# Chemistry Statistical Assessments

Section 303(d) of the Clean Water Act requires states to identify bodies of water not meeting water quality standards. Identification of these water bodies relies on limited environmental data to distinguish between two alternative conditions about the true state of nature (i.e., whether or not water bodies are meeting water quality standards to support their critical uses). In theory, these two conditions are mutually exclusive, while the practical decision process involves the potential for error due to data limitations.

## A. Statistical Considerations – Bias, Imprecision, and Decision Error

Decisions based on finite sets of sample data are subject to bias and imprecision. Consequently, these decisions are prone to error. Distinguishing between the two alternative water quality conditions reflects two mutually exclusive hypotheses:

(1) The water is meeting water quality criteria
(2) The water is not meeting water quality criteria

In order to statistically test these hypotheses, one is designated as the null hypothesis and the other as the alternative hypothesis. The null hypothesis ($\mathbf{H_o}$) is tested and can either be rejected or not rejected – which is not necessarily the same as accepting the null hypothesis – based on information provided by the sample. The potential exists for two types of decision errors:

**Type I errors** occur when the null hypothesis is true and rejected. This type of error is often referred to as a false positive. Type I error rates are customarily represented by the symbol $\alpha$.

**Type II errors** occur when the null hypothesis is false and not rejected. This type of error is often referred to as a false negative. Type II error rates are customarily represented by the symbol $\beta$.

Two other decisions are possible, which represent ideal scenarios: not rejecting the null hypothesis when it is true and rejecting it when it is false. The confidence level of the test is the probability of not rejecting the null hypothesis when it is true and equals one minus the type I error rate. The power of the test is the probability of rejecting the null hypothesis when it is false and equals one minus the type II error rate.

In water quality assessments, if the null hypothesis were that the water does not violate criteria, a type I error would result in a decision that a water body does not meet water quality criteria when, in reality, the water body does meet water quality criteria. Alternatively, a type II error with the same null hypothesis would result in a decision that we cannot reject that the water body is meeting criteria when, in reality, the water body does not meet criteria. With the null hypothesis that a water body does not violate water quality criteria, the decision matrix becomes:

| | | REALITY | |
|---|---|---|---|
| | | **Meets criteria**<br>Null hypothesis | **Does not meet criteria**<br>Alternative hypothesis |
| LISTING DECISION | **Meets criteria**<br>Cannot reject the null hypothesis | Correct decision<br>Probability = 1 - $\alpha$ = Confidence Level | Type II error<br>Probability = $\beta$ |
| | **Does not meet criteria**<br>Reject the null hypothesis | Type I error<br>Probability = $\alpha$ | Correct decision<br>Probability = 1 - $\beta$ = Power |

The two types of decision errors are interrelated. For a given sample size, when the probability of one error type decreases, the probability of the other increases. As a result, with limited data, a subjective decision must be made to minimize one type of error at the cost of the other. The only reliable way to decrease both types of error is to increase the number of sampling units. If the number of sampling units in the sample were increased until the sample contained the entire target population, then both types of error would be eliminated. In water quality assessment, sampling units can be expensive to collect and process. Consequently, simultaneous reduction of both error types to low levels often requires hundreds or thousands of sampling units per sample and is usually cost-prohibitive. Therefore, decisions often are based on a sample size that is extremely small compared to the target population size.

For any statistical test the type I error rate can be set *a priori* at any chosen level. The type II error rate varies with four factors: (1) the sample size; (2) the sample variance; (3) the effect size – the difference between the sample estimate of the population parameter and the criterion value to which the sample estimate is compared; and (4) the specified type I error rate. When each of the other three factors is held constant: decreasing the variance decreases the type II error rate; increasing the type I error rate decreases the type II error rate; increasing the effect size decreases the type II error rate; and increasing the sample size decreases the type II error rate. Thus, statistical tests with low type I error rates on small sample sizes with high variances that have population parameters close to the criterion value will have the highest type II error rates. Alternatively, statistical tests with high type I error rates on large sample sizes with low variances that have population parameters which differ greatly from the criterion value will have the lowest type II error rates.

## B. Evaluation Procedures

Title 25, Chapter 96, Section 3 of the Pennsylvania Code states that water quality criteria described in Chapter 93 of the same title "shall be achieved in all surface waters at least 99% of the time, unless otherwise specified." The underlying concept in the "at least 99% of the time" language is that there is some acceptable, albeit relatively small, frequency at which water quality criteria can be violated without harming the aquatic environment; the 10% rule (described below) also relies on this concept. The procedures described here translate this conceptual regulatory language into a practical statistical evaluation methodology for water quality data that can be used to inform use attainment decisions. It should be noted that individual criteria are expressed in a variety of ways in Chapter 93 (e.g., instantaneous and daily maxima, and minima, geometric means, 30-day averages, allowable violation

frequencies). For toxic substances, Title 25, Chapter 16 of the Pennsylvania Code sets both criteria maximum concentrations (CMC, acute criteria) and criteria continuous concentrations (CCC, chronic criteria). Criteria for toxic substances are defined in terms of magnitude, duration, and maximum frequency of occurrence. Given appropriate data, most of these criteria can be adapted for evaluation using the statistical evaluation methods outlined below.

