Article

Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act

Eric P. Smith, Keying Ye, Chris Hughes, and Leonard Shabman

Environ. Sci. Technol., **2001**, 35 (3), 606-612• DOI: 10.1021/es001159e • Publication Date (Web): 30 December 2000

Environ. Sci. Technol. **2001**, *35,* 606−612

Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act

ERIC P. SMITH,*,† KEYING YE,† CHRIS HUGHES,† AND LEONARD SHABMAN‡

Department of Statistics and Department of Agricultural Economics, Virginia Tech, Blacksburg, Virginia 24061-0439

Section 303(d) of the Clean Water Act requires states to assess the condition of their waters and to implement plans to improve the quality of waters identified as impaired. U.S. Environmental Protection Agency guidelines require a stream segment to be listed as impaired when greater than 10% of the measurements of water quality conditions exceed numeric criteria. This can be termed a "raw score" assessment approach. Water quality measurements are samples taken from a population of water quality conditions. Concentrations of pollutants vary naturally, measurement errors may be made, and occasional violations of a standard may be tolerable. Therefore, it is reasonable to view the assessment process as a statistical decision problem. Assessment of water quality conditions must be cognizant of the possibility of type I (a false declaration of standards violation) and type II (a false declaration of no violation) errors. The raw score approach is shown to have a high type I error rate. Alternatives to the raw score approach are the Binomial test and the Bayesian Binomial approach. These methods use the same information to make decisions but allow for control of the error rates. The two statistical methods differ based on consideration of prior information about violation. Falsely concluding that a water segment is impaired results in unnecessary planning and pollution control implementation costs. On the other hand, falsely concluding that a segment is not impaired may pose a risk to human health or to the services of the aquatic environment. An approach that recognizes type I and type II error in the water quality assessment process is suggested.

Introduction

The Total Maximum Daily Load (TMDL) process now dominates water quality policy discussions. Policy reviews (*1*), lawsuits (*2*), regulations (*3*), and congressional interest (*4*, *5*) all have been directed to what had, until recently, been an obscure provision of the Clean Water Act. The TMDL process originates with Section 303(d) of the Clean Water Act (*6*). That section requires states to conduct an assessment of and then report on the condition of their waters. In practice, this means that the states review the water quality conditions in specific segments in a water body (a lake, bay, or river) using a specific water quality monitoring location within the segment.

Each state's 303(d) impaired waters list identifies segments where anthroprogenic loads of pollutants are leading to violation of water quality standards. The listed segments must remain on the list until the identified pollution problem has been addressed or until evaluation of subsequent monitoring data or other information suggests that the segment was misclassified or the problem remediated. Addressing an identified water quality problem for a Section 303(d) listed water is a complicated and potentially expensive process. First, a watershed study is initiated to establish the maximum quantity of each pollutant that can be discharged to a segment if the segment is to meet water quality standards. Once the maximum load is defined, there are a series of steps to allocate responsibility for load reduction, to identify pollution sources, and to secure those reductions over time. These steps constitute the TMDL watershed study and implementation plan (*7*).

Planning alone can be costly. In comments to the U.S. Environmental Protection Agency (U.S. EPA), states agencies concluded that 25% of TMDLs will be simple and will cost $50 000−200 000, 65% of TMDLs will be of moderate difficulty and will cost $300 000−400 000, and 10% of TMDLs will be complex and will cost $600 000−1 000 000 (*5*). A state may have hundreds of segments on its impaired waters list (*8*). Then, implementation of a TMDL plan imposes additional and perhaps substantial pollution control costs. Given limited resources available for programs of water quality improvement planning and implementation, it is important that waters that are truly impaired be identified. Also, water listed as impaired may cause people to avoid use of that water and benefits to society may be forgone. For these reasons, it is appropriate to review how the list of impaired waters is constructed during the water quality assessment process.

A review is especially warranted because water quality standards, monitoring protocols, and guidelines for assessing data were developed before the TMDL program took on its current significance and may have been developed for different purposes. A review of the Section 303(d) assessment process might examine the basis and intended purpose of the water quality standards themselves. Also, such a review might evaluate the monitoring protocols that secure the data used to make the listing determination. In this paper, we review the guidelines for interpreting the monitoring data that are collected. Specifically, we evaluate the U.S. EPA assessment guidelines for comparing sample measurements of water quality conditions with numeric ambient water quality standards.

Numeric water quality standards are measurable criteria for dissolved oxygen, temperature, pH, and fecal coliform bacteria counts. Critical to the Section 303(d) assessment is the monitoring data collected by a state's environmental department to assess whether stream conditions meet standards. Cost realities, given the need for statewide monitoring and the fact that most monitoring is for enforcement of point source discharge permits, results in a limited number of stations and samples for each station. For example, Virginia waters are among the most monitored in the nation with over 17 000 mi of monitored waterways. Virginia's significant monitoring program collects data at each station on a quarterly basis. The Section 303(d) assessment occurs every 2 yr, so the Section 303(d) assessment might be based on 2 yr of data at a particular station (approximately eight

* Corresponding author phone: (540)231-7929; fax: (540)231-3863, e-mail: epsmith@vt.edu.
† Department of Statistics.
‡ Department of Agricultural Economics.

