A Statistical Approach for Performing Water Quality Impairment Assessments Under the TMDL Program

Robert D. Gibbons Ph.D.
Professor of Biostatistics
Director of the Center for Health Statistics
University of Illinois at Chicago
1601 W. Taylor St., Chicago, IL 60612, USA.

ABSTRACT

A statistical approach for making impairment determinations in the Section 303(d)listing process is developed. The method is based on the $100(1-\alpha)$ percent lower confidence limit on an upper percentile of the concentration distribution. Advantages of the method include: (1) it provides a test of the null hypothesis that a percentage of the true concentration distribution fails to meet a regulatory standard, (2) it is appropriate for a variety of different concentration distributions (*i.e.*, normal, lognormal, nonparametric), (3) it directly incorporates the magnitude of the measured concentrations in the test of the hypothesis that a percentage of the true concentration distribution exceeds the standard, and (4) it has explicit statistical power characteristics that describe the probability of detecting a true impairment conditional on the number of samples, the concentration distribution, and the magnitude of the exceedance.

KEYWORDS

TMDL, 303(d) listing process, statistical intervals, upper percentiles, environmental monitoring.

INTRODUCTION

While comprehensive guidelines for water quality assessments in our nation's water bodies are now available (USEPA, 1997), they lack statistically sound procedures for evaluating the resulting data. For example, a percentage of

exceedances of a numeric water quality criterion for a given pollutant in a particular water body is often used to classify whether or not that water body is impaired (*e.g.*, no more than 10% of samples can exceed the standard). The problem with this approach is that it is based on the observed percentage and not an estimate of the true percentage of the concentration distribution that exceeds the criterion. As such, the confidence in such a statement is directly a function of the number of samples taken, for which the guidelines are insufficient (*e.g.*, a minimum of 10 samples over a three year period). The fewer the number of samples, the greater the uncertainty in the percentage of the true concentration distribution that exceeds the regulatory standard. Furthermore, by simply evaluating the percentage of exceedances, the actual concentrations have no bearing on the decision rule. Should not a concentration that is an order of magnitude above the standard be of greater concern than a concentration that exceeds the standard by 1% of its magnitude?

Finally, many environmental monitoring applications involve testing hypotheses regarding the probability that a true concentration or true percentage of concentrations exceeds a regulatory standard, not simply an observed measurement or percentage of measurements (see USEPA Statistical Guidance Documents, 1989, 1992, 2000). For purposes of making water quality impairment determinations, the appropriate null hypothesis is that the true percentage of the concentration distribution that fails to achieve the regulatory standard is less than or equal to 10% or 25% (or whatever the impairment threshold requirement is for a particular pollutant and water body). The alternative hypothesis, which could establish the presumption of impairment, is that the true percentage is greater than the required percentage. A simple tally of the observed percentage of exceedances, based on an arbitrary number of available samples, does not test this hypothesis in any statistically rigorous way. Such an approach provides us with no information regarding the confidence with which a percentage of the true concentration distribution fails to meet a regulatory standard, nor does it provide a sound basis for making impairment determinations.

Based on this review, a more rigorous statistical approach for making impairment determinations is clearly needed. In the following sections, a general statistical methodology for that purpose is developed, illustrated and fully evaluated. The statistical approach presented:

1. provides a test of the null hypothesis that a percentage of the true concentration distribution fails to meet a regulatory standard,

2. is appropriate for a variety of different concentration distributions (*i.e.*, normal, lognormal, nonparametric),

3. directly incorporates the magnitude of the measured concentrations in the test of the hypothesis that a percentage of the true concentration distribution exceeds the standard,

4. has explicit statistical power characteristics that describe the probability of detecting a true impairment conditional on the number of samples (m), the concentration distribution, and the magnitude of the exceedance.

