155

ENVIRONMETRICS
Environmetrics 2003; 14: 373–386 (DOI: 10.1002/env.592)

Evaluation of water quality using acceptance sampling by variables

Eric P. Smith*[†], Alyaa Zahran, Mahmoud Mahmoud and Keying Ye

Department of Statistics, Virginia Tech, Blacksburg, VA 24061-0439, U.S.A

SUMMARY

Under section 303(d) of the Clean Water Act, states must identify water segments where loads of pollutants are violating numeric water quality standards. Consequences of misidentification are quite important. A decision that water quality is impaired initiates the total maximum daily load or TMDL planning requirement. Falsely concluding that a water segment is impaired results in unnecessary TMDL 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. Because of the consequences, a method is desired that minimizes or controls the error rates. The most commonly applied approach is to use the Environmental Protection Agency (EPA)'s raw score approach in which a stream segment is listed as impaired when greater than 10 per cent of the measurements of water quality conditions exceed a numeric criteria. An alternative to the EPA approach is the binomial test that the proportion exceeding the standard is 0.10 or less. This approach uses the number of samples exceeding the criteria as a test statistic along with the binomial distribution for evaluation and estimation of error rates. Both approaches treat measurements as binary; the values either exceed or do not exceed the standard. An alternative approach is to use the actual numerical values to evaluate standard. This method is referred to as variables acceptance sampling in quality control literature. The methods are compared on the basis of error rates.

If certain assumptions are met then the variables acceptance method is superior in the sense that the variables acceptance method requires smaller sample sizes to achieve the same error rates as the raw score method or the binomial method. Issues associated with potential problems with environmental measurements and adjustments for their effects are discussed. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: TMDL; monitoring; standards; acceptance sampling by variables; binomial distribution

1. INTRODUCTION

In the United States, every state is required under Section 303(d) of the Clean Water Act to identify water segments where anthropogenic pollution is leading to violations of water quality standards. A standard is a numerical criterion that is set as an indication of a healthy water body. Water segments that are in violation are listed and may be required to undergo a further assessment known as the Total Maximum Daily Load (TMDL) process. The TMDL process is a study of the watershed near the site to

*Correspondence to: Eric P. Smith, Department of Statistics, Virginia Tech, Blacksburg, VA 24061-0439, U.S.A.
[†]E-mail: epsmith@vt.edu

Contract/grant sponsor: U.S. Environmental Protection Agency's Science to Achieve Results (STAR) contract/grant number: R82795301.
Contract/grant sponsor: Virginia Water Resources Research Center.

Published online 10 April 2003
Received 5 October 2002
Revised 12 April 2002

determine the amount of pollutant that may be tolerated by the watershed without violating the standard. Part of the study involves identification of pollution sources, steps for reducing the pollution load to the system and a plan for reducing pollution inputs. Clearly this can be a costly process and state agencies have expressed concerns about the monitoring and listing process. For example, the State of Virginia monitors over 17 000 miles of waterways that are divided into segments. A set of segments within the waterways is sampled on a (typically) monthly or quarterly basis. Since the law requires an assessment on a two-year cycle the monitoring program results in a large amount of information. However, decisions are made on a site-by-site basis and the sample size for an individual site is often small. For example, quarterly sampling typically results in eight observations to make the decision to list or not list the segment.

Critical to the listing process are the data that are collected and the decision process used to list a segment. The challenge to the assessor is to use the limited data that is available to determine if the stream is violating standards, given that samples might be affected by variation and natural or background conditions. The typical approach for making a decision involves both objective and subjective methods. All information that is collected is to be used in the process, and this might include anecdotal information, testimony, and citizen monitoring data as well as samples collected by the agency. The objective approach involves a simple test of the data that we refer to as the 'raw score' approach. Specifically if more than 10 per cent of the samples exceed the standard then the site is declared to not meet the usability criteria.

From a statistical perspective, the conceptual model for evaluation of water quality is presented in Figure 1. In Figure 1 the measurement is a concentration of some contaminant in the ambient water. The distribution of the concentration represents a possible range of values that might be observed at a site over the time period of interest and the standard is one possible concentration. The standard may have been chosen based on laboratory data, previous field data or expert opinion. Suppose the water quality guidelines require that a concentration of 8.0 or less should be met 90 per cent or more of the time. In the raw score approach the proportion of samples that exceed the standard is calculated and, if this is greater than 10 per cent, the site is listed. For example, if there are five samples and one or more exceeds the standard, the site is declared impaired. The same is true for all sample sizes between one and nine. For sample sizes between 10 and 19, one sample is allowed to exceed the standard but not more. The diagram suggests several other approaches to the analysis of data arising under this scheme.