10.1021/es001159e CCC: $20.00
Published on Web 12/30/2000

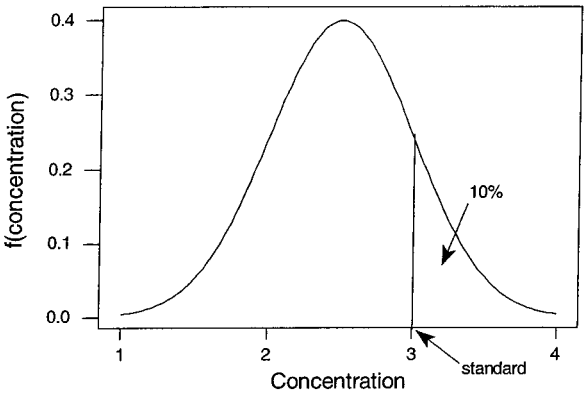

FIGURE 1. Plot of distribution of hypothetical chemical concentration. The standard allows for exceeding a concentration of 3 10% of the time.

observations; *9*). The reality of limited data must be recognized in the Section 303(d) assessment process.

The assessment challenge is to interpret the limited amount of sample data to determine whether an apparent violation of standards warrants listing a segment as impaired. Likewise, limited data must be relied upon to determine whether actions taken to address water quality degradation have had the desired results. The samples taken are affected by variability in human activity and natural or background conditions. Also, there are certain acceptable tolerances for violations. For example, an occasional violation of a dissolved oxygen standard, even if by anthropogenic sources, may not be critical for the aquatic environment. In addition, measurement errors in the analysis of the samples collected could be yet another reason the numeric standard might be violated in a sample. It appears that the U.S. EPA guidelines recognize these arguments because the guidelines require a water to be listed only if more than 10% of the samples violate the standard (*10*). In effect, the assessment guidelines imply that a violation of the numeric criterion is acceptable in 10% of the samples taken.

If the number of samples at a stream location greatly increases in frequency, conceptually approaching one for each hour (for example), the U.S. EPA guidelines suggest that it is acceptable for a standard to be violated 10% of the time. A statistical representation of this perspective is shown in Figure 1. In Figure 1, the measurement is a concentration of some contaminant in the ambient water. The distribution of the water quality parameter may be drawn to represent the likelihood of ranges of values. As displayed, the water quality standard requires that a concentration of 3.0 or less should be met 90% of the time, although some measurements may exceed the standard naturally.

The U.S. EPA guidelines suggest what can be called a raw score test to decide if a segment is impaired. The test statistic is the number of measurements that exceed the standard. The critical value is 10% of the sample size. Because the number of samples is typically not a multiple of 10, the approach requires truncation. If there are five samples and one or more exceed the standard, the site is declared impaired. The same is true for all sample sizes between 1 and 9. For sample sizes between 10 and 19, one sample is allowed to exceed the standard but not more. However, the raw score approach does not include consideration of the likelihood and costs of making an erroneous listing. Suppose eight samples are taken, and a raw score analysis is completed. If one of the samples (>10%) exceeds the standard, the site would be declared impaired. However, the one sample that violates the standard might be attributed to natural variability or an unusual human activity. In this case, the site may be classified as impaired when in fact this is not the case. This error is referred to as a type I error. Another error may occur when a site is truly impaired, but the sampled measurements from the site do not exceed the standard, and the site is not declared impaired. This error is referred to as a type II error.

In this paper, the error rates associated with the raw score approach and two statistical approaches are evaluated. The comparisons are made in terms of type I and type II error rates. One alternative to the raw score approach is the Binomial test. Both the raw score and the Binomial methods treat the sample observations as binary values, either exceeding the standard or not exceeding the standard. Another alternative to the raw score approach is the Bayesian version of the Binomial test. This method uses prior information about violation probability with sampled information to calculate a probability of violation that may then be used to make a decision. The three methods are evaluated in terms of their error rates. This evaluation of alternative approaches leads to a recommendation for improving water quality assessments in the Section 303(d) process.

Statistical Approaches

The Section 303(d) water quality assessment process is a statistical decision problem. Specifically, from a sample of water quality measurements the water quality assessor must decide if the site is impaired. Given uncertainty in the measurement and sampling process, one may use hypothesis testing to help with the decision process. In the statistical approach to impairment, the null hypothesis is that the site is not impaired. The alternative hypothesis is that the site is impaired. The hypothesis may be framed in terms of a parameter p describing the true degree or probability of impairment and p_0, the "safe level" or hypothesized probability of impairment under safe conditions. The impairment decision is based on the test H_0: $p \leq p_0$ versus H_1: $p > p_0$ where p_0 is a constant between 0 and 1 (in the current problem, it is 0.10). Under this framework, the two error rates [declare segment impaired when it is not (type I error or a false positive) or designate the segment as not impaired when in fact it is (type II error or false negative)] may be evaluated. The error rates are bounded between 0 and 1, with 0 indicating no error. However, given the sample sizes likely to be available, both errors will not be close to zero.

Because both type I and type II errors always will be present, water quality managers must choose (directly or indirectly) the tolerable amount of error. In principle, this choice should be based on an explicit consideration of the consequences (costs) of being wrong. Costs may be financial outlays made by governments or private individuals. Costs might be forgone public values that may not be reflected in markets. In the following sections, the tradeoff among error types is considered without regard to the cost of being wrong. Costs are considered in the Discussion section of the paper.

