smallest value, and so forth. If there are ties, assign the average of the ranks that otherwise would have been assigned to the tied observations (e.g., if two equal values occur after rank 5, then assign them each a rank of 6.5 = (6 + 7)/2).

L-6.4.2.5. Assign the sign for each observation to create the signed rank. The sign is positive if the deviation d_i is positive; the sign is negative if the deviation d_i is negative.

L-6.4.2.6. Calculate *R*, the sum of the ranks with a positive sign.

L-6.4.2.7. Use Table B-24 of Appendix B to find the critical value $w_{\alpha,n}$.

L-6.4.2.8. Compare the calculated test statistic, *R*, to the critical value.

L-6.4.2.8.1. If $R > n(n+1)/2 - w_{\alpha,n}$ $\{R < w_{\alpha,n}\}$, H_0 may be rejected.

L-6.4.2.8.2. If $R \le n(n+1)/2 - w_{\alpha,n}$ $\{R \ge w_{\alpha,n}\}$, there is not enough evidence to reject H_0.

L-6.4.2.9. The results of the test may be:

L-6.4.2.9.1. H_0 is rejected; $\mu > C$ $\{\mu < C\}$.

L-6.4.2.9.2. H_0 is not rejected $\mu \le C$ $\{\mu \ge C\}$.

L-6.4.3. *Example of the Wilcoxon Signed Rank Test for Simple and Systematical Random Samples*. Suppose $n = 14$ surface soil samples (from 0 to 5 feet below ground surface) were collected at Site B to evaluate cadmium concentrations on site against a regulatory threshold value of 0.75 using a 95% level of confidence ($\alpha = 0.05$) and the following hypothesis test.

$H_0 : \mu \geq 0.75$, $H_A : \mu < 0.75$.

L-6.4.3.1. Table L-7 presents the analytical results from samples collected at the site. Three of the cadmium concentrations were non-detects, so proxy concentrations are defined as the detection limit and are presented in parentheses.

Table L-7.
Analytical Results from Samples Collected at the Site in Example L-6.4.3

Site B sample location	Top depth of sample	Bottom depth of sample	Flag (ND = not detected)	Cadmium Concentration (mg/kg), x_i	$d_i = x_i - C$	Rank associated with $\lvert d_i \rvert$	Sign of d_i
EPC-BB01	1	2		1.6	0.85	13.5	+
EPC-BB01	4	5		1.6	0.85	13.5	+
EPC-BB02	1	2		1.55	0.8	12	+
EPC-BB02	4	5	ND	(0.242)	–0.508	9	–
EPC-BB03	1	2		0.624	–0.126	1	–
EPC-BB03	4	5		0.276	–0.474	7	–
EPC-BB04	1	2		1.5	0.75	11	+
EPC-BB04	4	5		0.301	–0.449	6	–
EPC-BB05	1	2		0.588	–0.162	3	–
EPC-BB05	4	5		0.264	–0.486	8	–
EPC-BB06	0.5	1		0.899	0.149	2	+
EPC-BB06	2	2.5		0.332	–0.418	4	–
EPC-BB07	1	2		1.42	0.67	10	+
EPC-BB07	4	5		0.326	–0.424	5	–

L-6.4.3.2. Steps 1, 2, and 3 are contained in the three right-hand columns, in order.

L-6.4.3.3. Step 4: From the six cases where the sign of d_i is positive,

$R = 13.5 + 13.5 + 12 + 11 + 2 + 20 = 62$.

L-6.4.3.4. Step 5: Table B-24 of Appendix B gives a critical value of $w_{0.05,14} = 26$.

L-6.4.3.5. Step 6: Compare the calculated test statistic and the critical value, $62 \geq 26$, so H_0 was not rejected.

L-6.4.3.6. Prior to performing the test, a histogram was created to check the symmetry of the data, which appear symmetrical, as shown below.

L-6.4.4. *Directions for the Large Sample Approximation to the Wilcoxon Signed Rank Test*. The following describes the steps for applying the large sample approximation of the Wilcoxon signed rank test for: Case 1 ($H_0 : \mu \le C, H_A : \mu > C$); and Case 2 ($H_0 : \mu \ge C, H_A : \mu < C$). Modifications for Case 2 are given in braces {}.

