Error Analysis
for Water Quality Policy 1-11:
Pertaining to the
Water Quality Assessment
December 2016

Purpose of the Error Analysis

Water Quality Policy 1-11 is the policy that guides listing decisions for Washington's Water Quality Assessment to meet Clean Water Act requirements for sections 303(d) and 305(b). During updates to Policy 1-11 in 2012, several comments were received about the risk of listing errors for waters that were placed on the 303(d) list based on limited data.

The comments related to this issue fell into three areas:

1. Concerns that the Policy 1-11 guidance does not minimize false positives (which result in unnecessary TMDL costs), or false negatives (which result in continued environmental degradation). Some commenters suggested using the binomial distribution statistical approach as a basis for determining impairment.

2. Concerns that the listing policy requires more data to move from Category 5 to Category 1 than to get listed in Category 5. Commenters felt that the assessment policy should require the same level of data to list areas in Category 1 as it does to determine the initial Category 5 impairment.

3. Concerns that Ecology should reconsider use of an instantaneous single grab-sample value to represent an average value, especially a four-day average (for example chronic aquatic life criteria for metals), which could lead to listings where there is no demonstrated exceedance of the water quality standard.

Ecology responded to the above issues in the response to public comments on the 2012 Policy 1-11 update and committed to conducting a "Type I and Type II Error Analysis" to provide information on the risk of false positives (Type I error) and false negatives (Type II error) relating to guidance in Policy 1-11.

The Water Quality Program requested that Environmental Assessment Program (EAP) technical staff analyze the risk of error associated with the three types of comments listed above. The analyses required the establishment of simplifying assumptions due to the underlying environmental and policy complexities. Therefore the conclusions of the analyses must be considered in this limited context. Nonetheless, the results are being shared as a basis for further discussion and review on potential revisions to improve Policy 1-11. The analyses can be found in the following three chapters attached:

1. Type I and Type II error probabilities for pH, temperature, and dissolved oxygen listings
2. Unequal data requirements for Category 5 and Category 1
3. Use of instantaneous measurements to represent multi-day averages associated with chronic criteria for toxic parameters

Each of the three issues and the respective analyses are discussed below.

Type I and Type II Error Probabilities

Background

In the 2002-2004 Assessment listing cycle, Ecology used a binomial distribution method in an effort to minimize false positives. Unfortunately, the approach did not work uniformly among different types of pollutant parameters and resulted in significant inconsistencies. EPA's 2006 Integrated Report guidance[1] states that when the percent threshold of a pollutant is clearly expressed in the water quality criteria (such as the geometric mean and 10 percent exceedance rule for bacteria) then the methodology written in the criteria should be used. Hence, Ecology discontinued use of the binomial method for the next listing cycle (2006-2008). EPA and others supported removal of this methodology from our listing process because the Type II error rate for small sample sizes is higher for the binomial distribution in comparison to the EPA "raw score" method[2], and therefore it was believed that the binomial method would not be protective enough given the high frequency of small datasets for waterbodies across the state.

To better deal with specific parameter characteristics in the 2006 Policy 1-11 revisions, Section 8 of the Policy was created to include specific listing methodologies based on the different pollutant parameters. The binomial distribution method currently is not used for any parameter in the Water Quality Assessment process. The ten percent exceedance rate for listing as suggested in EPA guidance for assessing several conventional parameters under the Aquatic Life Use criteria is used as well as a requirement that a minimum of three exceedances be observed before placing a waterbody on the 303(d) list.

[1] *2006 integrated report guidance.* Washington, DC: U.S. Environmental Protection Agency. Available: http://water.epa.gov/lawsregs/lawsguidance/cwa/tmdl/2006IRG_index.cfm.

[2] The EPA 2002 guidance suggested a simple rule which is called the "10% rule" or, in some publications, the "raw scores" method. This method is to determine that a waterbody is impaired if 10% or more of the sample measurements exceed the applicable water quality standard.

Basis for Analysis and Underlying Assumptions

This analysis focused on the probability of listing waters based on meeting a two-part decision rule for "grab sample" or instantaneous measurements of pH, temperature, and dissolved oxygen as specified in Policy 1-11. Category 5 requires: (a) at least one year within the past 10 years with at least 10% of samples (set of measurements within a calendar or water year) exceeding standards and (b) at least three exceedances (measurements exceeding standards) over the most recent 10-year period.

The first part of the above decision rule for listing comes from the U.S. Environmental Protection Agency (EPA) 2002 guidance that suggested a simple rule which is called the "10% rule" or, in some publications, the "raw scores" method. The "10% rule" method is to determine that a waterbody is impaired if 10% or more of the sample measurements exceed the applicable water quality standard. More than one year's worth of data is treated as a single dataset.

EPA has stated that no true exceedances of a criterion are allowable, unless one can show that human activities did not cause or contribute to the exceedance. Furthermore, when numeric criteria contain a built-in frequency of exceedance component, as is the case with dissolved oxygen and temperature criteria in Washington State, then the evaluation of compliance with the criteria cannot use a different frequency of exceedance. The only "allowable exceedances" are those attributable to error in measurement, analysis, and reporting. EPA's "raw score" method indicates that 10% of the samples may be attributed to error. **It is important to note that this is entirely different than allowing 10% of the population to exceed a criterion.**

The first part of the decision rule in Policy 1-11 is similar to the "raw scores" method but assumes each year (calendar year or water year, depending on the parameter) is independent of the others and that the exceedances counter starts again at zero at the beginning of the parameter year.

Because the Type I error rate of the "raw scores" method is both high (especially for small sample sizes) and uncontrolled, Policy 1-11 included the second part of the decision rule in an effort to reduce the Type I error rate.

A complicating factor is that the number of samples (total or per year) to be used for the Water Quality Analysis is not known in advance. To determine the number of ways that exceedances can be observed for the likely case in which the number of samples varies by year would require advanced mathematics beyond what was feasible for this analysis. However, it was possible to calculate listing probabilities for the special case in which the sample size and population proportion are the same for all ten years.

For this analysis, a general equation for the probability of listing based on the two-part decision rule was derived, as well as the "safety margin" resulting from the addition of the second part of the decision rule. Listing probabilities were calculated for a range of population proportions out of compliance and for all sample sizes from 1 per year to 1 per day. From the listing probabilities, it is possible to calculate the theoretical Type I and Type II error rates.

The treatment of sampling years as independent in order to mitigate against the effects of extraordinary conditions such as drought is another complication of the policy. The effects of this assumption of independence have not yet been studied.

Conclusions

Some generalizations about the Type I and II error rates for the current two-part decision rule can be made.

- The current requirement for at least three exceedances to list a waterbody as impaired affects only cases in which the sample size for all years is 20 or less, and effectively only for sample sizes 10 or less. This requirement strongly reduces the chances of a Type I error for very small sample sizes (e.g. fewer than 10 samples per year) when compared to the "10% rule", but strongly increases the risk of Type II error

- The closer the actual proportion out of compliance is to the hypothesized population proportion, the higher the Type I error rate will be, because it is more difficult to tell if the observed proportion is different than the hypothesized proportion. For example, if testing the hypothesis that the population exceedance rate is 10% or more, then the Type I error rate (false

positive – determining that the waterbody is impaired when it is in fact not impaired) for the two-part decision rule is always high (e.g. >20%) above a samples size of 20.

- Type II error rates are primarily a function of sample size and the size of the effect that one is seeking to detect. The Type II error rate (false negative– determining that the waterbody is not impaired when it is in fact impaired) for the two-part decision rule is high at low samples sizes and declines rapidly as sample sizes increase.

See Appendix 1: Type I and Type II Error Probabilities for pH, Temperature, and Dissolved Oxygen Listings.

Unequal Data Requirements for Category 5 and Category 1

Background

Policy 1-11 includes guidance for using data to place waterbody segments into the 5 categories for the different pollutant parameters in Section 8 of the policy. Determining that a waterbody is not meeting standards requires much less monitoring data because relatively few measurements can provide a high degree of statistical confidence that criteria are not being met. However, to determine that a waterbody is meeting standards requires much more data to confidently determine that a criterion is met under all conditions. Pollutants that are highly variable such as bacteria, or other parameters that naturally vary throughout the day and season such as temperature, dissolved oxygen, and pH, require a greater sampling effort and an appropriate sample design to show that the waterbody is meeting standards during the critical period typical of that waterbody. A lack of criteria exceedances alone in a dataset does not necessarily equate to meeting water quality standards. A waterbody may be in compliance with standards during specific times of a day, season, or outside of a critical period for a given condition but may not be in compliance at other times. For example, if a dissolved oxygen dataset for a waterbody contains 500 measurements collected between 10 a.m. and 6 p.m. and shows no criterion exceedances, one cannot conclude that dissolved oxygen criteria are being met because the dataset does not include measurements from the early morning when dissolved oxygen typically reaches its lowest point during the day. For this analysis EAP staff analyzed why, from a statistical perspective, the sample size required to "de-list" a waterbody is significantly higher than the sample size required to initially place a waterbody on the 303(d) list.

Basis for Analysis and Underlying Assumptions

The numbers of samples required for listing a waterbody as impaired and delisting a no-longer-impaired waterbody are different. An analogy to this process would be a medical diagnosis for cancer. It takes only a few tests to confirm the presence of cancer. After going through treatments, a number of tests over a long period of time are needed to confirm, to a high degree of confidence, that the cancer has been cured. The same applies to pollutants in the water – only a few samples are needed to confirm the presence; however, many more samples are needed to confirm that the pollutant no longer exists in the same waterbody.

The difference in the numbers of samples is explained by statistical theory. The most commonly used model for the occurrence of exceedances is a binomial probability distribution, which has two parameters, n = the sample size (number of measurements) and p = the true proportion of the population which is out of compliance. We cannot know p, but we can estimate it by dividing the observed number of exceedances into the number of samples. We can also calculate a confidence interval based on the number of exceedances found in the sample size. The hypothesis is tested using these values.

Conclusions

Appendix 2 demonstrates, from a statistical perspective, why the sample size required to delist a no-longer-impaired waterbody is significantly higher than the sample size required for listing an impaired waterbody. For example, when using an allowable exceedance rate of 10% for a population, the minimum combination of sample size and number of exceedances to be able to conclude with 95% confidence that a waterbody is impaired is if you had only two measurements and both of them were exceedances. On the other hand, it would take a minimum of 29 measurements and 0 exceedances to be able to say with 95% confidence that a waterbody is not impaired

The reason that a larger sample size is required to delist a no-longer-impaired waterbody than to list an impaired waterbody is a function of the hypotheses being tested, the statistical distribution type assumed for the population, and the statistical significance level used, as well as the mathematical characteristics of a ratio.

See Appendix 2: Unequal Data Requirements for Category 5 and Category 1.