The raw score approach uses limited, binary information to make the impairment determination. An alternative to the raw score, the Binomial testing approach focuses on the probability of violation using the same information. The Bayesian approach varies the Binomial method by using information from other sources about the probability of violation.

Binomial Method. When applying the Binomial approach, observations exceeding the numeric criterion are assigned the value 1, and those that do not are assigned the value 0. Then if n independent samples are collected, the number of observations exceeding the criterion (the number of 1's) may be viewed as a Binomial random variable with parameters p and n (*11*). Using the Binomial model, one may then test the hypothesis that the probability of exceeding the standard is less than or equal to 0.10 (H_0: $p \leq 0.10$, not impaired)

versus the alternative that the probability is greater than 0.10 (H_1: $p > 0.10$, impaired). With this approach, error rates associated with impairment declarations may be evaluated, and a process to limit the error rates can be described.

In a typical statistical analysis, the type I error rate is chosen by the assessor, perhaps in consideration of costs of being wrong. If the rate chosen is 0.10, then there is a 10% chance of making a type I error. For the Binomial method, the choice of the type I error rate determines the "cutoff" value. For a given sample size n, the cutoff is selected as the number of violations to make the probability of this many or fewer violations to be as large as possible but less than the type I error rate, assuming that the null hypothesis of no impairment is true. Given the cutoff and the alternative for the frequency of violation, the type II error rate for sample size n can then be calculated. The type II error rate may be reduced by choosing a greater type I error rate (for example 0.20), by increasing sample size and/ or by decreasing measurement uncertainty. With statistical procedures, it is common to select the type I error rate at 0.05 or 0.10 and to control the type II error rate through sample size.

Bayesian Approach to the Binomial Test. In the above analysis, the probability of exceeding the standard is treated as fixed and the data (i.e., does the sample exceed the standard) are treated as random. A Bayesian approach (*12*) computes the probability that the site exceeds the standard by treating the impairment probability as a random variable that has an associated distribution. Initially the form of this distribution is based on previous information and is referred to as the prior distribution. After data are collected, the prior is updated, and the data and prior are used to compute the posterior distribution of the impairment probability using Bayes rule. Based on this posterior distribution, a decision may be made using either a cutoff approach or an odds-ratio approach (Bayes factor). This process and the mathematical details are described in more detail in the Supporting Information and ref *13*.

Suppose there is a Binomial random variable with associated sample size n and parameter p. Suppose now that a prior distribution of p, $\pi(p)$, can be specified. A prior distribution for p might be developed by introducing additional information to the analysis. One possibility is to use samples from other similar sites that are not impaired. For the unimpaired sites, information would be collected, and the prior probability of exceeding the standard calculated.

Given observations and a prior distribution, Bayesian criteria can be used to make an inference about p. Using the prior and data, the posterior distribution of p may be written as

$$\pi(p|x) = \frac{f(x|p)\pi(p)}{\int_0^1 f(x|p)\pi(p)\ \mathrm{d}p}$$

where $f(x|p)$ is the density of the data, x, given p.

This new distribution represents current knowledge about the probability of a violation found by updating the prior information. Using the above distribution, the posterior probability of the null and alternative hypotheses may be calculated. For the null hypothesis (H_0) that the site is not exceeding standards, the probability is computed as $\alpha_0 = P(H_0|\text{data}) = P(p \le p_0|x)$. For the alternative ($H_1$) that the site is exceeding standards, the posterior may be calculated as $\alpha_1 = P(H_1|\text{data}) = P(p > p_0|x)$. Two approaches for evaluating these probabilities and making decisions are the cutoff method and the ratio method.

The cutoff method uses the posterior probability to determine the rejection rule. To do this, predetermine a probability q (analogous to the Binomial method type I error rate, q might be specified as 0.10). If the posterior probability that the alternative hypothesis is true exceeds q, then we reject the null hypothesis and conclude that the water is impaired, i.e., $P(H_1|\text{data}) > q$. The quantity q is referred to as the posterior cutoff.

The odds-ratio method uses the Bayes factor to determine the rejection rule. The Bayes factor of H_1 against H_0 is the odds ratio of the posterior probability of H_1 against H_0 divided by the odds ratio of the prior probability of H_1 against H_0. It can be expressed as

$$B_{10} = \frac{P(H_1|x)}{P(H_0|x)} \Big/ \frac{P(H_1)}{P(H_0)}$$

A large value of the Bayes factor would indicate that the null hypothesis is not correct. Kass and Raftery (*14*) (see also ref *15*) suggest that when B_{10} is between 3 and 20, the evidence of H_1 against H_0 is strong. Bayes factor cutoffs of 3 and 10 were used in our examples.

The difference between the cutoff and odds-ratio methods is in the importance given to the prior. The influence of the prior is usually diminished if the Bayes factor method is used. Because of the possible subjectivity of the prior, decision-makers may want to choose to use the Bayes factor approach. If the available prior information is empirical, the cutoff method might be adopted.

Both methods require evaluation of the prior probability of the null and alternative hypotheses. Using a weighting factor v (between 0 and 1) that balances the prior distribution between null and alternative hypotheses may extend the method. A value of v that is near 1 would indicate a stronger belief in the null hypothesis. In the figures comparing the methods, we refer to this value as $p(H_0)$ or $\text{prior}(H_0)$. Details of the computations are given in the Supporting Information.