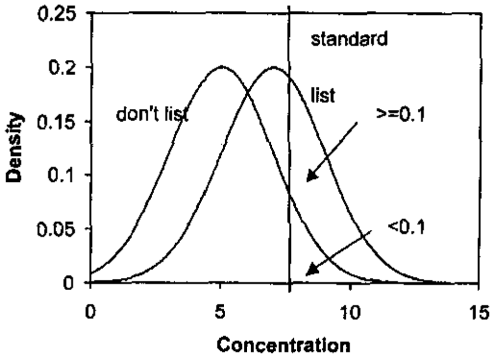

Figure 1. Hypothetical distributions of measurements for listing and not listing a site as impaired based on measurements

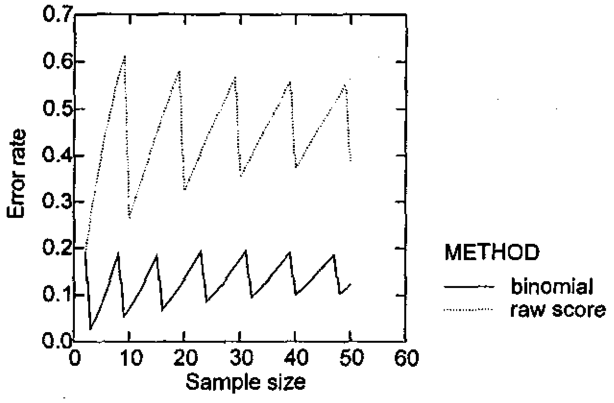

Figure 2. Probability of listing a site that is not impaired (Type I error rate) for different sample sizes for the raw score method and the binomial method choosing a cut point to bound the error at 0.20

One alternative described in Smith *et al.* (2001) uses a binomial testing approach for evaluation of a standard (see also McBride and Ellis, 2001, for a different approach). Smith *et al.* (2001) viewed the assessment problem as a decision process and recognized that error rates should be considered. Specifically, let p be the probability of exceeding the standard. Then the problem may be set up using hypothesis testing with $H_0: p \leq p_0$ (*no impairment, don't list*) versus $H_1: p > p_0$ (*impairment, list*). The value of p_0 would be set to 0.10 and the test based on $\hat{p}$, the proportion of measurements exceeding the standard.

The hypothesis testing scenario suggests evaluation of tests based on error rates. From the environmental perspective, the assessor needs to be concerned about falsely listing a site and failing to list a site that is in poor condition. Falsely listing a site may trigger the TMDL process and incur unnecessary constraints on agriculture or industry. Failing to list a site that has poor water quality may result in increased risk to human and ecological health. If error rates are considered, Smith *et al.* (2001) show that the raw score method has a strong tendency to falsely list a site. The binomial method as might be typically applied (i.e. using a Type I error rate of 0.05) will have a tendency to not list sites that should be listed (Figures 2 and 3). However, the binomial approach allows for control of both errors, and sampling plans may be developed to set error rates at satisfactory levels for sufficient sample sizes. Both methods may be criticized, as the actual numerical value is not accounted for. In the above example, a value of 8.1 is treated the same as a value of 801; it simply exceeds the standard of 8. An alternative to these approaches is based on acceptance sampling by variables.

Acceptance sampling is a major field in statistical quality control where the view is that a group or lot of manufactured items is to be accepted or rejected. Based on a sample from the lot, a manager makes a decision. Evaluation of the method is based on the risk of accepting lots of given quality, and sampling plans are developed to try to minimize error rates.

There are two approaches to acceptance sampling in the literature. The first approach is *acceptance sampling by attribute*, in which the product is specified as defective or non-defective based on a certain cutoff point. The number of defectives is used as the test statistic. This is the same as the binomial test.

The other type is *acceptance sampling by variables*, where the value of the measurements is taken into account when calculating the test statistic. There are two approaches to the problem. One is to calculate, given the probability model, the area under the probability density that is more extreme than the standard and evaluate if this exceeds the desired impairment probability (in our case 10 per cent).

The second approach is to calculate the value of a parameter, typically the mean, that results in the distribution having a standard with associated probability of 10 per cent. The mean of the observed data is then compared with the derived mean. As the method uses actual values, it may be more informative. This is reflected in the fact that acceptance sampling by variables requires a smaller sample size than the attributes plans to have the same operating characteristic curve (OC), i.e. the probability of accepting a lot of items as a function of the proportion of defectives. However, in acceptance sampling by variables the distribution of the measurements must be known.

Since World War II, variables-acceptance sampling plans have been in use, although primarily in an industrial quality control context. Under the assumption that the characteristic being measured is normally distributed, many plans have been developed. Lieberman and Resnikoff (1955) developed an OC-matching collection of plans based on known variance, unknown variance and average range. These plans were indexed by the acceptable quality level (AQL, i.e. producer's risk). Owen (1966) introduced one-sided variables sampling plans that are indexed by the AQL as well as the lot tolerance percent defective (LTPD, i.e. consumer's risk) for the unknown variance case. A year later, Owen (1967), developed a variable sampling plan for two specification levels. Variables-acceptance sampling plans were covered in many books. Among these are: Bowker and Goode (1952), Duncan (1974), Guenther (1977), Grant and Leavenworth (1980), and Schilling (1982).

Plans for non-normal distributions have been considered in the literature, as long as the form of the underlying distribution is known and the proportion of defectives could be calculated just from knowing the parameters of the distribution, usually the mean. Srivastava (1961) and Mitra and Das (1964) investigated the effect of non-normality for one-sided plans. Zimmer and Burr (1963) developed variables sampling plans using measures of skewness and kurtosis to adjust for non-normality. Guenther (1972) introduced variables sampling plans when the underlying distribution is Poisson or binomial. Duncan (1974), in his book, developed a plan under the assumption that the distribution of the observations is a member of the Type III Pearson family, which in its standardized form depends only on the coefficient of skewness. Guenther (1977), in his book, considered the exponential family as the underlying distribution. Lam (1994) developed Bayesian variables sampling plans for the exponential distribution with Type I censoring. Suresh and Ramanathan (1997) derived a plan for symmetric underlying distributions. Plans for autocorrelated observations are described in Hapuarachchi and Macpherson (1992).

