PART IV: COMPLIANCE/ASSESSMENT AND CORRECTIVE ACTION TESTS

This last part of the Unified Guidance addresses statistical methods useful in compliance/assessment and corrective action monitoring, where single-sample testing is required against a fixed groundwater protection standard [GWPS]. These standards include not only health- or risk-based limits, but also those derived from background as a fixed standard. The full subject of background GWPS testing is treated in **Section 7.5,** but any of the procedures in the following chapters might be applied to single-sample background tests.

The primary tool for both stages of monitoring is the confidence interval. Several varieties of confidence intervals are presented in **Chapter 21**, including confidence intervals around means, medians, and upper percentiles for stationary populations, and confidence bands around a trend for cases where groundwater concentrations are actively changing.

Strategies to implement confidence interval tests are discussed in **Chapter 22**. In particular, the focus is on designing tests with reasonable statistical performance in terms of power and per-test false positive rates.

Chapter 7 of **Part I** provides a discussion of the overall compliance/assessment and corrective action monitoring network design. Program elements such as the appropriate hypothesis structure, selecting the appropriate parameter for comparison to a fixed limit GWPS, sampling frequency, statistical power, and confidence levels are covered. These final two chapters present the tests in greater detail.

This page intentionally left blank

CHAPTER 21. CONFIDENCE INTERVALS

Confidence intervals are the recommended general statistical strategy in compliance/assessment or corrective action monitoring. Groundwater monitoring data must typically be compared to a fixed numerical limit set as a GWPS. In compliance/assessment, the comparison is made to determine whether groundwater concentrations have increased above the compliance standard. In corrective action, the test determines whether concentrations have decreased below a clean-up criterion or compliance level. In compliance/assessment monitoring, the lower confidence limit [LCL] is of primary interest, while the upper confidence limit [UCL] is most important in corrective action. For single-sample background GWPS testing, the hypothesis structures are the same as for fixed-limit health-based standards. Where a GWPS is based on two- or multiple sample testing, a somewhat different hypothesis structure is used (**Section 7.5**) and detection monitoring test procedures in **Part III** are applicable.

General strategies for using confidence intervals in compliance/assessment or corrective action monitoring are presented in **Chapter 7**, including discussion of how regulatory standards should be matched to particular statistical parameters (*e.g.*, mean or upper percentile). More specific strategies and examples are detailed in **Chapter 22**. In this chapter, basic algorithms and equations for each type of confidence interval are described, along with an example of the calculations involved.

21.1 PARAMETRIC CONFIDENCE INTERVALS

Confidence intervals are designed to estimate statistical characteristics of some parameter of a sampled population. *Parametric* confidence intervals do this for known distributional models, *e.g.*, normal, lognormal, gamma, Weibull, *etc*. Given a statistical parameter of interest such as the population mean (μ), the lower and upper limits of a confidence interval define the most probable concentration range in which the true parameter ought to lie.

Like any estimate, the true parameter may not be located within the confidence interval. The frequency with which this error tends to occur (based on repeated confidence intervals on different samples of the same sample size and from the same population) is denoted α, while its complement ($1-\alpha$) is known as the *confidence level*. The confidence level represents the percentage of cases where a confidence interval constructed according to a fixed algorithm or equation will actually contain its

intended target, *e.g.*, the population mean. **Section 7.2** discusses the difference between one- and two-sided confidence intervals and how the α error is assigned.

A point worth clarifying is the distinction between α as the complement of the confidence level when constructing a confidence interval and the significance level (α) used in hypothesis testing. Confidence intervals are often used strictly for estimation of population quantities. In that case, no test is performed, so α does not represent a false positive rate. Rather, it is simply the fraction of similar intervals that do not contain their intended target.

The Unified Guidance focuses on confidence interval limits compared to a fixed standard as a formal test procedure. In this case, the complement (α) of the confidence level used to generate the confidence interval is equivalent to the significance level (α) of the test. This assumes that the true population parameter under the null hypothesis is no greater than the standard in compliance/assessment monitoring or not less than the standard in corrective action.[1]

The parametric confidence intervals presented in the Unified Guidance share some common statistical assumptions. The most basic is that measurements used to construct a confidence interval be independent and identically distributed [*i.i.d.*]. Meeting this assumption requires that there be no outliers (**Chapter 12**), a stationary mean and variance over the period during which observations are collected (**Chapters 3** and **14**), and no autocorrelation between successive sampling events (**Chapter 14**). In particular, sampling events should be spaced far enough apart so that approximate statistical independence can be assumed (at many sites, observations should not be sampled more often than quarterly). Sample data should also be examined for trends. The mean is not stationary under a significant trend, as assumed in applying the other methods of this section. An apparent trend may need to be handled by computing a confidence band around the trend line (**Section 21.3**).

Another common assumption is that the sample data are either normal in distribution or can be normalized via a transformation (**Chapter 10**). Normality can be difficult to check if the sample contains a significant number of left-censored measurements (i.e., non-detects). The basic options for censored samples are presented in **Chapter 15**. If the non-detect percentage is no more than 10-15%, it may be possible to assess normality by first substituting one-half of the reporting limit [RL] for each non-detect. For higher non-detect percentages up to 50%, the Unified Guidance recommends computing a censored probability plot using either the Kaplan-Meier or Robust Regression on Order Statistics [Robust ROS] techniques (both in **Chapter 15**).

If a censored probability plot suggests that the sample (or some transformation of the sample) is normal, either Kaplan-Meier or Robust ROS can be used to construct estimates of the mean ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) adjusted for the presence of non-detects. These estimates should be used *in place of* the sample mean ($\bar{x}$) and standard deviation (s) in the parametric equations below.

[1] Technically, α represents the *maximum* possible false positive rate associated with the composite null hypothesis H_0: $\mu \leq$ GWPS or H_0: $\mu \geq$ GWPS.

21.1.1 CONFIDENCE INTERVAL AROUND NORMAL MEAN

BACKGROUND AND PURPOSE

When compliance point data is to be compared to a fixed standard (*e.g.*, a maximum concentration limit [MCL]) and the standard in question is interpreted to represent an average or true mean concentration, a confidence interval around the mean is the method of statistical choice. A confidence interval around the mean is designed to estimate the true average of the underlying population, while at the same time accounting for variability in the sample data.

REQUIREMENTS AND ASSUMPTIONS

Confidence intervals around the mean of a normal distribution should only be constructed if the data are approximately normal or at least are reasonably symmetric (*i.e.*, the skewness coefficient is close to zero). An inaccurate confidence interval is likely to result if the sample data are highly non-normal, particularly for right-skewed distributions. If the observations are better fit by a lognormal distribution, special equations or methods need to be used to construct an accurate confidence interval on the arithmetic mean with a specified level of confidence (**Section 21.1.3**). *Therefore, checking for normality is an important first step*.

A confidence interval should not be constructed with less than 4 measurements per compliance well, and preferably 8 or more. The equation for a normal-based confidence interval around the mean involves estimating the population standard deviation via the sample standard deviation (s). This estimate can often be imprecise using a small sample size (*e.g.*, $n \leq 4$). The equation also involves a Student's t-quantile based on n–1 degrees of freedom [df], where n equals the sample size. The t-quantile is large for small n, leading to a much wider confidence interval than would occur with a larger sample size. For a 99% confidence level, the appropriate t-quantile would be $t = 31.82$ for $n = 2$, $t = 4.54$ for $n = 4$, and $t = 3.00$ for $n = 8$.

This last consideration is important since statistically significant evidence of a violation during compliance/assessment or success during corrective action is indicated only when the entire confidence interval is to one side of the standard (*i.e.*, it does not *straddle* the fixed standard; see **Chapter 7**). For a small sample size, the confidence interval may be so wide that a statistical difference is unlikely to be identified. This can happen *even if* the true mean groundwater concentration *is* different from the compliance or clean-up standard, due to the statistical uncertainty associated with the small number of observations. More specific recommendations on appropriate sample sizes are presented in **Chapter 22**, where the *statistical power* of the confidence interval tests is explored.

PROCEDURE

Step 1. Check the basic statistical assumptions of the sample as discussed above. Assuming a normal distributional model is acceptable, calculate the sample mean ($\bar{x}$) and standard deviation (s).

Step 2. Given a sample of size n and the desired level of confidence ($1-\alpha$), for each compliance well calculate either the lower confidence limit (for compliance/assessment monitoring) with the equation:

$$LCL_{1-\alpha} = \bar{x} - t_{1-\alpha,n-1}\frac{s}{\sqrt{n}} \quad [21.1]$$

or the upper confidence limit (for corrective action) with the equation:

$$UCL_{1-\alpha} = \bar{x} + t_{1-\alpha,n-1}\frac{s}{\sqrt{n}} \quad [21.2]$$

where $t_{1-\alpha,n-1}$ is obtained from a Student's *t*-table with (n–1) degrees of freedom (**Table 16-1** in **Appendix D**). To construct a two-sided interval with overall confidence level equal to (1–α), substitute α/2 for α in the above equations.

Step 3. Compare the limit calculated in **Step 2** to the fixed compliance or clean-up standard (*e.g.*, the MCL or alternate concentration limit [ACL]. For compliance/assessment monitoring, the LCL in equation [21.1] should be used to compute the test. For corrective action, the UCL in equation [21.2] should be used instead.

►EXAMPLE 21-1

The table below lists concentrations of the pesticide Aldicarb in three compliance wells. For illustrative purposes, the health-based standard in compliance monitoring for Aldicarb has been set to 7 ppb. Determine at the α = 5% significance level whether or not any of the wells should be flagged as being out of compliance.

	Aldicarb Concentration (ppb)		
Sampling Date	Well 1	Well 2	Well 3
January	19.9	23.7	5.6
February	29.6	21.9	3.3
March	18.7	26.9	2.3
April	24.2	26.1	6.9
Mean	23.10	24.65	4.52
SD	4.93	2.28	2.10
Skewness (γ_1)	0.506	–0.234	0.074
Shapiro-Wilk (W)	0.923	0.943	0.950

SOLUTION

Step 1. First test the data for non-normality and/or significant skewness. Based on four samples per well, the skewness coefficients and Shapiro-Wilk statistics have been computed and are listed above. None of the skewness coefficients are significantly different from zero. In addition, the α = .10 critical point for the Shapiro-Wilk test with n = 4 (as presented in **Chapter 10**) is 0.792, less than each of the Shapiro-Wilk statistics; consequently, there is no significant evidence of non-normality. Construct a normal-based confidence interval around the mean.

Step 2. Calculate the sample mean and standard deviation of the Aldicarb concentrations for each compliance well. These statistics are listed above.

Step 3. Since $\alpha = 0.05$, the confidence level must be set to $(1-\alpha) = 0.95$. Obtain the upper 95th percentile of the *t*-distribution with $(n-1) = 3$ degrees of freedom from **Table 16-1** in **Appendix D**, namely $t_{.95,3} = 2.353$. Then calculate the lower confidence limit [LCL] for each well's mean concentration, using equation [21.1]:

$$\text{Well 1: } LCL_{.95} = 23.10 - \left(2.353 \times 4.93\right)\Big/\sqrt{4} = 17.30 \text{ ppb}$$

$$\text{Well 2: } LCL_{.95} = 24.65 - \left(2.353 \times 2.28\right)\Big/\sqrt{4} = 21.97 \text{ ppb}$$

$$\text{Well 3: } LCL_{.95} = 4.52 - \left(2.353 \times 2.10\right)\Big/\sqrt{4} = 2.05 \text{ ppb}$$

Step 4. Compare each LCL to the compliance standard of 7 ppb. The LCLs for Well 1 and Well 2 lie above 7 ppb, indicating that the mean concentration of Aldicarb in both of these wells significantly exceeds the compliance standard. However, the LCL for Well 3 is below 7 ppb. providing insufficient evidence at the $\alpha = 0.05$ level that the mean in Well 3 is out of compliance. ◄

21.1.2 CONFIDENCE INTERVAL AROUND LOGNORMAL GEOMETRIC MEAN

PURPOSE AND BACKGROUND

For many groundwater monitoring constituents, neither the assumption of normality nor approximate symmetry holds for the original concentration data. Often the underlying population is heavily right-skewed, characterized by a majority of lower level concentrations combined with a long right-hand tail of infrequent but extreme values. A model such as the lognormal distribution is commonly used to analyze such data.

The lognormal is traditionally designated by the notation $\Lambda(\mu, \sigma)$ (Aitchison and Brown, 1976), where μ and σ denote parameters controlling the *location* and *scale* of the population. Typically designated as $N(\mu, \sigma)$, a normal distribution also has parameters μ and σ which denote the true mean and standard deviation. These two parameters play different roles in lognormal distributions. The key distinction is between the *arithmetic* domain (or the original measurement scale of the data) and the *logarithmic* domain. The latter denotes the mathematical space following a logarithmic transformation. Transformed lognormal data are *normally-distributed* in the logarithmic domain. In this new domain, μ represents the true mean of the log-transformed measurements--- that is, the *log-mean*. Likewise, σ represents the true standard deviation of the log-transformed values or the *log-standard deviation*.

A common misperception is to assume that a standard equation for a normal-based confidence interval can be applied to log-transformed data, with the interval endpoints then back-transformed (*i.e.*, exponentiated) to the arithmetic domain to get a confidence interval around the lognormal *arithmetic* mean. Invariably, such an interval will underestimate the true mean. The Student t- confidence interval applies to a *geometric* mean of the lognormal population when back-transformed, rather than the higher-valued *arithmetic* mean. The reason is that the sample log-mean gives an estimate of the lognormal parameter μ. When this estimate is back-transformed to the arithmetic domain, one has an estimate of

$\exp(\mu)$ — the lognormal geometric mean — not an estimate of the lognormal arithmetic mean, which is expressed as $\exp(\mu + .5\sigma^2)$.

Although a confidence interval around the lognormal geometric mean is *not* an accurate estimate of the arithmetic mean, there are instances where such an interval may be helpful. While many GWPSs are interpreted to represent long-term arithmetic averages, some (as detailed in **Chapter 7**) can better represent medians or percentiles of the underlying distribution. Because the lognormal geometric mean is equivalent to the median, a geometric mean may in some cases be a better statistical parameter of comparison than the lognormal arithmetic mean. Furthermore, when the lognormal coefficient of variation is large, the arithmetic mean is substantially larger than the geometric mean, mostly due to infrequent but extreme individual measurements. The bulk of individual observations are located much closer to the geometric mean. It may be that a comparison of the GWPS to the geometric mean rather than to the arithmetic mean will provide a more reasonable test of long-term concentration levels.

Special equations or computational methods are used to construct an accurate confidence interval with a specified level of confidence (**Section 21.1.3**) when an estimate of the *arithmetic* mean is needed and the observations are approximately normal. There is another factor to consider when estimating an *upper* confidence limit on the lognormal arithmetic mean using Land's procedure (described in **Section 21.1.3**) or other possible procedures (see for instance Singh et al., 1997). When used with highly variable data, it can lead to severely-biased, high estimates of the confidence limit. This can make it very difficult to evaluate the success of corrective action measures.

In these cases, precise parametric estimation of the arithmetic mean may have to be foregone in favor of an alternate statistical procedure. One such alternative is a non-parametric confidence interval around the median (**Section 21.2**). Another alternative when the sample is approximately lognormal is an estimate around the geometric mean which is equivalent to the population median. A third more computationally intensive option is a *bootstrap confidence interval* around the lognormal arithmetic mean (see discussion in **Section 21.1.3**). Unlike the first two options, this last alternative allows a direct estimate of the arithmetic mean.

REQUIREMENTS AND ASSUMPTIONS

Confidence intervals around the geometric mean of a lognormal distribution should only be constructed if the log-transformed data are approximately normal or at least reasonably symmetric (*i.e.*, the skewness coefficient in the logarithmic domain is close to zero). The methods of **Chapter 10** can be used to test normality of the log-transformed values. If the log-transformed sample contains non-detects, normality *on the log-scale* should be assessed using a censored probability plot. Adjusted estimates of the mean and standard deviation on the log-scale can then be substituted for the log-mean ($\overline{y}$) and log-standard deviation (s_y) in the equations below. Like a normal arithmetic mean, a confidence interval around the lognormal geometric mean should not be constructed without a minimum of 4 measurements per compliance well, and preferably with 8 or more.

PROCEDURE

Step 1. Take the logarithm of each measurement, denoted as y_i, and check the n log-transformed values for normality. If the log-transformed measurements are approximately normal, calculate

the log-mean ($\overline{y}$) and log-standard deviation (s_y). If the normal model is rejected, consider instead a non-parametric confidence interval (**Section 21.2**).

Step 2. Given the desired level of confidence (1–α), calculate either the LCL (for compliance/assessment monitoring) with the equation:

$$LCL_{1-\alpha} = \exp\left(\overline{y} - t_{1-\alpha,n-1} \frac{s_y}{\sqrt{n}} \right) \quad [21.3]$$

or the UCL (for corrective action) with the equation:

$$UCL_{1-\alpha} = \exp\left(\overline{y} + t_{1-\alpha,n-1} \frac{s_y}{\sqrt{n}} \right) \quad [21.4]$$

where $t_{1-\alpha,n-1}$ is obtained from a Student's *t*-table with (n–1) degrees of freedom (**Table 16-1** in **Appendix D**). In order to construct a two-sided interval with the overall confidence level equal to (1–α), substitute α/2 for α in the above equations.

Step 3. Compare the limits calculated in **Step 2** to the fixed compliance or clean-up standard (*e.g.*, the MCL or ACL). For compliance/assessment, use the LCL in equation [21.3]. For corrective action, use the UCL in equation [21.4].

Note in either case that the regulatory authority will have to approve the use of the geometric mean as a reasonable basis of comparison against the compliance standard. In some cases, there may be few other statistical options. However, stakeholders should understand that the geometric and arithmetic means estimate two distinct statistical characteristics of the underlying lognormal population.

►EXAMPLE 21-2

Suppose the following 8 sample measurements of benzene (ppb) have been collected at a landfill that previously handled smelter waste and is now undergoing remediation efforts. Determine whether or not there is statistically significant evidence at the α = 0.05 significance level that the true geometric mean benzene concentration has fallen below the permitted MCL of 5 ppb.

Sample Month	Benzene (ppb)	Log Benzene log(ppb)
1	0.5	–0.693
2	0.5	–0.693
3	1.6	0.470
4	1.8	0.588
5	1.1	0.095
6	16.1	2.779
7	1.6	0.470
8	<0.5	–1.386

SOLUTION

Step 1. To estimate an upper confidence bound on the geometric mean benzene concentration with 95% confidence, first test the skewness and normality of the data set. Since the one non-detect concentration is unknown but presumably between 0 ppb and the RL of 0.5 ppb, a reasonable compromise is to impute this value at 0.25 ppb, half the RL. The skewness is computed as γ_1 = 2.21, a value too high to suggest the data are normal. In addition, a Shapiro-Wilk test statistic on the raw measurements works out to SW = 0.521, failing an assumption of normality at far below a significance level of α = 0.01.