L-6.4.4.1. Let $x_1, x_2, \ldots, x_n$ represent the *n* data points where *n* is greater than or equal to 20. If possible, assign values to any measurements below the detection limit with procedures described in Appendix H.

L-6.4.4.2. Subtract *C* from each observation, x_i, to obtain the differences $d_i = x_i - C$. If any of the differences are zero delete them and correspondingly reduce the sample size (*n*).

L-6.4.4.3. Assign ranks from 1 to *n* based on ordering the absolute deviations $|d_i|$ (i.e., magnitude of differences ignoring the sign) from smallest to largest. Rank 1 is assigned to the smallest value, rank 2 to the second smallest value, and so forth. If there are ties, assign the average of the ranks that would otherwise have been assigned to the tied observations.

L-6.4.4.4. Assign the sign for each observation to create the signed rank. The sign is positive if the deviation, d_i, is positive; the sign is negative if the deviation, d_i, is negative.

L-6.4.4.5. Calculate the test statistic *R*, the sum of the ranks with a positive sign.

L-6.4.4.6. Calculate the critical value

$$w_p = n(n+1)/4 + Z_p \sqrt{n(n+1)(2n+1)/24}$$

where $p = 1 - \alpha$ $\{p = \alpha\}$ and Z_p is the $100\,p^{th}$ percentile of the standard normal distribution (Table B-15 of Appendix B).

L-6.4.4.7. Compare the test statistic to the critical value. If $R > w_p$ $\{R < w_p\}$, H_0 may be rejected. Otherwise, there is not enough evidence to reject H_0.

L-6.4.4.8. The results of the test may be:

L-6.4.4.8.1 H_0 is rejected; $\mu > C$ { $\mu < C$ }.

L-6.4.4.8.2 H_0 is not rejected; $\mu \leq C$ { $\mu \geq C$ }.

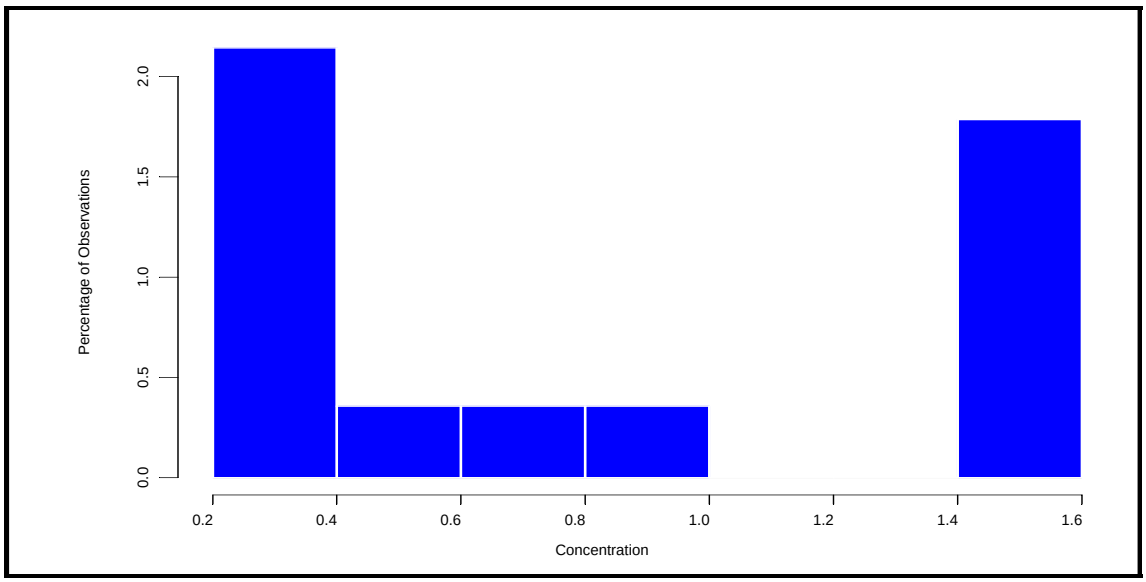

Figure L-2. Histogram Plot of Wilcoxon Signed Rank Test for random samples.

L-6.4.5. *Example for the Large Sample Approximation to the Wilcoxon Signed Rank Test for Simple and Systematic Random Samples*. Suppose additional surface soil samples (from 0 to 5 feet below ground surface) were collected at Site B to further delineate contamination. Additional samples were analyzed for cadmium and so the test performed earlier (see Paragraph L-6.4.3) for cadmium must be redone. The test was set up to compare cadmium concentrations on site to a regulatory threshold value of 0.75 using a 95% level of confidence ($\alpha = 0.05$) and the following hypothesis test.