In this article, the variables acceptance approach is used to study the problem of listing sites under the Clean Water Act. Methods based on use of actual values are shown to have error rates superior to methods based on binary information. Positive autocorrelation in data is a potential problem and is shown to increase the Type I error rates. Two methods for adjusting for autocorrelation are presented and error rates are calculated for these methods. Estimates of sample size are given that will achieve specific error rates for independent and autocorrelated situations. The approach taken is similar to that in Barnett and Bown (2002), who develop a testing approach for composite samples.

2. VARIABLES PLAN FOR THE NORMAL DISTRIBUTION

A variety of tests are discussed in the acceptance sampling literature for normally distributed data. All involve calculation of a critical value for a test based on the mean and are summarized in Table 1. Owen (1966, 1967) (see also Duncan, 1974) presented the acceptance sampling by variables technique when the distribution of the measurements is normal with either known or unknown standard deviation. This section summarizes that material in the context of environmental sampling.

Table 1. Cutoffs for different methods of testing $H_0 : p \leq p_0$ versus $H_1 : p > p_0$ using the test statistic for a lower standard L of the form $test = (\bar{x} - L)/s$ or a test statistic of the form $test = (U - \bar{x})/s$ for an upper standard. The value in the table is the cutoff k that defines the rejection region. For example, with a lower standard we reject if $test = (\bar{x} - L)/s < k$

Method	Plotting acronym	Cutoff
Normal, variance known		$k = \frac{z_\alpha}{\sqrt{n}} - z_{p0}$
Wallis approximation	Wallis	$k = \frac{z_\alpha\sqrt{2nz_{p0}^2 + 4n - 2z_\alpha^2} - 2nz_{p0}}{2n - z_\alpha^2}$
Normal, variance unknown (non-central t)	NCT	$k = \frac{t_{n-1,\lambda,\alpha}}{\sqrt{n}}$ where $\lambda = -\sqrt{n}z_{p_0}$
AR(1) using normal approximation (Hapuarachchi plan)	HAP	$k = z_\alpha\sqrt{\frac{1+\theta}{(1-\theta)n}} - z_{p0}$
AR(1) using non-central t	AR1	$k = \frac{t_{n-1,\lambda,\alpha}}{\sqrt{n\frac{(1-\theta)}{(1+\theta)}}}$ where $\lambda = -\sqrt{n\frac{(1-\theta)}{(1+\theta)}}z_{p0}$

Suppose that the environmental measurements have a normal distribution with mean μ and standard deviation σ. The criterion is assumed to be some upper (lower) specified limit $U(L)$. Then the proportion of defective items p can be calculated as $p_U = \Phi[(\mu - U)/\sigma]$, where Φ is the cumulative distribution function for a standard normal random variable (analogously $p_L = \Phi[(L - \mu)/\sigma]$). Specifying μ, σ and U will specify p. Moreover, if σ is constant, then p will depend only on U and μ. In this case, instead of testing hypotheses about p we can rather evaluate hypotheses about μ and then base our analysis on the sample mean $\bar{x}$. Therefore, to test $H_0 : p \leq p_0$ *(no impairment, don't list)* versus $H_1 : p > p_0$ *(impairment, list)*, we can proceed in two ways. First, if there is a single upper limit U, use the sample mean to compute $z_U = (U - \bar{x})/\sigma$ and then reject H_0 if $z_U \leq k$, where $k = z_\alpha/\sqrt{n} - z_{p0}$. In the second procedure, the proportion of impaired samples is estimated by

$$\hat{p}_u = \Phi\left[\left(\frac{\bar{x} - U}{\sigma}\right)\sqrt{\frac{n}{n-1}}\right]$$

and the null hypothesis is rejected if $\hat{p}_u > m$, where $m = \Phi\left(k\sqrt{n/(n-1)}\right)$. It is notable that the second procedure may be preferable since it gives a decision based on the probability of a defective item. If we have a lower standard instead of an upper one, then $z_L = (\bar{x} - L)/\sigma$ and k will be the same. The estimated proportion of impaired sites is

$$\hat{p}_L = \Phi\left[\left(\frac{L - \bar{x}}{\sigma}\right)\sqrt{\frac{n}{n-1}}\right]$$

Owen also presented an analysis for the case of double specification limits, i.e. lower and upper limits.

2.1. *Normal distribution with unknown variance*

Lieberman and Resnikoff (1955), developed acceptance sampling plans for data from a normal distribution with unknown variance. A reasonable test statistic for the hypothesis using a lower limit is

$$t = \frac{\bar{x} - L}{s/\sqrt{n}} \tag{1}$$

which has a non-central t distribution with $n-1$ degrees of freedom and non-centrality parameter $\lambda = -z_p\sqrt{n}$. The hypothesis is rejected if $t < t_{n-1,\lambda,\alpha}$. The power curve could be obtained by calculating $\Pr(t < \sqrt{n}k) = \Pr(t < t_{n-1,\lambda,1-\beta})$ for specified values of μ or p and Type II error β. Here $k = t_{n-1,\lambda,\alpha}/\sqrt{n}$. An alternative is to use

$$t = \frac{\bar{x} - L}{s} \tag{2}$$

The hypothesis is rejected if $t < k$.