**The 10% Rule**

Past guidance from EPA (U.S. EPA, 1997) distinguishes between conventional pollutants (e.g., alkalinity, dissolved oxygen, pH, nitrite-nitrate nitrogen) and toxic pollutants (e.g., metals and pesticides). This guidance from EPA suggests that when greater than 10% of the sampling units in a sample collected at a given station violate water quality criteria for both conventional and toxic pollutants, the water body is not meeting criteria sufficiently to support critical uses. This decision method is referred to here as the 10% rule.

The idea underlying the 10% rule is to compare the upper or lower 90$^{th}$ percentile of an estimated water quality parameter to a criterion. The associated hypothesis is that, for waters meeting criteria, the upper or lower 90$^{th}$ percentile of the parameter distribution will not violate the criterion. However, this method is based on the observed proportion of violations in the sample and does not estimate the proportion of violations in the target population. Thus, this method does not account for sampling error and does not test the underlying hypothesis in a statistically rigorous manner (Smith et al. 2001; Gibbons 2003).

Furthermore, the 10% rule is a nonparametric method and, as such, only concerns violation or non-violation of a specified threshold, not the actual values of the data. Data are treated as binary failure or success values and the magnitude of the failure or success is not taken into account; a value that exceeds the threshold by one has the same bearing on the decision as a value that exceeds the threshold by orders of magnitude (Smith et al. 2001; Gibbons 2003).

**Statistical Approaches**

Statistical intervals quantify uncertainty inherent in sample data. Just as confidence intervals can be calculated around the mean value of sample data, similar intervals can be calculated around percentiles of sample data (Hahn and Meeker 1991; Helsel and Hirsch 1993). A tolerance interval is a type of statistical interval that one can claim to contain at least a specified proportion of the population with a specified degree of confidence (Hahn and Meeker 1991). Tolerance intervals that contain a proportion (e.g., 90%) of the population can be used to calculate a limit that contains the parameters of at least that proportion of samples from the target population at a specified level of confidence. This method explicitly accounts for uncertainty associated with regulatory decisions about the true state of nature that are based on information from samples of a population (Gibbons and Coleman 2001).

Calculations of one-sided statistical tolerance intervals are applicable to many environmental measurements, including decisions as to if water bodies are meeting water quality criteria because they produce single values that can be compared to regulatory criteria. These values are referred to as tolerance limits or tolerance bounds. One-sided tolerance limits are equivalent to one-sided confidence limits on distribution percentiles (Hahn and Meeker 1991). For example, a one-sided upper 95% tolerance limit to be met by at least 99% of the population values is the same as an upper 95% confidence limit on the 99$^{th}$ percentile of the population distribution (Hahn and Meeker 1991). Here these values are referred to as confidence limits. Hahn and Meeker (1991) provide a more in depth discussion of the distinctions between these types of statistical intervals.

Three alternative procedures (two parametric and one nonparametric) to calculate confidence limits on percentiles are considered here: (1) normal confidence limits; (2) lognormal confidence limits; and (3) nonparametric confidence limits. Gibbons (2003) provides further details on these procedures.

## Parametric Confidence Limits

### Normal Confidence Limits

Computation of normal confidence limits assumes that the population parameter follows, or at least approximates, a normal distribution.

This method involves computing a one-sided confidence limit on the 10$^{th}$ or the 90$^{th}$ percentile of the target population distribution for conventional pollutants or the 95$^{th}$ percentile for toxic pollutants. The 95$^{th}$ percentile is used for toxic pollutants to provide a higher level of protection when dealing with potentially lethal compounds that pose greater threats than conventional pollutants to humans and other forms of life. The calculated percentile is compared to the water quality criterion.

The appropriate percentile for all confidence limit calculations in the present context depends on the nature of the pollutant and the associated criterion for each pollutant. Conventional pollutants with minimum threshold criteria, such as dissolved oxygen concentrations, concern the upper confidence limit (UCL) for the 10$^{th}$ percentile while conventional pollutants with maximum threshold criteria, such as nitrate-nitrite nitrogen concentrations, concern the lower confidence limit (LCL) for the 90$^{th}$ percentile. The choice of the 10$^{th}$ and 90$^{th}$ percentiles for conventional pollutants reflect the idea underlying the 10% rule that 90% of samples taken from a body of water for a particular parameter should not violate regulatory criteria. Toxic pollutants, such as lead or copper, which all have maximum threshold criteria, concern the LCL for the 95$^{th}$ percentile.

Two terms are often used to describe confidence limits. The first term identifies the percentage of the population distribution of interest. This term is often written as "the p(100)$^{th}$ percentile." For example, if an investigator is interested in the lower 90% of the distribution (p = 0.90), the 90$^{th}$ percentile (0.90 * 100

percentile) of the distribution should be compared to the criterion. The second term identifies the level of confidence in the estimation of the limit. This term is often written as "100(1 – α)% confidence," where α = the type I error rate. For example, if the type I error rate is set at 0.05, then one can be 95% confident (100 * (1 - 0.05)% confidence) that the limit estimate is correct. Hahn and Meeker (1991) provide a more detailed discussion of statistical intervals and associated terminology.