$H_0 : \mu \geq 0.75$, $H_A : \mu < 0.75$.

L-6.4.5.1. Table L-8 presents all analytical results from samples collected from both sampling events. Non-detected cadmium concentrations were present in the data set; therefore, proxy concentrations are defined as the detection limit and are presented in parentheses.

L-6.4.5.2. Steps 1, 2, and 3 are contained in the three right-hand columns, in order.

L-6.4.5.3. Step 4: The test statistic, which is the sum of the ranks associated with the positive signs, is equal to

Table L-8.
All Analytical Results from Samples Collected from Both Sampling Events

Site B sample location	Top depth of sample	Bottom depth of sample	Flag ND = not detected	Cadmium concentration (mg/kg), x_i	$d_i = x_i - C$	Rank associated with $\lvert d_i \rvert$	Sign of d_i
EPC-BB01	1	2		1.6	0.85	21.5	+
EPC-BB01	4	5		1.6	0.85	21.5	+
EPC-BB02	1	2		1.55	0.8	20	+
EPC-BB02	4	5	ND	(0.242)	–0.508	14	–
EPC-BB03	1	2		0.624	–0.126	2	–
EPC-BB03	4	5		0.276	–0.474	12	–
EPC-BB04	1	2		1.5	0.75	19	+
EPC-BB04	4	5		0.301	–0.449	10	–
EPC-BB05	1	2		0.588	–0.162	4	–
EPC-BB05	4	5		0.264	–0.486	13	–
EPC-BB06	0.5	1		0.899	0.149	3	+
EPC-BB06	2	2.5		0.332	–0.418	5	–
EPC-BB07	1	2		1.42	0.67	17	+
EPC-BB07	4	5		0.326	–0.424	8	–
EPC-BG08	1	2		1.48	0.73	18	+
EPC-BG08	4	5		0.302	–0.448	9	–
EPC-BG09	1	2		1.39	0.64	15	+
EPC-BG09	4	5		0.33	–0.42	6	–
EPC-BG10	0.5	1		0.812	0.062	1	+
EPC-BG10	2	2.5		0.287	–0.463	11	–
EPC-BG11	1	2		1.41	0.66	16	+
EPC-BG11	4	5		0.327	–0.423	7	–

$$R = 21.5 + 21.5 + 20 + 19 + 3 + 17 + 18 + 15 + 1 + 16 = 152 \ .$$

L-6.4.5.4. Step 5: The critical value is

$$w_p = 22(22+1)/4 - 1.645\sqrt{22(22+1)(2\times 22+1)/24} = 75.83$$

where $n = 22$ and by linear interpolation $Z_{0.05} = (-1.64 - 1.65)/2 = -1.645$.

L-6.4.5.5. Step 6: Comparing the test statistic to the critical value, 152 > 75.83, $(R > w_p)$, so H_0 is not rejected.

L-6.4.5.6. Therefore, there is no evidence to suggest that the true mean for cadmium in Site B surface soil is less than the regulatory threshold of 0.75 mg/kg.

L-6.4.5.7. A histogram was created to check the symmetry of the data. The data appear symmetrical, as indicated in Figure L-3.

L-6.4.6. *Directions for Calculating Sample Size for the Wilcoxon Signed Rank Test to Achieve a Specified Power*. Noether (1987) discusses determining an adequate sample size based on a defined level of power to apply the Wilcoxon signed rank test for the following hypothesis test: Case 1 ($H_0 : \mu \le C, H_A : \mu > C$); and Case 2 ($H_0 : \mu \ge C, H_A : \mu < C$). Modifications for Case 2 are given in braces {}.