STATISTICAL METHODS

In the present context, we are interested in comparing the true concentration for a particular constituent(s) in a particular water body to a regulatory standard. Of course, given a finite set of m samples, we can never know the true concentration with certainty. We can however, determine an interval that will contain a particular percentile of the true concentration distribution with a given level of confidence. For example, in evaluating water body monitoring data using EPA's 305(b) Guidelines, no more than 10% of the samples obtained from the water body are allowed to exceed a regulatory standard. Statistically, this amounts to a comparison of the upper 90th percentile of the distribution to the regulatory standard. As previously noted, with a finite number of measurements, we never know the 90th percentile of the distribution with certainty. However, just as we can compute a confidence interval for the mean of a distribution, we can compute a confidence interval for an upper percentile of the distribution as well. The confidence interval allows us to incorporate our uncertainty in the true parameters of the distribution into our comparison to the regulatory standard (Gibbons and Coleman, 2001).

In evaluating water body monitoring data we can then use this confidence interval for the upper 90th percentile of the distribution to determine if a particular pollutant has exceeded the regulatory standard with a reasonable level of confidence. This determination may be made if the entire confidence

interval exceeds the regulatory standard. More conservatively, we can compute a one-sided lower bound on the true 90th percentile of the concentration distribution as a $100(1-\alpha)\%$ lower confidence limit (LCL), where for 95% confidence, $\alpha = .05$. In doing so, we are testing the null hypothesis that the true 90th percentile of the concentration distribution is less than or equal to the regulatory standard. If we reject the null hypothesis the pollutant in the water body is deemed to be at an unacceptable level.

Based on the distributional form of the data and the frequency with which the pollutant has been detected, alternate parametric and nonparametric forms of the LCL are available. In the following sections procedures for deriving normal, lognormal and nonparametric LCLs are presented. Many of these computations can be performed by hand. Alternatively, all of these computations can be computed automatically for an unlimited number of constituents and water bodies using the CARStat computer program (www.discerningsystems.com).

Normal Confidence Limits for a Percentile

To compute a normal lower confidence limit for a percentile of the distribution, we use factors similar to those that are used in computing one-sided normal tolerance limits (see Gibbons, 1994). Here we seek the $(1-\alpha)100\%$ lower bound on the $p(100)$th percentile of the distribution, which is computed as

$$LCL_{1-\alpha,p} = \bar{x} + K_{\alpha,p}s \ , \tag{1}$$

where $\bar{x}$ is the sample mean of the m measurements,

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

and s is the observed sample standard deviation,

$$s = \sqrt{\sum_{i=1}^{m} \frac{(x_i - \bar{x})^2}{m-1}}$$

and $K_{\alpha,p}$ is the one-sided normal tolerance limit factor for $(\alpha)100\%$ confidence and $p(100)\%$ coverage (Hahn and Meeker, 1991). Table 1 presents values of K

TABLE 1 - One-Sided Factors for 95% Confidence LCLs for the 75th and 90th Percentiles of the Distribution m = 4 to 1000

m	75th Percentile	90th Percentile
4	-0.155	0.444
5	-0.063	0.519
6	0.002	0.575
7	0.052	0.619
8	0.091	0.655
9	0.123	0.686
10	0.150	0.712
11	0.174	0.734
12	0.194	0.754
13	0.212	0.772
14	0.228	0.788
15	0.243	0.802
16	0.256	0.815
17	0.267	0.827
18	0.278	0.839
19	0.288	0.849
20	0.298	0.858
21	0.306	0.867
22	0.314	0.876
23	0.322	0.884
24	0.329	0.891
25	0.335	0.898
26	0.342	0.904
27	0.348	0.911
28	0.353	0.917
29	0.358	0.922
30	0.363	0.928
35	0.385	0.951
40	0.403	0.970
50	0.431	1.000
60	0.451	1.022
120	0.514	1.093
240	0.560	1.146
480	0.592	1.184
1000	0.617	1.282

useful for computing 95% confidence LCLs for the 75th and 90th percentiles of the distribution.