The formula used to calculate a 100(1 – α)% LCL is for the $p(100)^{th}$ percentile is:

$$LCL_{1-\alpha,p} = x_{avg} + (K_{\alpha,p})(s)$$

where, $x_{avg}$ = the sample mean of the n measurements, $K_{\alpha,p}$ = the one-sided normal confidence limit factor for 100(1 – α)% confidence at the $p(100)^{th}$ percentile, and s = the observed sample standard deviation. Table 1 presents values of K useful for computing one-sided 95% confidence LCLs for the $90^{th}$ and $95^{th}$ percentiles of the distribution (Stedinger 1983; Gibbons 2003).

The formula used to calculate a 100(1 – α)% UCL for the $p(100)^{th}$ percentile is:

$$UCL_{1-\alpha,p} = x_{avg} + (K_{1-\alpha,p})(s)$$

where, $K_{1-\alpha,p}$ = the one-sided normal confidence limit factor for 100(1 – α)% confidence at the $p(100)^{th}$ percentile. Table 2 presents values of K useful for computing one-sided 95% confidence UCLs for the $10^{th}$ percentiles of the distribution (Stedinger 1983).

Equations presented in Link (1985) are used to calculate values of K that are not included in Table 1 or Table 2. For further explanation of the K-factor, which is derived from the non-central t-distribution, see Odeh and Owen (1980), Stedinger (1983), Link (1985), and Bagui (1993).

**Data Below the Detection Limit**

If data are normally distributed and nondetects (data points below the detection limit) are present, the adjusted mean of the n samples is calculated as follows:

$$x_{avg} = (1 - (n_o/n)) * x_{avg}'$$

where $n_o$ = the number of samples below the detection limit and $x_{avg}'$ = the average of the $(n – n_o)$ detected values.

Similarly, the adjusted standard deviation is calculated as follows:

$$s = [(1 - (n_o/n)) * (s')^2 + (n_o/n) * (1 - ((n_o - 1)/(n - 1))) * (x_{avg}')^2]^{0.5}$$

where s' = the standard deviation of the $(n - n_o)$ detected measurements. The adjusted values of $x_{avg}$ and s can then be used to calculate the normal confidence limit as described above.

**Lognormal Confidence Limits**

Computation of lognormal confidence limits assumes that the data follows, or at least approximates, a lognormal distribution. In addition to Gibbons (2003), Helsel and Hirsch (1993) and Stedinger (1983) describe this method in more detail.

This method involves computing a one-sided confidence limit on the $10^{th}$ or the $90^{th}$ percentile of the target population distribution for conventional pollutants or the $95^{th}$ percentile for toxic pollutants. The natural exponentiation of the calculated percentile is compared to the water quality criterion.

The formula used to calculate a $100(1 - \alpha)\%$ LCL $p(100)^{th}$ percentile is:

$$LCL_{1-\alpha,p} = \exp[y_{avg} + (K_{\alpha, p})(s_y)]$$

where, $y_{avg}$ = the mean of the natural log transformed data and $s_y$ = the standard deviation or the log transformed sample data. The equation $y = \log_e(x)$ describes the natural log transformation of the data.

The formula used to calculate a $100(1 - \alpha)\%$ UCL for the $p(100)^{th}$ percentile is:

$$UCL_{1-\alpha,p} = \exp[y_{avg} + (K_{1-\alpha, p})(s_y)]$$

**Data Below the Detection Limit**

The statistical adjustment to account for censored values in normally distributed data described above can be applied to lognormally distributed data, replacing $x_{avg}'$ with $y_{avg}'$ and s' with $s_y'$ in the equations for $x_{avg}$ and s. The adjusted values of $y_{avg}$ and $s_y$ can then be used to calculate the lognormal confidence limits as described previously. This adjustment only applies to positive random variables and cannot be applied to values less than one, which have natural logarithms less than zero. A simple solution is to add one to each value (i.e., $\log_e(x_i + 1) \geq 0$), compute the confidence limit on a log scale and then subtract one from the antilog of the confidence limit.

**Nonparametric Confidence Limits – The Binomial Method**

This method assumes nothing about the distribution of the population parameter. In addition to Gibbons (2003), Smith et al. (2001) describe this method in more detail.

Like the 10% rule, the binomial method considers measured values of continuous variables as binary successes or failures (i.e., non-violations or violations) in a statistical binomial population. The binomial method determines the critical number of violations needed to reject the null hypothesis for a given number of samples. The probability of the sample coming from a population with the specified violation rate is computed as:

$$P = {}_nC_x\,(\theta)^x\,(1-\theta)^{n-x}$$

where, ${}_nC_x = n! / (x!(n-x)!)$, n = the number of samples, x = the number of samples that violate the criterion, $\theta$ = the probability that a given sample violates the criterion, P = the probability that a population with the given $\theta$ will have x violations in n samples.