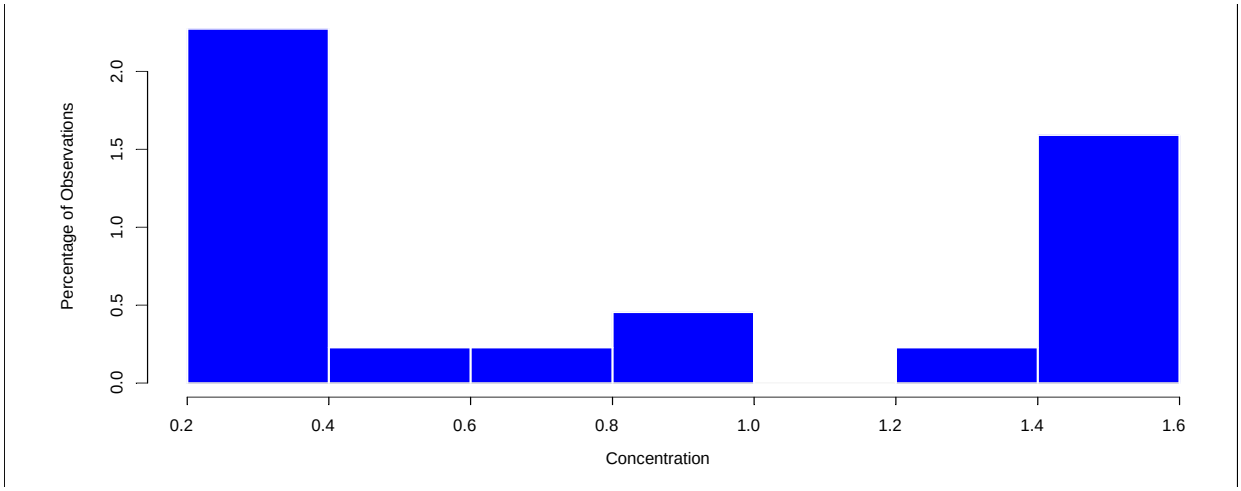

Figure L-3. Histogram plot of Wilcoxon Signed Rank Test for large random samples.

L-6.4.6.1. If the null hypothesis is not rejected, and the number of samples n' required to achieve some desired power $1 - \beta$ could be calculated, the power would be adequate if $n \ge n'$. If $n \ge 20$ samples are collected, a conservative estimate of the sample size required for a power of $1 - \beta$ is:

$$n' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{3\left(p' - 0.5\right)^2}$$

where Z_q is the q quantile of the standard normal distribution (from Table B-15), α is the significance level of the test, $1 - \beta$ is the desired power for the test, and p' is the true probability that the average of any two independent observations

$$\frac{x_i + x_j}{2}$$

where $i \ne j$, exceeds {is less than} C.

L-6.4.6.2. The equation for n' assumes that n is large enough for the test statistic R to be normally distributed (which is generally valid if the sample size exceeds 20). If the suggested sample size does not exceed 20, consult a statistician.

L-6.4.6.3. The value of p' can be determined from past information, a pilot sample, or chosen to represent a meaningful shift in the data (Noether, 1987). On the basis of what is considered to be a meaningful shift, one would assign p' equal to some probability greater than 0.5.

L-6.4.7. *Example of Calculating Sample Size for the Wilcoxon Signed Rank Test to Achieve a Specified Power*. Let us calculate the power for the hypothesis test performed in Paragraph L-6.4.5. In this example, n = 22 samples were collected to evaluate cadmium concentrations against a regulatory threshold value of 0.75 mg/kg at the 95% level of confidence ($\alpha = 0.05$) using the hypothesis test.

$$H_0 : \mu \geq 0.75,\ H_A : \mu < 0.75\ .$$

The null hypothesis was not rejected. We wish to ensure that n is large enough to find a meaningful decrease in the mean with 80% probability (power).

L-6.4.7.1. The objective is to ensure that the sample size is large enough to find a meaningful decrease in the mean with 80% probability. Let us assume that seven samples had been collected for a prior "pilot" study. Table L-9 presents the analytical results from samples collected for the pilot study in the left-most column and along the top. The independent pair wise averages are calculated in the body of the table. Averages that fall below the regulatory threshold of 0.75 mg/kg are shaded.

Table L-9.
Analytical Results from Samples Collected for the Pilot Study and Independent Pair Wise Averages

Cadmium concentration (mg/kg)	1.220	0.301	0.624	0.276	0.588	0.264	0.332
1.220	—	0.761	0.922	0.748	0.904	0.742	0.776
0.301	—	—	0.463	0.289	0.445	0.283	0.317
0.624	—	—	—	0.450	0.606	0.444	0.478
0.276	—	—	—	—	0.432	0.270	0.304
0.588	—	—	—	—	—	0.426	0.460
0.264	—	—	—	—	—	—	0.298
0.332	—	—	—	—	—	—	—

L-6.4.7.2. Of the initial 7 results, 17 of the 21 independent averages are less than 0.75. The observed probability that the average of any two observed observations is less than C is

17/21 = 0.8095. Therefore, on the basis of this estimated (pilot study) probability, assume that it was determined that a power of 80% is required for $p' = 0.809$.