Use of Instantaneous Measurements to Represent Multi-day Averages

Background

Comments were received that expressed concerns that Ecology should reconsider use of an instantaneous single grab-sample value to represent an average value, especially a four-day average (for example chronic aquatic life criteria for metals), which could lead to listings where there is no demonstrated exceedance of the water quality standard.

Basis for Analysis and Underlying Assumptions

Based on comments concerning the use of instantaneous samples to represent toxics substance criteria (TSC) exceedances for aquatic life, this analysis explored how representative a single “grab” sample is of multi-day averages of toxics contamination. The analysis addresses only a single aspect of the complex situation of 303(d)-listing criteria and the data available. Specifically, the analysis focused on whether single samples can be used to evaluate toxics contamination for which the criteria are based on 4-day running averages.

There were no actual datasets representing the two measurements to be able to work with. Therefore the general approach involved simulating hypothetical “observed” contaminant concentrations that corresponded to a waterbody **just meeting** the chronic water quality standard and determining how often the standard was not met. The basis for the simulation was the 1991 EPA technical guidance on derivation of acute and chronic water quality standards for toxic contaminants.

Large numbers of random values were generated from a probability distribution defined by the long-term average set at the chronic water quality standard for a given contaminant to represent single "grab" samples. Running averages of four single values for the entire sequence were calculated to represent "4-day running average" concentrations. The reason for using averages set at the standards is to simulate the worst-case scenario for waterbodies actually in compliance.

Both the individual "1-day" values and the "4-day average" values were compared to the chronic water quality standard for that particular contaminant, and the percent of the single and averaged values exceeding the standard was calculated. Such a simulation was repeated for many different toxic contaminant standards. Finally, the exceedance rates (percent exceedance) of the "1-day" and "4-day average" observations for the collection of all the contaminants simulated were statistically compared. It should be noted that the underlying assumptions of lognormality and coefficient of variation value have not been tested with real data; therefore, the results from the simulation are provisional. The simulation also did not take into consideration how the criteria were established in the first place, and so the lognormal distribution used in the simulation may not be the same distribution used to develop the standards. Several other caveats on the limitations of the simulation results are listed in Appendix 3.

This analysis necessitated the application of a simplifying assumption that the observed toxics values are relatively constant. Based on this assumption, the analysis shows how single samples have a much higher chance of exceeding a criterion than a 4 day average. However, toxic parameters in the environment often do not display a relatively constant distribution in time and space. Recent studies[1,2,3] have shown clear patterns of diel cycling (and therefore serial correlation) for certain metals and

metalloids in streams. If an underlying diel cycle in a parameter exists, then a single sample value may be higher or lower than a 4-day average depending on the time at which sampling occurs. The time of day at which a single sample of metals is collected can affect how representative it is of a 4-day average value since some metals appear to peak at night while others appear to peak during the afternoon. For example, if the concentration of a metal undergoing diel cycling tends to be lowest between 8am and 5pm when most sampling tends to occur, then a single sample would consistently be lower than the 4-day average concentration in the waterbody. If the concentration of a metal peaks between 8am and 5pm, then a single sample would consistently be greater than to the 4-day average concentration in the waterbody. Limitations to the application of statistical theory to toxics data must be recognized as we continue to assess compliance with water quality criteria despite not having a complete understanding of diel cycling in toxic parameters for different waterbody types within Washington State.

Conclusions

The model used in this analysis indicates that individual daily observations have a much greater chance of exceeding the chronic TSC than 4-day averages do. The exceedance rate is a function of the assumed lognormal percentile on which the long-term average (LTA) is based. On average:

- For LTAs based on 90th percentiles, 1-day observations are twice as likely to exceed the chronic standards as are the 4-day running averages.
- For LTAs based on 95th percentiles, the 1-day exceedance rate is almost three times that of 4-day running averages.
- For LTAs based on 99th percentiles, 1-day observations are more than seven times as likely as the 4-day running averages to exceed the chronic standards.

These results are based on a constant distributional model excluding autocorrelation. Real-world exceedance rates of single observations may differ due to autocorrelation and changing conditions.

[1]Nimick, D.A., Gammons, C.H., Parker, S.R., 2011, Diel biogeochemical processes and their effect on the aqueous chemistry of streams: A review. Chemical Geology, v. 283, p. 3-17.

[2]Nimick, D.A., Cleasby, T.E., McCleskey, R.B., 2005, Seasonality of diel cycles of dissolved metal concentrations in a Rocky Mountain stream. Environmental Geology, v. 47, p. 603-614.

[3]Nimick, D.A., Gammons, C.H., Cleasby, T.E, Madison, J.P., Skaar, D., Brick, C.M., 2003, Diel cycles in dissolved metal concentrations in streams: Occurrence and possible causes. Water Resources Research, v. 39, no. 9, citation no. 1247, doi:10.1029/WR001571.

See Appendix 3: Use of instantaneous Measurements to represent Multi-day Averages (such as chronic metals).

Appendices

(page purposely left blank)

APPENDIX 1

Type I and Type II Error Probabilities for pH, Temperature, and Dissolved Oxygen Listing Policy

Methods

The first task was to understand 303(d) listing policies, the second was to develop statistical models, and the third was to compute probabilities.

This exercise focused on only a single rule which affects the listing of waters for pH, temperature, and dissolved oxygen. That rule lists a waterbody as impaired based on two criteria (Ecology, 2012):

a) At least one year within the past 10 years with at least 10% of samples (set of measurements within a calendar or water year) exceeding standards.

b) At least three exceedances (measurements exceeding standards) over the most recent 10-year period.

Statistical model

The statistical model must be able to handle situations in which the number of samples is unknown until the data are received. For a given waterbody or waterbody segment, the number of measurements submitted may vary from none in a given year to essentially continuous, the latter boiled down to one per day, i.e., from 0 to 365 (366 for a leap-year). Furthermore, to mitigate against anomalies such as drought years, Ecology treats each year separately in water quality assessments.

For this analysis, it was not necessary to define what constitutes an exceedance. Rather, this analysis was concerned only with what to do once one has samples and exceedances. The definition of an exceedance varies by parameter; e.g., single grab samples vs. 4-day averages vs. 7-day maxima or minima, etc., and is a separate matter.

Nothing could be found in the primary or grey literature which addressed such a compound problem. The U.S. Environmental Protection Agency (EPA) promulgated a simple rule which is called the "10% rule" or, in some publications, the "raw scores" method. The "10% rule" method is to determine that a waterbody is impaired if 10% or more of the sample measurements exceed the applicable water quality standard (EPA, 2002). That same document did introduce the reader to the "binomial method" (treating the number of exceedances within a set of sample measurements as a binomial random variate), but relied on normal approximations (EPA, 2002). In addition, more than one year's worth of data is treated as a single dataset, not separate datasets for each year.

The first criterion in the above Ecology policy is similar to the "raw scores" method but **assumes** each year (calendar year or water year, depending on the parameter) is **independent** of the others and that the exceedances counter starts again at zero at the beginning of the parameter year.

Because the Type I error rate of the "raw scores" method is both high (especially for small sample sizes) and uncontrolled (Smith et al., 2001), for the 2014 Water Quality Assessment, Ecology added the second part of the policy (at least 3 exceedances) in an effort to reduce the Type I error rate.

A general equation for the probability of listing based on these two criteria was derived (details in Appendix 1A). From this equation, it is also possible to quantify the "safety margin" resulting from the addition of the second criterion.

Modeling 10 years of data

If exceedances observed in a year's worth of sampling are modeled as following a binomial distribution with fixed distributional parameters n (sample size) and p (proportion of the population out of compliance), ten years' worth of data would be modeled as the product of 10 independent binomial distributions, each with unique n_i and p_i, $i = 1, 2, \ldots, 10$.

In general, if X_i is a random variable for the number of exceedances in year i, $i = 1, 2, \ldots, 10$, then

$$P(list) = P(total\ \#of\ exceedances \geq 3\ AND\ \#\ exceedances \geq 10\%\ in\ at\ least\ one\ year)$$

$$= P\left(\sum_{i=1}^{10} X_i \geq 3\ AND\ at\ least\ one\ X_i \geq 10\%\ of\ n_i\right)$$

$$= 1 - P\left(\sum_{i=1}^{10} X_i \leq 2\ OR\ no\ X_i \geq 10\%\ of\ n_i\right)$$

$$= 1 - \begin{bmatrix} P\left(\sum_{i=1}^{10} X_i \leq 2\right) + P(no\ X_i \geq 10\%\ of\ n_i) \\ - P\left(\sum_{i=1}^{10} X_i \leq 2\ AND\ no\ X_i \geq 10\%\ of\ n_i\right) \end{bmatrix}$$

Details are given in Appendix 1A.

Computation

To determine the number of ways that exceedances can be observed, especially given that the sample size is not known in advance, requires use of number theory combinatorics and is beyond what it was feasible to accomplish for this task. However, it was possible to calculate listing probabilities for the special case in which the sample size and population proportion are the same for all ten years. An Excel spreadsheet calculating Probability of Listing based on meeting two criteria was developed: (a) at least one year within the past 10 years with at least 10% of samples (set of measurements within a calendar or water year) exceeding standards and (b) at least three exceedances (measurements exceeding standards) over the most recent 10-year period, calculated for select values of n and p **for the special case in which all 10 years have the same sample size (n) and same population proportion out of compliance (p)**, using the formula derived in Appendix 1A. The Excel spreadsheet calculated P(list) for the special case in which the X_i are distributed as identical Binomial(n_i, p_i), i.e., all n_i are equal and all p_i are equal, for $n_i = 1\ to\ 366$ and $p_i = 0.005\ to\ 0.15\ by\ 0.005$. For a copy of the Excel spreadsheet, please send an email request to 303d@ecy.wa.gov.

For all values of n from 1 to 366 and for values of p from 0.005 to 0.15 by 0.005 (with all years having the same values of n and p), the BINOM.DIST function in Excel was used to compute listing probabilities (For a copy of the Excel spreadsheet, please send an email request to 303d@ecy.wa.gov) according to

the formulae derived in Appendix 1A. Those computed probability values are graphed in Figure 1 for select values of p.

Evaluation

The following sections illustrate the derived listing probabilities for a range of values of p, demonstrate the effect of adding the 2nd criterion, and show the Type I and Type II error rates for select values of n (number of samples) and p (hypothesized population proportions).

Probability of listing based on both criteria

Figure 1 illustrates the derived listing probabilities for hypothesized population proportion $p = 0.01, 0.02, \ldots, 0.10$ for all values of n from 1 to 366. Note the effect of the "10% rule" criterion in the minimization of P(list) at each "breakpoint" (multiple of 10), jump increase for the next-larger sample size (n = breakpoint + 1), and subsequent decrease to the next breakpoint.