On the other hand, transforming the data via natural logarithms gives a smaller skewness coefficient of γ_1 = 0.90 and a Shapiro-Wilk statistic of W = 0.896. Because these values are consistent with normality on the log-scale (the critical point for the Shapiro-Wilk test with n = 8 and α = 0.10 is 0.818), the data set should be treated as lognormal for estimation purposes. As a consequence, equation [21.4] can be used to construct a one-sided UCL on the geometric mean.

Step 2. Compute the sample log-mean and log-standard deviation. This gives $\bar{y}$ = 0.2037 log(ppb) and s_y = 1.2575 log(ppb).

Step 3. Apply the log-mean and log-standard deviation into equation [21.4] for a UCL with α = .05, n = 8, and 7 degrees of freedom. This gives an estimated limit of:

$$UCL_{.95} = \exp\left(\bar{y} + t_{.95,7} \frac{s_y}{\sqrt{8}} \right) = \exp\left(.2037 + 1.895 \times .4446\right) = 2.847 \text{ ppb}$$

Step 4. Compare the UCL to the MCL of 5 ppb. Since the limit is less than the fixed standard, there is statistically significant evidence that the benzene geometric mean, and consequently, the median benzene concentration, is less than 5 ppb. However, this calculation does *not* show that the benzene *arithmetic* mean is less than the MCL. Extreme individual benzene measurements could show up with enough regularity to cause the arithmetic mean to be higher than 5 ppb. ◄

21.1.3 CONFIDENCE INTERVAL AROUND LOGNORMAL ARITHMETIC MEAN

PURPOSE AND BACKGROUND

Estimation of a lognormal arithmetic mean is not completely straightforward. As discussed in **Section 21.1.2**, applying standard equations for normal-based confidence limits around the mean to log-transformed measurements and then exponentiating the limits, results in confidence intervals that are invariably underestimate the arithmetic mean.

Inferences on arithmetic means for certain kinds of skewed populations can be made either exactly or approximately through the use of special techniques. In particular, if a confidence interval on the arithmetic mean is desired, Land (1971; 1975) developed an exact technique along with extensive tables

for implementing it when the underlying population is lognormal. Land also developed a more complicated approximate technique (for a full description and examples see EPA, 1997) when the population can be transformed to normality via any other increasing, 1-1, and twice differentiable transformation (*e.g.*, square, square root, cube root, *etc*.).

Although the core of Land's procedure is a correction for the so-called 'transformation bias' that occurs when making back-transforming estimates from the logarithmic domain to the raw concentration domain, it can produce unacceptable results, particularly with UCLs. The Unified Guidance advises caution when applying Land's procedure, particularly when the lognormal population has a high coefficient of variation. In those cases, the user may want to consider alternate techniques, such as those discussed in Singh, *et al* (1997 and 1999). One option is to use EPA's free-of-charge **Pro-UCL** software Version 4.0 (www.epa.gov/esd/tsc/software.htm). It computes a variety of upper confidence limits, including a bootstrap confidence interval around the arithmetic mean. This technique can be applied to lognormal data to get a direct, non-parametric UCL that tends to be less biased and to give less extreme results than Land's procedure.

For cases or sample sizes not covered by **Tables 21-1** through **21-8** in **Appendix D** when using Land's procedure, Gibbons and Coleman (2001) describe a method of approximating the necessary *H*-factors. The same authors review other alternate parametric methods for computing UCLs.

REQUIREMENTS AND ASSUMPTIONS

Confidence intervals around the arithmetic mean of a lognormal distribution should be constructed only if the data pass a test of approximate normality *on the log-scale*. While many groundwater and water quality populations tend to follow the lognormal distribution, the data should first be tested for normality on the original concentration scale. If such a test fails, the sample can be log-transformed and re-tested. If the log-transformed sample contains non-detects, normality *on the log-scale* should be assessed using a censored probability plot (**Chapter 15**). If a lognormal model is tenable, adjusted estimates of the mean and standard deviation on the log-scale can be substituted for the log-mean ($\overline{y}$) and log-standard deviation (s_y) in the equations below.

As with normal-based confidence intervals, the confidence interval here should not be constructed with fewer than 4 measurements per compliance well, and preferably with 8 or more. The reasons are similar: the equation for a lognormal-based confidence interval around the arithmetic mean depends on the sample log-standard deviation (s_y), used as an estimate of the underlying log-scale population standard deviation. This estimate can be quite imprecise when fewer than 4 to 8 observations are used. A special factor (H) was developed by Land to account for variability in a skewed population. These factors are larger for smaller samples sizes, and need to be exponentiated to estimate the final confidence limits (see below). Consequently there is a significant penalty associated with estimating the arithmetic mean using a small sample size, occasionally seen in remarkably wide confidence limits. The effect is especially noticeable when computing an UCL for corrective action monitoring.

PROCEDURE

Step 1. Test the log-tranformed sample for normality. If the lognormal model provides a reasonable fit, denote the log-transformed measurements by y_i and move to Step 2.

Step 2. Compute the sample log-mean ($\overline{y}$) and log-standard deviation (s_y).

Step 3. Obtain the correct bias-correction factor(s) (H_α) from Land's (1975) tables (**Tables 21-1 through 21-8** in **Appendix D**), where the correct factor depends on the sample size (n), the sample log-standard deviation (s_y), and the desired confidence level ($1-\alpha$).

Step 4. Plug these factors into one of the equations given below for the LCL or UCL (depending on whether the comparison applies to compliance/assessment monitoring or to corrective action). Note that to construct a two-sided interval with an overall confidence level of ($1-\alpha$), the equations should be applied by substituting $\alpha/2$ for α.

$$LCL_{1-\alpha} = \exp\left(\overline{y} + .5s_y^2 + \frac{s_y H_\alpha}{\sqrt{n-1}} \right) \quad [21.5]$$

$$UCL_{1-\alpha} = \exp\left(\overline{y} + .5s_y^2 + \frac{s_y H_{1-\alpha}}{\sqrt{n-1}} \right) \quad [21.6]$$

Step 5. Compare the confidence limit computed in **Step 4** to the fixed compliance or clean-up standard. In compliance/assessment monitoring, use the LCL of equation [21.5]. In corrective action, use equation [21.6] for the UCL.

►EXAMPLE 21-3

Determine whether the benzene concentrations of **Example 21-2** indicate that the benzene arithmetic mean is below the permitted MCL of 5 ppb at the $\alpha = 0.05$ significance level.

SOLUTION

Step 1. From **Example 21-2**, the benzene data were found to fail a test of normality, but passed a test of lognormality (*i.e.*, they were approximately normal on the log-scale). As a consequence, Land's equation in [21.6] should be used to construct a one-sided UCL on the arithmetic mean.

Step 2. Compute the log-mean and log-standard deviation from the log-scale data. This gives $\overline{y}$ = 0.2037 log(ppb) and s_y = 1.2575 log(ppb).

Step 3. Using **Table 21-6** in **Appendix D**, pick the appropriate H-factor for estimating confidence limits around a lognormal arithmetic mean, noting that to achieve 95% confidence for a one-sided UCL, one must use ($1-\alpha$) = 0.95. With a sample size of n = 8 and a standard deviation on the log-scale of 1.2575 log(ppb), $H_{.95}$ = 4.069.

Step 4. Plug these values along with the log-mean of 0.2037 log(ppb) into equation [21.6] for the UCL. This leads to a 95% one-sided confidence limit equal to:

$$UCL_{.95} = \exp\left(.2037 + .5\left(1.5813\right) + \frac{\left(1.2575\right)\left(4.069\right)}{\sqrt{7}} \right) = 18.7 \text{ ppb}$$

Step 5. Compare the UCL against the MCL of 5 ppb. Since the UCL is greater than the MCL, evidence is not sufficient at the 5% significance level to conclude that the true benzene arithmetic mean concentration is now below the MCL. This conclusion holds despite the fact that all but one of the benzene measurements is less than than 5 ppb. In lognormal populations, it is not uncommon to see one or two seemingly extreme measurements coupled with a majority of much lower concentrations. Since these extreme measurements help determine the location of the arithmetic mean, it is not unreasonable to expect that the true mean might be larger than 5 ppb.

The contrast in this result to **Example 21-2** is noteworthy. In that case, the UCL on the geometric mean was only 2.85 ppb. The estimated lognormal coefficient of variation with these data (**Chapters 3** and **10**) is $CV = 1.965$, somewhat on the high side. It is no surprise that results for the arithmetic and geometric means on the same sample are rather different. Neither estimator is necessarily invalid, but a decision needs to be made as to whether the MCL for benzene in this setting should be better compared to an arithmetic mean or to a geometric mean/ median for lognormal distributions. ◄

21.1.4 CONFIDENCE INTERVAL AROUND UPPER PERCENTILE

BACKGROUND AND PURPOSE

Although most MCLs and ACLs appear to represent arithmetic or long-term averages (**Chapter 7**), they can also be interpreted as standards not to be exceeded with any regularity. Other fixed standards like nitrate/nitrite attempt to limit short-term risks and thus represent upper percentiles instead of means. In these cases, the appropriate confidence interval is one built around a specific upper percentile.

The particular upper percentile chosen will depend on what the fixed compliance standard represents or is intended to represent. If the standard is a concentration that represents the 90th percentile, the confidence interval should be built around the upper 90th percentile. If the standard is meant to be a *maximum*, 'not to be exceeded,' concentration, a slightly different strategy should be used. Since there is no maximum value associated with continuous distributions like normal and lognormal, it is not possible to construct a confidence interval around the population maximum. Instead, one must settle for a confidence interval around a sufficiently high percentile, one that will exceed nearly all of the population measurements. Possible choices are the upper 90th to 95th percentile. By estimating the location of these percentiles, one needs to determine whether a sufficiently small fraction (*e.g.*, at most 1 in 10 or 1 in 20) of the possible measurements will ever exceed the standard. For even greater protection against exceedances, the upper 99th percentile could be selected, implying that at most 1 in 100 measurements would ever exceed the standard. But as noted in **Chapter 7**, selection of very high percentiles using non-parametric tests can make it extremely difficult to demonstrate corrective action success.

REQUIREMENTS AND ASSUMPTIONS

The equations for constructing parametric confidence intervals around an upper percentile assume that the data are normally distributed, at least approximately. If the data can be normalized via a transformation, the observations should first be transformed before computing the confidence interval. Unlike confidence intervals around an arithmetic mean for transformed data, no special equations are required to construct similar intervals around an upper percentile. The same equations used for normal data can be applied to data in the transformed domain. The only additional step is that the confidence interval limits must be back-transformed prior to comparing them against the fixed standard.

The confidence interval presented here should not be constructed with fewer than 4 measurements per compliance well, and preferably with 8 or more. Too small a sample size leads to imprecise estimates of the sample standard deviation (s). Another reason is that the confidence interval equation involves a special multiplier τ, which depends on both the desired confidence level ($1-\alpha$) and the sample size (n). When n is quite small, the τ multiplier is much greater. This leads to a much wider confidence interval than that obtained with a larger n, and therefore much greater statistical uncertainty. For example, at a confidence level of 95%, the appropriate τ multiplier for an upper one-sided limit on the 95th percentile is $\tau = 26.260$ when $n = 2$, $\tau = 5.144$ when $n = 4$, and $\tau = 3.187$ when $n = 8$.

When determining the τ factor(s) needed for a confidence interval around an upper percentile, it should be noted that unlike the symmetric Student's t-distribution, separate τ factors need to be determined for the LCL and UCL. Since an upper percentile like the 95th is generally larger than the population mean, the equations for *both* the lower (*i.e.*, LCL) and upper (*i.e.*, UCL) limits involve *adding* a multiple of the standard deviation to the sample mean. The only difference is that a smaller multiple τ_{LCL} is used for the LCL, while a larger τ_{UCL} is used for the upper confidence limit. For certain choices of n, P and 1-α, the multiple τ_{LCL} can even be negative.

PROCEDURE

Step 1. Test the raw data for normality. If approximately normal, construct the interval on the original measurements. If the data can be normalized via a transformation, construct the interval on the transformed values.

Step 2. For a normal sample, compute the sample mean ($\bar{x}$) and standard deviation (s). If the data have been transformed, compute the mean and standard deviation of the transformed measurements.

Step 3. Given the percentile (P) to be estimated, sample size (n), and the desired confidence level ($1-\alpha$), use **Tables 21-9** and **21-10** in **Appendix D** to determine the τ factor(s) needed to construct the appropriate one-sided or two-sided interval. A one-sided LCL is then computed with the equation:

$$LCL_{1-\alpha} = \bar{x} + s \cdot \tau\left(P; n, \alpha\right) \qquad [21.7]$$

where $\tau(P; n, \alpha)$ is the lower α factor for the Pth percentile given n sample measurements. A one-sided UCL is given similarly by the equation:

$$UCL_{1-\alpha} = \bar{x} + s \cdot \tau\left(P; n, 1-\alpha\right) \quad [21.8]$$

Finally, a two-sided confidence interval is computed by the pair of equations for the LCL and UCL:

$$LCL_{1-\alpha/2} = \bar{x} + s \cdot \tau\left(P; n, \alpha/2\right) \quad [21.9]$$

$$UCL_{1-\alpha/2} = \bar{x} + s \cdot \tau\left(P; n, 1-\alpha/2\right) \quad [21.10]$$

Step 4. If the data have been transformed, the equations of Step 3 would be used but with two changes: 1) the mean and standard deviation of the transformed values are substituted for $\bar{x}$ and s; and 2) the resulting limits back-transformed to get final confidence limits in the concentration domain. If a logarithmic transformation has been employed, the log-mean and log-standard deviation would be substituted for the sample mean and standard deviation. The resulting limit(s) must be exponentiated to get the final confidence limits, as in the equations below:

$$LCL_{1-\alpha} = \exp\left[\bar{y} + s_y \cdot \tau\left(P; n, \alpha\right)\right] \quad [21.11]$$

$$UCL_{1-\alpha} = \exp\left[\bar{y} + s_y \cdot \tau\left(P; n, 1-\alpha\right)\right] \quad [21.12]$$

Step 5. Compare the confidence limit(s) computed in Step 3 (or Step 4) versus the fixed compliance or clean-up standard. In compliance/assessment, use the LCL of equation [21.7]. In corrective action, use equation [21.8] for the UCL.

Note that although the above equations differentiate between the α-error used with the LCL and 1-α for the UCL, **Tables 21-9** and **21-10** in **Appendix D** are constructed identically. The α-error is represented by its confidence complement 1-α in **Table 21-10** of **Appendix D**.

►EXAMPLE 21-4

Assume that a facility permit has established an ACL of 30 ppb that should not be exceeded more than 5% of the time. Use the Aldicarb concentrations and diagnostic statistical information from **Example 21-1** to evaluate data from the three compliance wells. Determine whether any of the wells should be flagged as being out of compliance.

SOLUTION

Step 1. From **Example 21-1**, all of the wells pass a normality test. Use the sample mean and standard deviation for each compliance well, from the tabular information in **Example 21-1**.

Step 2. Select the correct τ factor from **Table 21-10** of **Appendix D** to construct a 99% LCL on the upper 95th percentile. The upper 95th percentile is needed because the permitted ACL cannot be exceeded more than 5% of the time, implying that 95% of all the Aldicarb measurements should fall below the fixed standard. With $n = 4$ observations per well, this leads to $\tau\ (P;\ n,\ \alpha) = \tau\ (.95;\ 4,\ .01) = 0.443$.

Step 3. Compute the LCL for each well as follows using equation [21.7]:

$$Well1: LCL_{.99} = 23.10 + (0.443)(4.93) = 25.28\, ppb$$
$$Well2: LCL_{.99} = 24.65 + (0.443)(2.28) = 25.66\, ppb$$
$$Well3: LCL_{.99} = 4.52 + (0.443)(2.10) = 5.45\ ppb$$

Step 4. Compare each LCL against the ACL of 30 ppm. Since each well LCL is less than the ACL, there is insufficient statistical evidence that the upper 95th percentile of the Aldicarb distribution exceeds the fixed standard. Consequently, there is no conclusive evidence that more than 5% of the Aldicarb concentrations will exceed the ACL.

If the site were in corrective action instead of compliance/assessment monitoring, UCLs around the 95th percentile would be needed instead of LCLs. In that case, with n = 4 observations per well, τ (P; n, 1–α) = τ (.95; 4, .99) = 9.083 from **Table 21-9** of **Appendix D**. Then, the respective well UCLs would be:

$$Well1: UCL_{.99} = 23.10 + (9.083)(4.93) = 67.88\, ppb$$
$$Well2: UCL_{.99} = 24.65 + (9.083)(2.28) = 45.36\, ppb$$
$$Well3: UCL_{.99} = 4.52 + (9.083)(2.10) = 23.59\ ppb$$

In this case, two of the three wells would not meet the corrective action limit of 30 ppb. ◄

21.2 NON-PARAMETRIC CONFIDENCE INTERVALS

BACKGROUND AND PURPOSE

A non-parametric confidence interval should be considered when a sample is non-normal and cannot be normalized, perhaps due to a significant fraction of non-detects. Non-parametric confidence interval endpoints are generally chosen as *order statistics* of the sample data. The specific order statistics selected will depend on the sample size (n), the desired confidence level (1–α), and the population characteristic being estimated.

Since the data are not assumed to follow a particular distribution, it is generally not possible to construct a confidence interval around the population mean. One fairly rare exception would be if it were already known that the distribution is symmetric (where the mean is also the median). Sample order statistics represent, by definition, concentration levels exceeded by a certain number and hence a *fraction* of the sample values. They are excellent estimators of the *percentiles* of a distribution, but *not* of quantities like the arithmetic mean. The latter entails summing the data values and averaging the result. In positively-skewed populations, not only is the arithmetic mean greater than the median, it also may not correspond to any particular percentile.

Non-parametric confidence intervals can be developed either around a measure of the center of the population (*i.e.*, the population median or 50th percentile) or around an upper or lower percentile (*e.g.*, the upper 90th). The choice of percentile affects which order statistics are selected as interval endpoints.

The sample median is generally estimated using a smaller order statistic than that used for an upper 95th percentile.

Despite the distinction between non-parametric confidence intervals around the median and similar intervals around an upper or lower percentile, the mathematical algorithm used to construct both types is essentially identical. Given an unknown $P \times 100$th percentile of interest (where P is between 0 and 1) and a sample of n concentration measurements, the probability that any randomly selected measurement will be less than the $P \times 100$th percentile is simply P. Then the probability that the measurement will exceed the $P \times 100$th percentile is $(1–P)$. Hence the number of sample values falling below the $P \times 100$th percentile out of a set of n should follow a *binomial distribution* with parameters n and *success probability* P, where 'success' is defined as the event that a sample measurement is below the $P \times 100$th percentile.

Because of this connection, the binomial distribution can be used to determine the probability that the interval formed by a given pair of order statistics will contain the percentile of interest. This kind of probability calculation makes repeated use of the cumulative binomial distribution, often denoted $Bin(x;n,p)$. It represents the probability of x or fewer successes occurring in n trials with success probability p. The computational equation for this expression[2] can be written as:

$$Bin\left(x;n,p\right)=\sum_{i=0}^{x}\binom{n}{i}p^{i}\left(1-p\right)^{n-i} \qquad [21.13]$$

To make statistical inferences about the $P \times 100$th percentile, P (expressed as a fraction) would be substituted for p in equation [21.13]. It can be seen why the same basic algorithm applies both to confidence intervals around the median and around upper percentiles like the 95th. If an interval around the median is desired, one would set $P = 0.50$. For an interval needed around the upper 95th percentile, one would set $P = 0.95$ and perform similar calculations.