To compare the error rates, the acceptable probability of violation is set at 10%. The analysis assumes that the water quality parameter of interest has a distribution that does not change over time and that the samples collected are independent of each other. On the basis of these assumptions, the variable that indicates if a sample exceeds the standard may be modeled as a random variable, with an associated probability of violation. The listing decision process may be viewed as a test of the null hypothesis that the probability of violation is less than or equal to 10% versus the alternative that it is greater than 10%. The type I error rate may then be computed. To compute a type II error rate for this illustration (given the site is impaired, how likely is it that we do not detect impairment), the true probability of exceeding the standard must be specified; this percentage is set at 25%. This value was selected as indicating severe problems and represents the minimum violation percentage we would almost always want to detect. Using this framework, the distribution may be used to calculate the error rate for the raw score method by calculating the probability of not rejecting the null hypothesis (i.e., getting less than a statistically significant number of violations). To evaluate decision rules based on the Bayesian method, we considered three situations for method 1 with a uniform prior for p ($v = 0.50$, 0.90 and 0.99, $q = 0.1$) and two values of cutoff for method 2 (using Bayes factors of 3 and 10).

Results

Type I error rates for the raw score, Binomial, and Bayesian methods are presented in Figure 2, and type II error rates are presented in Figure 3. The type I error rates are compared using calculations of Binomial probabilities under different sample size scenarios where p was set to 0.10. The probability that a site is declared as impaired when in fact it is not (false positive) is displayed in Figure 2. Note that the graphs are jagged, with each spike corresponding to a change in the

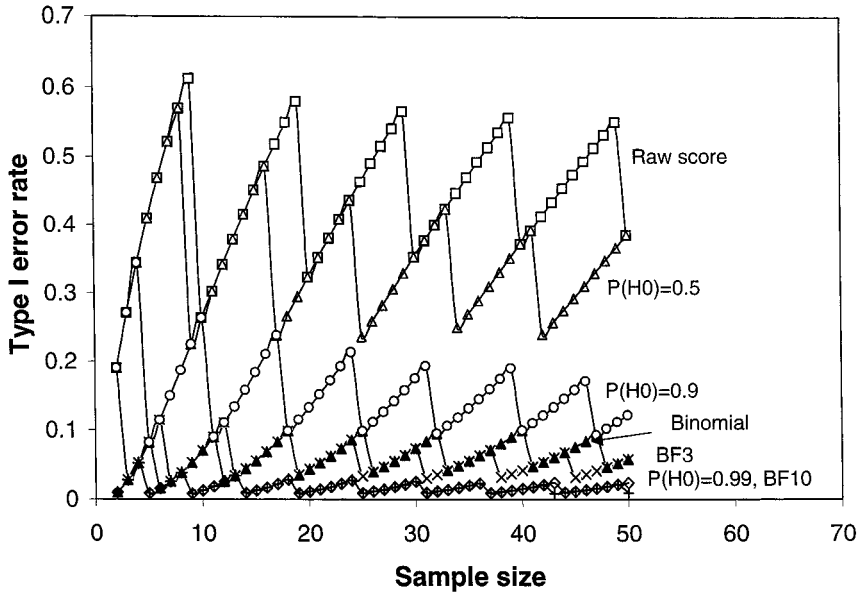

FIGURE 2. Type I probabilities for various methods. The Binomial method is based on setting the type I error rate at 0.1. Symbols: □, raw score; ▲, Binomial; △, p(H0) = 0.5; ○, p(H0) = 0.9; +, p(H0) = 0.99; ◇, BF 10; ×, BF3.

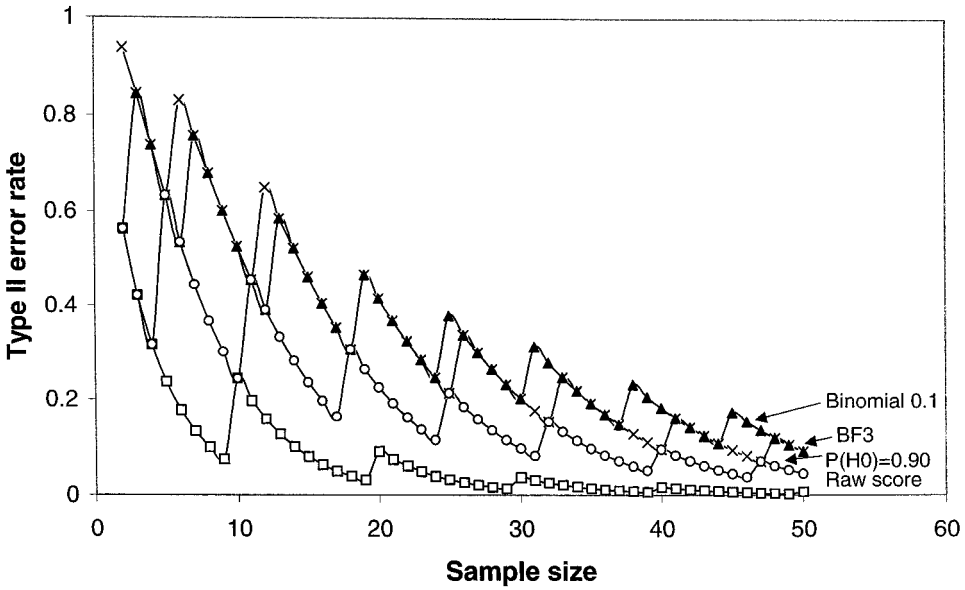

FIGURE 3. Type II probabilities for various methods. P(H0) = 0.9 represents the Bayesian method with a prior of the null hypothesis set at 0.9; BF3 refers to the Bayes Factor method using 3 as a cutoff. The symbols are the same as in Figure 2.