Wallis (1947) suggested an alternative approach. This approach is approximate and was designed to avoid the use of the non-central t distribution. The test statistic is still the same and the value of k is given by

$$k = \frac{z_\alpha\sqrt{2nz_{p0}^2 + 4n - 2z_\alpha^2} - 2nz_{p0}}{2n - z_\alpha^2} \tag{3}$$

2.2. *Tests with autocorrelated data*

A potentially important problem in using measurements collected over time is the possibility that the measurements are correlated. Hapuarachchi and Macpherson (1992) studied the effect of serial correlation on acceptance sampling plans by variables assuming the measurements follow an autoregressive of order p [AR(p)] process. They assumed that the t^{th} measurement x_t is modeled as

$$x_t = y_t + \mu, \qquad y_t = \theta_1 y_{t-1} + \theta_2 y_{t-2} + \cdots + \theta_p y_{t-p} + e_t \tag{4}$$

with μ a constant (usually the mean of the process) and the e_t's assumed to be independently normally distributed with mean of zero and variance of σ_e^2. In the model (4), y_t is said to follow an AR(p) process. A special case is the AR(1) process specified by

$$x_t = y_t + \mu, \qquad y_t = \theta_1 y_{t-1} + e_t \tag{5}$$

This model is more common to water quality issues.

For an AR(1) process it can be shown that the approximate sample variance of the mean is (Darken *et al.*, 2000)

$$\sigma_{\bar{x}}^2 = \left(\frac{1+\theta_1}{1-\theta_1}\right)\frac{\sigma^2}{n} \tag{6}$$

An approximate test for $\mathrm{H}_0 : \mu \geq \mu_0$ versus $\mathrm{H}_1 : \mu < \mu_0$ would again be based on the t statistic. The critical value k would be based on a non-central t distribution with $n-1$ degrees of freedom and non-centrality parameter $\lambda = -z_p\sqrt{[n(1-\hat{\theta})]/(1+\hat{\theta})}$, where $\hat{\theta}$ is the estimated autocorrelation. So, we are using the same test statistic when assuming independence; however, the cutoff point of the test is modified to take the dependency among the measurements into account. Note that the distribution of the test statistic depends on the autoregressive parameter θ.

Also note that, with a positive autocorrelation structure, the non-centrality parameter and hence the cutoff point is less than the independence cutoff point. Thus, rejection is not as easy as it is in the independence case.

3. SAMPLE SIZE

One use of the above methodology is to set up a sampling plan that attempts to control the error rates. In particular suppose that it is desired to set up a variables plan which has a probability α of falsely declaring a site impaired for a site with an acceptable proportion p_0 of measurements actually more extreme than the standard and a probability β for falsely not listing an impaired site that has an unacceptable proportion p_1 of measurements more extreme than the standard. Typically, p_1 is a value chosen to be a worst case in the sense that we want a site listed with high probability if this proportion of measurements does not meet the water quality standard. Under the first procedure (normal distribution with known variance) with an upper standard we will not declare impairment if $Z_u = [(U - \bar{x})/\sigma] \geq k$. It can be shown (Duncan, 1974) that the suitable sample size n is given by

$$n \geq \left(\frac{z_\alpha + z_\beta}{z_{p_0} - z_{p_1}}\right)^2 \tag{7}$$

where $z_\alpha, z_\beta, z_{p_0}$ and z_{p_1} represent respectively the α, β, p_0 and p_1 quantiles of the standard normal distribution.

3.1. Sample size with unknown variance

When the standard deviation is unknown, almost the same procedure for the known standard deviation is used to calculate sample size except that we replace the population standard deviation σ by the sample standard deviation s. Three methods are available to derive a variables plan in this case. In the first method, special formulas for finding n and k, according to certain values of p_0, p_1, α and β, are introduced by Wallis (1947). They are

$$k = \frac{z_\alpha z_{p1} + z_\beta z_{p0}}{z_\alpha + z_\beta} \tag{8}$$

and

$$n \geq \left(1 + \frac{k^2}{2}\right)\left(\frac{z_\alpha + z_\beta}{z_{p0} - z_{p1}}\right)^2 \tag{9}$$

These formulas are based on the assumption that $\bar{x} \pm ks$ is approximately normally distributed with a mean of $\mu \pm k\sigma$ and variance $\sigma^2(1/n + k^2/2n)$.

Table 2. Estimated lower values ($\hat{L}$) and cutoffs for test for impairment based on different statistics and Type I error rates for dissolved oxygen data. If the lower value is smaller than 5.0, this indicates listing the site. The autocorrelation was estimated to be 0.6699 and the calculated test statistic is 1.135

Method	Type I error rate	Cutoff (k)	Estimated lower value ($\hat{L}$)
t	0.05	0.891	5.44
Normal		0.885	5.45
AR(1)		0.522	6.10
t	0.10	0.972	5.29
Normal		0.965	5.31
AR(1)		0.687	5.80
t	0.20	1.076	5.11
Normal		1.066	5.12
AR(1)		0.890	5.44

The second method is from a monograph due to L. J. Jacobson. Using the Jacobson monograph and specified values of p_0, p_1, α and β, we can calculate n and k. The Jacobson approach is presented in Duncan (1974, Ch. 12, p. 271). However, the results obtained by this method are not precise and Owen mentioned a third method that gives more precise results.