When constructing non-parametric confidence intervals, the type of confidence interval needs to be matched against the kind of fixed standard to which it will be compared. Since the arithmetic mean cannot be estimated directly, a confidence interval around the *median* should be used for those cases where the compliance standard represents an average. Some fixed standards can, of course, be directly interpreted as *median* concentration levels, but even for those standards representing arithmetic averages, the confidence interval on the median will give the 'next best' comparison when a non-parametric method is used.

The interpretation of a confidence interval on the median is similar to that of a parametric confidence interval around the mean. In compliance/assessment monitoring, if the LCL with confidence level $(1–\alpha)$ exceeds the compliance standard, there is statistically significant evidence that the true

---

[2] The mathematical expression $\binom{n}{i}$ refers to the combination of n events taken i at a time. It can be calculated as: $n!/(i!\times[n-i]!)$, where $n! = \{n \times (n-1) \times \ldots \times 2 \times 1\}$. By convention, $0! = 1$.

population *median* is higher than the standard. In corrective action monitoring, if the UCL is *below* the clean-up standard, one can conclude that the true population median is less than the standard with α-level significance.

REQUIREMENTS AND ASSUMPTIONS

Because a non-parametric confidence interval does not assume a specific distributional form for the underlying population, there is no need to fit a probability model to the data. If a significant portion of the data are non-detect, it may be impossible to adequately fit such a model.. The non-parametric confidence interval method only requires the ability to rank the sample data values and pick out selected order statistics as the interval endpoints. Unfortunately, this ease of construction comes with a price. As opposed to parametric intervals, non-parametric confidence intervals tend to be wider and generally require larger sample sizes to achieve comparably high confidence levels. To compute the LCL around the median with 99% confidence, at least 7 compliance point measurements are needed in the non-parametric case. Therefore, sample data should be fit to a specific probability distribution whenever possible.

The general method for constructing non-parametric confidence intervals involves an *iterative testing procedure*, where potential endpoints are selected from the sorted data values (*i.e.*, order statistics) and then tested to determine what confidence level is associated with those endpoints. If the initial choice of order statistics gives an interval with insufficient confidence, the interval needs to be widened and tested again. Clearly, the greatest confidence will be associated with an interval defined by the minimum and maximum observed sample values. But if the sample size n is small, even the largest possible confidence level may be less than the desired target confidence (*e.g.*, $(1-\alpha) = 0.99$). As such, the *actual* or *achieved* confidence level needs to be listed when reporting results of a non-parametric confidence interval test.

It may be especially difficult to achieve target confidence levels around upper percentiles even when the sample size is fairly large. An instructive example is when estimating an upper 95th percentile with a sample size of $n = 20$. In that case, the highest possible two-sided confidence level is approximately 64%, achieved when the minimum and maximum data values are taken as the interval endpoints. The confidence level is substantially less than the usual targets of 90% or more, and has very limited value as a decision basis.

The *width* of a confidence interval (which expands as the level of confidence increases) should be balanced against the desire to construct an interval *narrow enough* to provide useful information about the probable location of the underlying population characteristic (*e.g.*, the P = 95th percentile in the above example). A reasonable goal is to construct the shortest interval possible that still approaches the highest confidence level. In the example, a confidence level of almost 63% could be achieved by setting the 17th and 20th ordered sample values as the confidence interval endpoints. The 20th ordered value is obviously the maximum observation and cannot be changed. However, if any ranked value less than the 17th is taken as the lower endpoint, the confidence level will increase only slightly, but the overall interval will be unnecessarily widened.

An iterative process is used to construct non-parametric confidence limits. It is recommended that a *stopping rule* be used to decide when the improvement in the confidence level brought about by picking more extreme order statistics is outweighed by the loss of information from making the interval

too wide. A reasonable stopping rule might be to end the iterative computations if the confidence level changes by less than 1 or 2 percent when a new set of candidate ranks is selected.

Repeated calculation of cumulative binomial distribution probabilities $Bin(x;n,p)$ are quite tedious when performed manually. One can make use of either an extensive table of binomial probabilities or a software package that computes them. Almost all commercial statistical packages will compute binomial probabilities. For small sample sizes up to $n \leq 20$, **Table 21-11** in **Appendix D** provides achievable confidence levels for various choices of the sample order statistic endpoints such as the median and common upper percentiles.

Tied values do not affect the procedure for constructing non-parametric confidence intervals. All tied values (including any non-detects treated as ties) should be regarded as distinct measurements. Because of this, ties can be arbitrarily broken when ranking the data. For example, a list of 6 values including 3 non-detects would be ordered as [<5, <5, <5, 8, 12, 20] and given the set of ranks [1, 2, 3, 4, 5, 6]. Note that it is possible for the LCL to be set equal to the RL used for non-detects.

PROCEDURE FOR A CONFIDENCE INTERVAL AROUND THE MEDIAN

Step 1. Given a sample of size n, order the measurements from least to greatest. Denote the ordered values by $x_{(1)}, x_{(2)},\ldots, x_{(n)}$, where $x_{(i)}$ is the ith concentration value in the ordered list and numbers 1 through n represent the data ranks.

Step 2. Given $P = .50$, pick candidate interval endpoints by choosing ordered data values with ranks as close to and as symmetrical as possible around the product of $(n+1) \times 0.50$. If this last quantity is a fraction (an even-numbered sample size), the ranks immediately above and below it can be selected as candidate endpoints. If the product $(n+1) \times 0.50$ is an integer (an odd-numbered sample size), add 1 and subtract 1 to get the upper and lower candidate endpoints. Once the candidate endpoints have been selected, denote the ranks of these endpoints by L^* and U^*.

Step 3. For a two-sided confidence interval, compute the confidence level associated with the tentative endpoints L^* and U^* by taking the difference in the cumulative binomial probabilities given by the equation:

$$1-\alpha = Bin\left(U^*-1;\ n,.50\right) - Bin\left(L^*-1;\ n,.50\right) = \sum_{x=L^*}^{U^*-1}\binom{n}{x}\left(\frac{1}{2}\right)^n \qquad [21.14]$$

For a one-sided LCL, compute the confidence level associated with endpoint L^* using the equation:

$$1-\alpha = 1 - Bin\left(L^*-1;\ n,.50\right) = \sum_{x=L^*}^{n}\binom{n}{x}\left(\frac{1}{2}\right)^n \qquad [21.15]$$

For a one-sided UCL, compute the confidence level associated with endpoint U^* using the equation:

$$1-\alpha = Bin\left(U^* - 1;\ n,.50\right) = \sum_{x=0}^{U^*-1}\binom{n}{x}\left(\frac{1}{2}\right)^n \qquad [21.16]$$

To minimize the amount of direct computation needed, these equations have been used to compute selected cases over a range of sample sizes for the median in **Table 21-11** of **Appendix D**.

Step 4. If the candidate endpoint(s) do not achieve the desired confidence level, compute new candidate endpoints (L^*-1) and (U^*+1) and re-calculate the achieved confidence level. Repeat this process until the target confidence level is achieved. If one candidate endpoint already equals the data minimum or maximum, only change the rank of the other endpoint. If neither endpoint rank can be changed, set either: 1) the minimum concentration value as a one-sided LCL; 2) the maximum concentration value as a one-sided UCL; or 3) the interval spanned by the range of the sample as a two-sided confidence interval around the median. In each case, report the achieved confidence level associated with the chosen confidence limit(s).

Step 5. Compare the confidence limit(s) computed in **Step 4** versus the fixed compliance or clean-up standard. In compliance/assessment monitoring, use the LCL derived as the order statistic with rank L^*. In corrective action monitoring, use the UCL derived as the order statistic with rank U^*.

►EXAMPLE 21-5

Use the following four years of well beryllium concentrations, collected quarterly for a total of n = 16 measurements, to compute a non-parametric LCL on the median concentration with (1–α) = 99% confidence.

SAMPLE DATA		ORDERED DATA	
Date	**Beryllium (ppb)**	**Be**	**Rank**
2002, 1st Q	3.17	2.32	(1)
2002, 2nd Q	2.32	3.17	(2)
2002, 3rd Q	7.37	3.39	(3)
2002, 4th Q	4.44	3.65	(4)
2003, 1st Q	9.50	3.74	(5)
2003, 2nd Q	21.36	4.44	(6)
2003, 3rd Q	5.15	5.15	(7)
2003, 4th Q	15.70	5.58	(8)
2004, 1st Q	5.58	6.15	(9)
2004, 2nd Q	3.39	6.94	(10)
2004, 3rd Q	8.44	7.37	(11)
2004, 4th Q	10.25	8.44	(12)
2005, 1st Q	3.65	9.50	(13)
2005, 2nd Q	6.15	10.25	(14)
2005, 3rd Q	6.94	15.70	(15)
2005, 4th Q	3.74	21.36	(16)

SOLUTION

Step 1. Order the 16 measurements from least to greatest and determine the rank associated with each value (listed above in the last two columns). The smallest observation, 2.32 ppb, receives the smallest rank, while the largest value, 21.36 ppb, receives a rank of 16.

Step 2. Since a confidence interval on the median must be constructed, the desired percentile is the 50th (*i.e.*, $P = 0.50$). Therefore the quantity $(n+1) \times P = 17 \times 0.50 = 8.5$. The data ranks closest to this value are $L^* = 8$ and $U^* = 9$, so these are used as initial candidate endpoints.

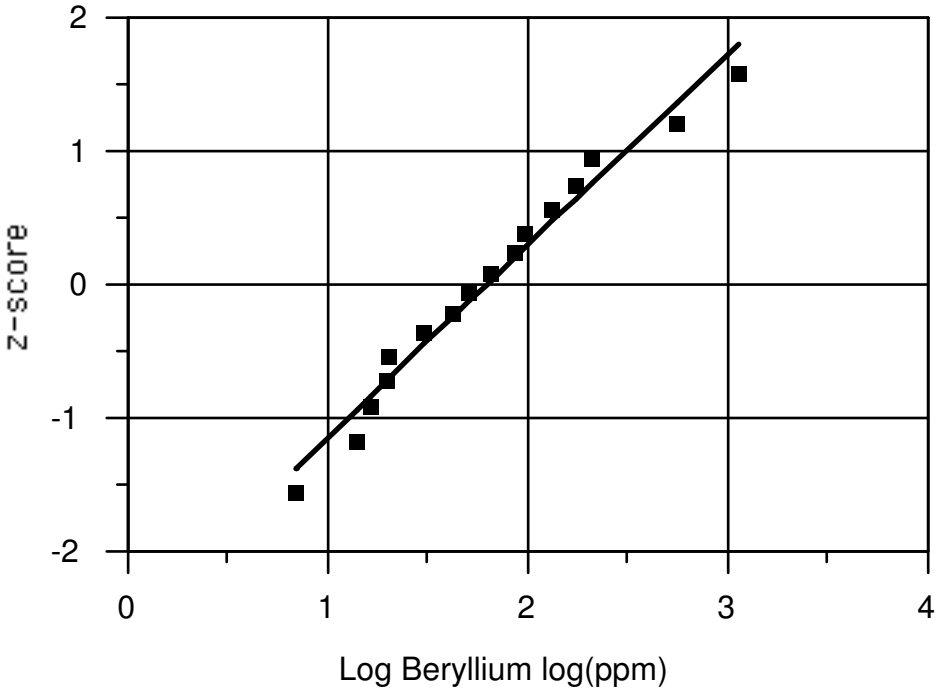
Step 3. Using the cumulative binomial distribution, and recognizing that only a lower confidence limit is needed, use equation [21.15] to calculate the actual confidence level associated with the order statistic $x_{(8)}$:

$$1-\alpha = 1 - Bin\left(L^* - 1;\ \text{n,P}\right) = 1 - Bin\left(7;\ 16,.50\right) = 1 - \sum_{x=0}^{7}\binom{16}{x}(.50)^{16} = 0.4018$$

Since the achieved confidence level is much less than 99%, subtract 1 from L^* and recompute the confidence level. Repeat this process until the confidence level is at least 99%. Since the achieved confidence when $L^* = 4$ is equal to .9894 or approximately 99%, the LCL should be selected as $x_{(4)}$ (*i.e.*, the 4th order statistic in the data set, also equal to the fourth smallest measurement), which equals 3.65 ppm. With statistical confidence of 98.94%, one can assert that the true median beryllium concentration in the underlying population is no less than 3.65 ppm.

Step 4. In this example, a lognormal model could also have been fit to the sample. Indeed the probability plot in **Figure 21-2** below indicates good agreement with a lognormal fit, enabling a comparison between the non-parametric LCL with that derived from assuming a parametric model for the same data.

Figure 21-2. Probability Plot on Logged Beryllium Data



Step 5. Since the non-parametric LCL was constructed around the population median, the fairest comparison is to construct a lognormal-based confidence interval around the *median* and not the arithmetic mean. As discussed in **Section 21.1.2**, this is equivalent to constructing a confidence interval around the lognormal *geometric mean*. This can be built via a normal-based confidence interval around the mean using the log-transformed measurements and then exponentiating the interval limits. Thus, using equation [21.3] with the log-mean and log-standard deviation given by $\bar{y}$ = 1.8098 log(ppm) and s_y = 0.60202 log(ppm) respectively, one can compute the 99% LCL as:

$$LCL_{.99} = \exp\left(\bar{y} - t_{.99,n-1}\frac{s_y}{\sqrt{n}} \right) = \exp\left[1.8098 - (2.602)(.60202)/\sqrt{16} \right] = 4.13 \text{ ppm}$$

The non-parametric LCL around the median is slightly lower than the limit computed by assuming an underlying lognormal distribution. Given the apparent lognormal fit, the parametric LCL is probably a slightly better estimate, but the non-parametric method performs well nonetheless.

The chief virtue of using a parametric confidence interval is the ability to generate estimates at any confidence level even with small sample sizes. On the other hand, if the data are lognormally-distributed, a confidence interval on the arithmetic mean may be preferred for comparisons to a fixed standard, depending on the type of standard. The advantage of a non-parametric interval around the median is its greater flexibility to define confidence intervals on non-normal data sets. ◄

PROCEDURE FOR A CONFIDENCE INTERVAL AROUND A PERCENTILE

Step 1. Given a sample of size n, order the measurements from least to greatest. Denote the ordered values by $x_{(1)}$, $x_{(2)}$, …, $x_{(n)}$, where $x_{(i)}$ is the ith concentration value in the ordered list and numbers 1 through n represent the data ranks.

Step 2. Given the desired percentile P, pick candidate interval endpoints by choosing ordered data values with ranks as close to and as symmetrical as possible around the product $(n+1) \times P$, where n is the sample size and P is expressed as a fraction. If this last quantity is a fraction (even-numbered sample size), the ranks immediately above and below it can be selected as candidate endpoints (unless the fraction is larger than n, in which case the maximum rank n would be chosen as the upper endpoint). If the product $(n+1) \times P$ is an integer (odd-numbered sample size), add 1 and subtract 1 to get the upper and lower candidate endpoints. Once the candidate endpoints have been selected, denote these by L^* and U^*.

Step 3. For a two-sided confidence interval, compute the confidence level associated with the tentative endpoints L^* and U^* by taking the difference in the cumulative binomial probabilities given by the equation:

$$1-\alpha = Bin\left(U^{*}-1;\ n,P\right) - Bin\left(L^{*}-1;\ n,P\right) = \sum_{x=L^{*}}^{U^{*}-1}\binom{n}{x}P^{x}\left(1-P\right)^{n-x} \qquad [21.17]$$

For a one-sided LCL, compute the confidence level associated with the endpoint L^* using the equation:

$$1-\alpha = 1 - Bin\left(L^{*}-1;\ n,P\right) = \sum_{x=L^{*}}^{n}\binom{n}{x}P^{x}\left(1-P\right)^{n-x} \qquad [21.18]$$

For a one-sided UCL, compute the confidence level associated with the endpoint U^* using the equation:

$$1-\alpha = Bin\left(U^{*}-1;\ n,P\right) = \sum_{x=0}^{U^{*}-1}\binom{n}{x}P^{x}\left(1-P\right)^{n-x} \qquad [21.19]$$

To minimize the amount of direct computation, these equations have been used to compute selected cases over a range of sample sizes and for certain percentiles in **Table 21-11** of **Appendix D**.

Step 4. If the candidate endpoint(s) do not achieve the desired or target confidence level, compute new candidate endpoints, (L^*–1) and (U^*+1), and re-calculate the achieved confidence level. Repeat this process until the target confidence level is achieved. If one candidate endpoint already equals the data minimum or maximum, only change the rank of the other endpoint. If neither endpoint rank can be changed, set either: 1) the minimum concentration value as a one-sided LCL; 2) the maximum concentration value as a one-sided UCL; or 3) the interval spanned by the range of the sample data as a two-sided confidence interval around the *P*th percentile. In each case, report the achieved confidence level associated with the chosen confidence limit(s).

Step 5. Compare the confidence limit(s) computed in **Step 4** versus the fixed compliance or clean-up standard. In compliance/assessment monitoring, use the LCL derived as the order statistic with rank L^*. In corrective action monitoring, use the UCL derived as the order statistic with rank U^*.

►EXAMPLE 21-6

Use the following 12 measurements of nitrate at a well used for drinking water to determine with 95% confidence whether or not the infant-based, acute risk standard of 10 mg/L has been violated. Assume that the risk standard represents an upper 95th percentile limit on nitrate concentrations.

Sampling Date	Nitrate (mg/L)	Rank
7/28/99	<5.0	(1)
9/3/99	12.3	(11)
11/24/99	<5.0	(2)
5/3/00	<5.0	(3)
7/14/00	8.1	(7)
10/31/00	<5.0	(4)
12/14/00	11.0	(10)
3/27/01	35.1	(12)
6/13/01	<5.0	(5)
9/16/01	<5.0	(6)
11/26/01	9.3	(8)
3/2/02	10.3	(9)

SOLUTION

Step 1. Half of the sample concentrations are non-detects, making a test of normality extremely difficult. One could attempt to fit these data via the *Kaplan-Meier* or *Robust ROS* adjustments (see **Chapter 15**), but here a non-parametric confidence interval around the upper 95th percentile will be constructed.

Step 2. Order the data values from least to greatest and assign ranks as in the last column of the table above. Note that the apparent ties among the non-detects have been arbitrarily broken in order to give a unique rank to each measurement.

Step 3. Using **Table 21-11** in **Appendix D** for n = 12, there is approximately 88% confidence associated with using L^* = 11 as the rank of the lower confidence bound and approximately 98% confidence associated with using L^* = 10. Since the target confidence level is 95%, it can only be achieved by using a rank of 10 or less. Thus the non-parametric LCL needs to be set to the 10th smallest observation or $x_{(10)}$. Scanning the list of nitrate measurements, the LCL = 11.0 ppm.