L-6.4.7.3. The required sample size to meet the power requirement is:

$$n' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{3\left(p' - \frac{1}{2}\right)^2} = \frac{\left(1.645 + 0.842\right)^2}{3\left(0.8095 - 0.5\right)^2} = 21.5 \ .$$

L-6.4.7.4. The required sample size is rounded up to 22. Because $n \geq n'$, the required power of 80% was achieved.

L-7. Tests for a Median. A population median ($\tilde{\mu}$) is another measure of the center of the population distribution. This population parameter is less sensitive than the sample mean to extreme values and non-detected results. Therefore, this parameter sometimes is used instead of the mean when the data contain a large number of non-detects or extreme values.

L-7.1. *The Binomial Sign Test for the Median*. Given a random sample of size *n* of continuous or discrete samples, the sign test may be used to test hypotheses regarding a population median for a distribution from which the data were drawn. The only assumption required for the sign test is that it be a random sample. The procedures are also robust to outliers, as long as they do not represent data errors. Directions for the sign test are given below in Paragraph L-7.2, followed by an example in Paragraph L-7.3.

L-7.2. *Directions for the Sign Test for the Median.* The following describes the steps for applying the sign test for a sample size (*n*).

Case 1 ($H_0 : \tilde{\mu}_x \leq C$ versus $H_A : \tilde{\mu}_x > C$); and

Case 2 ($H_0 : \tilde{\mu}_x \geq C$ versus $H_A : \tilde{\mu}_x < C$).

Modifications for Case 2 are given in braces {}. *C* is the hypothesized median or critical threshold value and $\tilde{\mu}_x$ is the median for the variable *X*. The level of significance is α.

L-7.2.1. Note that $\tilde{\mu}$ can also be defined as the median value for the variable *D,* where $D = X - C$ and so the hypotheses tests are written in terms of the difference.

Case 1 ($H_0 : \tilde{\mu}_D \leq 0$ versus $H_A : \tilde{\mu}_D > 0$); and

Case 2 ($H_0 : \tilde{\mu}_D \geq 0$ versus $H_A : \tilde{\mu}_D < 0$).

L-7.2.2 The hypotheses can also be written in terms of the probability of exceeding 0.

Case 1 ($H_0 : P(D \leq 0) \geq 0.5$ versus $H_A : P(D \leq 0) < 0.5$); and

Case 2 ($H_0 : P(D \geq 0) \geq 0.5$ versus $H_A : P(D \geq 0) < 0.5$).

Equivalently,

Case 1 ($H_0 : P(D > 0) \leq 0.5$ versus $H_A : P(D > 0) > 0.5$); and

Case 2 ($H_0 : P(D < 0) \leq 0.5$ versus $H_A : P(D < 0) > 0.5$).

This formulation suggests the use of the binomial distribution with p = 0.5 to test the null hypothesis.

L-7.2.3. Noether (1987) discusses determining an adequate sample size based on a defined level of power to apply the sign test for the median. Under the assumption that the test statistic (in this case the number of samples that exceed {are less than} C) is normally distributed, a conservative sample size, n', is calculated as:

$$n' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{4\left(p - \frac{1}{2}\right)^2}$$

where Z_q is the q quantile of the standard normal distribution (from Table B-15), α is the significance level of the test, $1-\beta$ is the desired power for the test, and p is the true probability that an observation exceeds {is less than} C. The value of p can be taken from past information, a pilot sample, or chosen to represent a meaningful shift in the data (Noether, 1987). The normality of the test statistic under the null hypothesis rests on the normal approximation to the binomial distribution. As discussed in Appendix E, this approximation works well when the sample size is at least 20 ($np \geq 10$, $p = 0.5$). If the suggested sample size does not exceed 20, consult a statistician.

L-7.2.4. Let $x_1, x_2, \ldots, x_n$ represent the n data points. Define a new variable $D = X - C$.