In the third method, the analysis is based on the exact distribution rather than the approximate distribution proposed by Wallis. The exact distribution when using the sample standard deviation instead of the population standard deviation is the non-central t distribution with $n-1$ degrees of freedom. To find n and k in this case, a special table specifically designed for this purpose has been prepared by Owen (1963, Table 2). Lieberman and Resnikoff (1955) provide special tables to get $\hat{p}_u$ and m according to a given sample size n and accepted quality level (i.e. p_0). These tables are found in (Lieberman and Resnikoff, 1955). Another set of tables was produced in Statistical Research Group (1947, pp. 22–25).

Hapuarachchi and Macpherson (1992) describe several approaches for estimating sample size when data are autocorrelated. In the case of unknown variance, the authors note that the distributions of needed quantities are difficult to obtain. For the AR(1) case, they suggest using an approximate sample size given by

$$n \approx \left(\frac{z_\alpha + z_\beta}{z_{p_0} - z_{p_1}}\right)^2 \left(\frac{1+\hat{\theta}}{1-\hat{\theta}}\right) \tag{10}$$

Table 3. Sample sizes calculated to achieve pre-determined error rates for different methods with autocorrelation set to zero. Values were rounded up to the nearest integer

Type I	Type II	Normal (σ known)	Wallis approx.	Non-central t
0.05	0.05	30	58	44
	0.1	24	44	34
	0.2	17	30	24
0.1	0.05	24	48	36
	0.1	18	35	27
	0.2	13	23	18
0.2	0.05	17	37	27
	0.1	13	26	20
	0.2	8	16	12

Table 4. Sample sizes calculated to achieve pre-determined error rates for different methods with varying autocorrelation. Values were rounded up to the nearest integer. The Type I and II error rates were both set to 0.2

Autocorrelation	HAP	NCT AR(1)
0	8	12
0.1	10	14
0.2	12	16
0.3	15	19
0.4	18	22
0.5	24	27
0.6	31	35
0.7	44	48
0.8	70	73
0.9	147	150

which is just the normal based approach (Equation 7) adjusted for autocorrelation. One might also adapt the sample sizes calculated using one of the other methods for the correlation by multiplying by a factor $[(1+\hat{\theta})/(1-\hat{\theta})]$. Another approach would use the non-central t distribution with the adjustment.

4. COMPARISON OF METHODS

Given the variety of methods a comparison is warranted. It is possible to analytically compare some of the methods under certain conditions. In situations where the error rates could be exactly calculated for all methods (Figures 2 and 3) we used the analytical approach. When analytical solutions were not possible for all the methods we relied on a simulation for the comparisons. In the simulations, normally distributed data were generated with different levels of autocorrelation ($\theta = 0.0$, 0.25, 0.5,

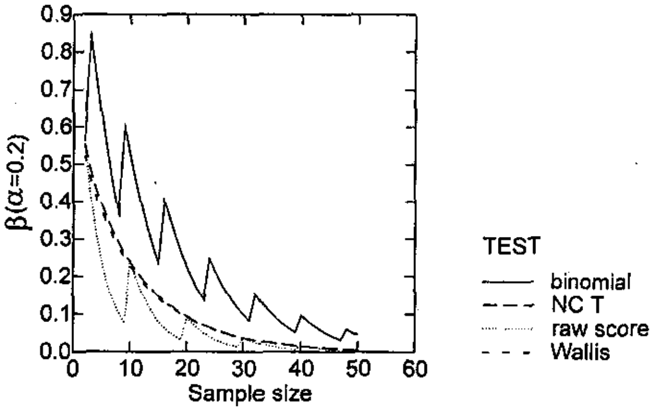

Figure 3. Plot of probability of not listing an impaired site (β) using binomial, raw score and acceptance sampling methods. Note that the probability an unimpaired site is listed (α) is set to 0.2 for the methods. This error rate is not set for the raw score method. Values of this error rate are given in Figure 2. The parameters of the normal distribution were set so the probability that a measurement was more extreme than the standard. The binomial probability of impairment was set to 0.25

0.75) and evaluated using cutoff points associated with different Type I error rates ($\alpha = 0.05, 0.1, 0.2$). For estimation of Type I error rates the site was assumed to meet standards ($p_0 = 0.1$), while for the estimation of Type II error rates the site was assumed to not meet standards ($p_1 = 0.25$). For estimation of the error rate for a given case, 10 000 times series were generated with sample sizes ranging from 5 to 50. A 'burn-in' of 20 points was used to stabilize the time series (e.g. for the case $N = 5$, 25 observations were generated but only the last 5 were used). For purposes of space just the results for the cases $\alpha = 0.2$ and $\theta = 0.0$ and 0.5 are discussed.