The simplest way to calculate the probability of interest is to use a cumulative binomial probabilities table. Such tables can easily be generated using computer software functions, such as the BINOMDIST function in Microsoft Excel. A cumulative binomial distribution probability table is shown in Table 3. The probability of interest is determined from the table by locating the appropriate column using the number of samples (n) and value of $\theta$. This test concerns a 10% violation rate for conventional pollutants ($\theta = 0.10$) and a 5% violation rate for toxic pollutants ($\theta = 0.05$). For the cumulative probability as close to, but not exceeding the $100(1 - \alpha)$% confidence level, the corresponding value of x represents the maximum number of violating samples out of n samples that can occur in a dataset and still list the water as not impaired. This is equivalent to calculating as close to possible of a $100(1 - \alpha)$% LCL on the chosen percentile with data represented by binary variables.

**Strengths and Weaknesses of Each Method**

Smith et al. (2001) and Gibbons (2003) show that the 10% rule consistently achieves the highest statistical power (i.e., lowest type II error rate) of any method presented here. However, these same studies also demonstrate that the 10% rule tends to have a high type I error rate, regardless of the sample size. Gibbons (2003) recorded type I error rates for the 10% rule ranging from $\alpha = 0.27$ to 0.42. Even with as many as 50 samples, the type I error rate recorded by Gibbons (2003) for the 10% rule was 0.42. Smith et al. (2001) recorded type I error rates for the 10% rule ranging from nearly 0.20 to above 0.60. These results indicate that the 10% rule tends to reject the null hypothesis more often than the other methods presented here, regardless of reality. However, the 10% rule performs consistently better that the other methods presented here in terms of type II error rate (Smith et al., 2001; Gibbons, 2003). Thus, the strength of the 10% rule is its ability to correctly reject the null hypothesis when it is false and its weakness is its tendency to mistakenly reject the null hypothesis when it is true. It is interesting to note

that for $3 < n \leq 25$, the 10% rule requires the same number of violations to reject the null hypothesis as the binomial method with a type I error rate between 27.5% and 30.0% assuming an violation rate of 10% ($\theta = 0.10$).

Gibbons (2003) demonstrated that all three statistical approaches presented here produced type I error rates near the intended nominal level. Smith et al. (2001) showed similar results for the binomial method. Therefore, the rate at which the presented statistical procedures mistakenly reject the null hypothesis falls very near the stated $\alpha$ value. Gibbons (2003) also showed that of all three presented statistical procedures, the normal interval estimation procedure achieved the lowest type II error rate, followed by the lognormal procedure, followed by the nonparametric procedure. Gibbons (2003) reported type II error rates for the nonparametric procedure greater than 0.80 with sample sizes less than 10. Compared to the 10% rule, the presented statistical procedures are more prone to not reject the null hypothesis regardless of reality. Assuming a fairly low $\alpha$ level, the interval estimation procedures tend to correctly not reject the null hypothesis more often than the 10% rule. The confidence in this decision lies in the fact that the type I error rate (i.e., the probability of incorrectly rejecting the null hypothesis) can be set at a predetermined level. Thus, the strength of the interval estimation procedures is their ability to correctly not reject the null hypothesis when it is true and their weakness is their tendency to mistakenly not reject the null hypothesis when it is false.

| | | **REALITY** | | |
|---|---|---|---|---|
| | | **Meets criteria**<br>Null hypothesis | **Does not meet criteria**<br>Alternative hypothesis | |
| **LISTING DECISION** | **Meets criteria**<br>Cannot reject the null hypothesis | Correct decision<br>Probability = 1 - $\alpha$ = Confidence Level | Type II error<br>Probability = $\beta$ | The statistical procedures are more prone to not reject the null hypothesis regardless of reality. |
| | **Does not meet criteria**<br>Reject the null hypothesis | Type I error<br>Probability = $\alpha$ | Correct decision<br>Probability = 1 - $\beta$ = Power | The 10% rule is more prone to reject the null hypothesis regardless of reality. |

As expected, the type II error rate drops for all procedures as the difference between the true value of the population parameter and the regulatory standard – the effect size – increases (Gibbons 2003). In addition, the type II error rate of all the presented statistical procedures decreases with increasing sample size (Smith et al. 2001; Gibbons 2003). Gibbons (2003) shows that the type II error rate of the 10% rule also decreases with more samples in almost all cases. Smith et al. (2001) also show that the type II error rate of the binomial method and the 10% rule generally improve with more samples.

### C. Use Support Decision Process

For water bodies presumed to be meeting water quality criteria, the methodologies utilized by DEP test the null hypothesis that the candidate water does not violate regulatory criteria.

This approach allows explicit control of the risk of mistakenly deciding a water body does not meet criteria when it really does at the risk of mistakenly deciding a water body meets criteria when it really does not. Therefore, the type I error rate is controlled at a specified level, but the type II error rate is not known because the true value of the parameter in the target population, and therefore the effect size, is unknown.

The type I error rate used by DEP in all statistical methods is 5.0% ($\alpha = 0.05$). The choice of type I error rate is a policy decision designed to minimize the risk of mistakenly identifying water bodies that are really meeting criteria as not meeting criteria. DEP recognizes that costs are associated with both type I and type II errors (Smith et al., 2001) and that the choice of type I error rate is ultimately subjective. Based on the studies of Smith et al. (2001) and Gibbons (2003), this type I error rate is believed to be prudent for the decision process outlined below.