Step 4. Since the order statistic $x_{(10)}$ achieves a confidence level of 98%, one can conclude that the true upper 95th percentile nitrate concentration is no smaller than 11.0 ppm with 98% confidence. Even by this more stringent confidence level, the acute risk standard for nitrate is violated and there is statistically significant evidence that at least 1 of every 20 nitrate measurements from the well will exceed 10 mg/L.

Step 5. If the well was being remediated under corrective action monitorign, the fixed standard would be compared against a one-way UCL around the upper 95th percentile. In that case, for n = 12, **Table 21-11** of **Appendix D** indicates that the maximum observed value of 35.1 mg/L taken as the UCL achieves a confidence level of only 46%. 95% confidence could not be achieved unless at least 59 sample measurements were available and the UCL was set to the maximum of those values. The remedial action would be considered successful only if *all* 59 measurements were below the fixed standard of 10 mg/L. ◄

21.3 CONFIDENCE INTERVALS AROUND TREND LINES

It was assumed that the underlying population is stable, (*i.e.*, characteristics like the mean, median, or upper percentiles are stationary over the period of sampling) for the confidence intervals so far presented in this chapter. In some cases, however, the concentration data will exhibit a trend. Examples might include successful remediation efforts that serve to gradually drive down a well's concentration levels, or interception of an intensifying plume of contaminated groundwater.[3]

The problem with ignoring a discernible trend when building a confidence interval is that the interval will incorporate not only the natural variability in the underlying population, but also additional variation induced by the trend itself. The net result is a confidence interval that can be much wider than expected for a given confidence level and sample size (n). A wider confidence interval makes it more difficult to demonstrate an exceedance or return to compliance versus a fixed standard in compliance/assessment or corrective action monitoring. The confidence interval will have less statistical power to identify compliance violations, or to judge the success of remedial efforts.

When a linear trend is present, it is possible to construct an appropriate confidence interval built around the estimated trend. A continuous series of confidence intervals is estimated at each point along the trend, termed a *simultaneous confidence band*. An upper or lower confidence band will tend to follow the estimated trend line whether the trend is increasing or decreasing. It is computed once the trend line has been estimated.

Construction of a confidence interval around a trend line presumes that a trend actually exists. The algorithms presented in this section *assume* that a trend is readily discernible on a time series plot of the measurements and that it is essentially linear. Otherwise, the results may be less than credible.

21.3.1 PARAMETRIC CONFIDENCE BAND AROUND LINEAR REGRESSION

BACKGROUND AND PURPOSE

A standard method for estimating a linear trend is *linear regression*, introduced in **Chapter 17**. In this section, equations for constructing a linear regression are extended to form a *confidence band* around the trend. Although a parametric technique, there is no requirement that the concentration measurements be normal or transformable to normality. Instead, the *residual concentrations* after subtracting out the estimated trend line should be roughly normal in distribution or at least symmetric.

By way of interpretation, each point along the trend line is an estimate of the true mean concentration *at that point in time*. As the underlying population mean either increases or decreases, the confidence band similarly increases or decreases to reflect this change.

Although the equations presented below can be used to simultaneously construct a confidence interval around each point on the trend line, in practice, the user will want to compute a confidence

[3] This might occur if the well screen first intercepts the leading edge of the plume, followed by the more heavily contaminated core.

interval for a few or several of the most recent sampling events. Because the individual confidence intervals comprising the simultaneous confidence band have a joint confidence level of (1–α), no matter how many confidence intervals are constructed, the overall false positive rate associated with the entire set of tests against the fixed standard will be no greater than a pre-specified α.

REQUIREMENTS AND ASSUMPTIONS

To accurately estimate a confidence band, the sample variance should be stationary or constant as a function of time. Although the mean level may be increasing or decreasing with time, the level of variation about the mean should be essentially the same.

Once a linear regression is fitted to the data, the residuals around the trend line should be tested for normality and apparent skewness. Inferences concerning a linear regression are generally appropriate when two conditions hold: 1) the residuals from the regression are approximately normal or at least reasonably symmetric in distribution; and 2) a plot of residuals versus concentrations indicates a scatter cloud of essentially uniform *vertical thickness or width*. That is, the scatter cloud does not tend to increase in width with the level of concentration or exhibit any kind of regular pattern other than looking like a random scatter of points.

If one or both of these conditions is seriously violated, it may indicate that the basic trend is either non-linear, or the size of the variance is not independent of the mean level. If the variance is roughly proportional to mean concentrations, one possible remedy is to try a transformation of the measurements and re-estimate the linear regression. This will change the interpretation of the estimated regression from a linear trend of the form $y = a + bt$, where y and t represent concentration and time respectively, to a non-linear pattern. As an example, if the concentration data are transformed via logarithms, the regression equation will have the form $\log y = a + bt$. On the original concentration scale, the trend function will then have the form $y = \exp\left(a + bt\right)$.

When the regression data are transformed in this way, the estimated trend in the concentration domain (after back-transforming) no longer represents the original mean. The transformation induces a *bias* in the confidence intervals comprising the confidence band when converted back to the original scale as in the case of samples with no trend. If a log transformation is used, for instance, the back-transformed confidence band around the trend line represents confidence intervals around the original-scale *geometric means* and not the *arithmetic means*. If a comparison of an estimated geometric mean or similar quantity to the fixed standard makes sense, computing a trend line on the transformed data should be acceptable. However, if a confidence interval around an arithmetic mean is required, consultation with a professional statistician may be necessary.

The technique presented here produces a confidence interval around the *mean* as a function of time and not an *upper percentile*. Thus, we recommend that the use of this method be restricted to cases where the fixed standard represents a mean concentration and not an explicit upper percentile or a 'not-to-exceed' limit.

At least 8 to 10 measurements should be available when computing a confidence band around a linear regression. There must be enough data to not only estimate the trend function but also to compute the variance around the trend line. In the simplest case when no trend is present, there are (n–1) *degrees*

of freedom [*df*] in a sample of size *n* with which to estimate the population variance. With a linear trend, however, the available degrees of freedom *df* is reduced to (*n*–2). For moderate to large samples, loss of one or two degrees of freedom makes little difference. But for the smallest samples, the impact on the resulting confidence limits can be substantial.

One last assumption is that there should be few if any non-detects when computing the regression line and its associated confidence band. As a matter of common sense, a readily discernible trend in a data set (either increasing or decreasing) should be based on quantified measurements. Changes in detection and/or RLs over time can appear as a declining trend, but may actually be an artifact of improved analytical methods. Such artifacts of plotting and data reporting should generally *not* be considered real trends.

PROCEDURE

Step 1. Construct a time series plot of the measurements. If a discernible trend is evident, compute a linear regression of concentration against sampling date (time), letting x_i denote the *i*th concentration value and t_i denote the *i*th sampling date. Estimate the linear slope with the equation:

$$\hat{b} = \sum_{i=1}^{n} (t_i - \bar{t}) \cdot x_i / (n-1) \cdot s_t^2 \qquad [21.20]$$

This estimate leads to the regression equation, given by:

$$\hat{x} = \bar{x} + \hat{b} \cdot (t - \bar{t}) \qquad [21.21]$$

where $\bar{t}$ denotes the mean sampling date, s_t^2 is the variance of the sampling dates, $\bar{x}$ is the mean concentration level, and $\hat{x}$ represents the estimated mean concentration at time *t*.

Step 2. Compute the regression residual at each sampling event with the equation:

$$r_i = x_i - \hat{x}_i \qquad [21.22]$$

Check the set of residuals for lack of normality and significant skewness using the techniques in **Chapter 10**. Also, plot the residuals against the estimated regression values ($\hat{x}_i$) to check for non-uniform vertical thickness in the scatter cloud. If the residuals are non-normal and substantially skewed and/or the scatter cloud appears to have a definite pattern (*e.g.*, funnel-shaped; 'U'-shaped; or, residuals mostly positive on one end of the graph and mostly negative on the other end, instead of randomly scattered around the horizontal line $r = 0$), repeat Steps 1 and 2 after first transforming the concentration data.

Step 3. Calculate the estimated variance around the regression line (also known as the *mean squared error* [MSE]) with the equation:

$$s_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} r_i^2 \qquad [21.23]$$

Step 4. Given confidence level (1–α) and a point in time (t_0) at which a confidence interval around the trend line is desired, compute the lower and upper confidence limits with the respective equations:

$$LCL_{1-\alpha} = \hat{x}_0 - \sqrt{2s_e^2 \cdot F_{1-2\alpha,\,2,\,n-2} \cdot \left[\frac{1}{n} + \frac{(t_0 - \bar{t})^2}{(n-1)\cdot s_t^2}\right]} \quad [21.24]$$

$$UCL_{1-\alpha} = \hat{x}_0 + \sqrt{2s_e^2 \cdot F_{1-2\alpha,\,2,\,n-2} \cdot \left[\frac{1}{n} + \frac{(t_0 - \bar{t})^2}{(n-1)\cdot s_t^2}\right]} \quad [21.25]$$

where $\hat{x}_0$ is the estimated mean concentration at time t_0 from the regression using equation [21.21], and $F_{1-2\alpha,\,2,\,n-2}$ is the upper (1–2α)th percentage point from an *F*-distribution with 2 and (*n*–2) degrees of freedom. Values for *F* can be found in **Table 17-1** of **Appendix D**.

Step 5. Depending on whether the regulated unit is in compliance/assessment or corrective action monitoring, compare the appropriate confidence limit against the GWPS. Multiple confidence limits can be computed at a single compliance point well without increasing the significance level (α) of the comparison. It is possible to estimate at what point in time (if ever) the confidence limit first lies completely to one side of the fixed comparison standard, without risking an unacceptable false positive rate increase for that well.

►EXAMPLE 21-7

Trichloroethylene [TCE] concentrations are being monitored at a site undergoing remediation. If the GWPS for TCE has been set at 20 ppb, test the following 10 measurements collected at a compliance point well over the last two and a half years to determine if the clean-up goal has been reached at the α = 0.05 level of significance.

Month Sampled	TCE Concentration (ppb)	Regression Estimates	Residuals
2	54.2	51.735	2.465
4	44.3	48.530	-4.230
8	45.4	42.119	3.281
11	38.3	37.311	0.989
13	27.1	34.106	-7.006
16	30.2	29.298	0.902
20	28.3	22.888	5.412
23	17.6	18.080	-0.480
26	14.7	13.272	1.428
30	4.1	6.861	-2.761

SOLUTION

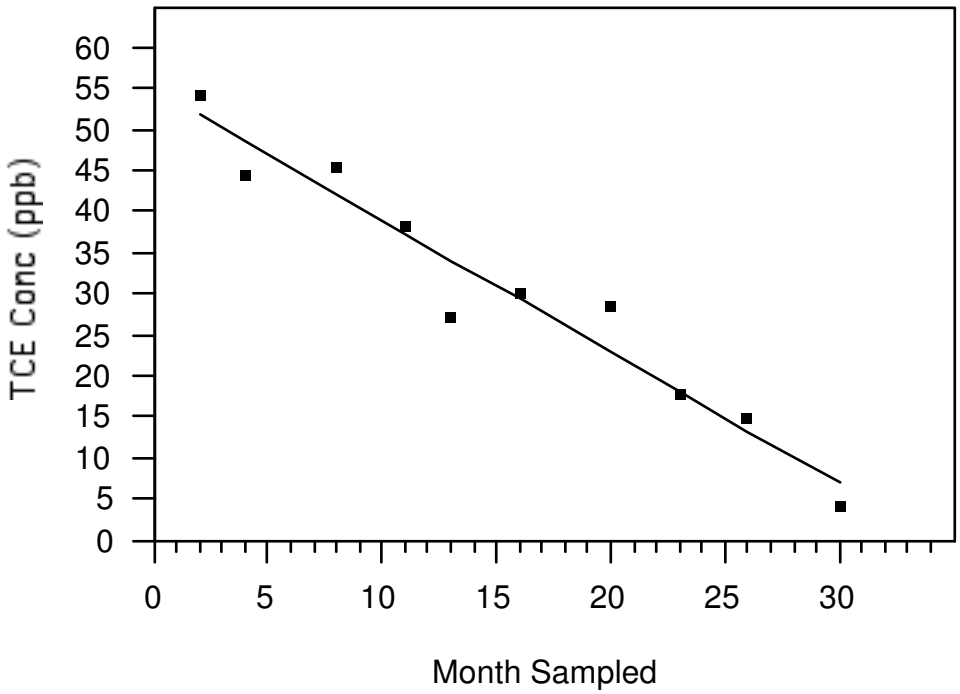
Step 1. Construct a time series plot of the TCE measurements as in the graph below (see **Figure 21-3**). A general downward, linear trend is evident. Then compute the estimated regression line

using equations [21.20] and [21.21], first determining that the mean time value is $\bar{t} = 15.3$, the variance of time values is $s_t^2 = 88.2333$, and the mean TCE measurement is $\bar{x} = 30.42$ ppb:

$$\hat{b} = \left[(2-15.3)\cdot 54.2 + (4-15.3)\cdot 44.3 + \ldots + (30-15.3)\cdot 4.1\right]/(9\times 88.2333) = -1.603\ ppp/month$$

$$\hat{y} = 30.42 - 1.603\cdot(t-15.3)$$

Figure 21-3. Time Series Plot and Regression Line of TCE Measurements



Step 2. Compute the regression residuals using equation [21.22] (listed in the table above). Note that the residuals are found by first computing the regression line estimate for each sampled month (*i.e.*, t = 2, 4, 8, *etc.*) and then subtracting these estimates from the actual TCE concentrations. A probability plot of the regression residuals appears reasonably linear (**Figure 21-4**) and the Shapiro-Wilk statistic computed from these data yields $SW = 0.962$, well above the $\alpha = 0.05$ critical point for $n = 10$ of $sw_{.05,10} = 0.842$. Thus, normality of the residuals cannot be rejected.

In addition, a plot of the residuals versus the regression line estimates (**Figure 21-5**) exhibits no unusual pattern, merely random variation about the residual mean of zero. Therefore, proceed to compute a confidence interval around the trend line.

Figure 21-4. Probability Plot of TCE Residuals

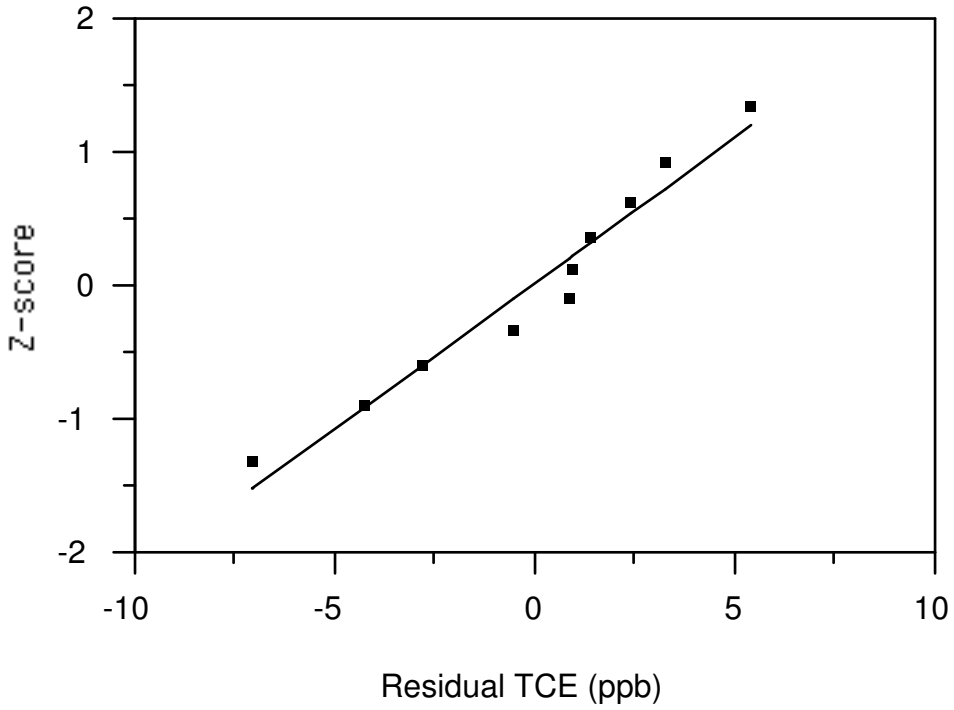
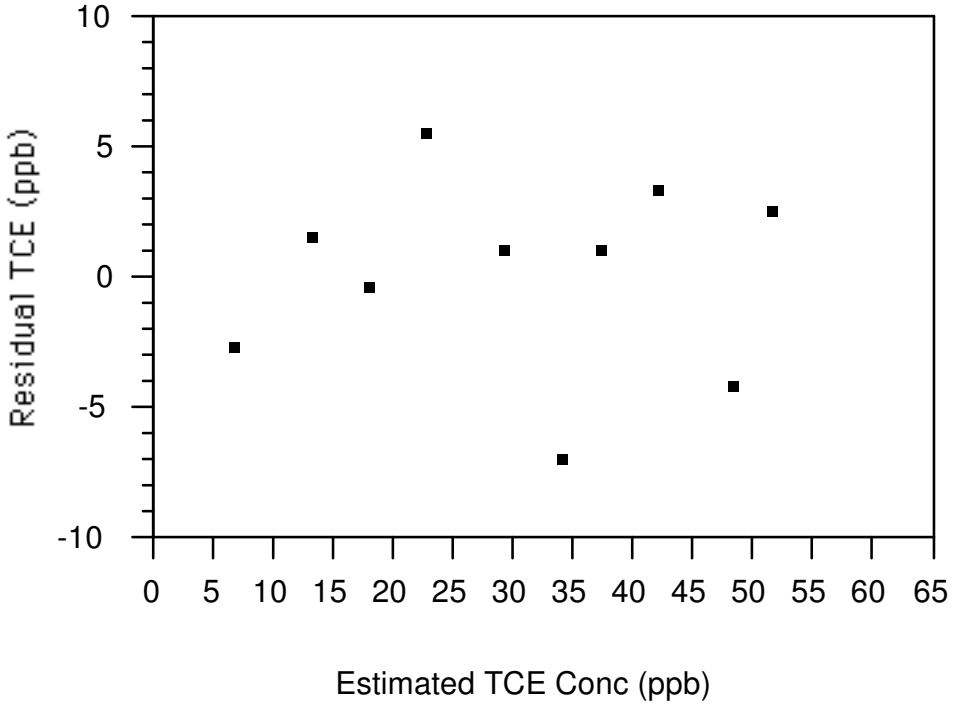


Figure 21-5. Scatterplot of TCE Residuals vs. Regression Line Estimates



Step 3. Compute the variance around the estimated trend line using equation [21.23]:

$$s_e^2 = \frac{1}{8} \cdot \left[(2.465)^2 + (-4.230)^2 + \ldots + (-2.761)^2 \right] = 15.60$$

Step 4. Since the comparison to the GWPS of 20 ppb is to be made at the $\alpha = 0.05$ significance level, the confidence limit is $(1-\alpha) = 95\%$ confidence. Since the remediation effort aims to demonstrate that the true mean TCE level has dropped below 20 ppb, a one-way UCL needs to be determined using equation [21.25]. A logical point along the trend to examine is the last sampling event at $t_0 = 30$. Using the estimated regression value at $t_0 = 30$, and the fact that $F_{.90,2,8} = 3.1131$, the UCL on the mean TCE concentration at this point becomes:

$$UCL_{95} = 6.861 + \sqrt{2 \times 15.60 \times 3.1131 \times \left[\frac{1}{10} + \frac{(30-15.3)^2}{9 \times 88.2333} \right]} = 12.87 \; ppb$$

Since this upper limit is less than the GWPS for TCE, conclude that the remediation goal has been achieved by $t_0 = 30$. In fact, other times can also be tested using the same equation. At the next to last sampling event ($t_0 = 26$), the UCL is:

$$UCL_{95} = 13.272 + \sqrt{2 \times 15.60 \times 3.1131 \times \left[\frac{1}{10} + \frac{(26-15.3)^2}{9 \times 88.2333} \right]} = 18.14 \; ppb$$

which also meets the remediation target at the $\alpha = 0.05$ level of significance.