critical value (i.e., number of violations required to declare impairment). The Binomial method controls for type I error (i.e., it is always less than or equal to a preset value of 0.10), and the raw score approach does not. With the Binomial method, the type I error rate is fixed at some value (referred to as α) that is an upper bound on the error. The actual error rate for the Binomial method is determined by computing the (cumulative) probability of getting less than "x" samples exceeding the standard. The actual type I error rate is calculated as the greatest cumulative probability that does not exceed α. Figure 2 shows that the type I error rate (a false declaration of impairment) for the raw score method is quite high relative to the Binomial. For example, with a sample size of 9 the type I error rate for the raw score approach is around 61%. With one more sample, it drops to 26% (an example of the effect of truncation) but is still roughly 3 times the type I error rate of the Binomial approach. Error rates this high are not used in standard statistical practices. As sample size increases, the type I error rates for the different methods do not converge. Thus, relative to the Binomial approach, the raw score approach is prone to type I error (a false declaration of impairment). Type I errors for the Bayesian method decrease with increasing p(H0). Priors for H_0 near 0.5 are similar to the raw score approach while priors near 0.9 are closer to the Binomial approach. Having a high prior opinion that there is no impairment leads to making fewer decisions that there is impairment when there is none. The Bayes factor methods produce results that have smaller type I error rates than the Binomial method. Using a higher factor for rejection leads to smaller type I errors.

Figure 3 presents type II error rates. We assume for the computations that the actual level of impairment is 25%, so the segment violates standards; however, the violation is not detected. In statistical terms, this represents failure to reject the hypothesis that the violation rate is equal to 0.10 when in fact the violation probability is 0.25. In this case, Figure 3 is reversed from Figure 2. The Binomial method is prone to type II error relative to the raw score method. For example, with a sample size of 9, the type II error rate for the Binomial is about 8 times the rate for the raw score approach (60%

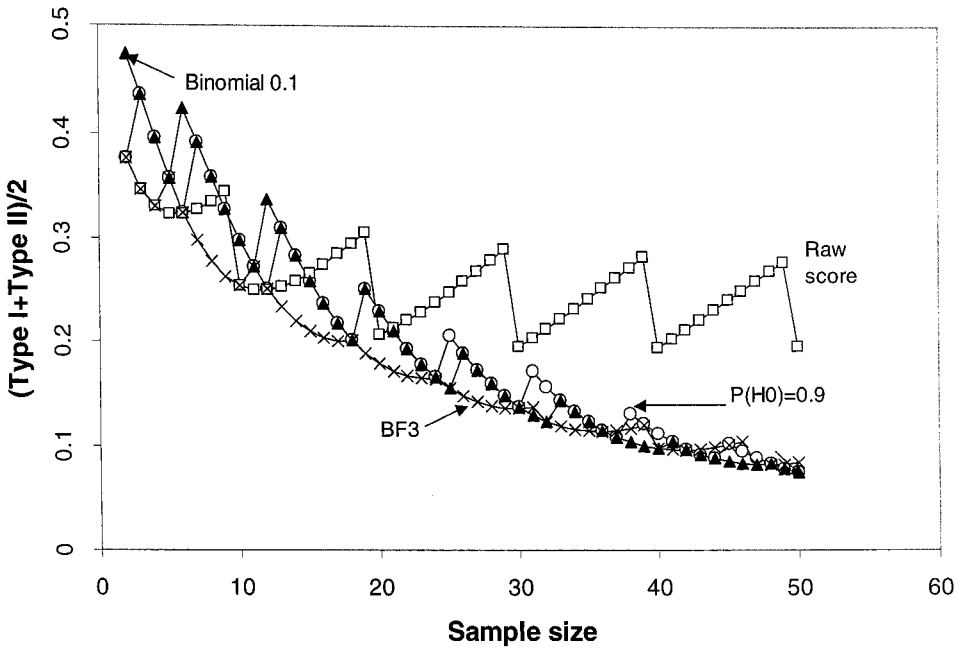

FIGURE 4. Average error rate of the different methods using different sample sizes. P(H0) = 0.9 represents the Bayesian method with a prior of the null hypothesis set at 0.9; BF3 refers to the Bayes Factor method using 3 as a cutoff. The symbols are the same as in Figure 2.