Handling Data Below the Detection Level

Note that if the data are normally distributed and nondetects are present, we can compute the adjusted mean of the m samples as:

$$\bar{x} = \left(1 - \frac{m_0}{m}\right) \bar{x}'$$

where $\bar{x}'$ is the average of the $m - m_0$ detected values, and m_0 is the number of samples in which the compound was not detected. The adjusted standard deviation is:

$$s = \sqrt{\left(1 - \frac{m_0}{m}\right)(s')^2 + \frac{m_0}{m}\left(1 - \frac{m_0 - 1}{m - 1}\right)(\bar{x}')^2}$$

where s' is the standard deviation of the $m - m_0$ detected measurements. The normal confidence limit can then be computed as previously described. This method is due to Aitchison (1955) - (also see USEPA 1992 section 2.2.2).

Lognormal Confidence Limits for a Percentile

In the lognormal case, confidence limits for percentiles are obtained by computing LCLs as described in the previous section on the natural logarithms of the measured values and exponentiating the resulting limits. Since the limits are for percentiles of the distribution, and not the mean, the simple transformation estimator applies directly. For example,

$$LCL_{1-\alpha,p} = exp\left[\bar{y} + K_{\alpha,p} s_y\right] \ , \qquad (2)$$

where $\bar{y}$ and s_y are the mean and standard deviation of the natural log transformed data $y = log_e(x)$. The factors used for computing these limits are the same as those given in Table 1.

Note that if nondetects are present, we can use the previously described statistical adjustment, replacing $\bar{x}'$ with $\bar{y}'$ and s' with s'_y in the equations for $\bar{x}$

and s. The lognormal prediction limit may then be computed as previously described. Note that this adjustment only applies to positive random variables. The natural logarithm of concentrations less than 1 are negative and therefore the adjustment does not apply. A simple solution is to add 1 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 for a Percentile

When data are neither normally or lognormally distributed, or the detection frequency is too low for a meaningful distributional analysis (*e.g.*, $< 50\%$), nonparametric confidence limits become the method of choice. The nonparametric confidence limit is defined by an order statistic (*i.e.*, ranked observation) of the m water body measurements. Note that in the nonparametric case, we are restricted to computing confidence limits on percentiles of the distribution, for example, the 90th percentile of the distribution. Unless the distribution is symmetric (*i.e.*, the mean and median are equivalent), there is no direct nonparametric way of constructing a confidence limit for the mean concentration.

To construct a nonparametric confidence limit for the 90th percentile of the concentration distribution, we use the fact that the number of samples falling below the $p(100)$th percentile of the distribution (*e.g.*, $p = .9$, where p is between 0 and 1) out of a set of m samples will follow a binomial distribution with parameters m and success probability p, where success is defined as the event that a sample measurement is below the $p(100)$th percentile. The cumulative binomial distribution ($Bin(x; m, p)$) represents the probability of getting x or fewer successes in m trials with success probability p, and can be evaluated as

$$Bin(x; m, p) = \sum_{i=1}^{x} \binom{m}{i} p^i (1-p)^{m-i} . \tag{3}$$

The notation $\binom{m}{i}$ denotes the number of combinations of m things taken i at a time, where

$$\binom{m}{i} = \frac{n!}{i!(n-i)!} ,$$

and $k! = 1 \cdot 2 \cdot 3 \ldots k$ for any counting number, k. For example, the number of ways in which 2 things can be selected from 3 things is:

$$\binom{3}{2} = \frac{3!}{2!(1)!} = \frac{1 \cdot 2 \cdot 3}{(1 \cdot 2)(1)} = \frac{6}{2} = 3.$$

To compute a nonparametric confidence limit for the 90th percentile we begin by rank ordering the m measurements from smallest to largest as $x_{(1)}, x_{(2)}, \ldots, x_{(m)}$. Denoting the candidate LCL as L^*, begin with $L^* = m$, and compute the probability

$$1 - Bin(L^* - 1; m, .9) .$$

If the probability is less than the desired confidence level, $1 - \alpha$, select a new value of $L^* = L^* - 1$, and repeat the process until the desired confidence level (*e.g.*, 95%) is achieved.