Step 5. If the linear trend is ignored, a one-way UCL of the mean might have been used. The overall TCE sample mean $\bar{x} = 30.42$, the TCE standard deviation s = 15.508, and the upper 95th percentage point of the t-distribution with 9 degrees of freedom is $t_{.95,9}$ = 1.8331. Using equation [21.2] with the same data yields the following:

$$UCL_{.95} = 30.42 + (1.8331)(15.508) / \sqrt{10} = 39.41 \text{ ppb}$$

Had the linear trend been ignored when computing the UCL, the remediation target would not have been achieved. The downward trend induces the largest part of the variation observed over the two and a half years of sampling and needs to be taken into account. ◄

21.3.2 NON-PARAMETRIC CONFIDENCE BAND AROUND THEIL-SEN LINE

BACKGROUND AND PURPOSE

The Theil-Sen trend line is introduced in **Section 17.3.3** as a non-parametric alternative to linear regression. Whether due to the presence of non-detects or trend residuals that cannot be normalized, the

Theil-Sen method can usually construct a trend estimate without some of the assumptions needed by linear regression.

The Theil-Sen trend line is non-parametric because it combines the median pairwise slope (**Section 17.3.3**) with the median concentration value and the median sample date to construct the trend. Because of this construction, the Theil-Sen line estimates the change in *median* concentration over time and not the *mean* as in linear regression.

There are no simple formulas to construct a confidence band around the Theil-Sen line. However, a more computationally-intensive technique — bootstrapping — can be employed instead. The conceptual algorithm is fairly simple. First consider the set of n pairs of measurements used to construct the Theil-Sen trend. Each pair consists of a sample date (t_i) and the concentration value measured on that date (x_i) as a statistical sample. Next, repeatedly draw samples of size n *with replacement* from the original sample of pairs. These artificially constructed samples are known as *bootstrap samples*. At least 500 to 2,000 bootstrap samples are generated in order to improve the accuracy of the final confidence band. Note that a bootstrap sample is not precisely the same as the original because pairs are sampled with replacement. This means that a given pair might show up multiple times in any particular bootstrap sample.

For each bootstrap sample, use the Theil-Sen algorithm to construct an associated trend line (**Section 17.3.3**). Each of these trend lines is known as a *bootstrap replicate*. Finally, determine the distribution of the bootstrap replicates and select certain percentiles of this distribution to form lower and upper confidence limits. These limits can be constructed to represent a non-parametric simultaneous confidence band around the Theil-Sen trend line with $(1-\alpha)$ confidence.

REQUIREMENTS AND ASSUMPTIONS

The key requirements for constructing a confidence band around a Theil-Sen trend are the same as for the Theil-Sen procedure itself (**Section 17.3.3**). As a non-parametric procedure, the trend residuals do not have to be normal or have equal variance across the data range. But the residuals are assumed to be statistically independent. Approximate checks of this assumption can be made using the techniques of **Chapter 14**, after removing the estimated Theil-Sen trend and as long as there aren't too many non-detects. It is also important to have at least 8-10 observations from which to construct the bootstrap samples.

Non-detects can be accommodated by the Theil-Sen method as long as the detection frequency is at least 50%, and the censored values occur in the lower part of the observed concentration range. Then the median concentration value and the median pairwise slope used to compute the Theil-Sen trend will be based on clearly quantified values.

Since there are no simple mathematical equations which can construct the Theil-Sen confidence band, a computer software program is essential for performing the calculations. Perhaps the best current solution is to use the open-source, free-of-charge, statistical computing package **R** (www.r-project.org). A template program (or script) written in **R** to compute a Theil-Sen confidence band is listed in **Appendix C**. This script can be adapted to any site-specific data set and used as many times as necessary, once the **R** computing environment has been installed.

PROCEDURE

Step 1. Given the original sample of n measurements, form a sample of n pairs (t_i, x_i), where each pair consists of a sample date (t_i) and the concentration measurement from that date (x_i).

Step 2. Form B bootstrap samples by repeatedly sampling n pairs at random with replacement from the original sample of pairs in Step 1. Typically, set $B \geq 500$.

Step 3. For each bootstrap sample, construct a Theil-Sen trend line using the algorithm in **Section 17.3.3**. Denote each of these B trend lines as a bootstrap replicate.

Step 4. Determine a series of equally spaced time points (t_j) along the range of sampling dates represented in the original sample, $j = 1$ to m. At each time point, use the Theil-Sen trend line associated with each bootstrap replicate to compute an estimated concentration ($\hat{x}_j^B$). There will be B such estimates at each of the m equally-spaced time points when this step is complete.

Step 5. Given a confidence level (1–α) to construct a two-sided confidence band, determine the lower (α/2)th and the upper (1–α/2)th percentiles, denoted $\hat{x}_j^{[\alpha/2]}$ and $\hat{x}_j^{[1-\alpha/2]}$ from the distribution of estimated concentrations at each time point (t_j). The collection of these lower and upper percentiles along the range of sampling dates (t_j, $j = 1$ to m) forms the bootstrapped confidence band. To construct a lower confidence band, follow the same strategy. But determine the lower αth percentile $\hat{x}_j^{[\alpha]}$ from the distribution of estimated concentrations at each time point (t_j). For an upper confidence band, compute the upper (1–α)th percentile, $\hat{x}_j^{[1-\alpha]}$ at each time point (t_j).

Step 6. Depending on whether the regulated unit is in compliance/assessment or corrective action monitoring, compare the appropriate confidence band against the GWPS. Estimate at what point in time (if ever) the confidence band first sits completely to one side of the fixed comparison standard.

►EXAMPLE 21-8

In **Example 17-7**, a Theil-Sen trend line was estimated for the following sodium measurements. Note that the sample dates are recorded as the year of collection (2-digit format), plus a fractional part indicating when during the year the sample was collected. Construct a two-sided 95% confidence band around the trend line.

Sample Date (yr)	Sodium Conc. (ppm)
89.6	56
90.1	53
90.8	51
91.1	55
92.1	52
93.1	60
94.1	62
95.6	59
96.1	61
96.3	63

SOLUTION

Step 1. Designate the $n = 10$ (sample date, concentration) pairs as the original sample for purposes of bootstrapping. Set the number of bootstrap samples to $N_B = 500$.

Step 2. Sample at random and with replacement $N_B = 500$ times from the original sample to form the bootstrap samples. Compute a bootstrap replicate Theil-Sen trend line for each bootstrap sample. This gives 500 distinct linear trend lines.

Step 3. Divide the observed range of sampling dates from 89.6 to 96.3 into $m = 101$ equally-spaced time points, t_j (note: choice of m is arbitrary, depending on how often along the time range an estimate of the confidence band is needed). At each time point, compute the Theil-Sen concentration estimate using each bootstrap replicate trend. This leads to 500 estimates of the form:

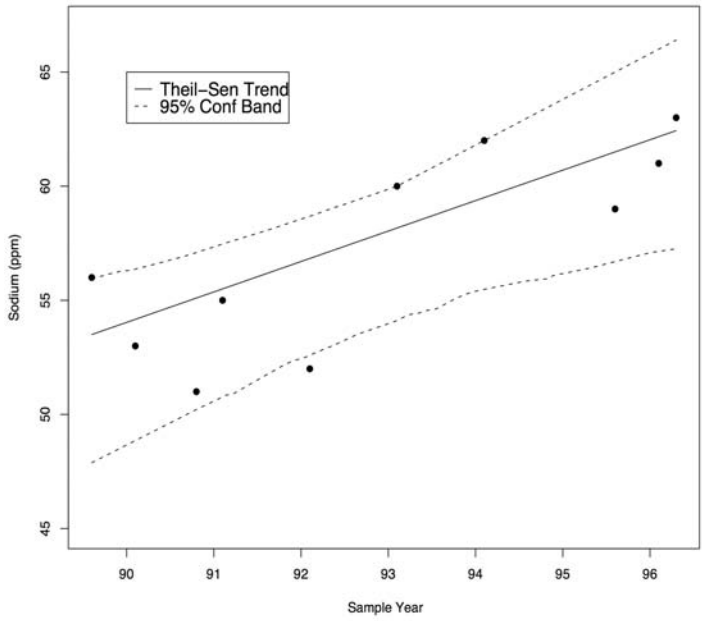
$$\hat{x}_j^B = \tilde{x}^B + Q^B \cdot \left(t_j - \tilde{t}^B\right)$$

where $\tilde{x}^B$ is the median concentration of the Bth bootstrap sample, Q^B is the Theil-Sen slope of the Bth bootstrap sample, and $\tilde{t}^B$ is the median sampling date of the Bth bootstrap sample.

Step 4. Given a two-way confidence level of 95%, compute the lower $\alpha/2 = 0.05/2 = 0.025$ and upper $(1-\alpha/2) = (1-0.05/2) = 0.975$ sample percentiles (**Chapter 3**) for the set of 500 concentration estimates associated with each time point (t_j). This entails sorting each set and finding the value closest to rank $(n+1) \times p$, where p = desired percentile. In a list of $n = 500$, find the sorted values closest to the ranks $501 \times 0.025 = 12.525$ for the lower percentile and $501 \times 0.975 = 488.475$ for the upper percentile. Collectively, the lower and upper percentiles plotted by the time points give an approximation to the 95% two-sided confidence band.

Step 5. Plot the lower and upper confidence bands as well as the original Theil-Sen trend line and the raw sodium measurements, as in **Figure 21-6**. The fact that the trend is increasing over time is confirmed by the rising confidence band. ◄

Figure 21-6. 95% Theil-Sen Confidence Band on Sodium Measurements



This page intentionally left blank

CHAPTER 22. COMPLIANCE/ASSESSMENT AND CORRECTIVE ACTION TESTS

Chapter 7 lays out general strategies for statistical testing in compliance/assessment and corrective action monitoring via the use of confidence intervals. Procedures for constructing confidence intervals are described in **Chapter 21**. This chapter discusses potential methods for developing confidence interval tests so that adequate statistical power is maintained in compliance/assessment monitoring and false positive rates are minimized in corrective action monitoring.

22.1 CONFIDENCE INTERVAL TESTS FOR MEANS

As discussed in **Chapter 7**, EPA's primary concern in compliance/assessment and corrective action monitoring is the identification and remediation of contaminated groundwater. The basic statistical hypotheses are reversed in these two phases of monitoring as described in **Chapter 21** and earlier. The lower confidence limit [LCL] is of most interest in compliance/assessment, while the upper confidence limit [UCL] is used in corrective action. Statistical power is also of greater concern to the regulatory agency in compliance/assessment— representing the probability that contamination above a fixed standard will be identified. A sufficiently conservative false positive rate during corrective action is important from a regulatory standpoint, since a false positive implies that contaminated groundwater has been falsely declared to meet a compliance standard. The reverse of these risks is generally true for a regulated entity.

To ensure that contaminated groundwater is treated in ways that are statistically sound, the two specific strategies which follow separately address compliance/assessment monitoring and formal testing in corrective action. The latter occurs after the completion of remedial activities or when potential compliance can be anticipated. Each strategy is designed to allow stakeholders on both sides of the regulator/regulated divide to understand the expected statistical performance of a given confidence interval test.

The two strategies which follow are based on the behavior of the *normal mean* confidence interval. They especially assume that the monitoring data are stationary over the period of record. Other important assumptions were discussed in **Chapter 21**. In the discussion which follows, consideration is given to data that is normal following a logarithmic transformation and the possible tests which can be applied.

22.1.1 PRE-SPECIFYING POWER IN COMPLIANCE/ASSESSMENT

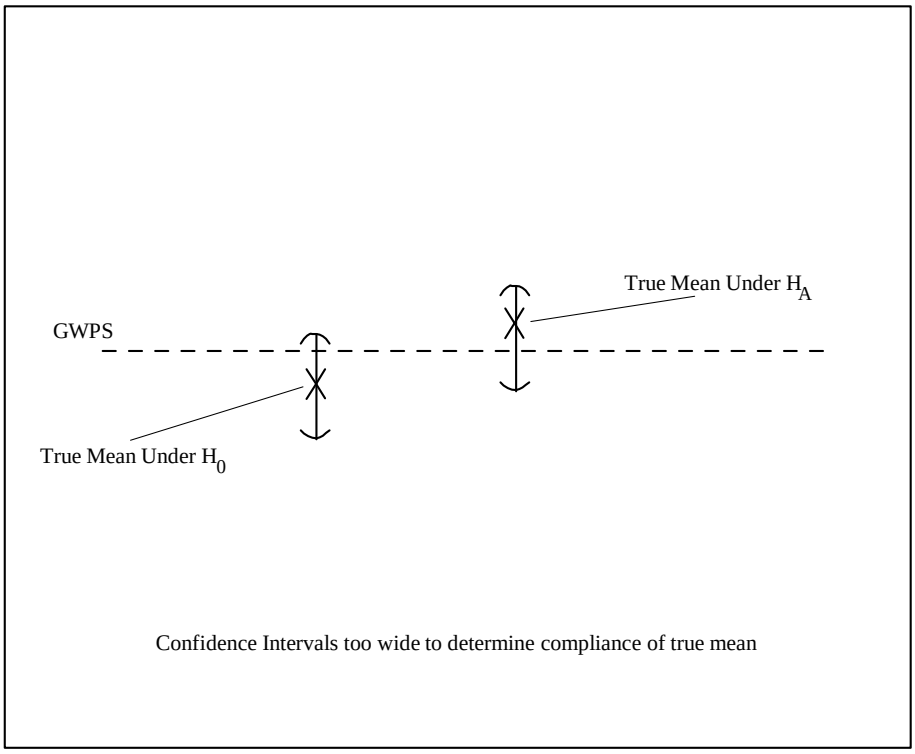
In most statistical literature including Gibbons & Coleman (2001) comparing a confidence interval against a fixed standard, a low false positive error rate (α) is chosen or recommended without respect to the power of the test. However, the power to detect increases above a fixed standard using a lower confidence limit around the mean can be negligible when contaminant variability is high and the sample size is small (**Chapter 7**). To remedy this problem, the Unified Guidance suggests an alternate strategy. That is, instead of pre-specifying the false positive rate α prior to computing confidence interval limits, a desired level of power ($1-\beta$) should be set as an initial target.

Ideally one would like to simultaneously minimize α and maximize power by also minimizing β (*i.e.*, the false negative rate). However, this is generally impossible given a fixed sample size (**Chapter 3**), since there is a trade-off between power and the false positive rate. Especially for small sample sizes, fixing a low α often leads to less than desirable power. Conversely, pre-specifying a high power necessitates a higher than typical false positive rate. Larger sample sizes are needed if both power and α are pre-specified. High variability at a fixed sample size both lowers power and/or increases the need for a larger false positive error rate.

A number of considerations are relevant when constructing mean confidence limits to achieve adequate statistical power. In most Agency risk assessment evaluations, chronic risk levels are generally *proportional* to the average concentration. Development of MCLs followed similar proportional risk methodologies. Fixed health-based limits which can serve as groundwater protection standards [GWPS] also cover an enormous concentration range when both carcinogenic and non-carcinogenic constituents are included.

Another relevant factor pertains to those situations where the true mean concentrations lie quite close to either side of a compliance standard. The difference between complying and not complying with the GWPS in terms of the true mean concentration level may be so small as to make a clear determination of compliance very difficult (**Figure 22-1**). Only sufficiently large differences relative to a standard are likely to be determined with a high level of certainty (*i.e.*, statistical power).

Figure 22-1. True Means Too Close to Standard to Clearly Identify Violation



With the wide range of GWPS in place and recognizing that risk factors are proportional or multiplicative rather than additive (***e.g.***, a 10^{-6} cancer risk), it would be appropriate to use a consistent measure of increased risk that is ***independent*** of the actual GWPS concentration level. While ultimately the decision of the regulatory authority, the Unified Guidance suggests a proportional increase (***i.e.***, a ratio) above the GWPS, which is identified at some predetermined level of statistical power to judge the appropriateness of any specific mean confidence interval test.

For compliance/assessment monitoring purposes, increases in the true concentration mean of 1.5 and 2 times a fixed standard are evaluated at a range of confidence levels. While this is not quite the same as evaluating an ***absolute*** mean increase for a given constituent, the use of a risk ratio (***R***) does in fact define a specific increase in concentration level. For example, a risk ratio of 1.5 would identify a critical increase above the 15 µg/l MCL standard for lead of 22.5 – 15 = 7.5 µg/l, while for chromium with an MCL = 100 ug/l, the absolute increase would be 50 ug/l. Each represents a 50% increase in risk relative to the GWPS.

Two approaches for assessing statistical power in compliance/assessment monitoring are provided using these critical risk ratios, based on different assumptions regarding sample variability. In the first approach, a constant population variance is assumed, equal to the standard (***i.e.***, GWPS) being tested. Under the null hypothesis that the true population mean is no greater than the GWPS, this assumption corresponds to having a coefficient of variation [***CV***] of 1 when the true mean equals the standard. Although observed sample variability is ignored, this case can be considered a relatively conservative approach.

Assuming *CV* = 1, the relationship between the risk ratio (*R*), statistical power (1–β), sample size (*n*), and the false positive rate (α) can be obtained using the following equation:

$$1-\beta = G_{T,n-1}\left(t_{1-\alpha,n-1}\middle|\Delta=\sqrt{n}\left(R-1\right)\right) \qquad [22.1]$$

where $t_{1-\alpha,n-1}$ is the (1–α)th Student's *t*-quantile with (*n*–1) degrees of freedom and $G_{T,n-1}(\bullet|\Delta)$ represents the cumulative *non-central t*-distribution with (*n*–1) degrees of freedom and non-centrality parameter Δ. By fixing a desired or target power level, equation [22.1] can be used to choose the necessary α based on the available sample size *n*. Alternatively, the equation can be used to determine the sample size (*n*) needed to allow for a pre-determined choice of α.

Numerical tabulations of equation [22.1] are found in **Tables 22-1** and **22-2** in **Appendix D**. These tables cover a practical range of *n* = 3 to 40 and α = .001 to .20, and offer combinations of the minimum false positive rate (α) and sample size (*n*) for several fixed levels of power. These can be used to construct lower confidence limits having a pre-specified level of power. It is important to note that the listed combinations are the ***smallest*** α-values resulting in the targeted power. For a fixed *n*, use of an α-value ***larger*** than that listed in the tables will provide even greater power than the target. Similarly, for given α, use of a larger sample size than that listed in the tables will also result in greater power than the target.