versus 7.5%). With one more sample, the ratio decreases to about 2 times (a result of the effect of truncation). As sample sizes get larger, the type II error rates do converge to zero, which is to be expected. These results are appropriate for the case of a critical error being associated with a violation probability of 0.25 and a preset type I error rate of 0.10. The results indicate that the chance of a type II error using the Binomial method decreases with an increase in the type I error rate and with increased sample sizes. For sample sizes of $n = 8$, the type II error is 0.37 for a type I error of 0.20 while for a type I error of 0.10, the type II error is 0.68. For $n = 20$, the error rates are 0.23 versus 0.41. The pattern for the Bayesian approaches is similar, and only two of the Bayesian approaches are displayed in Figure 3. Type II error rates decrease as the prior probability that the null is true decreases. The curve for p(H0) = 0.5 is closer to the raw score method than is the curve for p(H0) = 0.9. When our belief that the null is true is higher, we are more likely to decide an impaired site is not impaired. Similarly, if the Bayes factor criterion is small (e.g., 3.0) then we are more likely to declare impaired than if we use a large Bayes factor criteria (e.g., 10). This leads to higher type I and smaller type II for smaller criteria. In terms of type II error, we have

$$p(H0) = 0.99 \geq BF10 \geq \text{Binomial} \geq BF3 \geq p(H0) = 0.9 \geq p(H0) = 0.75 \geq \text{raw score}$$

Figure 4 displays the average error rate for different sample sizes. This display is interesting in that the average error rate diminishes and approaches the same value for the statistical approaches but not for the raw score approach. This results from the type II error rate decreasing as a function of sample size and low type I error rates (for methods other than the raw score). Again it indicates that the error rates for the statistical methods have controllable error rates that may be made reasonably small while the raw score method has a large error rate.

One possible approach to addressing the different error rates is to seek to make type I and type II error rates the same for each sample size (*16*). In effect, this implies that the cost of type I and type II errors are the same. Another argument for balancing the error rates is that the errors are less affected by switching the null and alternative hypothesis. Instead of considering H_0: $p \leq p_0$ versus H_1: $p > p_0$, it may be better to use the hypotheses H_0: $p \geq p_0$ versus H_1: $p < p_0$. With balanced error rates, the choice of the null and alternate hypotheses is less important. In Figure 5, the error rates are plotted against sample size using a Binomial test with the null $p = 0.1$ and the alternate $p = 0.25$, with cutoff values chosen to make the error rates as close as possible. If there are at least these numbers of samples exceeding the standard, the site is declared impaired. Cutoff values are plotted on a second vertical axis. Note that for small sample sizes it is difficult to equate the error rates although there are sample sizes where the error rate lines cross. Examples are $n = 10$, type I error = 0.26, type II error = 0.24, and cutoff = 2; $n = 16$, type I error = 0.21, type II = 0.20, and cutoff = 3; $n = 22$, type I error = 0.17, type II error = 0.16, and cutoff = 4. Note that if it is desired to have both error rates around 10%, then a sample of size 34 would be required (cutoff = 6, type I error = 0.12, and type II error = 0.11).

Relative to the EPA raw score approach, the Binomial method (with common choices for the type I error rate) is more prone to type II error and less prone to a type I error. The tendency toward type II errors in either approach is mitigated by increased sample size, although even at sample sizes over 20, type II error rates for the Binomial are around 2−3 times higher that the raw score approach. An advantage of the Binomial approach is that it is more flexible in the choice of cutoff through the selection of the type I error rate, with type II errors controlled through sample size. This means better control of error rates and the possibility of setting error rates to the same value. Specifically, at sample sizes of around 25 type I and type II error rates with the Binomial method can be made around 20% for each type of error. With the raw score approach, there is no control over the type I error rate. The Bayesian approach allows for control of the error rates through the choice of cutoff and prior opinion. While the results may be similar to the Binomial, the Bayesian method may be intuitively more appealing to managers. It allows managers to set prior belief about how likely sites are to be impacted. Sites with a high prior for impairment require fewer violations to declare impairment

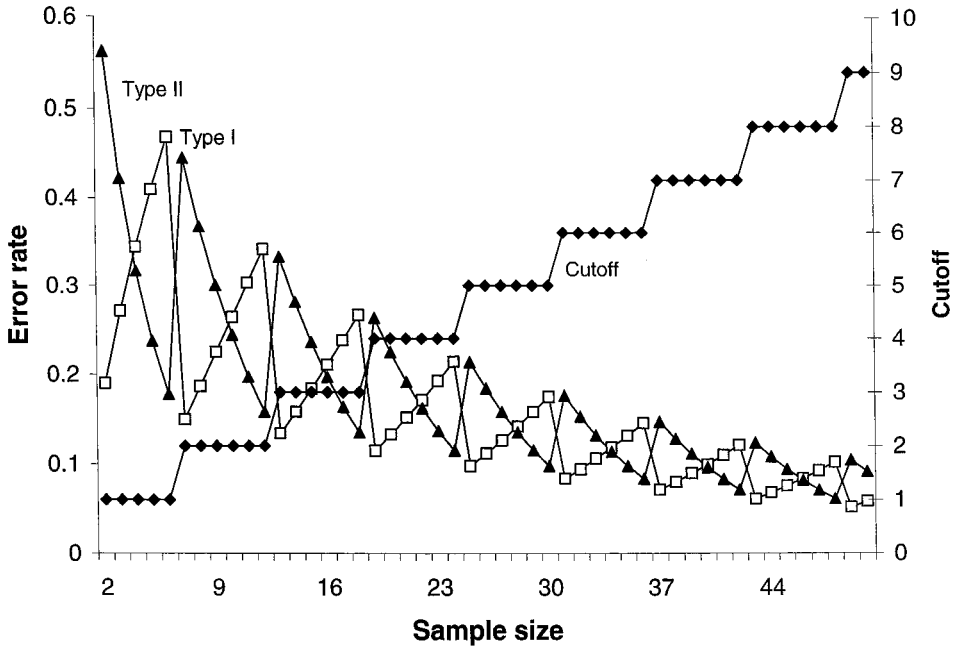

FIGURE 5. Error rates and cutoffs for different sample sizes, trying to make the type I and type II error rates as close as possible for the Binomial test. Cutoff values correspond to the minimum number of samples that may exceed the standard to declare the site as impaired.

while sites with a high prior for no impact would require more violations to declare impairment. Selecting priors can be difficult when there is little information, and the analysis becomes subjective and subject to criticism. However, support for these probabilities could come from previous Section 303(d) reports and surrounding sites. This would lead to more objective formulation of priors and would make the Bayesian approach a sound alternative.