The proposed methodology classifies samples into one of three groups based on the number of sampling units (n):

1. Less than eight sampling units ($n < 8$)
2. Between eight and 23 sampling units ($8 \leq n \leq 23$)
3. Greater than 24 sampling units ($n \geq 24$)

Based on these groupings, each sample will proceed through the use support decision process as illustrated in Figure 1.

Group 1 ($n < 8$)

Additional samples must be collected. This decision is based on the low number of sampling units and the high probability for decision errors using any of the presented approaches.

Group 2 ($8 \leq n \leq 23$)

When deciding which interval estimation procedure is best to use for small datasets, a problem arises. The parametric normal and lognormal procedures have higher statistical power than the nonparametric alternative of the binomial method (Gibbons, 2003). However, to be properly applied, the parametric procedures require that the untransformed or natural log transformed data distributions meet assumptions of normality. Most tests of normality tend to indicate normality for small sample sizes regardless of the actual distribution. Thus, small sample sizes may pass normality tests when they are in fact not normally distributed, leading to incorrect conclusions drawn from the parametric procedures. In order to address this issue, DEP has adopted the following sequential approach.

Given the statistical strengths and weaknesses of the 10% rule and the binomial method, samples in this group are first tested using the binomial method. If the results of the binomial method support rejection of the null hypothesis, the water body is identified as not meeting water quality criteria. If the results of the binomial method cannot reject the

null hypothesis, the samples are then subjected to the 10% rule. If the results of the subsequent 10% rule evaluation support rejection of the null hypothesis, additional samples or information must be collected. If the results of the subsequent 10% rule cannot reject the null hypothesis, the water body is considered to be meeting criteria.

This sequential approach draws on the strengths of each method and reduces error by minimizing both the type I and type II error rates associated with the 10% rule and the binomial method for these small datasets.

To provide a higher level of protection for toxic pollutants the acceptable violation rate is lowered from 10% to 5%. Similarly, the 10% rule for toxic pollutants becomes the 5% rule.

Due to inherent uncertainty in small datasets and the greater risk for detrimental impacts related to toxic pollutants, for datasets in Group 2 the sequential binomial method and 10% rule are only used to reject the null hypothesis. If the binomial method and subsequent 10% rule cannot reject the null hypothesis with respect to a toxic pollutant, the water body must be evaluated further. For toxic pollutants, a decision that the water body is meeting criteria must be based on the results of a parametric confidence limit test from a dataset of 24 or more sampling units. In other words, sites with detectable toxics may be declared as not meeting criteria using the sequential approach outlined above if there are between eight and 23 samples. However, 24 or more samples are required to declare a water body is meeting criteria with respect to toxic pollutants.

Datasets with between eight and 23 sampling units having a detection frequency of ≤50% are subject to the same sequential procedure as other datasets in Group 2. Censored data below detection limits are treated as binary successes that do not violate criteria.

Group 3 ($n \geq 24$)

These samples are subjected to the parametric confidence limit tests. These tests assume that either the untransformed data or the natural log transformed data approximate a normal distribution. A number of different statistical tests exist to test distributional form; here the Anderson-Darling and the Ryan-Joiner tests are employed to test the sample data for normality and lognormality. Any generally accepted statistical test of normality might be used. The decision to use the normal or lognormal test is based on the value of the Anderson-Darling and Ryan-Joiner test results. Only datasets that have p-values $\geq 0.05$ for both tests are considered to be normally or lognormally distributed. If the results of the normality tests do not support a normal or lognormal data distribution, then the procedures for Group 2 datasets are employed. If the results of the appropriate parametric confidence limit test support rejection of the null hypothesis, the water body is considered to not be meeting criteria. If the results of the appropriate parametric confidence limit test cannot reject the null hypothesis, the samples are subjected to the 10% rule (or 5% rule for toxics). If the results of the subsequent 10% rule (or 5% rule) support rejection of the null hypothesis, the water body must be evaluated further. If the results of the subsequent 10% rule (or 5% rule) cannot reject

the null hypothesis, the water is considered to be meeting criteria. This sequential approach also draws on the strengths of each method and provides a less error prone method than using either test alone.

For datasets with >23 sampling units and with a detection frequency ≤50%, parametric tests are considered inappropriate and the procedures used for Category 2 samples are applied with censored data treated as binary successes that do not violate criteria. For datasets with >23 sampling units and with a detection frequency >50%, only uncensored data is subject to the normality tests.

Rare instances involving highly skewed datasets may result in a parametric confidence limit test producing a value that is beyond the range of the values in the dataset. In these cases, the procedures used for Category 2 samples are applied.

In an effort to ensure that samples are representative of the overall conditions of the water body from which they were collected, data in Group 2 and Group 3 must cover at least one year and must be collected quarterly, at a minimum, in order to be used in the use support decision process.