Minimum parameter values are presented in **Tables 22-1** and **22-2** of **Appendix D** to document how the desired power level can be achieved with as few observations and as small a false positive error rate as possible. It is also true that an assumption of *CV* = 1 should be somewhat conservative at many sites. Actual power will be higher than that listed in these tables if the coefficient of variation is smaller. Not every power level is achievable in every combination of *n* and α, so some of the entries in these two tables are left blank.

The second approach requires an estimate of the population coefficient of variation. In this case, the required (but approximate) false positive rate of the test can be directly obtained from equation [22.2], where *R* is the desired risk ratio, *n* is the sample size, $C\hat{V}$ is the estimated sample coefficient of variation, $t_{1-\beta,n-1}$ is the (1–β)th Student's *t*-quantile with (*n*–1) degrees of freedom, and $F_{T,n-1}(\bullet)$ is the cumulative (central) Student's *t*-distribution function:

$$a \cong 1 - F_{T,n-1}\left(\frac{(R-1)\cdot\sqrt{n}}{R\cdot C\hat{V}} - t_{1-\beta,n-1}\right) \qquad [22.2]$$

Equation [22.2] was evaluated for sample sizes varying from n = 4 to 12 and for *CV*s ranging from 0.1 to 3.0 at two target combinations of power and risk ratio — *R* = 1.5 at 50% power and *R* = 2 at 80% power. Results of these calculations are provided in **Table 22-3** of **Appendix D**. Similar to the critical power targets recommended by the Unified Guidance in detection monitoring (*i.e.*, 55-60% power at 3σ above background, and 80-85% power at 4σ over background), two high power targets at proportionally increasing risk ratios were also chosen for this setting.

Table 22-3 in **Appendix D** provides the approximate minimum false positive rate (α) necessary to achieve each power target in a single confidence interval test. The shaded and italicized entries in the table represent those cases where the minimum α is below the RCRA regulatory limitation of $\alpha = .01$ from §264.97(i)(2) for an individual test false positive error rate. For these situations, the user would need to set $\alpha = 0.01$, which in turn would provide even greater statistical power than the target.

For higher estimated *CV*s, many of the entries in this table exceed $\alpha = .5$ (bolded entries). These cases illustrate the difficulty of simultaneously attaining the recommended level of power while controlling the false positive rate, especially for small sample sizes and highly variable data. Setting a lower α, results in insufficient statistical power. On the other hand, setting $\alpha \geq .5$ amounts to a simple comparison of the sample mean against the fixed standard, with essentially no adjustment for sample variability or uncertainty. Similar to the first approach, a maximum false positive rate of $\alpha = .2$ is a reasonable upper bound which implies at most a 1-in-5 chance of an error.

Generally speaking, setting 80% power at a risk ratio of $R = 2$ in **Table 22-3** of **Appendix D** is more constraining (requiring higher α's) than 50% power at a risk ratio of $R = 1.5$, although the effect can be reversed for low *CV*s and sample sizes. To meet both targets simultaneously for a given *n*, the larger of the corresponding significance levels (α) should be selected. Guidance users may choose either of the two approaches described above. Other ratio and power options not covered in **Tables 22-1** through **22-3** of **Appendix D** can be handled by direct computation using either equation [22.1] or equation [22.2]. The first method makes an *a priori* assumption about the *CV*. The second method is approximate, depending on a sample *CV* estimate which might be erratic at small sample sizes and larger true population *CV*s especially if the compliance data are non-normal.

Both approaches are directly applicable to the normal mean LCL test in **Section 21.1.1**. While the *CV* can be directly estimated using $s/\bar{x}$ on the original concentration data, this statistic will underestimate the likely variability when data are lognormal. In that case, the logarithmic CV estimate in **Chapter 10**, **Section 10.4** should be used. If the data best fit a lognormal distribution, a number of considerations follow:

- It is possible to misapply the normal mean confidence interval test using the original concentration data, even when the data stem from a lognormal distribution. The mean is relatively robust with respect to departures from normality as long as the CV variability is not too great. If the predetermined false positive error α is selected based on the normal power criteria above, the resulting LCL test will be at least as powerful as the normal test. The actual false positive error rate will also differ.

- If a geometric mean test in **Section 21.1.2** is used, the LCL should be computed from the logarithmically transformed data. **Tables 22-1** to **22-3** in **Appendix D** are based on normal distribution assumptions and the error rates are very conservative with respect to the achievable power. As an example, given a data set from a lognormal distribution with n = 10, and an estimated CV = .8, an alpha value of .151 can be identified from **Table 22-3** in **Appendix D**. The actual power to detect a doubling above a GWPS at 80% confidence

would result in a power level of 94.5%. The false positive needed to detect a geometric mean doubling for this example to meet the above criteria would be α = .026.[1]

- The Land lower confidence interval test from **Section 21.1.3** can also be used. But since there are limited α-choices in the tables, the guidance option is to select a fixed limit of .01, .05, or .1. If data are truly lognormal, the power of this test is at least as great as would be predicted by equations [22.1] and [22.2]. Otherwise, professional statistical assistance may be necessary.

Since compliance data will often be pooled over time to increase the eventual sample size (**Chapter 7**), the two approaches can be combined by determining the false positive rate for a given risk ratio and power level during the first year with **Tables 22-1** and **22-2** of **Appendix D**. Tests in subsequent years might use the second power approach when a better *CV* estimate (using more data) can be derived. Overall, each approach should provide a reasonable manner of adjusting the individual test false positive rate (α) to ensure adequate power to detect real contaminant increases. As a general guide, the Unified Guidance suggest formulating power in terms of risk ratios no higher than $R = 2$. There should be at least 70-80% statistical power for detecting increases of that magnitude during compliance/assessment monitoring.

►EXAMPLE 22-1

Compliance monitoring recently began at a solid waste landfill. Measurements for vinyl chloride during detection monitoring are listed below for two compliance wells. If a value of 5 ppb vinyl chloride is used as the GWPS and a confidence interval test must have 80% power for detecting an increase in mean vinyl chloride levels of twice the GWPS, how should the confidence interval bounds be constructed and what do they indicate? Assume that compliance monitoring began with Year 2 of the sampling record and that annual groundwater evaluations are required.

Vinyl Chloride Concentrations (ppb)

Sample	GW-1	GW-2	Sample	GW-1	GW-2
Q1, Yr1	6.3	5.9	Q1, Yr3	8.4	13.8
Q2, Yr1	9.5	3.0	Q2, Yr3	6.4	5.6
Q3, Yr1	8.1	8.8	Q3, Yr3	8.9	11.0
Q4, Yr1	11.9	12.0	Q4, Yr3	4.9	9.8
Q1, Yr2	7.3	11.2	Q1, Yr4	9.6	6.3
Q2, Yr2	11.2	8.6	Q2, Yr4	9.7	10.4
Q3, Yr2	6.0	12.6	Q3, Yr4	8.7	7.5
Q4, Yr2	7.5	7.2	Q4, Yr4	8.7	9.7

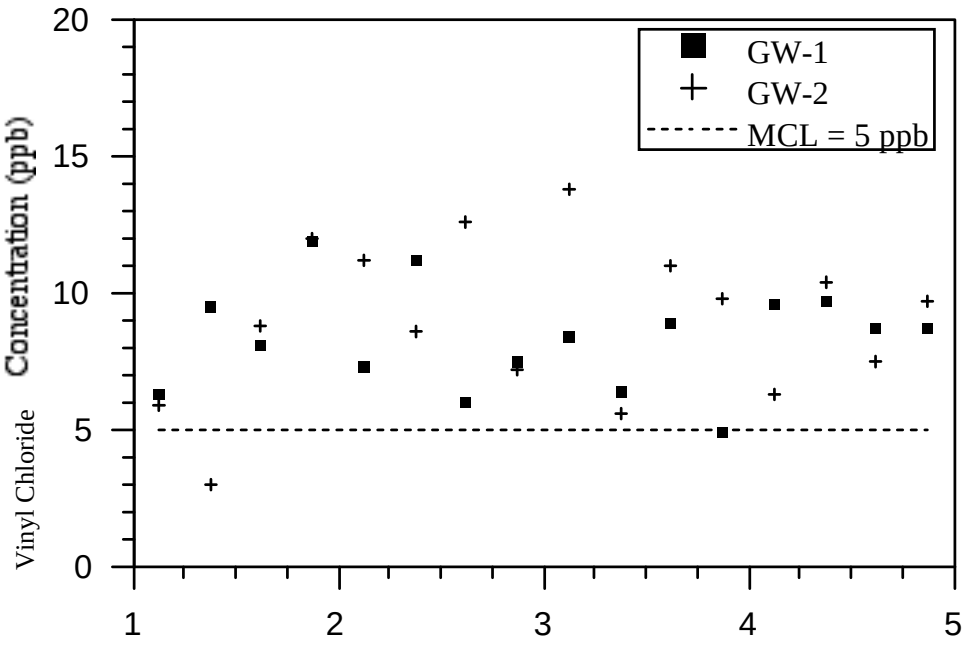
SOLUTION

Step 1. Assume for purposes of this example that the vinyl chloride data are approximately normal. In practice, this should be explicitly checked. Also evaluate potential trends in the vinyl chloride

[1] For users with access to statistical software containing the cumulative non-central t-distribution, the inverse non-central t CDF can be used to identify the appropriate false positive level. For sample size df =n -1= 9, and a non-central t parameter = $\delta = \sqrt{n} \cdot \log(R) / s_y$, the appropriate central t-value can be obtained from $F^{-1}(df, \beta, \delta)$. The confidence level of this t-value is 1 - α. For the example, df = 9, β= .2, δ = 3.115, and the central t-value is 2.23 with α = .0264.

measurements over time, as in the time series plot of **Figure 22-2**. Despite apparent fluctuations, no obvious trend is observed. So treat these data as if the population has a stable mean at least for the time frame indicated in the sampling record.

Figure 22-2. Vinyl Chloride Time Series Plot



Step 2. Given that compliance monitoring began in Year 2, use the four measurements available from each well to construct lower confidence limits. Since 80% power is desired for detecting vinyl chloride increases of two times the 5 ppb GWPS, **Table 22-2** in **Appendix D** indicates that for n = 4, a false positive rate of α = 0.163 must be used to guarantee the desired power. This corresponds to a Student's t-quantile of $t_{1-\alpha,n-1} = t_{.837,3} = 1.1714$. Then using the sample means and standard deviations of the Year 2 vinyl chloride measurements, the lower confidence limits can be computed as:

$$LCL_{GW-1} = \bar{x} - t_{1-\alpha,n-1}\frac{s}{\sqrt{n}} = 8.0 - 1.1714\left(2.2346/\sqrt{4}\right) = 6.7 \text{ ppb}$$

$$LCL_{GW-2} = \bar{x} - t_{1-\alpha,n-1}\frac{s}{\sqrt{n}} = 9.9 - 1.1714\left(2.4468/\sqrt{4}\right) = 8.5 \text{ ppb}$$

Step 3. Since both lower confidence limits exceed the GWPS, there is statistically significant evidence of an increase in vinyl chloride at these wells above the compliance limit. Such a conclusion also seems reasonable from **Figure 22-2**. However, the chance is better than 15% (*i.e.*, α =

16.3%) that the apparent exceedance is merely a statistical artifact. If power criteria are ignored and a fixed minimum rate of $\alpha = .01$ is used, the lower confidence limits would be:

$$LCL_{GW-1} = \bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}} = 8.0 - 4.541\left(2.2346/\sqrt{4}\right) = 2.9 \text{ ppb}$$

$$LCL_{GW-2} = \bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}} = 9.9 - 4.541\left(2.4468/\sqrt{4}\right) = 4.3 \text{ ppb}$$

Neither limit now exceeds the GWPS, so the vinyl chloride concentrations would be judged in compliance with this test, illustrating the lack of power in lowering the false positive rate (α).

Step 4. To increase the confidence level (*i.e.*, by lowering α) of the tests at the end of the first year of compliance monitoring (*i.e.*, Year 2 in the preceding table of vinyl chloride values) ***without*** losing statistical power, combine the measurements from Years 1 and 2, where Year 1 samples represent the final measurements from detection monitoring prior to the start of compliance monitoring. In this case, $n = 8$, and the minimum false positive rate from **Table 22-2** of **Appendix D** can be lowered to $\alpha = .046$ or approximately 4.5%. Then the re-computed lower confidence limits $LCL_{GW-1} = 7.0\ ppb$ and $LCL_{GW-2} = 6.4\ ppb$ again both exceed the GWPS, indicating significant evidence of a compliance violation.

Step 5. If the strategy presented in **Step 4** of combining measurements from detection monitoring and compliance monitoring is considered untenable, additional confirmation of the results can be made at the end of Year 3 by combining the first two years of compliance monitoring samples and ignoring the measurements from Year 1. Again with $n = 8$, the minimum false positive rate guaranteeing at least 80% power will be $\alpha = .046$. The lower confidence limits are then:

$$LCL_{GW-1} = \bar{x} - t_{1-\alpha,n-1}\frac{s}{\sqrt{n}} = 7.575 - 1.9512\left(1.9521/\sqrt{8}\right) = 6.2 \text{ ppb}$$

$$LCL_{GW-2} = \bar{x} - t_{1-\alpha,n-1}\frac{s}{\sqrt{n}} = 9.975 - 1.9512\left(2.7473/\sqrt{8}\right) = 8.1 \text{ ppb}$$

Step 6. An even lower false positive rate can be achieved after the first three years of compliance monitoring. Pooling these measurements gives $n = 12$. Then **Table 22-2** in **Appendix D** identifies a minimum false positive rate of $\alpha = .013$ or less than 1.5%. In this case, the lower confidence limits $LCL_{GW-1} = 6.8\ ppb$ and $LCL_{GW-2} = 7.6\ ppb$ again exceed the GWPS, confirming the previous vinyl chloride exceedances from either Year 2, Years 1 and 2 combined, or Years 2 and 3 combined. Furthermore, not only is the false positive rate quite low, but the power of the test still meets the pre-specified target. ◄

22.1.2 PRE-SPECIFYING FALSE POSITIVE RATES IN CORRECTIVE ACTION

As noted earlier, the primary regulatory concern in formal corrective action testing is false declarations of remedial success. If groundwater is truly contaminated above a regulatory standard yet a statistical test result indicates the concentrations are no longer so elevated, then on-going contamination has been missed and the remedial process should not be exited. Statistically, this idea translates into a desire to minimize the corrective action false positive rate (α). False positives in corrective action are precisely those decisions where the true concentration mean is falsely identified to be below the regulatory standard, when in fact it still exceeds the standard.

Constructing confidence interval tests by fixing a low target false positive rate is straightforward. All of the confidence interval tests presented in **Chapter 21** can be calibrated for choice of α. What is not straightforward is how best to incorporate statistical power in corrective action. As with any confidence interval test, selecting a low α when the sample size is small typically results in a confidence limit with low power. Power under corrective action monitoring represents the probability that the upper confidence limit [UCL] will fall below the fixed standard when in fact the true population mean is also less than the standard. Facilities undergoing remediation clearly have an interest in demonstrating the success of those clean-up efforts. They therefore may want to maximize the power of the confidence interval tests during corrective action, under the constraint that α must be kept low.

What statistical power criteria might a facility reasonably define in corrective action testing? Because of the orders of magnitude range found among various GWPS, a risk ratio approach similar to what is suggested in **Section 22.1.1;** only in this case, the target ratios (*R*) are *less than one*. While a true mean at a level of $R = 0.9$ times a given standard might be declared in compliance very infrequently, one at $R = 0.5$ times or $R = 0.25$ times the standard should meet the compliance requirements much more often. By using a consistent risk ratio across a variety of constituents, absolute decreases in the mean concentration are consistent with an assumed level of risk.

Unlike the risk ratio method detailed for compliance/assessment monitoring, where power was pre-specified but a combination of the false positive rate (α) and sample size (*n*) might be varied to meet that power level, in corrective action both power and α are likely to be pre-specified (power by the facility and α by the regulatory authority). The remaining component is how large a sample size is needed to attain the desired level of power, given a pre-specified false positive rate (α).

The normal distribution can be used to estimate sample size requirements for such risk ratios, given a specific false positive rate (α) and desired level of power (1–β). There is likely to be uncertainty, however, in the degree of sample variation, as expressed by the *CV*. Since the constituents in a contaminated aquifer may be modified by remedial actions, it can be difficult to estimate *future* variability (and the *CV*) from pre-treatment data. In some situations, a decrease in the mean over time might be paralleled by a decrease in total variation. If proportional, the *CV* would remain relatively constant. However, the *CV* could decrease or increase depending on aquifer conditions, constituent behavior, *etc*. The best that can be recommended is to develop an estimate of the expected future *CV* under conditions of aquifer stability.

As with compliance/assessment testing, future year estimates of the *CV* could be developed from the accumulated previous years' data. Sample sizes necessary to meet specific power targets (1–β) can

then be generated from the following approximate equation, where R = fractional risk ratio (less than 1.0), (1–α) is the desired confidence level, and $\tilde{CV}$ = estimated coefficient of variation:

$$n = \left[R \cdot \left(t_{1-\alpha,n-1} + t_{1-\beta,n-1}\right) \cdot \hat{CV} / (1-R)\right]^2 \qquad [22.3]$$

Since n appears on both sides of equation [22.3], it has to be solved iteratively for trial-and-error choices of n, making it difficult to calculate without a proper computing environment. **Tables 22-4** to **22-6** in **Appendix D** provide requisite sample sizes (n) based on equation [22.3] for three specific risk ratios (R = .75, .50, and .25) over a variety of inputs of α, β, and CV.

These tables can be consulted when designing a remedial program, especially when determining a sampling frequency adequate for generating the minimally needed sample size over a specific period of time. For example, to detect a drop in the true mean down to 0.75 × GWPS (*i.e.*, R = 0.75) with 80% power when CV = 0.6, **Table 22-4** in **Appendix D** indicates that a minimum of n = 16 observations are needed to have a false positive rate (α) no greater than 10%. Demonstrating such a reduction over the next two years might then require the collection of 8 measurements per year (or two per quarter) from the compliance well involved.[2]

While **Tables 22-4** to **22-6** in **Appendix D** identify the sampling requirements needed to simultaneously meet pre-specified targets for power (1–β) and the false positive rate (α), they come with some limitations. First, many of the minimum sample sizes are prohibitively large when sample variation as measured by the CV is substantial. Proving the success of any remedial program will be difficult when the compliance data exhibit significant relative variability. Less sampling is required to demonstrate a more substantial concentration drop below the compliance standard than to demonstrate a slight decrease (*e.g.*, compare the sample sizes for R = 0.75 to R = 0.25). This fact mirrors the statistical truth in both detection and compliance/assessment monitoring that highly contaminated wells are more easily identified (and require fewer observations to do so) than are only mildly contaminated wells.