STATISTICAL POWER

An essential feature of any statistical method is the power with which it can detect a real exceedance of a given magnitude. To determine the statistical power of the 95% LCL for the 90th percentile of the distribution, and the corresponding EPA method of no more than 10% exceedances of the regulatory standard, the following simulation study was conducted. Three types of LCLs were considered, normal, lognormal and nonparametric. Normally distributed data were simulated to examine statistical power of normal and nonparametric LCLs, whereas lognormally distributed data were simulated to examine statistical power for lognormal LCLs. To examine the effect of sample size, we studied sample sizes of m = 5, 10, 20, 30, 40 and 50. In terms of effects sizes, we examined cases in which the true mean was one-half the size of the standard 0.5(STD), 0.6(STD), 0.7(STD), 0.8(STD), 0.9(STD), and equal to the standard 1.0(STD). To this end we used a standard of 50 μg/L and a standard deviation (SD) for each condition of 10 μg/L. As such, the true mean concentrations of the six simulated conditions were 25, 30, 35, 40, 45 and 50 μg/L respectively. For each condition, 1000 simulated datasets were generated and the percentages of times the LCL exceeded the standard of 50 μg/L, or the EPA method found more than 10% exceedances were recorded.

The results are summarized in the following.

First, there are very few cases in which the LCL exceeded the standard of 50 μg/L when the true mean was 25, 30, or 35 μg/L. This finding indicates that there is very little chance of an exceedance when the true mean is more than one standard deviation unit below the standard (*i.e.*, SD = 10 μg/L). This is not true for the EPA approach, which indicated reasonably high percentages of failures, even when the true concentration was as low as 25 μg/L (*i.e.*, 2.5 standard deviation units below the standard). For example, with a lognormal distribution and $m = 5$ measurements, their were 10% failures even when the true concentration was half of the regulatory standard. At a true concentration of 35 μg/L (which is still well below the standard of 50 μg/L), the rate at which the EPA method indicates impairment is approximately 30% or more, regardless of the distributional form of the data. These false positive results for the EPA method were most pronounced for small sample sizes ($m < 20$) and lognormal distributions, both of which are characteristic of practice in this area. Second, for a true mean of 40 μg/L (*i.e.*, exactly one SD unit below the standard), exceedances are detected with some frequency for both parametric and nonparametric LCLs, however, at best, power was 63% at $m = 50$ samples. Third, when the true mean is one-half of a standard deviation unit below the standard (*i.e.*, 45 μg/L) power is approximately 80% or more for either parametric or nonparametric LCLs and a minimum sample size of $m = 20$. Fourth, when the true mean is equal to the standard, power is generally excellent in all cases. A single exception is the nonparametric LCL for $m = 5$. As such, the minimum sample size that should be used for the nonparametric LCL is $m = 10$ samples. Fifth, power is generally best for the normal LCL (when the data are normally distributed), there is a small loss of power for lognormally distributed data, and a more substantial loss of power for the nonparametric LCL, particularly for smaller sample sizes (*i.e.*, $m < 20$).

In terms of the EPA method (*i.e.*, no more than 10% exceedances), power was reasonably comparable for the higher concentrations; however, the very high false positive rates should preclude further use of the method.

ILLUSTRATION

Consider the following data for hexavalent chromium (μg/L) obtained from a water body.