## D. Delistings

In the case of delistings, there is reason to presume water bodies are already not meeting criteria. Therefore, the null hypothesis is that the water violates criteria in such cases. Changing the null hypothesis results in a switch of the type I and type II error rates from the error rates under the previous null hypothesis that water bodies are meeting criteria. Thus, for delisting scenarios, the type I error rate represents the probability of incorrectly deciding a water body that is not meeting criteria is in fact meeting criteria and the type II error rate represents the probability of incorrectly deciding a water body that is meeting criteria is in fact not. With the null hypothesis of impairment, the decision matrix becomes:

| | | **REALITY** | |
|---|---|---|---|
| | | **Meets criteria**<br>Alternative hypothesis | **Does not meet criteria**<br>Null hypothesis |
| **DELISTING DECISION** | **Meets criteria**<br>Reject the null hypothesis | Correct decision<br>Probability = 1 - β = Power | Type I error<br>Probability = α |
| | **Does not meet criteria**<br>Cannot reject the null hypothesis | Type II error<br>Probability = β | Correct decision<br>Probability = 1 - α = Confidence Level |

There is little literature documenting statistical performance of the approaches outlined here in delisting scenarios. U.S. EPA (2002) provides some analysis of delisting scenarios for the binomial method, which adopts the assumptions made by Smith et al. (2001) (i.e., an actual population violation rate of 0.25 and the desirability of balancing type I and type II error rates). The statistical rejection region for any delisting scenario (i.e., 0 – 0.05 for toxic pollutants or 0 – 0.10 for conventional pollutants) will always be much smaller than for the listing case (i.e., 0.05 – 1.00 for toxic pollutants and 0.10 – 1.00 for conventional pollutants).

Consequently, delisting decisions will always require more samples than listing decisions (U.S. EPA, 2002).

Since statistical considerations dictate that more samples are required to delist waters, this methodology only uses the procedures for the largest group referenced above (i.e., Group 3, $n \geq 24$) for delisting decisions. The same type I error rate of 0.05 applies to the delisting scenario as the listing scenario. When compared to datasets used to make the original listing decision, datasets used to inform delisting decisions must: (1) have been collected more recently; (2) have been collected as or more frequently; and (3) contain more samples.

## E. Clarifications

The intended application of this methodology concerns ambient water quality, as opposed to the quality of treated water. For water quality sampled at a fixed point, any use attainment decisions based on that data will be applied one mile upstream and one mile downstream of the sample collection point unless there are compelling reasons to apply the decision to a different length of stream.

Although this methodology aims to make use attainment decisions as objectively as possible, DEP recognizes that some decisions will inevitably involve subjective judgments regarding representativeness of the sample dataset, seasonal factors, and other considerations.

## F. Literature Cited


Bagui, S.C. 1993. *CRC Handbook of Percentiles of Non-central T-distributions*. Chemical Rubber Company, Boca Raton.

Gibbons, R.D. 2003. A statistical approach for performing water quality impairment assessments. Journal of the American Water Resources Association 39(4): 841-849.

Helsel, D.R. and R.M. Hirsch. 1993. *Statistical Methods in Water Resources*. Elsevier, Amsterdam.

Link, C.L. 1985. An equation for one-sided tolerance limits for normal distributions. Research Paper FPL 458. Madison, Wisconsin. United States Department of Agriculture, Forest Service, Forest Products Laboratory.

Odeh, R.E. and D.B. Owen. 1980. *Tables for Normal Tolerance Limits, Sampling Plans, and Screening*. Marcel Dekker, Inc., New York.

Smith, E.P., K. Ye, C. Hughes, and L. Shabman. 2001. Statistical assessment of violations of water quality standards under Section 303(d) of the Clean Water Act. Environmental Science and Technology 35(3): 606-612.

Stedinger, J.R. 1983. Confidence intervals for design events. Journal of Hydraulic Engineering 109(1): 13-27.

U.S. EPA. 1997. Guidelines for preparation of the comprehensive state water quality assessments (305(b) reports) and electronic updates. EPA-841-B-97-002A and 002B.

U.S. EPA. 2002. Consolidated Assessment and Listing Methodology – Toward a Compendium of Best Practices. First edition.

**Table 1 – One-sided $K_\alpha$ factors for 95% lower confidence limits on the 90th and 95th percentile of the distribution.**
After Stedinger (1983), Link (1985), and Gibbons (2003)

| n | 90th Percentile | 95th Percentile |
|---|---|---|
| | α | |
| | 0.05 | |
| 4 | 0.444 | 0.745 |
| 5 | 0.519 | 0.820 |
| 6 | 0.575 | 0.877 |
| 7 | 0.619 | 0.923 |
| 8 | 0.655 | 0.961 |
| 9 | 0.686 | 0.993 |
| 10 | 0.712 | 1.020 |
| 11 | 0.734 | 1.044 |
| 12 | 0.754 | 1.065 |
| 13 | 0.772 | 1.084 |
| 14 | 0.788 | 1.101 |
| 15 | 0.802 | 1.117 |
| 16 | 0.815 | 1.131 |
| 17 | 0.828 | 1.144 |
| 18 | 0.839 | 1.156 |
| 19 | 0.849 | 1.167 |
| 20 | 0.858 | 1.177 |
| 21 | 0.867 | 1.187 |
| 22 | 0.876 | 1.196 |
| 23 | 0.884 | 1.204 |
| 24 | 0.891 | 1.212 |
| 25 | 0.898 | 1.220 |
| 26 | 0.904 | 1.227 |
| 27 | 0.911 | 1.234 |
| 28 | 0.917 | 1.240 |
| 29 | 0.922 | 1.246 |
| 30 | 0.928 | 1.252 |
| 35 | 0.951 | 1.277 |
| 40 | 0.970 | 1.298 |
| 45 | 0.986 | 1.316 |
| 50 | 1.000 | 1.331 |
| 55 | 1.012 | 1.344 |
| 60 | 1.022 | 1.356 |
| 70 | 1.040 | 1.375 |
| 80 | 1.054 | 1.391 |
| 90 | 1.066 | 1.404 |
| 100 | 1.077 | 1.415 |
| 120 | 1.093 | 1.434 |
| 240 | 1.146 | 1.492 |
| 480 | 1.184 | 1.535 |
| 1000 | 1.282 | 1.568 |