Figure 2 displays the Type I error rates for the binomial and raw score methods with no autocorrelation. What is obvious from this display is that the binomial method bounds the Type I error by choice of cutoff to be below 0.20. The raw score method results in a rather large Type I error. The raw score method may be viewed as a binomial method with changing Type I error rate. Alternatively the binomial method may be viewed as a raw score method with varying cut point. The normal based methods with no autocorrelation are not presented as the error rate is set to a fixed value.

Figure 3 displays the Type II error rate for the two binary methods, the non-central *t* approach and Wallis' approach with at Type I error set at 0.1. The normal case with known variance was not considered. The Type II error rate is large for the binomial method relative to the raw score. This result makes sense since the raw score has a high Type I error rate and hence a low cutoff relative to the binomial. Therefore it will be more powerful. The normal-based approaches using the actual data produce Type II error rates similar to the raw score method. Thus if the data are consistent with the normal assumption, there is much to gain from using the normal based methods as the Type I error rate can be fixed. The method is consistent with the low Type I error rates for the binomial and has similar Type II error rates to the raw score. We note that there is little difference in terms of error rates for the non-central *t* and Wallis' method.

Figures 4 and 5 plots this error rate for the tests when the autocorrelation is 0.0 and 0.5. When the correlation is zero (Figure 4), we find that the Type I error rates are as expected with the exception of the approximation proposed by Hapuarachchi and Macpherson. For this method, the error rate is higher than expected. The inflation in the error rate is due to the use of the estimated standard deviation rather than the actual standard deviation in the test statistic. Type II error rates for the different methods based on the normal model yield better error rates than either the binomial or raw score approach. These results indicate that if the assumption of normality is met, there is considerable advantage to using a parametric method.

When the autocorrelation is 0.5, we find that not accounting for positive autocorrelation results in an increase in the Type I error rate (Figure 5) for most of the procedures. The error rates for the non-central *t*, Wallis and binomial methods exceed the preset level of 0.2. The error rate for the non-central *t* is close to the nominal value for small sample sizes but increases with sample size. Error rates for the autocorrelation-adjusted tests are less than the desired level for small sample sizes but approach the level as the sample size increases. Also note that the maximum error rate for the raw score method is slightly decreased over that of Figure 4. When Type II error rates are considered, we find that the autocorrelation increases the error rate relative to the uncorrelated case (compare Figures 4b and 5b). The Type II error rates for the AR(1) non-central *t* and the method of Hapuarachchi and Macpherson are higher than those of the non-central *t* and Wallis' method. Relative to the binomial, the autocorrelation-adjusted tests generally have smaller error rates.

Also apparent from the figures is that the Type II error rates will be large for small sample sizes. When the data are not correlated, samples of size 10 might yield reasonable results if one would be willing to balance Type I and Type II error rates at around 0.2. In the case of correlated data, additional

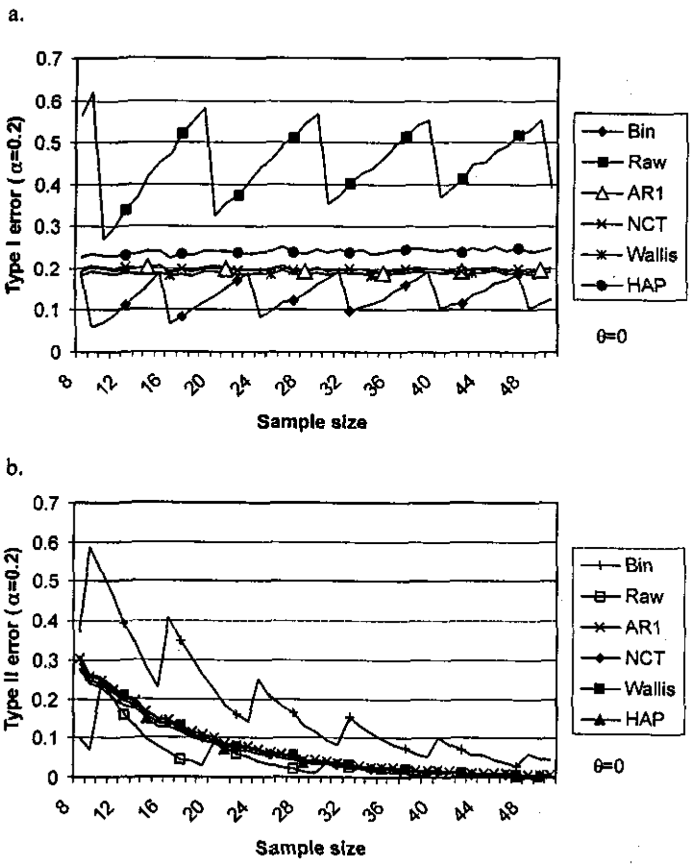

Figure 4. (a) Simulated actual Type I error rates for different methods using $\alpha = 0.2$ and autocorrelation of 0.0; (b) Type II error rates using $\alpha = 0.2$

samples would be required to produce the same error rates and a sample size of 20 is better although the number of additional samples depends on the size of the autocorrelation.