Another limitation of equation [22.3] is that it assumes all n measurements are statistically independent. This assumption puts practical limits on the amount of sampling at a compliance well that can reasonably be achieved over a specific time period. Samples obtained too frequently may be autocorrelated and thus violate statistical independence. Minimum sample sizes do not apply to data exhibiting an obvious trend, and are appropriate only when the aquifer is in a relatively steady-state. Alternate methods to construct confidence bands around trends are presented in **Chapter 21**. However, equation [22.3] cannot be used to plan sample sizes in this setting. Finally, **Tables 22-4** to **22-6** in **Appendix D** are based on an assumption of normally-distributed data. Although non-normal data sets might be approximated to some degree by the range of CVs considered, more sophisticated methods might be needed to compute sample size requirements for such data. This might entail consultation with a professional statistician.

---

[2] A slightly more approximate direct calculation using the standard normal distribution instead of Student t-values will also provide the needed sample size estimate as: $n = \left[R \cdot \left(z_{1-\alpha} + z_{1-\beta}\right) \cdot \hat{CV} / (1-R)\right]^2$. The recommended sample size in the example above is rounded to n = 15 using the z-normal equation. The estimate can be improved and made more conservative by adding an additional sample.

►EXAMPLE 22-2

Suppose elevated levels of specific conductance (μmho) shown in the table below must be remediated at a hazardous waste facility. If the clean-up standard has been set at L = 1000 μmho, at what point should remediation efforts be declared a success for the two compliance well data in the table below? Assume that the risk of false positive error needs to be no greater than $\alpha = 0.05$ at either well.

Well ID	Date	Spec. Cond.	Well ID	Date	Spec. Cond.
GW-12	10-16-87	2100	GW-13	10-16-87	2200
GW-12	01-28-88	2550	GW-13	01-27-88	1463
GW-12	04-13-88	2360	GW-13	04-13-88	935
GW-12	06-15-88	2405	GW-13	07-12-88	809
GW-12	10-12-88	2560	GW-13	10-12-88	469
GW-12	12-20-88	1163	GW-13	12-19-88	465
GW-12	04-19-89	1880	GW-13	01-31-89	374
GW-12	10-12-89	1650	GW-13	04-19-89	499
GW-12	04-25-90	2410	GW-13	07-10-89	503
GW-12	07-19-90	862	GW-13	10-10-89	590
GW-12	10-23-90	1114	GW-13	01-29-90	403
GW-12	02-13-91	1346	GW-13	04-25-90	527
GW-12	06-27-91	909	GW-13	07-23-90	513
GW-12	09-10-91	888	GW-13	10-24-90	451
GW-12	12-06-91	749	GW-13	02-13-91	622
GW-12	03-18-92	515	GW-13	06-27-91	495
GW-12	06-03-92	180	GW-13	09-12-91	420
GW-12	09-16-92	526	GW-13	12-04-91	634
GW-12	12-02-92	610	GW-13	03-20-92	526
GW-12	03-24-93	570	GW-13	06-04-92	472
			GW-13	09-17-92	442
			GW-13	12-01-92	530
			GW-13	03-24-93	625

SOLUTION

Step 1. First consider the data in well GW-12. A time series plot of the most recent 20 specific conductance values is shown in **Figure 22-3**. This plot indicates a fairly linear downward trend, suggesting that a trend line should be fit to the data, along with an upper confidence bound around the trend.

Figure 22-3. Time Series Plot of Specific Conductance Measurements at GW-12

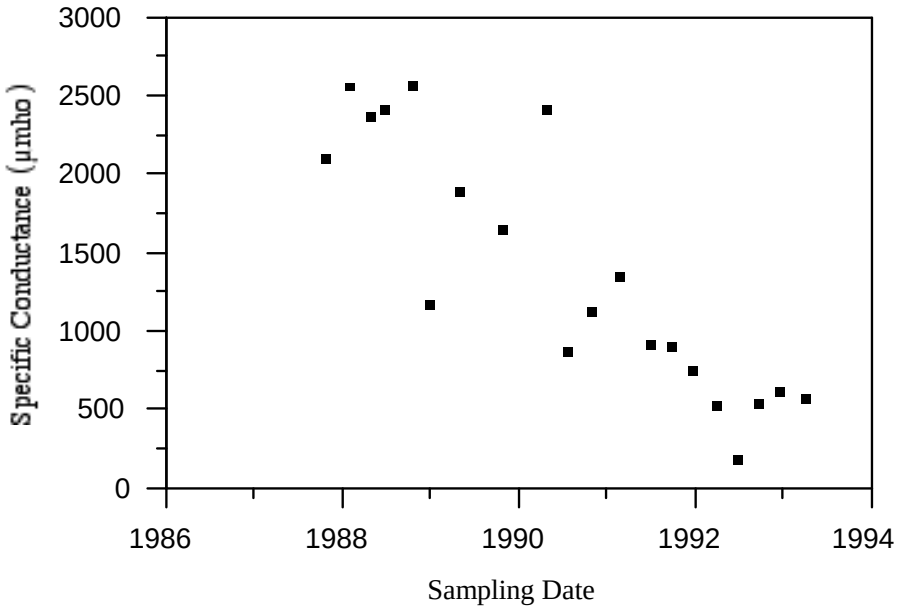
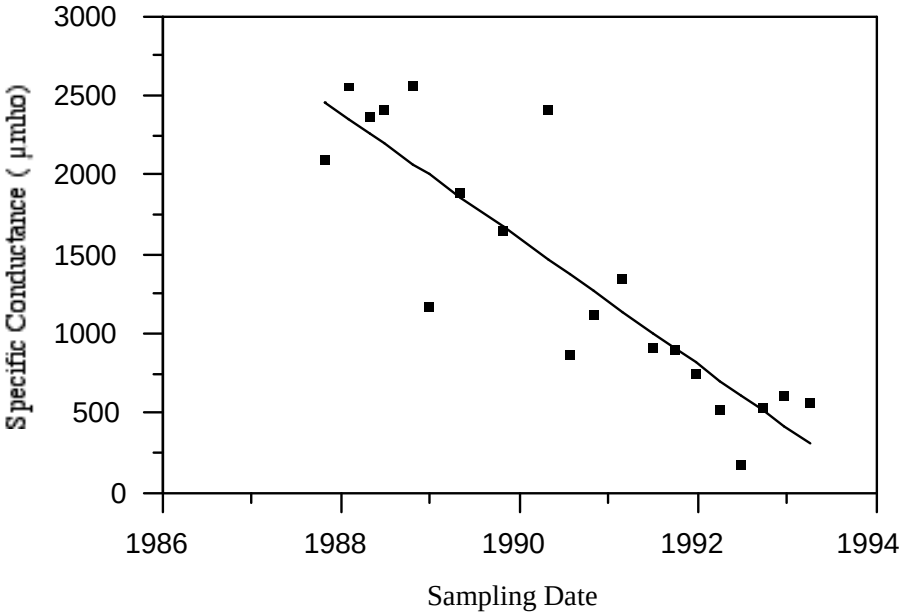


Figure 22-4. Regression of Specific Conductance *vs*. Sampling Date

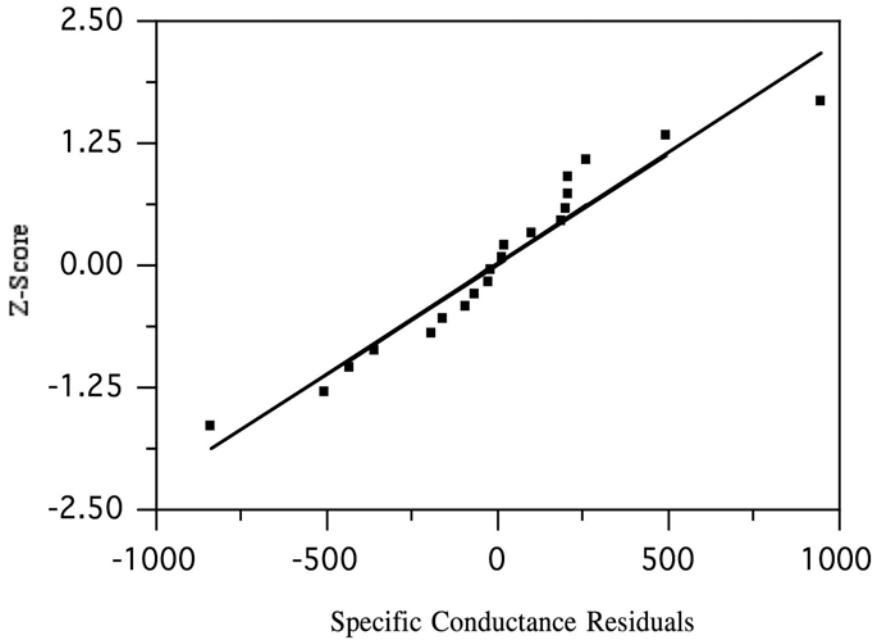


Step 2. Fit a regression line of specific conductance versus sampling date using the formulas in **Section 21.3**. The equation of estimated trend line shown on **Figure 22-4** is:

$$\hat{y} = 790360 - 396.36 \cdot t$$

Step 3. Examine the trend residuals. A probability plot of the residuals is given in **Figure 22-5**. Since this plot is reasonably linear and the Shapiro-Wilk test statistic for these residuals (SW = .9622) is much larger than the 1% critical point for n = 20 ($sw_{.01,\ 20}$ = 0.868), there is no reason to reject the assumption of normality.

Figure 22-5. Probability Plot of Specific Conductance Residuals at GW-12



Also plot the residuals against sampling date (**Figure 22-6**). As no unusual pattern is evident on this scatter plot (*e.g.*, trend, funnel-shape, *etc.*) and the variability of the residuals is reasonably constant across the range of sampling dates, the key assumptions of the linear regression appear to be satisfied.

Step 4. Since the false positive error rate must be no greater than 5%, use α = .05 when constructing an upper confidence band around the regression line. Using the formulas in **Section 21.3** at each observed sampling date, both a 95% *upper* confidence band and a 95% *lower* confidence band are computed and shown in **Figure 22-7**. Only the upper confidence band is needed to measure the success of the remedial effort. Note that the formula uses an F-confidence level of 1-α or .95 for a one-sided confidence interval. The lower 95% confidence band is shown for illustrative purposes and the confidence level between the upper and lower bands is actually 90%.

Figure 22-6. Plot of Specific Conductance Residuals *vs*. Sampling Date

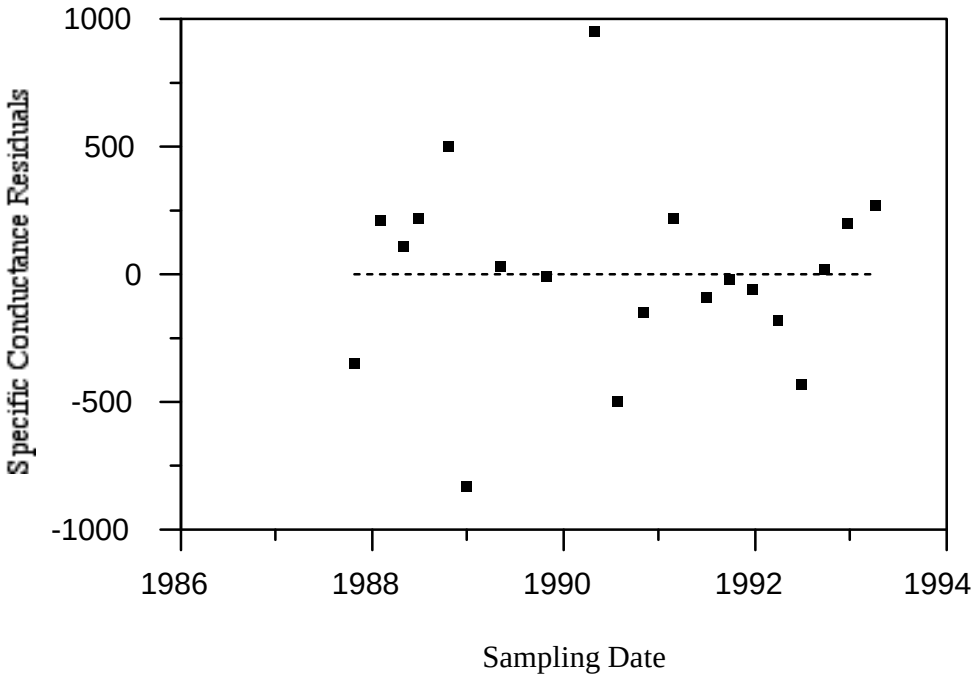
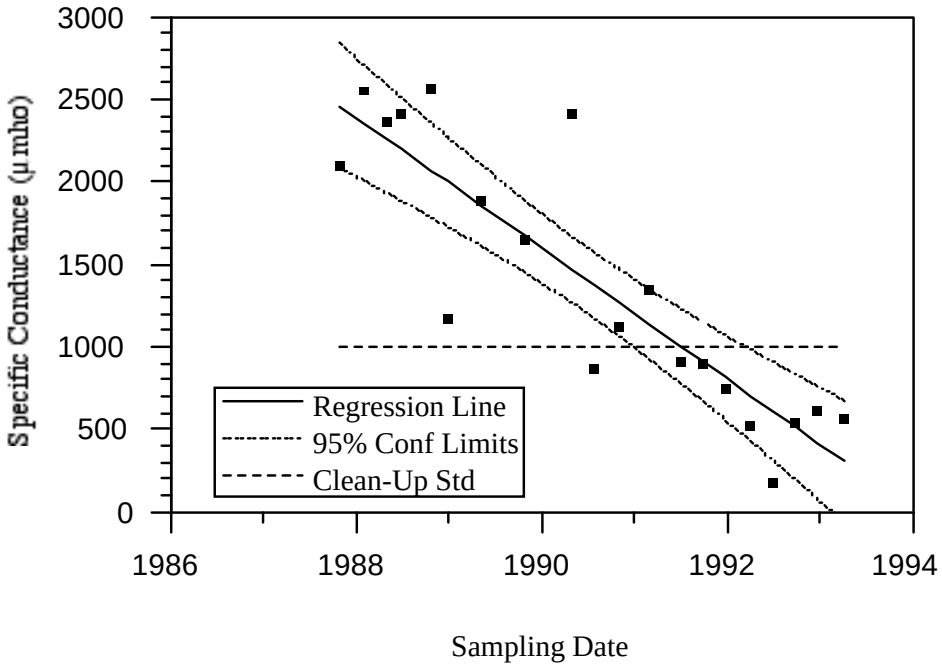


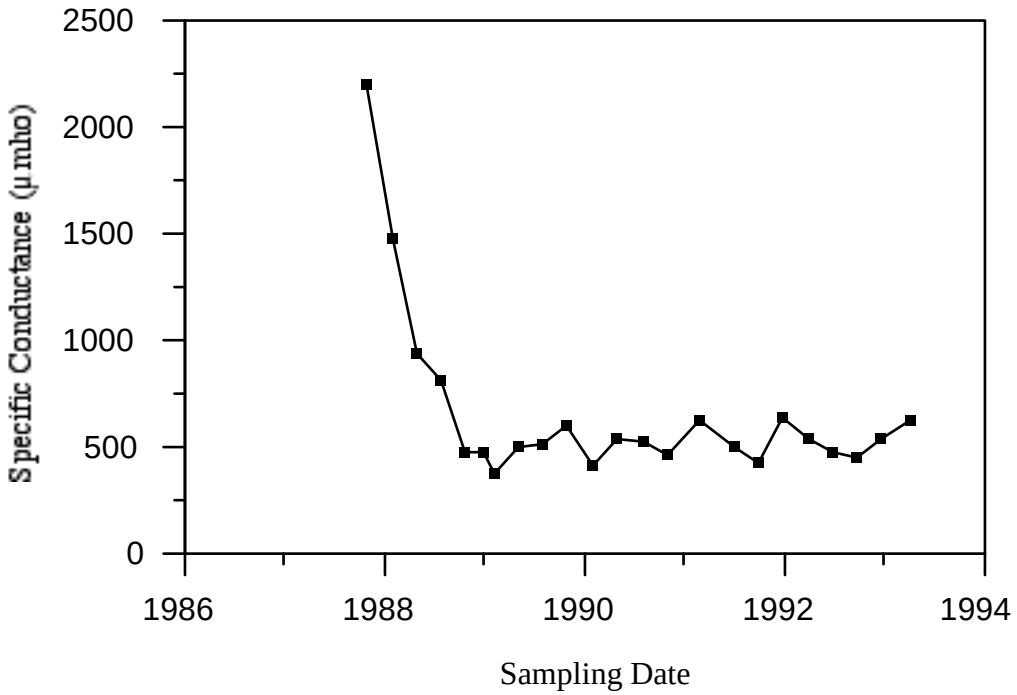
Figure 22-7. 95% Confidence Bounds Around Trend Line at GW-12



Step 5. Determine the first time point at which the remediation effort should be judged successful. In **Figure 22-7**, the upper confidence band drops below the clean-up standard of L = 1000 μmho in the second quarter of 1992, so well GW-12 could be declared in compliance at this point.

Step 6. Now consider compliance well GW-13. A time series plot of the specific conductance measurements in this case (**Figure 22-8**) shows an initially steep drop in conductance level, followed by a more or less stable mean for the rest of the sampling record. The best strategy in this situation is to remove the four earliest measurements and then compute an upper confidence limit on the remaining values.

Figure 22-8. Time Series Plot of Specific Conductance Measurements at GW-13



Step 7. Before computing an upper confidence limit, test normality of the data. If the entire sampling record is included, the Shapiro-Wilk test statistic is only .5804, substantially below the 1% critical point with n = 23 of $sw_{.01,23}$ = 0.881, indicating a non-normal pattern. Certainly, a transformation of the data could be attempted. But simply removing the first four values (representing the steep drop in conductance levels) gives a Shapiro-Wilk statistic equal to .9536, passing the normality test easily. Further confirmation is found by comparing the probability plots in **Figures 22-9** and **22-10**. In the first plot, all the data from GW-13 are included, while in the second the first four values have been removed.