L-7.2.4.1. If possible, assign values to any measurements below the detection limit with procedures described in Appendix H. Subtract C from each observation, x_i, to obtain the devia-

tions, $d_i = x_i - C$. If any of the deviations are zero, delete them and correspondingly reduce the sample size (n).

L-7.2.4.2. Count the number of positive {negative} deviations (d_i) and denote this number by y.

L-7.2.4.3. The number of positive {negative} differences is described by a binomial distribution. In terms of the notation and terminology used in Appendix E, the number of data points is the number of "trials," n. Under the null hypothesis, the probability, p, of a positive {negative} difference (a success) is 0.5. The total number of positive {negative} differences, y, is the successful occurrence of an event y times out of n. Therefore, bin(y; n, p = 0.5) is the probability of y positive {negative} differences for a set of n trials, where the probability of a positive {negative} difference p = 0.5 (when H_0 is assumed to be true). The probability of obtaining less than or equal to y positive {negative} differences,

$$P(Y \le y) = \sum_{i=0}^{y} bin(i, n, p)$$

is the value of the "cumulative binomial distribution." Table B-1 presents the probabilities of the cumulative binomial distribution for various values of n, p, and k where $k = y$.

L-7.2.4.4. If the probability of obtaining an equal or larger number of positive {negative} differences than the observed number y is small, that is, if $P(Y \ge y \mid n, p = 0.5) \le \alpha$, then it is unlikely that the null hypothesis is true and the null hypothesis is rejected. Equivalently,

L-7.2.4.4.1. If $P(Y < y \mid n, p = 0.5) = P(Y \le y - 1 \mid n, p = 0.5) \ge (1 - \alpha)$, H_0 may be rejected.

L-7.2.4.4.2. Otherwise, there is not enough evidence to reject H_0.

L-7.2.5. Use Table B-1 of Appendix B to find the probability value associated with n, $y - 1$, and $p = 0.5$, which is the cumulative binomial distribution probability,

$$P(Y \le y - 1 \mid n, p = 0.5)$$

to determine whether or not to reject the null hypothesis.

L-7.3. *Example of the Sign Test for the Median*. Suppose arsenic concentrations at a site are to be compared to a regulatory threshold value of 5 mg/kg using a 90% level of confidence ($\alpha = 0.10$). The median can be compared to this threshold using the following hypothesis test:

$H_0 : \tilde{\mu} \le 5$, $H_A : \tilde{\mu} > 5$.

L-7.3.1. Suppose we wish to know the adequate sample size necessary to be 80% certain that we can detect a meaningful difference from the null hypothesis. The meaningful difference for this site is defined to be when the probability of exceeding the regulatory threshold is twice as likely as being below the threshold, $P(\tilde{\mu} > 5) = 2/3$. The required sample size is 41:

$$n' = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{4\left(p - \frac{1}{2}\right)^2} = \frac{\left(Z_{0.9} + Z_{0.8}\right)^2}{4\left(\frac{2}{3} - \frac{1}{2}\right)^2} = \frac{\left(1.2816 + 0.8416\right)^2}{0.1111} = 40.6 \ .$$

L-7.3.2. Consider the data presented in Paragraph L-6.3.2 for arsenic concentrations in surface soil samples (from 0 to 5 feet below ground surface) at Site B. Table L-10 presents the analytical results from samples collected at the site. All arsenic concentrations were detected, so no proxy concentrations are needed to evaluate the data.

L-7.3.3. The number of positive deviations (d_i), $y = 3$.

L-7.3.4. Using Table B-1 in Appendix B, we find $P(Y \le 2 \mid n = 16, p = 0.5) = 0.002090$.

L-7.3.5. As 0.002090 < 0.9, H_0 may not be rejected. Therefore, it appears that the true median for arsenic is less than the regulatory threshold of 5 mg/kg. However, to achieve 80% power and satisfy the sample size requirement calculated earlier, an additional 25 randomly selected samples would be needed to increase the total sample size to 41.