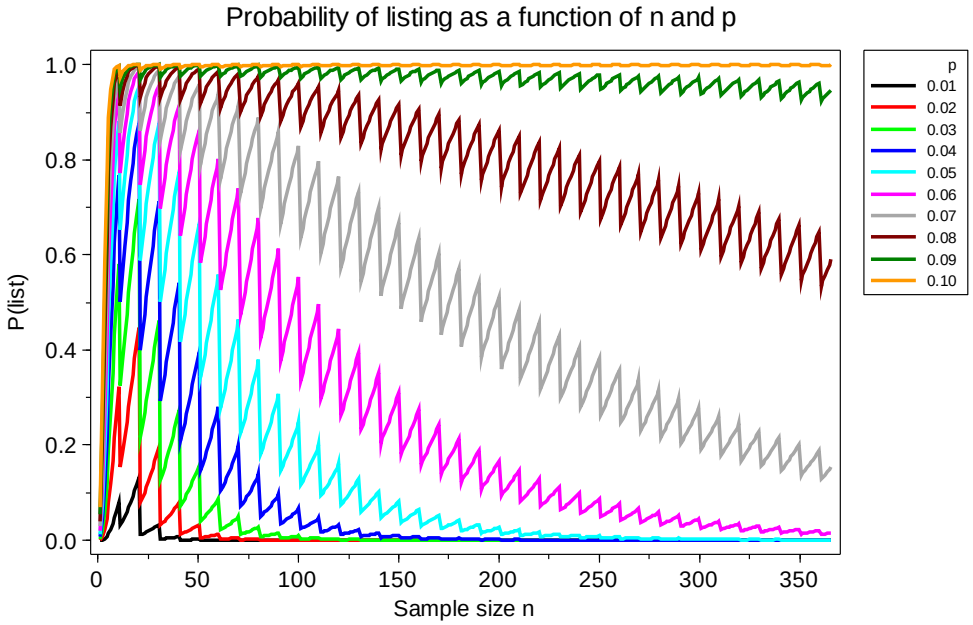


Figure 1. Probability of Listing based on meeting two criteria: (a) at least one year within the past 10 years with at least 10% of samples (set of measurements within a calendar or water year) exceeding standards and (b) at least three exceedances (measurements exceeding standards) over the most recent 10-year period, calculated for all values of n from 1 to 366 and select values of p for the case in which all 10 years have the same sample size (n) and same population proportion out of compliance (p), using the formulae derived in Appendix 1A.

To reiterate, these derived probabilities are for the case in which the sample sizes and population proportion out of compliance are the same for all 10 years, i.e. $p_i = p$ and $n_i = n$. To calculate listing probabilities for the more likely case of n_i varying by year would require combinatorics that are beyond what is feasible to do for this paper. To calculate listing probabilities for varying p_i by year – but same n_i for all years – would be easier but still computer-intensive and time-consuming.

Effect of Requiring at Least 3 Exceedances

The second criterion, that of at least three exceedances (measurements exceeding standards) over the most recent 10-year period, was added with the intention of reducing the Type I error rate of the rule

that when at least 10% of samples (set of measurements within a calendar or water year) exceed standards for at least one year within the past 10 years. This section examines the effect of that addition.

The solution is part of the derivation of the probability of listing based on both criteria (Appendix 1A).

It turns out that the effect of adding the requirement that there be at least three exceedances is applicable only in cases in which all $n_i \leq 20$ (Appendix 1A). The second criterion reduces the listing probability considerably for sample sizes 10 or less and very little for sample sizes 11 to 20 (Figures 2 and 3).

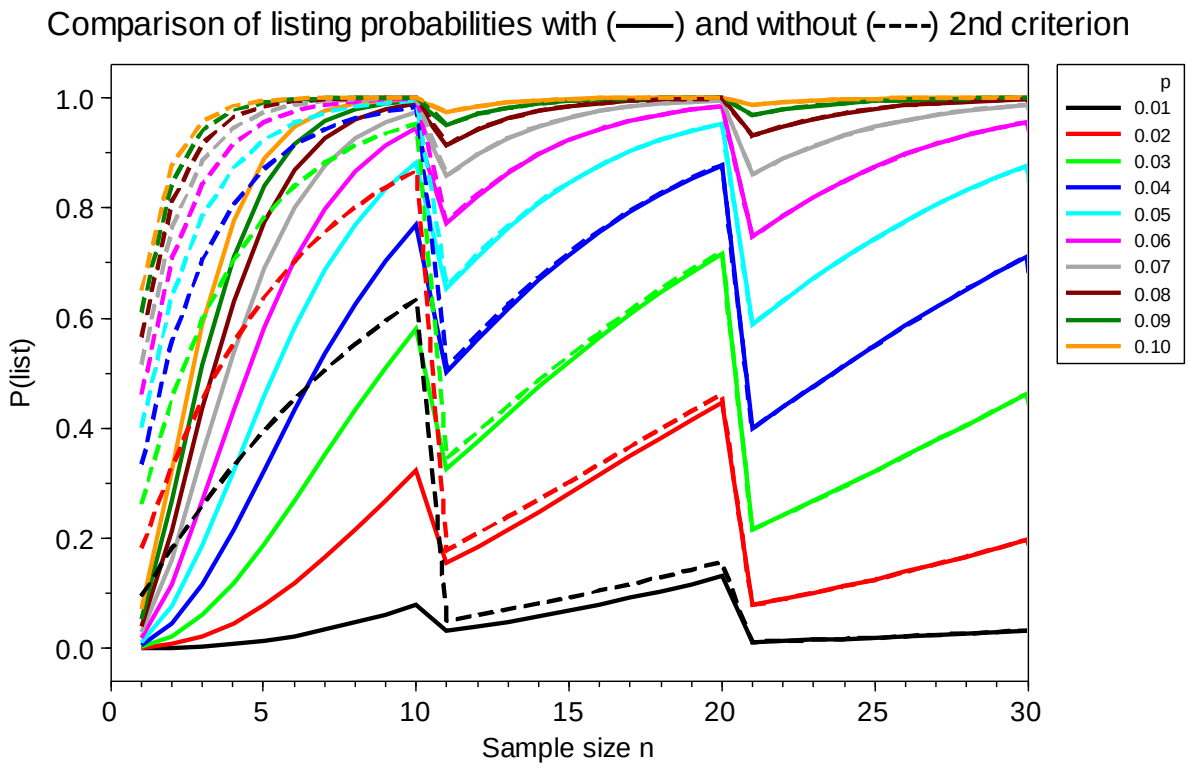


Figure 2. Listing Probabilities (solid line) with and without (dashed line) the second criterion of at least three exceedances (measurements exceeding standards) over the most recent 10-year period, for the case in which all 10 years have the same sample size (n) and same population proportion out of compliance (p), for select values of p. Probabilities were calculated with the formulae derived in Appendix 1A.

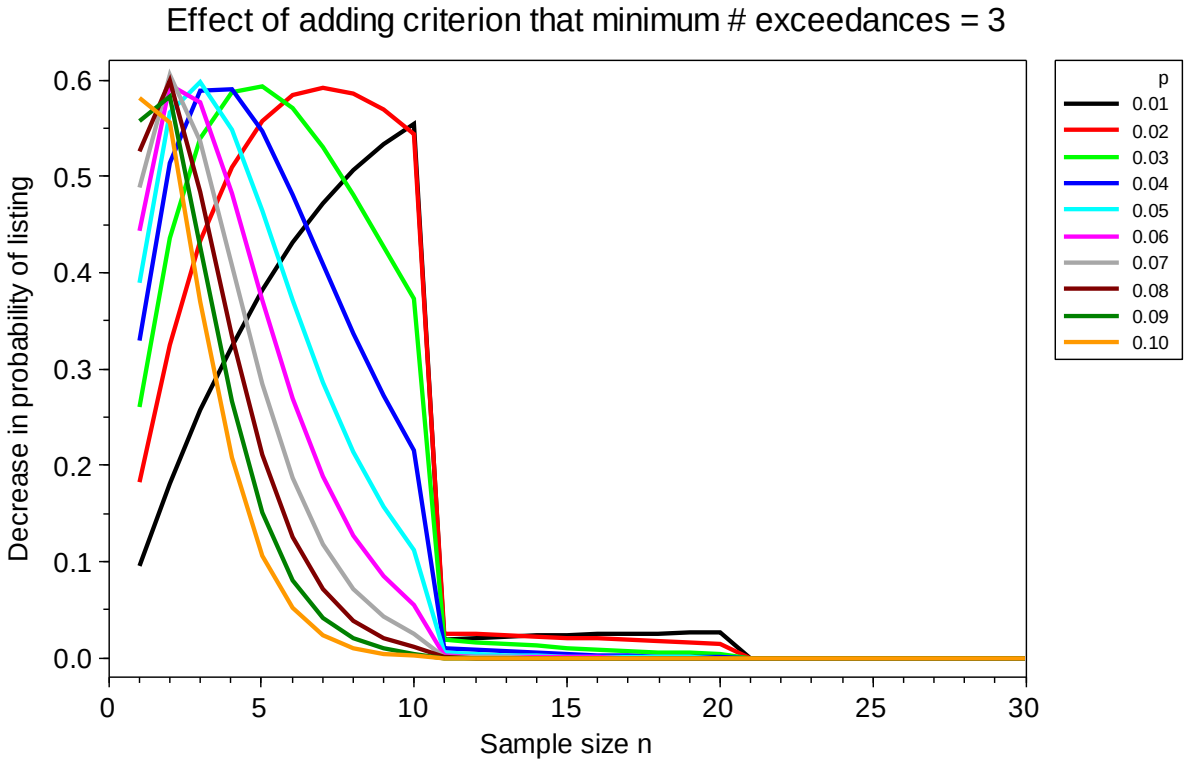


Figure 3. Decrease in Probability of Listing from that in Figure 1 based on the second criterion: at least three exceedances (measurements exceeding standards) over the most recent 10-year period, for the case in which all 10 years have the same sample size (n) and same population proportion out of compliance (p), for select values of p. Probabilities were calculated with the formulae derived in Appendix 1A.

Probabilities of Type I and Type II Error

From the probabilities calculated for all values of n from 1 to 366 and select values of p for the case in which all 10 years have the same sample size (n) and same population proportion out of compliance (p), it is possible to determine the probabilities of Type I and Type II error, i.e., of incorrectly listing a waterbody which is actually in compliance (Type I error) or incorrectly failing to list a waterbody which is actually out of compliance (Type II error).. These are the same probabilities calculated in the Excel spreadhseet as those graphed above, but plotted as a function of p instead of a function of n.

If one is testing the hypotheses H_0: $p \leq 0.05$ $vs.$ H_1: $p > 0.05$, then $P(Type\ I\ error) = P(list \mid p < 0.05)$ and $P(Type\ II\ error) = P(do\ not\ list \mid p > 0.05) = 1 - P(list \mid p > 0.1)$, which are illustrated in Figures 4 and 5, respectively, for select values of n.