**Table 2 – One-sided $K_{1-\alpha}$ factors for 95% upper confidence limits on the 10th percentile of the distribution.**
After Stedinger (1983) and Link (1985)

| n | α |
|---|---|
| | 0.05 |
| 4 | -4.162 |
| 5 | -3.407 |
| 6 | -3.006 |
| 7 | -2.756 |
| 8 | -2.582 |
| 9 | -2.454 |
| 10 | -2.355 |
| 11 | -2.275 |
| 12 | -2.210 |
| 13 | -2.156 |
| 14 | -2.109 |
| 15 | -2.068 |
| 16 | -2.033 |
| 17 | -2.002 |
| 18 | -1.974 |
| 19 | -1.949 |
| 20 | -1.926 |
| 22 | -1.886 |
| 25 | -1.838 |
| 27 | -1.811 |
| 30 | -1.777 |
| 35 | -1.732 |
| 40 | -1.697 |
| 45 | -1.669 |
| 50 | -1.646 |
| 55 | -1.626 |
| 60 | -1.609 |
| 70 | -1.581 |
| 80 | -1.559 |
| 90 | -1.542 |
| 100 | -1.527 |

## Table 3 – Cumulative binomial probability table.

| n | x | θ | |
|---|---|---|---|
| | | 0.10 | 0.05 |
| 8 | 0 | 0.4305 | 0.6634 |
| 8 | 1 | 0.8131 | 0.9428 |
| 8 | 2 | 0.9619 | 0.9942 |
| 8 | 3 | 0.9950 | 0.9996 |
| 8 | 4 | 0.9996 | 1.0000 |
| 8 | 5 | 1.0000 | 1.0000 |
| 9 | 0 | 0.3874 | 0.6302 |
| 9 | 1 | 0.7748 | 0.9288 |
| 9 | 2 | 0.9470 | 0.9916 |
| 9 | 3 | 0.9917 | 0.9994 |
| 9 | 4 | 0.9991 | 1.0000 |
| 9 | 5 | 0.9999 | 1.0000 |
| 9 | 6 | 1.0000 | 1.0000 |
| 10 | 0 | 0.3487 | 0.5987 |
| 10 | 1 | 0.7361 | 0.9139 |
| 10 | 2 | 0.9298 | 0.9885 |
| 10 | 3 | 0.9872 | 0.9990 |
| 10 | 4 | 0.9984 | 0.9999 |
| 10 | 5 | 0.9999 | 1.0000 |
| 10 | 6 | 1.0000 | 1.0000 |
| 11 | 0 | 0.3138 | 0.5688 |
| 11 | 1 | 0.6974 | 0.8981 |
| 11 | 2 | 0.9104 | 0.9848 |
| 11 | 3 | 0.9815 | 0.9984 |
| 11 | 4 | 0.9972 | 0.9999 |
| 11 | 5 | 0.9997 | 1.0000 |
| 11 | 6 | 1.0000 | 1.0000 |
| 12 | 0 | 0.2824 | 0.5404 |
| 12 | 1 | 0.6590 | 0.8816 |
| 12 | 2 | 0.8891 | 0.9804 |
| 12 | 3 | 0.9744 | 0.9978 |
| 12 | 4 | 0.9957 | 0.9998 |
| 12 | 5 | 0.9995 | 1.0000 |
| 12 | 6 | 0.9999 | 1.0000 |
| 12 | 7 | 1.0000 | 1.0000 |
| 13 | 0 | 0.2542 | 0.5133 |
| 13 | 1 | 0.6213 | 0.8646 |
| 13 | 2 | 0.8661 | 0.9755 |
| 13 | 3 | 0.9658 | 0.9969 |
| 13 | 4 | 0.9935 | 0.9997 |
| 13 | 5 | 0.9991 | 1.0000 |
| 13 | 6 | 0.9999 | 1.0000 |
| 13 | 7 | 1.0000 | 1.0000 |
| 14 | 0 | 0.2288 | 0.4877 |
| 14 | 1 | 0.5846 | 0.8470 |
| 14 | 2 | 0.8416 | 0.9699 |
| 14 | 3 | 0.9559 | 0.9958 |
| 14 | 4 | 0.9908 | 0.9996 |
| 14 | 5 | 0.9985 | 1.0000 |
| 14 | 6 | 0.9998 | 1.0000 |
| 14 | 7 | 1.0000 | 1.0000 |
| 15 | 0 | 0.2059 | 0.4633 |
| 15 | 1 | 0.5490 | 0.8290 |
| 15 | 2 | 0.8159 | 0.9638 |
| 15 | 3 | 0.9444 | 0.9945 |
| 15 | 4 | 0.9873 | 0.9994 |