5. EXAMPLE

As an example we consider monthly data collected over a two-year period on dissolved oxygen. The values are plotted in Figure 6. Interest is achieving a (lower) standard of 5.0 with $p_0 = 0.1$. The sample mean is $\bar{x} = 7.03$ with standard deviation $s = 1.787$, $n = 24$ and $\alpha = 0.05$. The calculated value of the test statistic is 1.135. Using a non-central t distribution we find the critical value $k = 0.890$. So, one does not reject the null hypothesis at $\alpha = 0.05$ and the site would not be listed. Assuming an AR(1) process, the autocorrelation coefficient is calculated as 0.6699. We find $k = 0.521$. Using the adjusted non-central t statistics, the decision would be to not reject the null hypothesis at $\alpha = 0.05$. An alternative approach is to base results on the estimated limit. Table 2 provides a summary of three different tests at different α levels. In all cases, the estimated lower value $(\hat{L})$ is above 5.0 so one would not reject.

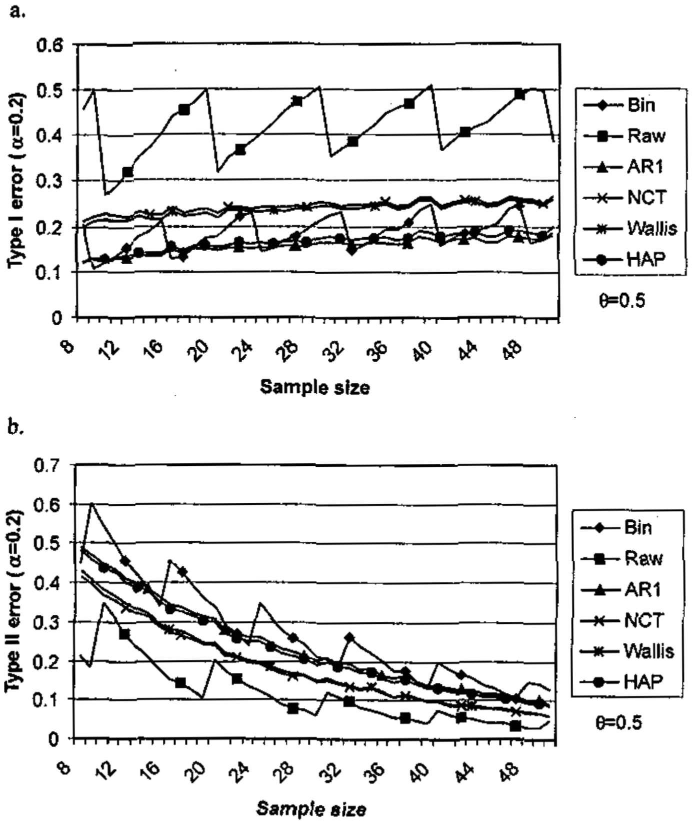

Figure 5. (a) Simulated Type I error rates using $\alpha = 0.2$ for autocorrelated data with correlation of 0.5; (b) simulated Type II error rates using $\alpha = 0.2$ for autocorrelated data with correlation of 0.5

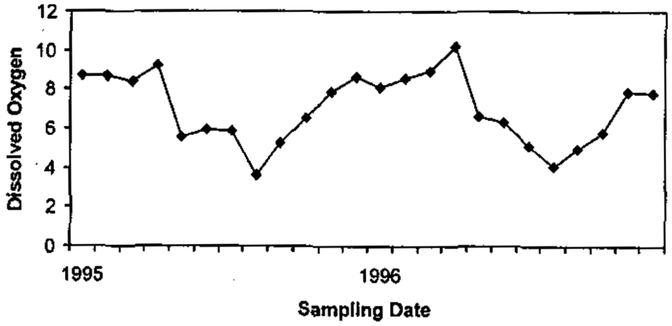

Figure 6. Plot of 24 observations on dissolved oxygen for a two-year period from 1995 to 1997

Sample size calculations are presented in Tables 3 and 4. Table 3 provides different sample sizes for varying Types I and II error rates with autocorrelation set to zero. For most cases, sample sizes are greater than those achieved by quarterly sampling in a two-year period. Unless error rates are set to be large (0.2), relatively large sample sizes are required. Table 4 gives estimates of sample sizes required under autocorrelation with Types I and II error rates set to 0.2. As indicated in the table, as the autocorrelation increases, so does the required sample size. With correlations that might be expected in water studies (0.5), roughly 6 years of quarterly data would be required.

6. DISCUSSION

The investigation of error rates reveals that the methods currently in use may be improved by adopting approaches based on acceptance sampling by variables. In particular, methods based on the non-central t distribution produce error rates better than the methods based on discretized data. When measurements are made over time correlation is to be expected and adjustments to the test are available. Positive autocorrelation effectively reduces the sample size. Current protocols use around eight observations over a two-year period to make inferences. Based on the analysis of error rates and dissolved oxygen data, the sample size seems inadequate to make strong inferences, even if error rates around 0.2 are desired.

Our approach is based on testing and the use of methods from acceptance sampling. Another approach may be based on the use of tolerance intervals (Gibbons, 1994; Hahn and Meeker, 1991). The approach based on tolerance intervals using a non-central t distribution is equivalent to the acceptance sampling approach using the non-central t. The tolerance limit approach was suggested by Lin *et al.* (2000) for use with 303(d) listing. Smith (2002) describes a generalized approach in which samples from different times or spatial locations are used with a tolerance interval for setting standards.