The sample sizes selected for illustration, and the rationale for choosing them, are: 10 (first breakpoint for 10% rule), 12 (one sample per month), 30 (one sample/day for 1 month), 52 (one sample/week), 75 (one sample/day for 2.5 months), 90 (one sample/day for 3 months), 120 (10 samples/month, or one sample/day for 4 months), 180 (one sample/day for 6 months), 270 (one sample/day for 9 months), and 365 (one sample/day). Note that the relative positions of the curves reflect the effects of both the "10% rule" criterion and the minimum-3-exceedances criterion. For example, the curve for n=10 is more similar to the curve for n=30 than for n=12, but also crosses the curve for n=30.

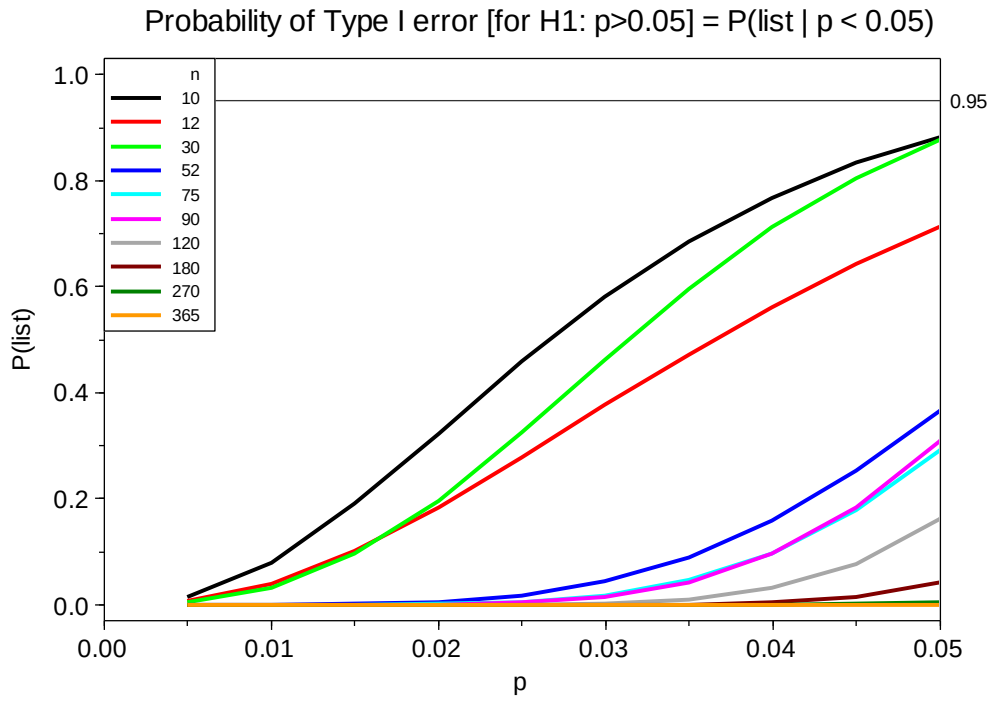


Figure 4. Probability of Type I error (listing unimpaired waters), for the test of hypotheses $H_0: p \leq 0.05$ $vs.$ $H_1: p > 0.05$, for the case in which all 10 years have the same sample size (n) and same population proportion out of compliance (p), for select values of n. Probabilities were calculated with the formulae derived in Appendix 1A.

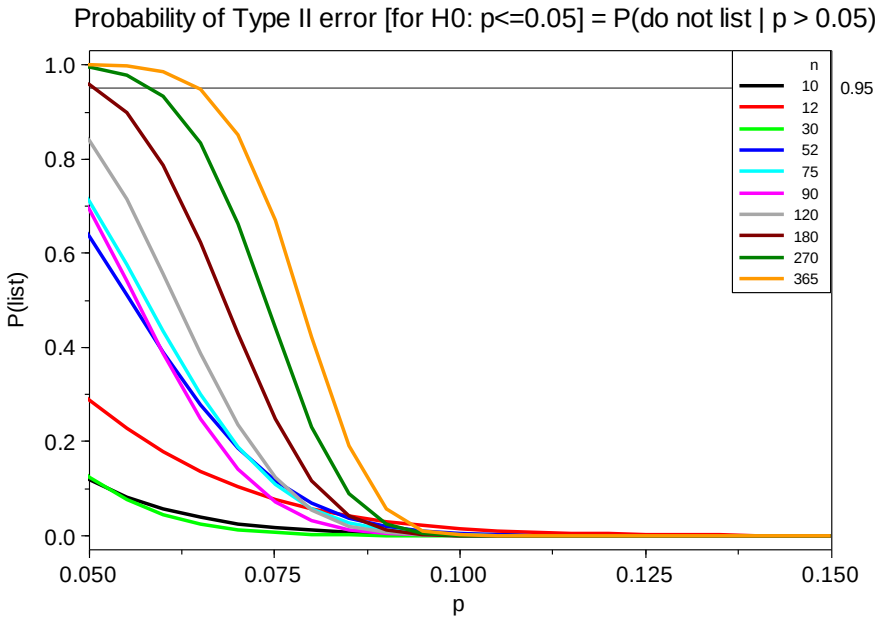


Figure 5. Probability of Type II error (not listing impaired waters), for the test of hypotheses $H_0: p \leq 0.05$ $vs.$ $H_1: p > 0.05$, for the case in which all 10 years have the same sample size (n) and same population proportion out of compliance (p), for select values of n. Probabilities were calculated with the formulae derived in Appendix 1A.

Regarding assumptions

The equation for the probability of listing a waterbody as impaired was derived assuming the use of 10 years of data. The same derivation can be used for any number of years, which would change only the maximum index value in the sums and products in Appendix 1A and Excel formulae (for a copy of the Excel spreadsheet, please send an email request to 303d@ecy.wa.gov).

(. In the case of a single year of measurement, the equation for the listing probability reduces to the simple binomial probability of observing 3 or more exceedances.

The probability equation derived for this technical memorandum assumes independence not only of years but also of individual days of measurement. Hence, the highly likely autocorrelation between measurements close in time in real life is not taken into consideration. The purpose of avoiding autocorrelation is to assure that the *samples* are independent measurements of the population.

The treatment of sampling years as independent in order to mitigate against the effects of extraordinary conditions such as drought complicates the derivation of the listing probabilities. The effects of this assumption of independence have not yet been studied.

In addition, the population proportion p is assumed to be constant within a given year (calendar or water), a simplifying assumption that would likely not be true, for example: a regulated entity remediating a waterbody.

Conclusions

- A theoretical equation for the listing probability for pH, temperature, and dissolved oxygen has been derived. However, calculation of such probability is practical only for the special case in which the sample size and population proportion out of compliance are the same for all 10 years and both years and samples are statistically independent.

- The addition of the requirement that there be at least three exceedances to list a waterbody as unimpaired affects only cases in which the sample size for all years is 20 or less, and effectively only for sample sizes 10 or less.

- In the ideal world, we would have comprehensive datasets from monitoring surveys designed specifically for determining compliance with water quality standards. The reality is that we have data from disparate sources, collected for various reasons and by inconsistent means. This requires assumptions to be made before applying any statistical test to 303(d)-listing decisions, and considering these assumptions when evaluating the results.

References

California Environmental Protection Agency, State Water Resources Control Board, Division of Water Quality. 2004. Final Functional Equivalent Document: Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List.

Khalil, B., and T.B.M.J. Ouarda. 2009. Statistical Approaches Used To Assess and Redesign Surface Water Quality Monitoring Networks. Journal of Environmental Monitoring 11(11):1915-1929.

Smith, E.P., K. Ye, C. Hughes, and L. Shabman. 2001. Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act. Environmental Science and Technology 35:606-612.

Smith, E.P., I. Lipkovich, and K. Ye. 2002. Weight-of-evidence (WOE): Quantitative estimation of probability of impairment for individual and multiple lines of evidence. Human and Ecological Risk Assessment. 8(7):1585-1596.

Smith, E.P., A. Zahran, M. Mahmoud, and K. Ye. 2003. Evaluation of water quality using acceptance sampling by variables. Environmetrics 24:373-386.

U.S. Environmental Protection Agency (EPA). 2002. Consolidated Assessment and Listing Methodology: Toward a Compendium of Best Practices, First Edition. U.S. Environmental Protection Agency Office of Wetlands, Oceans, and Watersheds, Washington, DC.

Washington State Department of Ecology (Ecology). 2012. Assessment of Water Quality for the Clean Water Act Section 303(d) and 305(b) Integrated Report. Washington State Department of Ecology Water Quality Program, Olympia, WA.

Ye, K., and E.P. Smith. 2002. A Bayesian approach to evaluating site impairment. Environmental and Ecological Statistics 9:379-392.

Appendix 1A

Derivation of Probability of Listing based on meeting two criteria:

a) At least two years within the past 10 years with at least 10% of samples (set of measurements within a calendar or water year) exceeding standards.
b) At least three exceedances (measurements exceeding standards) over the most recent 10-year period.

Let
n_i = number of samples in year $i, i = 1, 2, \ldots, 10$

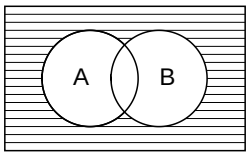
X_i = number of exceedances in year i

Then

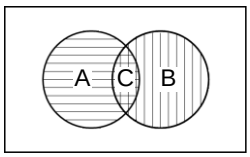
$$P(list) = P\left(\sum_{i=1}^{10} X_i \geq 3 \; AND \; number \; of \; exceedances \geq 10\% \; in \; at \; least \; one \; year\right)$$

$$= P\left(\sum_{i=1}^{10} X_i \geq 3 \; AND \; at \; least \; one \; X_i \geq 0.1n_i\right)$$

$$= 1 - P\left(\sum_{i=1}^{10} X_i \leq 2 \; OR \; no \; X_i \geq 0.1n_i\right)$$



$$= 1 - \left[P\left(\sum_{i=1}^{10} X_i \leq 2\right) + P(no \; X_i \geq 0.1n_i) - P\left(\sum_{i=1}^{10} X_i \leq 2 \; AND \; no \; X_i \geq 0.1n_i\right)\right]$$



If the X_i are distributed as Binomial(n_i, p_i), then

$$P(X_i \le k | n_i, p_i) = \sum_{j=0}^{k} \binom{n_i}{j} p_i^j (1-p_i)^{n_i - j}$$

$$= \sum_{j=0}^{m_i} \binom{n_i}{j} p_i^j (1-p_i)^{n_i - j} + \sum_{j=m_i+1}^{k} \binom{n_i}{j} p_i^j (1-p_i)^{n_i - j},$$

where $m_i = floor\left(\frac{n_i - 1}{10}\right)$

$$= P(X_i \le m_i \; and \; X_i < 0.1 n_i) + P(m_i < X_i \le k \; and \; X_i \ge 0.1 n_i)$$

$$= P(X_i \le m_i \; and \; exceedances < 10\%) + P(m_i < X_i \le k \; and \; exceedances \ge 10\%)$$