| 15 | 5 | 0.9978 | 0.9999 |
| 15 | 6 | 0.9997 | 1.0000 |
| 15 | 7 | 1.0000 | 1.0000 |
| 16 | 0 | 0.1853 | 0.4401 |
| 16 | 1 | 0.5147 | 0.8108 |
| 16 | 2 | 0.7892 | 0.9571 |
| 16 | 3 | 0.9316 | 0.9930 |
| 16 | 4 | 0.9830 | 0.9991 |
| 16 | 5 | 0.9967 | 0.9999 |
| 16 | 6 | 0.9995 | 1.0000 |
| 16 | 7 | 0.9999 | 1.0000 |
| 16 | 8 | 1.0000 | 1.0000 |
| 17 | 0 | 0.1668 | 0.4181 |
| 17 | 1 | 0.4818 | 0.7922 |
| 17 | 2 | 0.7618 | 0.9497 |
| 17 | 3 | 0.9174 | 0.9912 |
| 17 | 4 | 0.9779 | 0.9988 |
| 17 | 5 | 0.9953 | 0.9999 |
| 17 | 6 | 0.9992 | 1.0000 |
| 17 | 7 | 0.9999 | 1.0000 |
| 17 | 8 | 1.0000 | 1.0000 |
| 18 | 0 | 0.1501 | 0.3972 |
| 18 | 1 | 0.4503 | 0.7735 |
| 18 | 2 | 0.7338 | 0.9419 |
| 18 | 3 | 0.9018 | 0.9891 |
| 18 | 4 | 0.9718 | 0.9985 |
| 18 | 5 | 0.9936 | 0.9998 |
| 18 | 6 | 0.9988 | 1.0000 |
| 18 | 7 | 0.9998 | 1.0000 |
| 18 | 8 | 1.0000 | 1.0000 |
| 19 | 0 | 0.1351 | 0.3774 |
| 19 | 1 | 0.4203 | 0.7547 |
| 19 | 2 | 0.7054 | 0.9335 |
| 19 | 3 | 0.8850 | 0.9868 |
| 19 | 4 | 0.9648 | 0.9980 |
| 19 | 5 | 0.9914 | 0.9998 |
| 19 | 6 | 0.9983 | 1.0000 |
| 19 | 7 | 0.9997 | 1.0000 |
| 19 | 8 | 1.0000 | 1.0000 |
| 20 | 0 | 0.1216 | 0.3585 |
| 20 | 1 | 0.3917 | 0.7358 |
| 20 | 2 | 0.6769 | 0.9245 |
| 20 | 3 | 0.8670 | 0.9841 |
| 20 | 4 | 0.9568 | 0.9974 |
| 20 | 5 | 0.9887 | 0.9997 |
| 20 | 6 | 0.9976 | 1.0000 |
| 20 | 7 | 0.9996 | 1.0000 |
| 20 | 8 | 0.9999 | 1.0000 |
| 20 | 9 | 1.0000 | 1.0000 |
| 21 | 0 | 0.1094 | 0.3406 |
| 21 | 1 | 0.3647 | 0.7170 |
| 21 | 2 | 0.6484 | 0.9151 |
| 21 | 3 | 0.8480 | 0.9811 |
| 21 | 4 | 0.9478 | 0.9968 |
| 21 | 5 | 0.9856 | 0.9996 |
| 21 | 6 | 0.9967 | 1.0000 |
| 21 | 7 | 0.9994 | 1.0000 |
| 21 | 8 | 0.9999 | 1.0000 |
| 21 | 9 | 1.0000 | 1.0000 |
| 22 | 0 | 0.0985 | 0.3235 |
| 22 | 1 | 0.3392 | 0.6982 |
| 22 | 2 | 0.6200 | 0.9052 |
| 22 | 3 | 0.8281 | 0.9778 |
| 22 | 4 | 0.9379 | 0.9960 |
| 22 | 5 | 0.9818 | 0.9994 |
| 22 | 6 | 0.9956 | 0.9999 |
| 22 | 7 | 0.9991 | 1.0000 |
| 22 | 8 | 0.9999 | 1.0000 |
| 22 | 9 | 1.0000 | 1.0000 |
| 23 | 0 | 0.0886 | 0.3074 |
| 23 | 1 | 0.3151 | 0.6794 |
| 23 | 2 | 0.5920 | 0.8948 |
| 23 | 3 | 0.8073 | 0.9742 |
| 23 | 4 | 0.9269 | 0.9951 |
| 23 | 5 | 0.9774 | 0.9992 |
| 23 | 6 | 0.9942 | 0.9999 |
| 23 | 7 | 0.9988 | 1.0000 |
| 23 | 8 | 0.9998 | 1.0000 |
| 23 | 9 | 1.0000 | 1.0000 |

**Figure 1 – Flowchart of the Process for Listing Decisions Based on Statistical Evaluation of Chemical or Bacteriological Data**