Step 8. Another instructive comparison is to compute the upper confidence limits on the same data with and without the first four values. Consider the initial 8 conductance measurements, representing the first two years of quarterly data under corrective action. If all 8 values are used to compute the upper 95% confidence bound (taking 95% so that $\alpha = .05$) and the formula for a confidence interval around a normal mean from **Section 21.1** is applied, the limit becomes:

$$UCL_{.95} = \bar{x} + t_{1-\alpha,n-1}\frac{s}{\sqrt{n}} = 901.75 + 1.8946 \times \frac{635.7126}{\sqrt{8}} = 1327.6\ \mu\text{mho}$$

While this limit exceeds the clean-up standard of L = 1000 µmho, the same limit excluding the first four measurements is easily below the compliance standard:

$$UCL_{.95} = 451.75 + 2.3534 \times \frac{54.0085}{\sqrt{4}} = 515.3\ \mu\text{mho}$$

Figure 22-9. Probability Plot at GW-13 Using Entire Sampling Record

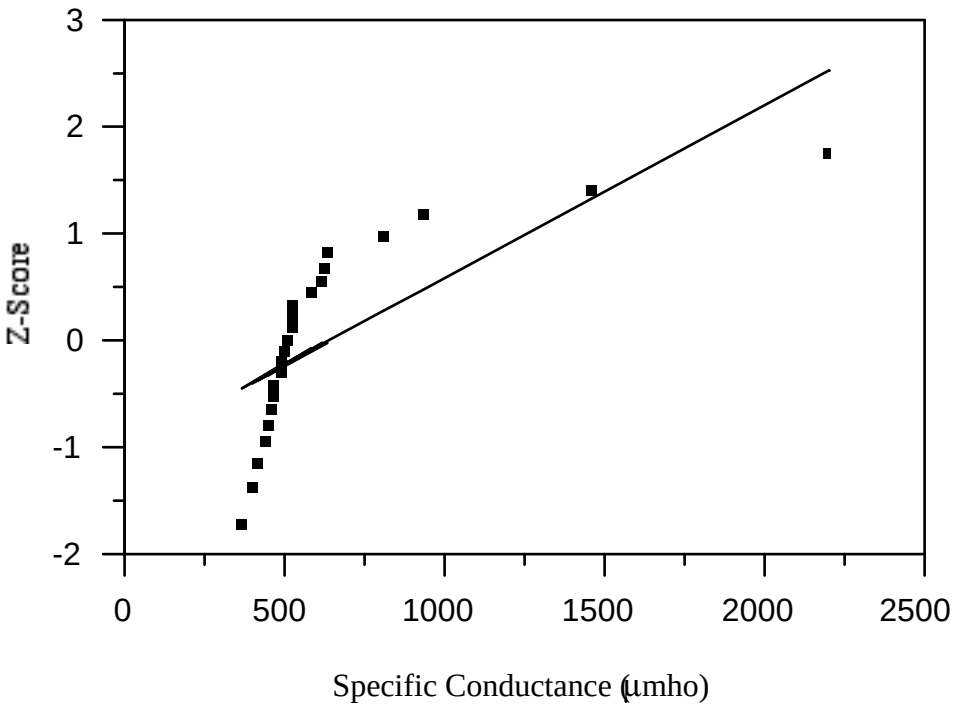
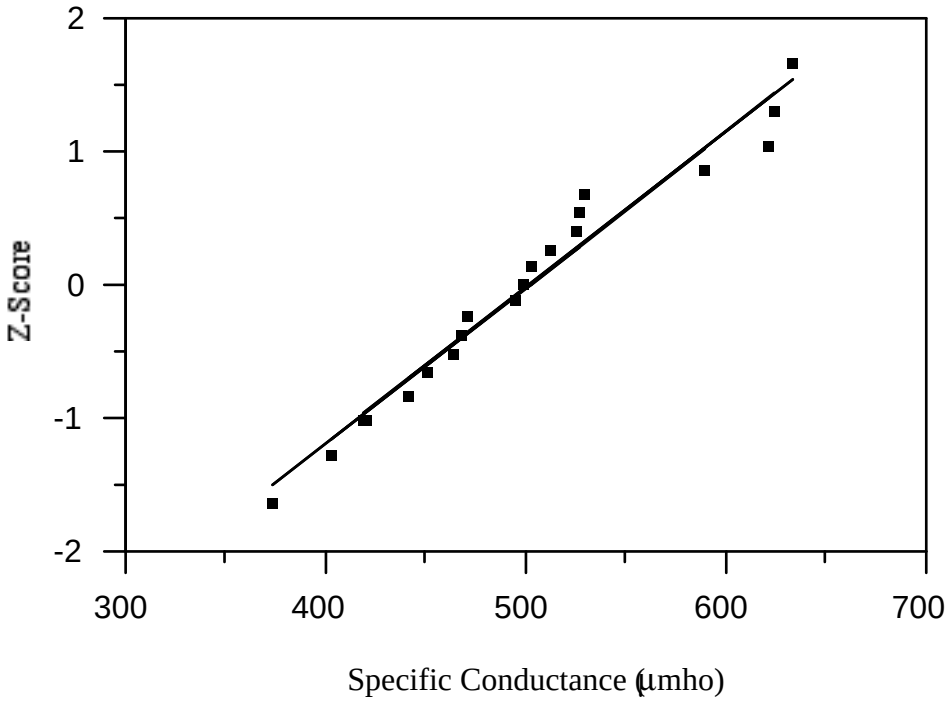


Figure 22-10. Probability Plot at GW-13 Excluding First Four Measurements



Step 9. Based on the calculation in **Step 8**, the clean-up standard is certainly met by early 1990 at GW-13. However, it is also instructive to examine the confidence bounds on larger sets of data from the stable portion of the sampling record. For instance, if the initial 4 measurements are excluded and then the next 8 values are used, the upper 95% confidence bound is:

$$UCL_{.95} = 478.75 + 1.8946 \times \frac{68.3452}{\sqrt{8}} = 524.5\ \mu\text{mho}$$

If all of the last 19 specific conductance values are used, a similar 95% confidence bound becomes:

$$UCL_{.95} = 503.158 + 1.7341 \times \frac{74.2760}{\sqrt{19}} = 532.7\ \mu\text{mho}$$

Step 10. Both of the limits in **Step 9** easily meet the clean-up standard of L = 1000 µmho. However, the amount of data used in the latter case is more than double than that of the former, which can impact the relative statistical power of the upper confidence limit for detecting decreases below the fixed standard. Given that the specific conductance seems to level off at close to 500 µmho, or one-half the clean-up standard, and given that the CV is approximately equal to .15, **Table 22-5** in **Appendix D** (looking under CV = 0.2) indicates that at least 6 measurements are needed to have a 95% chance of detecting a drop in conductance level to half the standard. So in this example, both UCLs are sufficiently powerful for detecting such a decrease. ◄

22.2 CONFIDENCE INTERVAL TESTS FOR UPPER PERCENTILES

For fixed standards which represent an upper percentile or maximum, the proper comparison in compliance/assessment monitoring utilizes the *lower* confidence limit around an *upper* percentile tested against the GWPS. In formal corrective action testing, the appropriate comparison employs an *upper* confidence limit around an upper percentile. Parametric and non-parametric confidence intervals around percentiles are presented in **Chapter 21**.

While the basic comparison is similar to confidence intervals around a mean, two points should be noted. First, any numerical standard identified as a maximum concentration 'not to be exceeded' needs to be treated statistically as an upper percentile. The reason is that while every observed data set has a finite maximum, there is no way to estimate the confidence bounds around the maximum of a continuous population. The true 'maximum' is always positive infinity, illustrating a point of breakdown between mathematical models and physical reality. Nonetheless, confidence limits around an upper 90th to 99th percentile can be used as a close approximation to a maximum or some limit likely to only be infrequently exceeded.

Secondly, computing statistical power for an interval around an upper percentile is similar to but not quite the same as, statistical power for an interval around the mean. Statistical power for a compliance/assessment test of the upper 90th percentile is derived by considering whether more than 10% of all the population measurements exceed the GWPS. If so, the 90th percentile must also exceed the standard. In corrective action testing, the equivalent question is whether *less* than 10% of the measurements exceed the GWPS. In that case, the true 90th percentile must also be less than the standard.

Statistically, each observation is set equal to 0 or 1 depending on whether the measured concentration is less than or greater than the fixed standard. Then the percentage of measurements exceeding the GWPS is given by the *average* of the set of zeros and ones. In other words, the problem is similar to estimating an arithmetic *mean*.

The similarity ends, however, when it comes to setting power targets. For mean-based evaluations, power at true mean concentration levels is equivalent to a fixed multiple or fraction of the GWPS (*e.g.*, 1.5 or 2 times the standard; 0.25 or 0.5 times the standard). But for upper percentile power, the alternative hypothesis is defined in terms of the *actual percentage of measurements* either *exceeding the standard* in compliance/assessment monitoring (*e.g.*, 20% or 30% instead of the null hypothesis value of 10%) or *exceeding the clean-up level* in corrective action monitoring (*e.g.*, 2% or 5% instead of 10%). In both hypothesis frameworks, the actual fraction of measurements above the standard can be denoted by p. Furthermore, the power formulas rely on a normal approximation to the binomial distribution. If p is the probability that an individual observation exceeds the GWPS, and p_0 is the percentage of values exceeding the GWPS when the $(1–p_0)$th upper percentile concentration equals the standard, the quantity:

$$Z = \left(p - p_0\right) \Big/ \sqrt{p_0\left(1 - p_0\right)/n} \qquad [22.4]$$

has an approximately standard normal distribution under either the compliance/assessment null hypothesis $p \le p_0$ or the corrective action null hypothesis $p \ge p_0$.

Under the compliance/assessment alternative hypothesis (H_A), the true fraction exceeding the standard is greater than the null value ($p = p_1 > p_0$). With the corrective action alternative hypothesis, the true fraction is less than the null value ($p = p_1 < p_0$). p_1 can be specified as a multiple of p_0, say $p_1 = k \cdot p_0$, where k can either be greater or less than one. Then it is possible to compute the sample size (n) necessary to simultaneously achieve a pre-specified level of power (1–β) and false positive rate (α) with the equation:

$$n = \left[\frac{z_{1-\alpha}\sqrt{p_0\left(1-p_0\right)} + z_{1-\beta}\sqrt{kp_0\left(1-kp_0\right)}}{\left(k-1\right)p_0} \right]^2 \qquad [22.5]$$

where z_c represents the cth percentile from a standard normal distribution.

Equation [22.5] can be used for designing and constructing confidence interval tests around upper percentiles in both compliance/assessment and corrective action monitoring. However, the interpretation and practical approach to its use differ depending on the stage of monitoring.

22.2.1 UPPER PERCENTILE TESTS IN COMPLIANCE/ASSESSMENT

In compliance/assessment monitoring, the alternative hypothesis (*i.e.*, that the well is contaminated above the compliance standard) is expressed in terms of the relative percentage of concentration values that will exceed the GWPS compared to an uncontaminated well. To illustrate, if the compliance standard represents the 95th percentile so that no more than p_0 = 5% of the individual measurements exceed this level, the percentage exceeding under the alternative hypothesis might be taken as $p_1 = 2 \times p_0$ = 10%. Then a power level would be targeted so that exceedances of the standard occurring as frequently as p_1 would be identified with a probability equal to (1–β).

If n measurements are used to construct a lower confidence bound on the upper percentile ($1–p_0$) of interest (*e.g.*, the 95th), there will be a (1–β) × 100% chance of showing that the lower confidence limit exceeds the GWPS when in fact at least $kp_0 \times 100\%$ of the measurements actually exceed the standard. Furthermore, equation [22.5] also implies that the LCL will *falsely* exceed the GWPS with probability α. That is, when the true percentage of measurements exceeding the standard is actually p_0 or less, the test will identify a compliance violation $\alpha \times 100\%$ of the time.

Because EPA's primary concern in compliance/assessment monitoring is having adequate statistical power to detect groundwater contaminated above the regulatory standard, a high power level (1–β) should first be pre-specified. Then α can be varied in equation [22.5] until the resulting minimum sample size (n) matches the available sample or a feasible sample size for future sampling is found. In other words, power should always be kept high (*e.g.*, at least 70-75%), even at the expense of the false positive rate (α). However, there may be sites where a feasible sample size can be calculated such that *both* β and α are minimized.

Values of n for various choices of power level (1–β), Type I error rate (α), and upper percentile (1–p_0) are tabulated in **Table 22-7** in **Appendix D**. These can be used to maintain a specific level of power when employing a confidence interval around an upper percentile in compliance/assessment monitoring. The percentiles covered in this table include the 90th, 95th, 98th, and 99th. Levels of statistical power (1–β) provided include .50, .60, .70, .80, .90, .95, and .99, while the false positive rate (α) ranges from .20 down to .01. Specific cases not covered by **Table 22-7** in **Appendix D** can be computed directly with equation [22.5].

►EXAMPLE 22-3

Suppose a compliance limit for the pressure under which chlorine gas is stored in a moving container (for instance, a rail car) is designed to protect against acute, short-term exposures due to ruptures or leaks in the container. If the compliance limit represents an upper 90th percentile of the possible range of pressures that might be used to seal a series of such containers, how many containers should be sampled/tested to ensure that if in fact 30% or more of the container pressures exceed the limit, violation of the standard will be identified with 90% probability and exhibit only a 5% chance of false positive error?

SOLUTION

Step 1. Since the compliance limit on chlorine gas pressure represents the 90th percentile, at most 10% of the container pressures should exceed this limit under normal operations. In statistical notation, $p_0 = 0.10$ and $(1–p_0) = 0.90$. If there is a problem with the process used to seal the containers and 30% of the pressures instead exceed the limit, this amounts to considering a multiple of $k = 3$ times the nominal exceedance amount.

Step 2. Since a violation of the pressure standard by at least $3p_0$ or 30% needs to be identified with 90% probability, the target power is $(1–\beta) = 0.90$. Also, the chance of constructing a lower confidence limit on the true 90th percentile gas pressure that *falsely* identifies an exceedance of the standard must be kept to $\alpha = .05$.

Step 3. Looking in **Table 22-7** in **Appendix D** under the 90th percentile and $k = 3$, the necessary minimum sample size is $n = 30$. Thus, 30 similarly-sealed containers should be tested for gas pressure so that a confidence interval around the 90th percentile can be constructed on these 30 measurements using either the parametric or non-parametric formulas in **Chapter 21**. ◄

22.2.2 UPPER PERCENTILE TESTS IN CORRECTIVE ACTION

Equation [22.5] can also be used in formal corrective action testing. In this setting, an upper confidence limit [UCL] around an upper percentile is of interest and the false positive rate (α) needs to be minimized to ensure a low probability of falsely or prematurely declaring remedial success. In practice, α should be pre-specified to a low value. Then, different values for power (1–β) can be input into equation [22.5] until the resulting minimum sample size (n) either matches the available amount of sampling data or is feasible to collect in future sampling.

Once the minimum sample size is computed and these n measurements are used to construct a UCL on the upper percentile (1–p_0) of interest (*e.g.*, the 95th), there will be a $(1–\beta) \times 100\%$ chance that

the UCL will be less than the clean-up standard when in fact no more than $kp_0 \times 100\%$ of the measurements actually exceed the standard. For instance, if $k = 1/2$, $(1-\beta)$ will be the power of the test when in fact half as many of the measurements exceed the standard as are nominally allowed.

Equation [22.5] also implies that the UCL will *falsely* drop below the clean-up standard with probability α. That is, when the true percentage of measurements exceeding the standard is actually p_0 or greater — indicating that the clean-up standard has not been met — the test will still declare the remedial effort successful $\alpha \times 100\%$ of the time.

Values of n for various choices of power level $(1-\beta)$, Type I error rate (α), and upper percentile $(1-p_0)$ are tabulated in **Table 22-8** in **Appendix D**. This table can be used to determine or adjust the feasible power level based on a pre-specified α when employing a confidence interval around an upper percentile in corrective action. Note that the minimum sample sizes in **Table 22-8** of **Appendix D** are generally quite large, especially for small error rates (α). Because of the regulatory interest in minimizing the risk of prematurely exiting remediation, statistical comparisons in corrective action are likely to initially have fairly low power. As the clean-up process continues, enough additional data can be accumulated to adequately raise the odds of declaring the remediation a success when in fact it is.

►EXAMPLE 22-4

Suppose excessive nitrate levels must be remediated in a rural drinking water supply. If the clean-up standard for infant nitrate exposure represents an upper 95th percentile of the concentration distribution, what sample size (n) should be selected to ensure that if true nitrate levels drop below the clean-up standard, the remediation effort will be judged successful with at least 80% probability?

SOLUTION

Step 1. Examining **Table 22-8** in **Appendix D** under the 95th percentile and power = $(1-\beta) = .80$, a choice of n cannot be made until two other statistical parameters are fixed: the false positive rate (α) and the relative fraction of exceedances (p). The false positive rate governs the likelihood that the upper confidence limit on nitrate will be below the clean-up standard, even though *more* than 5% of all nitrate measurements are above the compliance standard (so that the *true* 95th percentile for nitrate still exceeds the clean-up criterion). The relative fraction of exceedances (p) sets the true percentage of individual nitrate concentrations that exceed the clean-up standard under the alternative hypothesis (H_A); that is, what fraction of nitrate values are exceedances when the clean-up standard is truly met.

Unfortunately, no matter what choices of α and p are selected in **Table 22-8** of **Appendix D**, the smallest required sample size is $n = 55$, when $\alpha = .20$ and $p = .25$. Even if it is practical and affordable to test 55 samples of groundwater for nitrate, the chance of falsely declaring the remediation effort a success will still be 20%. To cut that probability in half to $\alpha = .10$, $n = 99$ samples needs to be tested.

Step 2. To lessen the required sampling effort, consider the alternatives. Lower sample sizes are needed if the percentile of interest is less extreme, for instance if the clean-up standard represents a 90th percentile instead of the 95th. In this case, only $n = 48$ samples are needed for 80% power and a 10% false positive rate with $p = .25$. Of course, more frequent exceedances of the compliance limit are then allowed (*i.e.*, 10% versus 5% of the largest nitrate concentrations).

Another less desirable option is to raise the α level of the test. This raises the risk of falsely declaring the remediation effort to be a success. One could also lower p. At $p = .25$ for the 95th percentile, 80% power is guaranteed only when the true nitrate exceedance frequency is one-fourth the maximum allowable rate--- i.e., when the true rate of exceedances is $.25 \times 5\% = 1.25\%$. Exceedance rates greater than this will be associated with *less* than 80% power. But while lowering p and keeping other parameters constant will indeed decrease n, it also has the effect of requiring a very low actual exceedance rate before the power of the test will be sufficiently high. At $p = .10$ for the 95th percentile, for instance, the true exceedance rate then needs to be only $.10 \times 5\% = 0.5\%$ to maintain the same level of power.

The final option is to lower the desired power. Power in this setting is the probability that the UCL on nitrate will be below the clean-up standard, when the groundwater is no longer contaminated above the standard. When the true nitrate levels are sufficiently low to meet the compliance standard, demonstrating this fact will only occur with high probability (*i.e.*, high power) when the sample size is fairly large. By taking a greater chance that the status of the remediation will be declared inconclusive (*i.e.*, when the UCL still exceeds the clean-up standard even though the true nitrate levels have dropped), power could be lowered to 70% or 60% for instance, with a corresponding reduction in the required n. To illustrate, if the power is set at 60% instead of 80% for the 95th percentile and the false positive rate is set at $\alpha = .10$, the required sample size would drop from $n = 99$ to $n = 68$.

Step 3. In many groundwater contexts, the minimum sample sizes of **Table 22-8** in **Appendix D** may seem excessive. Certainly, the sampling requirements associated with upper percentile clean-up standards are substantially greater than those needed to test mean-based standards. However, remediation efforts often last several years, so it may be possible to accumulate larger amounts of data for statistical use than is possible in, say, detection or compliance monitoring. In any event, it is important to recognize how the type of standard and the statistical parameters associated with a confidence interval test impact the amount of data necessary to run the comparison. Each parameter should be assessed and interpreted in the planning stages of an analysis, so that the pros and cons of each choice can be weighed.

Step 4. Once a sample size has been selected and the data collected, either a parametric or non-parametric upper confidence limit should be constructed on the nitrate measurements and compared to the clean-up standard. ◄