(A) $P\left(\sum_{i=1}^{10} X_i \le 2\right) = P\left(some \; X_i + X_l \le 2 \; and \; all \; other \; X_{h \ne i,l} = 0\right)$,

for all combinations of i and l

$$= P\left(some \; X_i + X_l = 2 \; and \; all \; other \; X_{h \ne i,l} = 0\right) + P(some \; X_i = 1 \; and \; all \; other \; X_{l \ne i} = 0) + P(all \; X_i = 0),$$ for all combinations of i and l

$$= \sum_{i=1}^{10} \sum_{l \ne i} \left[P(X_i + X_l = 2) \prod_{h \ne i,l} P(X_h = 0) \right] + \sum_{i=1}^{10} \left[P(X_i = 1) \prod_{l \ne i} P(X_l = 0) \right] + \prod_{i=1}^{10} P(X_i = 0)$$

$$= \sum_{i=1}^{10} \left[P(X_i = 2) \prod_{l \ne i} P(X_l = 0) \right] + \sum_{i=1}^{10} \sum_{l \ne i} \left[P(X_i = 1) P(X_l = 1) \prod_{h \ne i,l} P(X_h = 0) \right] + \sum_{i=1}^{10} \left[P(X_i = 1) \prod_{l \ne i} P(X_l = 0) \right] + \prod_{i=1}^{10} P(X_i = 0)$$

If the X_i are distributed as Binomial(n_i, p_i), then

$$= \sum_{i=1}^{10}\left[\binom{n_i}{2} p_i^2 (1-p_i)^{n_i-2} \cdot \prod_{l\neq i}(1-p_l)^{n_l}\right]$$
$$+ \sum_{i=1}^{10}\sum_{l\neq i}\left[n_i p_i (1-p_i)^{n_i-1} \cdot n_l p_l (1-p_l)^{n_l-1} \cdot \prod_{h\neq i,l}(1-p_h)^{n_h}\right]$$
$$+ \sum_{i=1}^{10}\left[n_i p_i (1-p_i)^{n_i-1} \cdot \prod_{l\neq i}(1-p_l)^{n_l}\right] + \prod_{i=1}^{10}(1-p_i)^{n_i}$$

(B)

$P(no\ X_i \geq 0.1n_i) = P(all\ X_i < 0.1n_i)$

$$= \prod_{i=1}^{10} P(X_i \leq m_i),$$

where $m_i = floor\left(\frac{n_i - 1}{10}\right)$

If the X_i are distributed as Binomial(n_i, p_i), then

(B)
$$= \prod_{i=1}^{10}\left[\sum_{j=0}^{m_i}\binom{n_i}{j} p_i^j (1-p_i)^{n_i-j}\right]$$

(C)

$$P\left(\sum_{i=1}^{10} X_i \leq 2\ AND\ no\ X_i \geq 0.1n_i\right)$$
$$= P\left(\sum_{i=1}^{10} X_i = 2\ AND\ all\ X_i < 0.1n_i\right) + P\left(\sum_{i=1}^{10} X_i = 1\ AND\ all\ X_i < 0.1n_i\right) + P\left(\sum_{i=1}^{10} X_i = 0\right)$$

$$= \sum_{i=1}^{10}\left[P(X_i = 2)\prod_{l\neq i} P(X_l = 0)\right] for\ n_i > 20$$
$$+ \sum_{i=1}^{10}\sum_{l\neq i}\left[P(X_i = 1)P(X_l = 1)\prod_{h\neq i,l} P(X_h = 0)\right] for\ n_i, n_l > 10$$

$$+\sum_{i=1}^{10}\left[P(X_i=1)\prod_{l\neq i}P(X_l=0)\right] for\ n_i>10$$

$$+\prod_{i=1}^{10}P(X_i=0)\ \ for\ all\ n_i$$

If the X_i are distributed as Binomial(n_i, p_i), then

(C) $$=\sum_{i=1}^{10}\left[\binom{n_i}{2}p_i^2(1-p_i)^{n_i-2}\cdot\prod_{l\neq i}(1-p_l)^{n_l}\right]\ \ for\ n_i>20$$

$$+\sum_{i=1}^{10}\sum_{l\neq i}\left[n_ip_i(1-p_i)^{n_i-1}\cdot n_lp_l(1-p_l)^{n_l-1}\cdot\prod_{h\neq i,l}(1-p_h)^{n_h}\right] for\ n_i,n_l>10$$

$$+\sum_{i=1}^{10}\left[n_ip_i(1-p_i)^{n_i-1}\cdot\prod_{l\neq i}(1-p_l)^{n_l}\right] for\ n_i>10$$

$$+\prod_{i=1}^{10}(1-p_i)^{n_i}\ \ for\ all\ n_i$$

APPENDIX 2

Unequal Data Requirements for Category 5 and Category 1

Background

This appendix explains from a statistical perspective why the sample size required to delist a no-longer-impaired waterbody is significantly higher than the sample size required for listing an impaired waterbody.

The numbers of samples required for listing a waterbody as impaired and delisting a no-longer-impaired waterbody are different. An analogy to this process would be a medical diagnosis for cancer. It takes only a few tests to confirm the presence of cancer. After going through treatments, a number of tests over a long period of time are needed to confirm, to a high degree of confidence, that the cancer has been cured. The same applies to pollutants in the water – only a few samples are needed to confirm the presence; however, many more samples are needed to confirm that the pollutant no longer exists in the same waterbody.

Theory

The difference in the numbers of samples is explained by statistical theory. What is known, once the sample (collection of water quality measurements) is in hand is the sample size, n, and the number of exceedances, x. The quantity which is unknown, and which we wish to estimate, is the population proportion, p, of a waterbody which is out of compliance.

Estimation and Confidence Intervals

The most commonly used model for the occurrence of exceedances is a binomial probability distribution, which has two parameters, n = the sample size (number of measurements) and p = the true proportion of the population which is out of compliance. We cannot know p, but we can estimate it by $\hat{p} = \frac{x}{n} = \frac{observed\ number\ of\ exceedances}{number\ of\ measurements}$. Because it is a single number, $\hat{p}$ is called a point estimate of p.

We can also calculate an interval estimate of p. So, based on the number of exceedances, x, found in the sample (set of measurements) of size n, we can calculate a 95% confidence interval for p, i.e., a range of values which has a 95% chance of covering the true, unknown value of p.[1]

[1] Note that a confidence interval is not a probability statement about p, such as "p has a 95% probability of being within this interval." p is fixed, but the value is unknown to us. A confidence interval is a statement about the procedure for calculating an interval estimate of p based on the sample data. What a 95% confidence level means is that if we repeatedly take samples and calculate these interval estimates for p from the sample data, in the long run, 95% of the time, the interval calculated will include the true value of p.

Presumably, the proportion of the population which is out of compliance is small, say 10% (i.e., $p = 0.1$), and thus we would expect the proportion of exceedances in the sample $\left(\frac{x}{n}\right)$ also to be small. And in fact, small values of x are far more likely to be observed when p is small, and large values of x will be unlikely.

Because the sample size is small relative to the population and sampling is not perfect, the observed number of exceedances, x, will vary from sample to sample, thus our estimate of p (i.e., $\hat{p}$) will vary, as will our calculated confidence interval. Also, the smaller the sample size is, the less reliably the sample reflects the true environment. Therefore, confidence intervals are wider for smaller sample sizes than for larger sample sizes for the same level of confidence.

The Link Between Confidence Intervals and Hypothesis Testing

So if a 95% confidence interval means that we are 95% sure, based on our data, that the calculated interval covers the true value of p, what do values **outside** the confidence interval mean?

It turns out that a 95% confidence interval is the flip side of a 5% test of hypothesis, i.e., a test of hypothesis with a 5% level of significance. In hypothesis testing, we decide based on our data whether the evidence supports the null or the alternative hypothesis with a 5% chance of being wrong if we decide in favor of the alternative. The start of the rejection region for a hypothesis test with significance level 5% corresponds to the end of the 95% confidence interval.

Just as hypothesis tests can be one-sided (i.e., the alternative hypothesis specifies only one direction), so too can confidence intervals be one-sided. That means that for H_0: $p \leq 0.1$ vs. H_1: $p > 0.1$, we can use our data to calculate a one-sided confidence interval with lower bound, $(p_L, 1]$, which we are 95% confident covers the true value of p. Likewise, for H_0: $p \geq 0.1$ vs. H_1: $p < 0.1$, we can calculate a one-sided confidence interval with upper bound, $[0, p_U)$, for which we have 95% confidence that it covers the true value of p.

Thus, for example, in a test of the null hypothesis H_0: $p \leq 0.1$ vs. the alternative hypothesis H_1: $p > 0.1$, if our data lead us to conclude that H_1 is more likely true and to decide to reject H_0, then it will also be true that <u>our 95% one-sided confidence interval for p will **not include** the value 0.1</u>.[2]

Putting that all together: If we are testing H_0: $p \leq 0.1$ vs. H_1: $p > 0.1$ at the 5% level of significance (such as for a listing decision), that is equivalent to calculating a 95% one-sided confidence interval based on our sample data (i.e., n and x) and looking to see whether it includes the value 0.1. If the **lower end** of the confidence interval $(p_L, 1]$ is **higher than** 0.1, in other words, **if the 95% confidence interval does <u>not</u> include 0.1**, that's equivalent to saying that we have enough evidence to reject H_0 in favor of H_1.

And if instead we are testing H_0: $p \geq 0.1$ vs. H_1: $p < 0.1$ (such as for a delisting decision) and the **upper end** of the one-sided confidence interval $[0, p_U)$ is **less than** 0.1 – again, **if the 95% confidence interval does not include 0.1** – that's equivalent to saying that we have enough evidence to reject H_0 and conclude that $p < 0.1$.

[2] Because the binomial distribution is not continuous, but has jumps in the values, there may be slight gaps or overlaps in the exact binomial probabilities, rejection regions, and confidence intervals, especially for small values of n.

Because the sample proportion $\hat{p} = {}^{x}/_{n}$ changes more rapidly with smaller n than larger n, it takes fewer measurements for the lower end of a confidence interval to end up being above 0.1 than for the upper end of a confidence interval to end up being below 0.1. And that is the answer to the question. An Excel spreadsheet was developed containing exact binomial 95% and 90% two-sided confidence intervals for population proportion p (Clopper-Pearson method) calculated for all values of sample size n from 1 to 366 and all values of number of exceedances x from 0 to 366, as well as the corresponding 95% one-sided confidence intervals. To see for yourself, look at the matrix of confidence intervals in the Excel spreadhseet. (for a copy, please send an email request to 303d@ecy.wa.gov).