Discussion

Ideally, the choice of an error rate should be a risk management decision based on explicit consideration of the consequences (costs) of being wrong. Cost may be financial outlays made by governments or private individuals. For planning and pollution control, costs also might be forgone public values that may not be reflected in markets as people avoid use of waters that are listed as impaired and calculation of these costs may be more or less certain. Consider as an example, a violation of a fecal contamination standard.

First, the assessor recognizes there is a cost of a false positive (type I) error that initiates the listing and the TMDL process. There is a cost to TMDL planning and modeling that is significant financial outlay. Each study is a claim on a limited agency budget, and so available resources are spread out more thinly as the number of segments listed as impaired increases. Therefore, in the face of limited budgets, a segment that is declared impaired when it is not impaired may divert limited resources from actual to false problems. Once the impairment is declared, there may be public avoidance of the segment and a loss of public use values. Once again, if the segment is not impaired, then those values forgone are an unnecessary cost. Next, planning moves forward and there are implementation costs (BMPs, etc.) imposed to change practices at the suspected source of the pollutant. Such implementation costs might be imposed on public agencies and the private sector at the end of the TMDL process. These considerations argue for selecting a decision process that might avoid type I error.

The assessor must also consider the possibility of declaring a segment as safe when in fact it is impaired (a type II error), especially when human health is at issue. Missing a fecal coliform problem may lead to an outbreak of infection with high costs to individuals. Low levels of dissolved oxygen may result in economic loss to fisheries and loss of species. Costs to human and environmental health may be great when a type II error is made and thus argue for selecting an decision process that might avoid a type II error.

Even when a site is correctly identified, there may be issues associated with action. For example, in the case of microbial contamination there is much uncertainty about the source and pathways for the pollutant and the effects on human health (*17*). There may be uncertainty about whether the measured contaminant poses a health risk, there may be uncertainty about the exposure to the pollutant (who swims in a creek and when for example), there may be uncertainty about whether the exposed population will in fact be affected by the contaminant even if it is in the segment, and finally the severity of the reaction to the exposure may be uncertain. These possible costs, despite—or perhaps because of—their uncertainty, might make the assessor willing to accept a higher type II error.

The significant consequences of a Section 303(d) listing or of a failure to list makes the interpretation of sample data especially critical. Therefore, the analytical approach that extracts the most information about water quality conditions from a data set should be employed. In particular, the approach used should allow the water quality assessor to explicitly recognize and consider the different errors that might be made, the consequences of those errors, and then assess water quality conditions in consideration of the errors and their possible costs. If a Binomial procedure is adopted, error rates can be explicitly managed by the water quality assessor by controlling the number of samples taken, by selecting the acceptable and unacceptable violation rates, and/or by selection of the cutoff values for declaration of impairment. Such choices might be governed by the concerns over the consequences of a type I versus type II error, considering the pollutant and the uses of the water segment.

The U.S. EPA mandated raw score approach to data analysis does not explicitly manage error rates. The raw score approach is conceptually similar to the Binomial test. Both methods use the number of violations as the test statistics. However, the raw score is a poorly designed test statistic. As the computational results document, the raw score approach

results in an unusually large type I error rate, regardless of sample size. As sample sizes increase, the type II error rate is reduced, but the average error rate is still large. Indeed, in other contexts, approaches to evaluating standards have been criticized for a number of reasons, including the inability to consider and manage error rates (*18*).

The results show that the Binomial method can be easily applied to address the balancing of error rates, using the same data that are now used to apply the raw score approach. The Bayesian approach changes the view of the error rates by focusing on prior probabilities and cutoffs and will require the assessor to have a basis for establishing a prior expectation about the condition of the water segment. One method for selecting the priors is to make use of information from surrounding sites or from previous reports. Given the familiarity most assessors will have with the conditions in watersheds under study, this may not be a significant additional information requirement.

Given the information routinely used in an assessment, the Binomial method should replace the raw score approach. When sample sizes are around 20−25, the assessment process can confidently rely on statistical procedures to manage and measure type I and type II errors. Such an increase in sample sizes might be readily obtained by extending the data record from 2 to 5 yr, assuming quarterly sampling. However, accounting for possible trends in the data (*9*) may be necessary.

It has also been recognized that type II errors are more likely to occur with the statistical methods than with the raw score approach. While the increased sample size will reduce the probability of type II error, water quality assessors may feel that the statistical approaches are still too prone to type II error. One strategy for reducing the type II error would be to increase the type I error rate. The desired error rates need to be set through discussions with interested parties and when agreement is not possible, we suggest balancing the error rates.