Table L-10.
Analytical Results From Samples Collected At The Site For Example L-7.3

Site B sample location	Top depth of sample	Bottom depth of sample	Arsenic concentration (mg/kg)	$d_i = x_i - C$	Sign of d_i
EPC–BG01	1	2	4.84	–0.16	–
EPC–BG01	4	5	4.15	–0.85	–
EPC–BG02	1	2	4.53	–0.47	–
EPC–BG02	4	5	4.72	–0.28	–
EPC–BG03	1	2	4.76	–0.24	–
EPC–BG03	4	5	4.93	–0.07	–
EPC–BG04	1	2	4.34	–0.66	–
EPC–BG04	4	5	4.51	–0.49	–
EPC–BG05	1	2	5.01	0.01	+
EPC–BG05	4	5	3.83	–1.17	–
EPC–BG06	1	2	4.8	–0.2	–
EPC–BG06	4	5	4.07	–0.93	–
EPC–BG07	0.5	1	7.43	2.43	+
EPC–BG07	2	2.5	4.6	–0.4	–
EPC–BG08	1	2	8.12	3.12	+
EPC–BG08	4	5	4.96	–0.04	–

L-8. Test for a Proportion or Percentile.

L-8.1. *The One-Sample Proportion Test*. Given a random sample of size n, the non-parametric, one-sample proportion test may be used to test hypotheses regarding a population proportion or population percentile for a distribution from which the data were drawn. The only assumption required for the one-sample proportion test is that it be a random sample. To verify this assumption, review the procedures and documentation used to select the sampling points and ascertain that proper randomization has been used in sample collection.

L-8.1.1. The null and alternative hypotheses for this test can be stated as:

$$H_0 : X_{P_o} \le C, \quad H_A : X_{P_o} > C$$

where X_{P_o} is the P_0 quantile of the variable X; that is,

$$P(X \le X_{P_o}) = P_0 \ .$$

L-8.1.2 If P is the "true" proportion of X that is less than or equal to $C = X_P$, then

$$P(X \le C) = P \ .$$

L-8.1.3. The hypothesis statement can be written as:

$H_0 : P_0 \le P$, $H_A : P_0 > P$.

L-8.1.4. Equivalently,

$H_0 : P \ge P_0$, $H_A : P < P_0$.

(Note that *P,* the true portion of the population less than *C*, should note be confused with the probability density function *P*(*X*) for the variable *X* discussed in Appendix E.)

L-8.1.5. Because the only assumption is that it be a random sample, the procedures are valid for any underlying distributional shape. The procedures are also robust to outliers, as long as they do not represent data errors. This test is recommended when fewer than 50% of the results are detected. The test may be used as long as the proportion of non-detects is smaller than the proportion, p_0, of interest, and *n* must be relatively large for the test to be reliable.

L-8.1.6. Directions for the one-sample proportion test for a simple random sample and a systematic random sample are given below in Paragraph L-8.2, followed by an example presented in Paragraph L-8.3.

L-8.2. *Directions for a Simple Random Sample and a Systematic Random Sample*. Directions to apply the one-sample proportion test for Case 1 and Case 2: Case 1 ($H_0 : P \le P_0, H_A : P > P_0$); and Case 2 ($H_0 : P \ge P_0, H_A : P < P_0$), which are given in braces { }.

L-8.2.1. Given a random sample $x_1, x_2, \ldots, x_n$ of measurements from the population, let *P* denote the proportion of *X*'s that do not exceed *C*. This true proportion can be estimated from the sample data by dividing the number (*k*) of sample points that are less than or equal to *C* by the sample size (*n*).

$P \approx p = k/n$.

L-8.2.2. Compute *np*, and $n(1 - p)$. If both *np* and $n(1 - p)$ are greater than or equal to 5, proceed.

L-8.2.3. Otherwise, consult a statistician as analysis may be complex. Calculate:

$$z = \frac{p - P_0}{\sqrt{P_0(1 - P_0)/n}}.$$

L-8.2.4. Use Table B-15 of Appendix B to find the critical value, $Z_{1-\alpha}$, such that $(1-\alpha)100\%$ of the normal distribution is below $Z_{1-\alpha}$. For example, if $\alpha = 0.05$ then $Z_{1-\alpha} = 1.645$.

L-8.2.4.1. If $z > Z_{1-\alpha}\{z < -Z_{1-\alpha}\}$, H_0 may be rejected.

L-8.2.4.2. If $z \le Z_{1-\alpha}\{z \ge -Z_{1-\alpha}\}$, there is not enough evidence to reject H_0. Therefore, the false acceptance error rate must be verified.

L-8.2.5. To calculate the power of the test, choose a proportion, P_1, that would constitute a meaningful difference from P_0, and use a statistical software package such as the DEFT software (EPA QA/G-4D) or the DataQUEST software (EPA QA/G-9D) to generate the power curve of the test.