For example, if the number of exceedances $x = 5$, any sample size $n = 20$ or smaller will result in a one-sided 95% confidence interval whose lower end is above 0.1; i.e., we would have 95% confidence that the true value of p is greater than 0.1 and would thus reject the null hypothesis H_0: $p \leq 0.1$, concluding that the waterbody is impaired. On the other hand, also for $x = 5$, only sample sizes $n = 103$ and greater will have a one-sided 95% confidence interval whose upper end is below 0.1, leading us to reject the null hypothesis H_0: $p \geq 0.1$ and conclude that the waterbody is unimpaired.

Methods

Binomial confidence intervals for every combination of sample size from 1 to 366 (for a leap year) and number of exceedances from 0 to 366 were calculated. Then the results were summarized, tabulated, and graphed for an illustrative example.

Computation

Using an Excel program available on StatPages.com (Laycock, *date unknown*), exact[3] binomial 95% and 90% two-sided confidence intervals were calculated for every combination of n from 1 to 366 (for a leap year) and x from 0 to 366. The limits of symmetrical 90% two-sided confidence intervals are the same as limits of one-sided 95% confidence intervals.

Evaluation

The entire 366 x 367 matrix of confidence intervals is given in the Excel spreadsheet, for each the 95% and 90% confidence levels. In addition, the spreadsheet contains the corresponding 95% one-sided confidence intervals. From these matrices, you can find all combinations of sample size and number of exceedances which do or do not cover your hypothesized proportion of the population (waterbody) which is out of compliance.

For example, the minimum combination of sample size and number of exceedances to be able to conclude with 95% confidence that a waterbody is impaired is if you had only two measurements and both of them were exceedances. On the other hand, it would take a minimum of 29 measurements and 0 exceedances to be able to say with 95% confidence that a waterbody is not impaired

Note that this method does not define what is considered to be an exceedance. Rather, it determines what to do once you have samples and exceedances. Whether exceedances are to be based on grab samples or running averages of continuous data, daily maxima/minima, or other measures is a separate matter.

[3] Clopper-Pearson method (Clopper and Pearson, 1934), not normal approximation.

For the example of hypothesized $p = 0.1$, Appendix 2A provides a table that lists the combinations of n and x which result in one-sided 95% confidence intervals which **do not cover** the value $p = 0.1$. List/delist decisions based on these confidence intervals are illustrated in Figure 1.

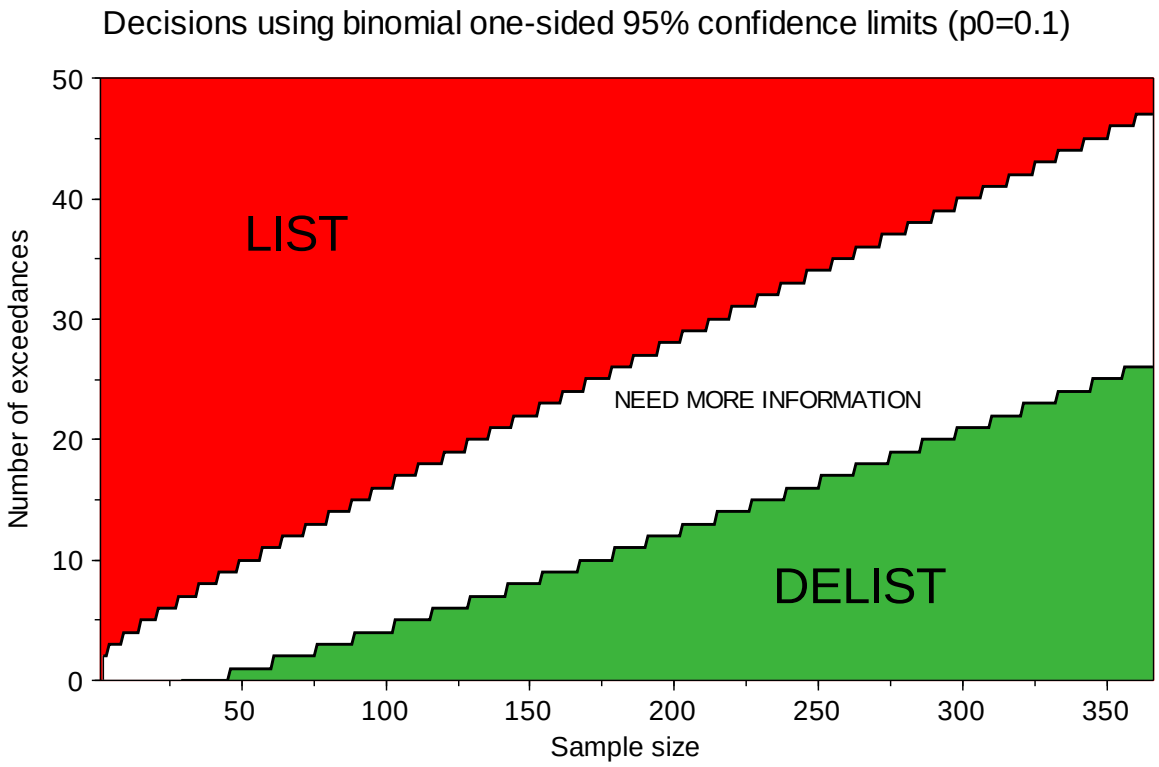


Figure 1. Sample sizes and numbers of exceedances required for listing/delisting waterbodies based on binomial one-sided 95% confidence intervals, assuming $p_0 = 0.1$.

Conclusions

The reason that a larger sample size is required to delist a no-longer-impaired waterbody than to list an impaired waterbody is a function of the hypotheses being tested, the statistical distribution type assumed for the population, and the statistical significance level used, as well as the mathematical characteristics of a ratio.

References

California Environmental Protection Agency, State Water Resources Control Board, Division of Water Quality. 2004. Final Functional Equivalent Document: Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List.

Clopper, C.J., and E.S. Pearson. 1934. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. Biometrika 26(4):404-413.

Khalil, B., and T.B.M.J. Ouarda. 2009. Statistical Approaches Used To Assess and Redesign Surface Water Quality Monitoring Networks. Journal of Environmental Monitoring 11(11):1915-1929.

Laycock, P.J. *Date unknown*. Table of exact binomial confidence limits. *In*: Pezzullo, John C. 2009. Exact C.I.'s for Binomial (observed proportion) and Poisson (observed count): Excel calculator confint.xls. Downloaded May 2016 from StatPages.net at http://statpages.info/#Confidence.

Smith, E.P., K. Ye, C. Hughes, and L. Shabman. 2001. Statistical Assessment of Violations of Water Quality Standards under Section 303(d) of the Clean Water Act. Environmental Science and Technology 35:606-612.

Smith, E.P., I. Lipkovich, and K. Ye. 2002. Weight-of-evidence (WOE): Quantitative estimation of probability of impairment for individual and multiple lines of evidence. Human and Ecological Risk Assessment. 8(7):1585-1596.

Smith, E.P., A. Zahran, M. Mahmoud, and K. Ye. 2003. Evaluation of water quality using acceptance sampling by variables. Environmetrics 24:373-386.

U.S. Environmental Protection Agency (EPA). 2002. Consolidated Assessment and Listing Methodology: Toward a Compendium of Best Practices, First Edition. U.S. Environmental Protection Agency Office of Wetlands, Oceans, and Watersheds, Washington, DC.

Ye, Keying, and E.P. Smith. 2002. A Bayesian approach to evaluating site impairment. Environmental and Ecological Statistics 9:379-392.

Appendix 2A

Table of minimum and maximum sample size and number of exceedances such that 95% exact binomial one-sided confidence intervals[4] for the population proportion **do not include 0.1**.

Minimum number of exceedances for 95% confidence that $p > 0.1$, **based on sample size**		**Maximum sample size** for 95% confidence that $p > 0.1$, **based on number of exceedance**s	
Sample Size, n	Minimum # exceedances	Number of exceedances, k	Maximum sample size
1	not possible	0 or 1	not possible
2-3	2	2	3
4-8	3	3	8
9-14	4	4	14
15-20	5	5	20
21-27	6	6	27
28-34	7	7	34
35-41	8	8	41
42-48	9	9	48
49-56	10	10	56
57-63	11	11	63
64-71	12	12	71
72-79	13	13	79
80-87	14	14	87
88-94	15	15	94
95-102	16	16	102
103-110	17	17	110

Maximum number of exceedances for 95% confidence that $p < 0.1$, **based on sample size**		**Minimum sample size** for 95% confidence that $p < 0.1$, **based on number of exceedances**	
Sample Size, n	Maximum # exceedances	Number of exceedances, m	Minimum sample size
1-28	not possible	0	29
29-45	0	1	46
46-60	1	2	61
61-75	2	3	76
76-88	3	4	89
89-102	4	5	103
103-115	5	6	116
116-128	6	7	129
129-141	7	8	142
142-153	8	9	154
154-166	9	10	167
167-178	10	11	179
179-190	11	12	191
191-202	12	13	203
203-214	13	14	215
215-226	14	15	227
227-238	15	16	239

[4] Clopper-Pearson binomial confidence intervals (Clopper and Pearson, 1934). Computed with Excel calculator programmed by Laycock (*date unknown*).

Minimum number of exceedances for 95% confidence that $p > 0.1$, **based on sample size**		**Maximum sample size** for 95% confidence that $p > 0.1$, **based on number of exceedance**s	
Sample Size, n	Minimum # exceedances	Number of exceedances, k	Maximum sample size
111-119	18	18	119
120-127	19	19	127
128-135	20	20	135
136-143	21	21	143
144-152	22	22	152
153-160	23	23	160
161-168	24	24	168
169-177	25	25	177
178-185	26	26	185
186-194	27	27	194
195-202	28	28	202
203-211	29	29	211
212-219	30	30	219
220-228	31	31	228
229-236	32	32	236
237-245	33	33	245
246-254	34	34	254
255-262	35	35	262
263-271	36	36	271
272-280	37	37	280
281-289	38	38	289
290-297	39	39	297
298-306	40	40	306
307-315	41	41	315
316-324	42	42	324
325-332	43	43	332
333-341	44	44	341
342-350	45	45	350

Maximum number of exceedances for 95% confidence that $p < 0.1$, **based on sample size**		**Minimum sample size** for 95% confidence that $p < 0.1$, **based on number of exceedances**	
Sample Size, n	Maximum # exceedances	Number of exceedances, m	Minimum sample size
239-250	16	17	251
251-262	17	18	263
263-274	18	19	275
275-285	19	20	286
286-297	20	21	298
298-309	21	22	310
310-320	22	23	321
321-332	23	24	333
333-344	24	25	345
345-355	25	26	356
356-366	26	27 or more	not possible

Minimum number of exceedances for 95% confidence that $p > 0.1$, **based on sample size**		**Maximum sample size** for 95% confidence that $p > 0.1$, **based on number of exceedance**s	
Sample Size, n	Minimum # exceedances	Number of exceedances, k	Maximum sample size
351-359	46	46	359
360-366	47	47 or more	all sample sizes

Maximum number of exceedances for 95% confidence that $p < 0.1$, **based on sample size**		**Minimum sample size** for 95% confidence that $p < 0.1$, **based on number of exceedances**	
Sample Size, n	Maximum # exceedances	Number of exceedances, m	Minimum sample size

Appendix 3

Use of instantaneous Measurements to represent Multi-day Averages (such as chronic metals)

Background

This analysis explores how representative a single "grab" sample is of multi-day averages of toxics contamination. The frequency and type of water quality data to be used for Water Quality Assessments are unknown until the data are received by Ecology. Therefore, Ecology is sometimes in the position of having to make 303(d)-listing decisions with limited data. The analysis documented in this paper addresses only a single aspect of the complex situation of 303(d)-listing criteria and the data available, specifically, whether single samples can be used to evaluate toxics contamination for which the criteria are based on 4-day running averages.