Many of decisions made about water quality are based on small sample sizes. For example, if it is required to produce a report every two years, the evaluation of water quality might be based on quarterly sampling, yielding eight observations. In cases with small sample sizes decisions may be affected by variation due to estimation of variance. One approach for improving estimation of variance is to use data from alternate sources. For example, there may be several locations within a watershed. The information from the multiple sites may be pooled to produce a variance estimate. Estimation of variance may also use a Bayesian approach.

The above literature does not cover multivariate data sets. There is not much work done for the multivariate setup. The first work published on this problem was of Baillie (1987a,b). He developed direct multivariate generalizations of the variables-acceptance sampling procedures under the assumption of multivariate normal distribution. A few years later, Hamilton and Lesperance (1991) dropped the assumption of multivariate-normality and proposed a method to deal with multivariate data using Wallis' (1947) advice. They transformed the bivariate data into one normal variable, and then applied the univariate techniques to their new variable.

ACKNOWLEDGEMENTS

This research has been supported through grants from the Virginia Water Resources Research Center and U.S. Environmental Protection Agency's Science to Achieve Results (STAR) (grant no. R82795301). Although the research described in the article has been funded wholly or in part by the U.S. Environmental Protection Agency's STAR programs, it has not been subjected to any EPA review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

REFERENCES

Baillie DH. 1987a. Multivariate acceptance sampling: some applications to defense procurement. *Statistician* **36**: 465–478.

Baillie DH. 1987b. Multivariate acceptance sampling. In *Frontiers in Statistical Quality Control 3*, Lenz HJ, Wetherill GG, Wilrich PT (eds). Physica Verlag: Wüerzbirg, Vienna.

Barnett V, Bown M. 2002. Statistically meaningful standards for contaminated sites using composite sampling. *Environmetrics* **13**: 1–14.

Bowker AH, Goode HP. 1952. *Sampling Inspection by Variables*. McGraw-Hill: New York.

Darken PF, Holzman GI, Smith EP, Zipper CE. 2000. Detecting changes in trends in water quality using modified Kendall's tau. *Environmetrics* **11**: 423–434.

Duncan AJ. 1974. *Quality Control and Industrial Statistics*. Irwin: Homewood, IL.

Gibbons RD. 1994. *Statistical Methods for Groundwater Monitoring*. John Wiley & Sons: New York.

Grant EL, Leavenworth RS. 1980. *Statistical Quality Control*, 5th edition. McGraw-Hill: New York.

Guenther WC. 1972. Variables sampling plans for the Poisson and Binomial. *Statistica Neerlandica* **26**: 17–24.

Guenther WC. 1977. *Sampling Inspection in Statistical Quality Control*. Macmillan Publishing Co., Inc: New York.

Hahn GJ, Meeker WQ. 1991. *Statistical Intervals: A Guide for Practitioners*. John Wiley & Sons: New York.

Hamilton DC, Lesperance ML. 1991. A consulting problem involving bivariate acceptance sampling by variables. *The Canadian Journal of Statistics* **19**: 109–117.

Hapuarachchi KP, Macpherson BD. 1992. Autoregressive processes applied to acceptance sampling by variables. *Communications in Statistics: Simulation* **21**(3): 833–848.

Lam Y. 1994. Bayesian variables sampling plans for the exponential distribution with Type I censoring. *Annals of Statistics* **22**: 696–711.

Lieberman GJ, Resnikoff GJ. 1955. Sampling plans for inspection by variables. *Journal of the American Statistical Association* **50**: 457–516.

Lin P, Meeter D, Niu X. 2000. A nonparametric procedure for listing and delisting impaired waters based on criterion exceedences. *Technical Report* submitted to the Florida Department of Environmental Protection.

McBride GB, Ellis JC. 2001. Confidence of compliance: a Bayesian approach for percentile standards. *Water Research* **35**: 1117–1124.

Mitra SK, Das NG. 1964. The effect of non-normality on sampling inspection. *Sankhya* **26**(A): 169–176.

Owen DB. 1963. *Factors for one-sided tolerance limits and for variables sampling plans*. Monograph No. SCR-607, Sandia Corporation.

Owen DB. 1966. One-sided variables sampling plans. *Industrial Quality Control* **22**: 450–456.

Owen DB. 1967. Variables sampling plans based on the normal distribution. *Technometrics* **9**: 417–423.

Schilling EG. 1982. *Acceptance Sampling in Quality Control*. Marcel Dekker: New York.

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

Smith R. 2002. The use of random-model tolerance intervals in environmental monitoring and regulation. *Journal of Agricultural, Biological, and Environmental Statistics* **7**: 74–94.

Srivastava ABL. 1961. Variables sampling inspection for non-normal samples. *Journal of Science and Engineering Research* **5**: 145–152.

Statistical Research Group, Columbia University. 1947. *Techniques of Statistical Analysis*. McGraw-Hill Book Co., Inc: New York.

Suresh RP, Ramanathan TV. 1997. Acceptance sampling plans by variables for a class of symmetric distributions. *Communications in Statistics: Simulation* **26**: 1379–1391.

Wallis WA. 1947. Use of variables in acceptance inspection for percent defective. In *Selected Techniques of Statistical Analysis for Scientific and Industrial Research and Production and Management Engineering*. McGraw Hill: New York; 3–93.

Zimmer WJ, Burr IW. 1963. Variables sampling plans based on non-normal populations. *Industrial Quality Control* **21**: 18–26.