L-8.2.6. If only one false acceptance error rate (β) has been specified (at P_1), it is possible to calculate the sample size that achieves the DQOs. To do this, calculate:

$$m = \left[\frac{Z_{1-\alpha}\sqrt{P_0(1-P_0)} + Z_{1-\beta}\sqrt{P_1(1-P_1)}}{P_1 - P_0} \right]^2 .$$

L-8.2.7. If $m \le n$, the false acceptance error rate has been satisfied. Otherwise, the false acceptance error rate has not been satisfied. It is usually more helpful to do this calculation before sampling, as all of the parameter values needed for the calculation are available before the sampling begins.

L-8.2.8. The results of the test could be:

L-8.2.8.1. H_0 is rejected, conclude that $P > P_0$ $\{P < P_0\}$.

L-8.2.8.2. H_0 is not rejected, the false acceptance error rate was satisfied, and conclude that $P \le P_0$ $\{P \ge P_0\}$.

L-8.2.8.3. H_0 is not rejected, the false acceptance error rate was not satisfied, and the conclusion that $P \le P_0$ $\{P \ge P_0\}$ is uncertain because the sample size was too small.

L-8.2.9. *Example of the One-Sample Test for Proportions of Simple and Systematic Random Samples.* Groundwater concentrations of gasoline at a site are compared to a regulatory threshold C = 35 micrograms per liter (µg/L). Suppose this site has only 13 detections out of 90

groundwater samples collected to date. Because more than 50% of the data are censored, the test of proportions is more appropriate than a *t*-test or Wilcoxon signed rank test. The test of proportions can be used to determine if more than 95% of the concentrations are less than the regulatory threshold at the 90% level of confidence. The null and alternative hypotheses are as follows:

$H_0 : X_{0.95} \geq 35$ µg/L, $H_A : X_{0.95} < 35$ µg/L .

L-8.2.9.1. Equivalently,

$H_0 : P \leq 0.95$, $H_A : P > 0.95$.

(This is Case 1 in Paragraph L-8.2.) Suppose 11 of the detected concentrations exceed this regulatory threshold; therefore, the proportion of samples with detected concentrations below the threshold is $p = (90 - 11)/90 = 0.8778$.

L-8.2.9.2. Determine whether $np \geq 5$ and $n(1 - p) \geq 5$:

$np = 90 \times 0.8778 = 79$

$n(1-p) = 90 \times (1 - 0.08778) = 11$.

L-8.2.9.3. Because $np \geq 5$ and $n(1 - p) \geq 5$, the test of proportions can be used. In this example, $P_0 = 0.95$ and $1 - \alpha = 0.90$.

$$z = \frac{p - P_0}{\sqrt{P_0(1-P_0)/n}} = \frac{0.8778 - 0.95}{\sqrt{0.95(1-0.95)/90}} = -3.143 .$$

L-8.2.9.4. Using Table B-15 of Appendix B, we find the critical value $Z_{0.90}$=1.282.

L-8.2.9.5. Compare the calculated value *z* with the critical value. The null hypothesis is rejected if $z > Z_{0..90}$. As $-3.143 \leq 1.282$ ($z \leq Z_{0.90}$), there is not enough evidence to reject H_0. Therefore, the false acceptance error rate has to be verified through a power curve or sample size calculation. Suppose a false acceptance error rate was specified at $P_1 = 0.99$ ($\beta = 0.20$); it is possible to calculate the sample size that achieves this error rate using the following equation:

$$m = \left[\frac{Z_{1-\alpha}\sqrt{P_0(1-P_0)} + Z_{1-\beta}\sqrt{P_1(1-P_1)}}{P_1 - P_0} \right]^2$$

$$= \left[\frac{1.282\sqrt{0.95(1-0.95)} + 0.8417\sqrt{0.99(1-0.99)}}{0.99 - 0.95} \right]^2 = 82.43 \approx 83.$$

L-8.2.9.6. Because $83 \le 90$ ($m \le n$), the false acceptance error rate has been satisfied. Therefore, H_0 was not rejected and the false acceptance error rate was satisfied. There is at least 90% confidence that the proportion of gasoline concentrations below the regulatory threshold is less than 0.95 (i.e., $P \le 0.95$, or, equivalently, $X_{0.95} \ge 35$).