Methods

The general approach involved simulating hypothetical “observed” contaminant concentrations that corresponded to a waterbody **just meeting** the chronic water quality standard and determining how often the standard was not met. The basis for the simulation was EPA technical guidance on derivation of acute and chronic water quality standards for toxic contaminants (EPA, 1991).

Large numbers of random values were generated from a probability distribution defined by the long-term average set at the chronic water quality standard for a given contaminant to represent single "grab" samples. Running averages of four single values for the entire sequence were calculated to represent "4-day running average" concentrations. The reason for using averages set at the standards is to simulate the worst-case scenario for waterbodies actually in compliance.

Both the individual "1-day" values and the "4-day average" values were compared to the chronic water quality standard for that particular contaminant, and the percent of the single and averaged values exceeding the standard was calculated. Such a simulation was repeated for many different toxic contaminant standards.

Finally, the exceedance rates (percent exceedance) of the "1-day" and "4-day average" observations for the collection of all the contaminants simulated were statistically compared.

Simulation

Two Excel spreadsheets were used for this analysis:

- *wqbp3.xls* calculates acute and chronic Wasteload Allocations (WLAs) and Long-Term Averages (LTA), and Daily Maximum Permit Limit (MDL) and Monthly Average Permit Limit (AML), using formulae given in EPA (1991) Section 5.4. The WLAs are calculated from the acute and chronic water quality standards, assuming no (0) upstream receiving water concentration and effluent

dilution factor of 1.[5] The acute and chronic LTAs are calculated from the WLAs; the 90th, 95th, or 99th percentile of a standard lognormal distribution; and the shape parameter of a lognormal distribution calculated from **assumed** coefficient of variation[6] (cv) of 0.6. The AML and MDL are based on 95th and 99th percentiles, respectively, of a standard lognormal distribution, the more limiting of the LTAs, and the same assumed shape parameter.

- *wqpb_303d_example_distributions.xlsx* does three things: (1) generates random "daily observations" from lognormal distributions representing chronic and acute conditions; (2) calculates running "4-day" averages of the "chronic" observations, and (3) determines the proportions of individual and averaged observations which exceed the respective water quality standards. The random lognormally-distributed numbers are generated by an Excel add-in called *YASAIw.xla* using as location and shape parameters the LTA and shape parameter computed by *wqbp3.xls*. The number of random observations generated were extended from 1000 to 10,000.

Using these two spreadsheets, 10,000 single "1-day" (or "1-hour") observations and 10,000 "4-day running averages" for chronic contamination were simulated, **for each of the toxics parameters with constant numerical Toxics Substances Criteria** (TSCs) (Table 1), i.e., criteria not dependent on specific values of pH, temperature, or hardness (Ecology, 2011). The spreadsheet also generated 10,000 single observations for comparison to the acute TSCs, but since it did not also calculate running averages, the results were not used further. The simulations were repeated for LTAs based on each the 90th, 95th, and 99th percentiles of the standard lognormal distribution.

Comparison

To estimate the frequency at which the single measurements exceed the chronic standards, compared to that of the 4-day running averages, the ratio of the exceedance rates (percent exceedance) of the "1-day" and "4-day average" observations for each of the contaminants were calculated and percentile-defined lognormal distributions simulated.

Table 1. Toxics Substances Criteria (Ecology, 2011) for which Monte Carlo simulations were performed. In all cases except the few noted, the chronic criterion was the limiting condition, meaning the more stringent standard for the particular lognormal distribution Long-Term Average. The acute TSC was the limiting condition only for: * = LTA based on 99th percentile; *** = LTA based on 90th, 95th, and 99th percentiles. Simulations for only the chronic criteria were used in the comparison analysis.

Parameter	Freshwater TSC Acute	Freshwater TSC Chronic	Marine Water TSC Acute	Marine Water TSC Chronic
Aldrin/Dieldrin (Dieldrin/Aldrin)	2.5	0.0019	0.71	0.0019
Ammonia (un-ionized NH3)			0.233	0.035
Arsenic	360	190	69	36

[5] With upstream receiving water concentration = 0 and dilution factor = 1, the WLA is therefore calculated to be equal to the water quality standard.

[6] The coefficient of variation is the ratio of the standard deviation to the mean.

Parameter	Freshwater TSC Acute	Freshwater TSC Chronic	Marine Water TSC Acute	Marine Water TSC Chronic
Cadmium			42	9.3
Chlordane	2.4	0.0043	0.09	0.004
Chloride (Dissolved)	860	230		
Chlorine (Total Residual)	19	11	13	7.5
Chlorpyrifos	0.083	0.041	0.011	0.0056
Chromium (Hex)	15	10	1100	50
Copper			4.8*	3.1
Cyanide Pt Roberts to Pt Wilson	22	5.2	9.1	2.8
Cyanide elsewhere			1***	1
DDT (and metabolites)	1.1	0.001	0.13	0.001
Endosulfan	0.22	0.056	0.034	0.0087
Endrin	0.18	0.0023	0.037	0.0023
Heptachlor	0.52	0.0038	0.053	0.0036
Hexachlorocyclohexane (Lindane)	2	0.08	0.16	
Lead			210	8.1
Mercury	2.1	0.012	1.8	0.025
Nickel			74	8.2
Parathion	0.065	0.013		
Pentachlorophenol (PCP)			13	7.9
Polychlorinated Biphenyls (PCBs)	2	0.014	10	0.03
Selenium	20	5	290	71
Silver			1.9	
Toxaphene	0.73	0.0002	0.21	0.0002
Zinc			90***	81

Evaluation

As expected, the output simulated the input with some degree of variability (Appendix 3A Tables 3A1-3A3). The output cv for the single measurements was close to 0.6 (Appendix 3A Figure 3A1). The output cv for the "4-day averages" was close to 0.3 (not shown), as to be expected because the standard deviation of collections of averages of 4 numbers is algebraically ½ the standard deviation of collections of single values.

The TSC-exceedance proportions for the "4-day averages" compared to the chronic criteria and for "1-day" observations compared to acute criteria were close to 90%, 95%, and 99%, accordingly for the bases of the input LTAs (Appendix 3A Table 3A4, Figure 3A2).

The output exceedance proportions for the "1-day" chronic simulations, however, were on average 1.96, 2.82, and 7.11 times the "4-day average" chronic exceedance proportions for the 90%, 95%, and 99% LTAs, respectively (Table 2, Figures 1-2, Appendix 3A Table 3A4). In other words, **individual daily observations have a much greater chance of exceeding the chronic TSC than 4-day averages do**. And the more extreme the percentile on which the Long-Term Average is based, the more extreme the exceedance rate for the single measurements. The latter is because of the greater variability in the upper tail of a right-skewed distribution such as the lognormal.

Table 2. Summary statistics for **ratios** of "1-day" to "4-day average" exceedance rates (percent of observations exceeding **chronic TSCs**) for each toxics parameter. Results are based on 10,000 randomly generated concentrations from lognormal distributions defined by Long-Term Averages based on 90th, 95th, and 99th percentiles, for each parameter.

LTA based on:	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
90th Percentile	41	1.96	0.11	1.72	1.87	1.95	2.05	2.21
95th Percentile	41	2.82	0.14	2.52	2.74	2.83	2.91	3.24
99th Percentile	41	7.11	0.75	5.19	6.65	7.02	7.54	8.91

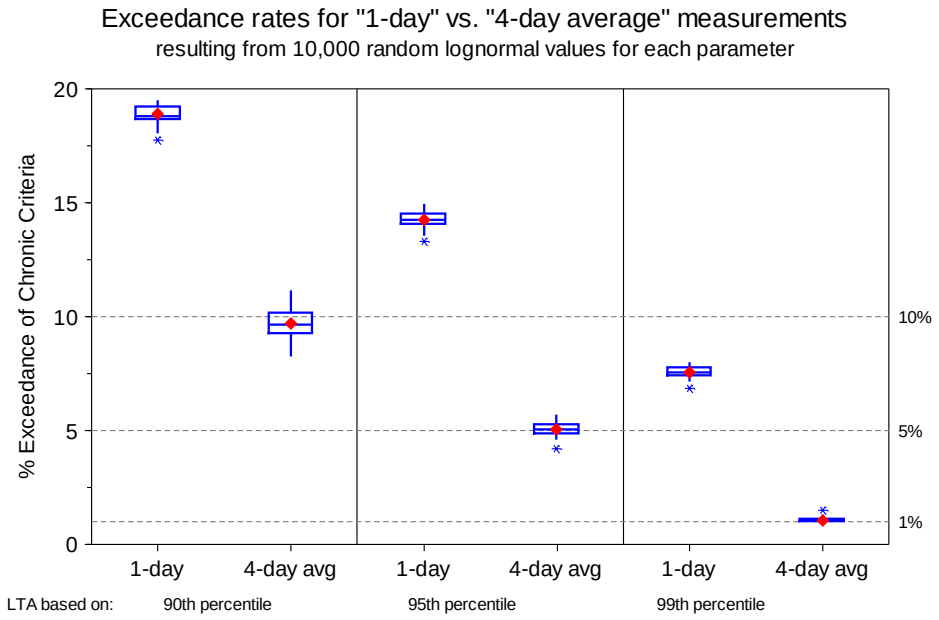


Figure 1. Exceedance rates (percent of observations exceeding **chronic TSCs**) of "1-day" and "4-day average" simulated observations for each toxics parameter. Results are based on 10,000 randomly generated concentrations from lognormal distributions defined by Long-Term Averages based on 90th, 95th, and 99th percentiles, for each parameter. Means are indicated by red diamonds.