Date	Result
08/02/99	29
08/04/99	14
08/06/99	13
08/08/99	14
08/10/99	19
08/12/99	9
08/14/99	<1
08/16/99	33
08/18/99	150
08/20/99	60
08/22/99	57

The adjusted mean concentration is 36.182 μg/L with standard deviation of 42.483 μg/L. The normal 95% LCL for the upper 90th percentile is 67.364 μg/L. Note that inspection of the raw data clearly does not provide evidence for normality of these data. Indeed distributional testing using the Shapiro-Wilk test (see following section) revealed $W_{raw} = .716$, $W_{log} = .936$ and a critical value of W_{crit}=.842 indicating that the data are consistent with a log-normal distribution. The lognormal 95% LCL for the upper 90th percentile is 53.119 μg/L The nonparametric LCL for the upper 90th percentile is 33.000 μg/L, which corresponds to the 8th ordered observation, and provides confidence of .981. Given that the distribution was shown to be consistent with lognormality, we have 95% confidence that the true 90th percentile of the hexavalent chromium distribution is no less than 53.119 μg/L. If the regulatory standard is 50 μg/L, then we would conclude that there is 95% confidence that the true 90th percentile of the distribution has exceeded the regulatory standard (*i.e.*, more than 10% of the concentrations in the water body are above the regulatory standard).

CONCLUSIONS

In summary, use of the LCL for upper percentiles of a pollutant distribution in water body sampling applications provides a rigorous solution to the problems associated with taking a simple tally of the number of observed measurements which happen to exceed the standard. Statistical power characteristics indicate that the approach provides sensible results in those cases in which there is clearly not an impact, as well as those cases where an exceedance is apparent. In fact, power is excellent even when the mean of the pollutant distribution is one-half of a standard deviation unit below the regulatory standard. The best results overall, are obtained when the number of samples is at least 20 or more. The statistical power computations also revealed that the nonparametric approach should never be used when fewer than 10 samples are available. By contrast, the false positive rate for the EPA method of no more than 10% exceedances was quite high, particularly for small sample sizes and lognormal distributions, conditions which typify routine practice.

By using the statistical approach, focus is now on estimating a lower bound of a percentile of the true pollutant distribution and not a collection of available environmental measurements (*e.g.*, the 10th percentile of the observed data computed as the 9th largest of 10 available samples). Furthermore, the LCL is sensitive to the actual measured concentrations and does not simply rely on a binary determination of whether or not the observed measurement exceeded the regulatory standard. As shown in this paper, exact confidence levels for environmental decisions are directly available with the statistical approach, whereas they cannot even be guessed at using the current non-statistical approach. In this paper, both normal and lognormal forms of the LCL have been developed, as well as a nonparametric alternative which is quite useful for nonnormally distributed data that cannot be suitably transformed. The common case of having "nondetects" in an environmental dataset, makes the nonparametric approach attractive as well.

ACKNOWLEDGEMENTS

I would like to acknowledge earlier funding by the Utility Water Act Group (UWAG) that led in part to the development of the statistical methodology described here, and to Steven Koorse for several helpful comments.

References

[1] Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *Journal of the American Statistical Association,* **50,** 901-908.

[2] Gibbons, R.D. (1994). *Statistical Methods for Groundwater Monitoring,* Wiley, New York.

[3] Gibbons, R.D. and Coleman, D.E. (2001). *Statistical Methods for Detection and Quantification of Environmental Contamination,* Wiley, New York.

[4] Hahn, G.J. and Meeker, W.Q. (1991). *Statistical Intervals : A Guide for Practitioners.*, Wiley, New York.

[5] Shapiro, S.S., and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika,* **52,** 591-611.

[6] USEPA, Interim Final Guidance Document *Statistical analysis of ground-water monitoring data at RCRA facilities* (April, 1989).

[7] USEPA, Addendum to Interim Final Guidance Document *Statistical analysis of ground-water monitoring data at RCRA facilities* (July, 1992).

[8] USEPA, Unified Guidance Document *Statistical analysis of ground-water monitoring data at RCRA facilities* (In preparation, 2000).

[9] USEPA, *Guidelines for the Preparation of the Comprehensive State Water Quality Assessments (305(b) Reports) and Electronic Updates: Supplement* (September, 1997).