Given the information routinely used in an assessment, the Binomial method should replace the raw score approach when sample sizes are greater than 20. With samples smaller than 20, neither the raw score or the Binomial method adequately control the error rates. Given sufficient prior information, Bayesian methods may be used with smaller sample sizes to help select the error rate of concern. Agencies should be encouraged and provided the resources to increase sample sizes for the assessment process to adequately control these error rates.

Although our focus is on the Binomial approach for evaluation of impairment, there are other statistical approaches available that make use of the actual measurements rather than if the measurement exceeds the standard. Acceptance sampling by variables (*19*) is a method based on using the mean and variance of the measurements rather than simply if they exceed a standard. The method converts questions about the proportion exceeding some value to questions about a mean. Tolerance intervals and prediction intervals also represent useful approaches (*20*−*22*). Tolerance intervals are intervals for a percentile of the samples. Another method is based on comparison of a reference site with that sampled (*23*). Such approaches are common in groundwater evaluation. These methods evaluate the information in a different manner and may be quite useful. As with all decision procedures, these methods also require consideration of error rates before implementing.

Acknowledgments

We are grateful for support from Virginia's Department of Environmental Quality to study this problem. Also thanks go to the University of Washington's National Research Center for Statistics and the Environment for support and encouragement in writing the manuscript during the summer of 1999. We are thankful for the careful reading and comments from the referees and Associate Editor.

Supporting Information Available

Details on the calculation of the posterior distribution for the null and alternative hypotheses and Bayes factor are presented in the Supporting Information for the case of a mixture prior and a uniform prior. This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Houck, O. A. *The Clean Water Act TMDL Program: Law Policy and Implementation*; Environmental Law Institute: 1999.
(2) Office of Water, U.S. Environmental Protection Agency. *Total Maximum Daily Load Program: TMDL Litigation by State*; March 16, 2000; http://www.epa.gov/owow/tmdl/lawsuit1.html.
(3) Office of Water, U.S. Environmental Protection Agency. *Total Maximum Daily Load Program: Proposed Rules*; March 16, 2000; http://www.epa.gov/OWOW/TMDL/proprule.html.
(4) U.S. General Accounting Office. *Water Quality: Identification and Remediation of Polluted Waters Impeded by Data Gaps*; U.S. Government Printing Office: Washington, DC, 2000.
(5) U.S. Congress. Subcommittee on Water Resources & Environment Hearing on EPA's Proposed Regulations Regarding Total Maximum Daily Loads (TMDLs), the National Pollutant Discharge Elimination System (NPDES), and the Federal Anti-Degradation Policy. February 10 and 15, 2000; http://www.house.gov/transportation/ctisub5.html and http://www.house.gov/transportation/ctisub5.html.
(6) Office of Water, U.S. Environmental Protection Agency. *Total Maximum Daily Load Program: National Overview*; March 16, 2000; http://www.epa.gov/OWOW/TMDL/status.html.
(7) Office of Water, U.S. Environmental Protection Agency. *Total Maximum Daily Load Program*; July 28, 2000; http://www.epa.gov/owow/tmdl/nutrient/nutrient.html.
(8) Office of Water, U.S. Environmental Protection Agency. *Total Maximum Daily Load Program: Major Pollutants Causing Impairment by State*; March 17, 2000; http://www.epa.gov/OWOW/TMDL/303dcaus.html.
(9) Zipper, C.; G. Holtzman; P. Darken; P. Thomas; J. Gildea; Younos, T. *Long-Term Water Quality Trends In Virginia's Waterways*; VWRRC Report SR11-1998; Virginia Water Resources Research Center: Blacksburg VA, 1998.
(10) Office of Water, U.S. Environmental Protection Agency. *Guidelines for Preparation of the Comprehensive State Water Quality Assessments*; Washington, DC, 1997.
(11) Zar, J. H. *Biostatistical Analysis*, 3rd ed.; Prentice Hall: Upper Saddle, NY, 1996.
(12) Berger, J. *Statistical Decision Theory and Bayesian Analysis*, 2nd ed.; Springer-Verlag: New York, 1985.
(13) Ye, K.; Smith, E. P. *Environ. Ecol. Stat.* In press.
(14) Kass, R. E.; Raftery, A. E. *J. Am. Stat. Assoc.* **1995**, *90*, 773−795.
(15) Jeffreys, H. *Theory of Probability*; Oxford University Press: Oxford, 1961.
(16) Bross, I. D. *Biometrics* **1985**, *41*, 785−793.
(17) Hagedorn, C. *Virginia Water Central* **2000**, *October*, 7−10.
(18) Barnett, V.; O'Hagan, A. *Setting Environmental Standards*; Chapman and Hall: London, 1997.
(19) Duncan, A. J. *Quality Control and Industrial Statistics*; Irwin: New York, 1974.
(20) Gibbons, R. D. *Ground Water* **1987**, *25*, 455−465.
(21) Gibbons, R. D. *Statistical Methods for Groundwater Monitoring*; John Wiley and Sons: New York, 1994.
(22) Whitmore, G. A. *Am. Stat.* **1986**, *46*, 193−197.
(23) Small, M. J. *Water Resour. Res.* **1997**, *33*, 957−969.

Received for review April 5, 2000. Revised manuscript received November 10, 2000. Accepted November 16, 2000.

ES001159E