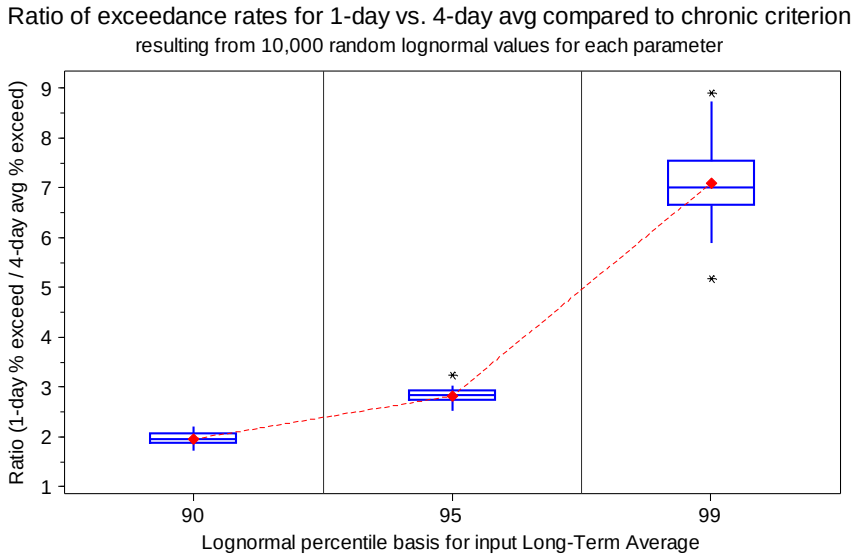


Figure 2. Ratios of "1-day" to "4-day average" exceedance rates (percent of observations exceeding **chronic TSCs**) for each toxics parameter. Results are based on 10,000 randomly generated concentrations from lognormal distributions defined by Long-Term Averages based on 90th, 95th, and 99th percentiles, for each parameter. Means are indicated by red diamonds.

Caveats and Discussion

- The underlying assumptions of lognormality and coefficient of variation value have not been tested with real data; therefore, these results are provisional.

- Although the individual observations were generated from the same lognormal distribution for a given parameter and percentile-based LTA, the fact that they were randomly generated means (if the underlying random-number generator in *YASAIw.xla* is sound) that the observations are independent, meaning that serial observations would be uncorrelated. In real life, however, serial observations would likely be temporally autocorrelated; in other words, observations taken close in time would likely be more similar to each other than if they were truly random.

  Therefore, if a waterbody is in compliance, with contaminant concentration truly below the relevant standard, it would be more likely for serial observations to reflect that. Similarly, if the true average concentration is higher than the standard, it would be more likely for serial individual measurements to exceed the criteria.

  Ideally, all measurements would be independent and truly random. Autocorrelation results from taking samples too close together in space or time for the samples to be independent. Autocorrelation is a function of sampling, not of the underlying population. Therefore, the results of these simulations represent an idealized situation.

  Furthermore, the situations in which single samples would be used would most likely be cases in which the samples are taken so far apart in time as to be uncorrelated. If daily samples are autocorrelated, the variability of the resulting 4-day averages would be artificially depressed, making the difference in exceedance rates potentially even greater.

- All of the random observations generated in a given run were based on fixed distributional location and shape. Such constancy, however, would not be true if the underlying true distribution were changing, such as when a regulated entity is actively remediating a waterbody or when some new source of contamination begins (e.g., oil train derailment).

- There is a circularity to the simulation, in that the particular lognormal distribution is based on the WLA (formula in EPA, 1991, Section 5.4), which is equal to the input water quality standards, and then the results are compared to the same standards.

- This simulation did not take into consideration how the TSCs were established in the first place, and so the lognormal distribution used in the simulation may not be the same distribution used to develop the standards.

Conclusions

- Individual daily observations have a much greater chance of exceeding the chronic TSC than 4-day averages do. The exceedance rate is a function of the assumed lognormal percentile on which the LTA is based. On average:
    - For LTAs based on 90th percentiles, 1-day observations are twice as likely to exceed the chronic standards as are the 4-day running averages.
    - For LTAs based on 95th percentiles, the 1-day exceedance rate is almost three times that of 4-day running averages.
    - For LTAs based on 99th percentiles, 1-day observations are more than seven times as likely as the 4-day running averages to exceed the chronic standards.

- These results are based on a constant distributional model excluding autocorrelation. Real-world exceedance rates of single observations may differ due to autocorrelation and changing conditions.

References

Brown, C. Undated. Determining the 303(d) listings based on toxics data in the water column. Unpublished technical memorandum. Washington State Department of Ecology, Olympia, WA.

Khalil, B., and T.B.M.J. Ouarda. 2009. Statistical Approaches Used To Assess and Redesign Surface Water Quality Monitoring Networks. Journal of Environmental Monitoring 11(11):1915-1929.

U.S. Environmental Protection Agency (EPA). 1991. Technical Support Document for water quality-based toxics control. Publication EPA/505/2-90-001. U.S. Environmental Protection Agency Office of Water, Washington, DC.

Washington State Department of Ecology (Ecology). 2011. Water quality standards for surface waters of the State of Washington. Publication 06-10-091. Washington State Department of Ecology Water Quality Program, Olympia, WA.

Appendix 3A

Table 3A1. Comparison of input Long-Term Averages based on 90th percentile of lognormal distribution (calculated per EPA, 1997) and output sample means of 10,000 randomly generated values

Fresh-water / Marine Water	Parameter	For LTA based on 90th percentile					
		Acute criteria			Chronic criteria		
		Input LTA	Sample Mean	Sample StDev	Input LTA	Sample Mean	Sample StDev
FW	Aldrin/Dieldrin	1.4324	1.4241	0.8488	0.0014	0.0014	0.0008
MW	Aldrin/Dieldrin	0.4068	0.4090	0.2461	0.0014	0.0014	0.0009
MW	Ammonia (un-ionized NH3)	0.1335	0.1333	0.0795	0.0251	0.0250	0.0154
MW	Arsenic	39.5349	39.6086	24.2074	25.8002	25.8683	15.5805
MW	Cadmium	24.0647	24.1739	14.7580	6.6650	6.7114	4.1056
FW	Chlordane	1.3751	1.3763	0.8199	0.0031	0.0031	0.0019
MW	Chlordane	0.0516	0.0512	0.0306	0.0029	0.0029	0.0017
FW	Chloride (Dissolved)	492.7544	494.1810	304.8315	164.8344	164.2523	97.2711
FW	Chloride (Total Residual)	10.8864	10.8994	6.5869	7.8834	7.8509	4.6440
MW	Chloride (Total Residual)	7.4486	7.4265	4.4056	5.3750	5.3561	3.1625
FW	Chlorpyrifos	0.0476	0.0473	0.0284	0.0294	0.0293	0.0175
MW	Chlorpyrifos	0.0063	0.0063	0.0037	0.0040	0.0041	0.0025
FW	Chromium (Hex)	8.5946	8.6749	5.3935	7.1667	7.1628	4.1332
MW	Chromium (Hex)	630.2672	633.5648	377.6326	35.8336	35.5917	21.1570
MW	Copper	2.7503	2.7671	1.7056	2.2217	2.2005	1.2987
FW	Cyanide	12.6053	12.5019	7.4811	3.7267	3.6979	2.1889
MW	Cyanide elsewhere	0.5730	0.5701	0.3418	0.7167	0.7170	0.4269
MW	Cyanide Pt Roberts to Pt Wilson	5.2140	5.1962	3.1282	2.0067	2.0190	1.2185
FW	DDT (and metabolites)	0.6303	0.6273	0.3766	0.0007	0.0007	0.0004
MW	DDT (and metabolites)	0.0745	0.0757	0.0451	0.0007	0.0007	0.0004
FW	Endosulfan	0.1261	0.1259	0.0744	0.0401	0.0396	0.0235

Fresh-water / Marine Water	Parameter	For LTA based on 90th percentile					
		Acute criteria			Chronic criteria		
		Input LTA	Sample Mean	Sample StDev	Input LTA	Sample Mean	Sample StDev
MW	Endosulfan	0.0195	0.0757	0.0451	0.0062	0.0007	0.0004
FW	Endrin	0.1031	0.1027	0.0617	0.0016	0.0016	0.0010
MW	Endrin	0.0212	0.0212	0.0129	0.0016	0.0017	0.0010
FW	Heptachlor	0.2979	0.2985	0.1786	0.0027	0.0027	0.0017
MW	Heptachlor	0.0304	0.0303	0.0183	0.0026	0.0026	0.0016
FW	Hexachlorocyclohexane (Lindane)	1.1459	1.1426	0.6779	0.0573	0.0563	0.0330
MW	Hexachlorocyclohexane (Lindane)	0.0917	0.0907	0.0536	NA	NA	NA
MW	Lead	120.3237	120.2093	71.3299	5.8050	5.7956	3.5248
FW	Mercury	1.2032	1.2039	0.7135	0.0086	0.0087	0.0052
MW	Mercury	1.0313	1.0351	0.6163	0.0179	0.0179	0.0108
MW	Nickel	42.3998	42.4911	25.7408	5.8767	5.8964	3.6226
FW	Parathion	0.0372	0.0375	0.0226	0.0093	0.0093	0.0055
MW	Pentachlorophenol (PCP)	7.4486	7.3987	4.4269	5.6617	5.6197	3.3495
FW	Polychlorinated Biphenyls (PCBs)	1.1459	1.1437	0.6768	0.0100	0.0100	0.0061
MW	Polychlorinated Biphenyls (PCBs)	5.7297	5.7565	3.4458	0.0215	0.0216	0.0126
FW	Selenium	11.4594	11.4796	6.9348	3.5834	3.5567	2.1468
MW	Selenium	166.1614	165.4830	102.2240	50.8837	50.6941	29.9958
FW	Silver	206.2693	206.1154	123.3480	136.1675	135.6015	80.0611
MW	Silver	1.0886	1.0957	0.6493	NA	NA	NA
FW	Toxaphene	0.4183	0.4228	0.2594	0.0001	0.0001	0.0001
MW	Toxaphene	0.1203	0.1208	0.0714	0.0001	0.0001	0.0001
MW	Zinc	51.5673	52.0253	30.3796	58.0504	57.6771	34.8144

Table 3A2. Comparison of input Long-Term Averages based on 95th percentile of lognormal distribution (calculated per EPA, 1997) and output sample means of 10,000 randomly generated values.

Fresh-water / Marine Water	Parameter	For LTA based on 95th percentile					
		Acute criteria			Chronic criteria		
		Input LTA	Sample Mean	Sample StDev	Input LTA	Sample Mean	Sample StDev
FW	Aldrin/Dieldrin	1.1710	1.1722	0.7083	0.0012	0.0012	0.0007
MW	Aldrin/Dieldrin	0.3326	0.3306	0.1995	0.0012	0.0012	0.0007
MW	Ammonia (un-ionized NH3)	0.1091	0.1091	0.0658	0.0225	0.0225	0.0135
MW	Arsenic	32.3196	32.4619	19.2610	23.1895	23.1518	13.6824
MW	Cadmium	19.6728	19.7910	11.9516	5.9906	5.9910	3.5698
FW	Chlordane	1.1242	1.1298	0.6751	0.0028	0.0028	0.0017
MW	Chlordane	0.0422	0.0422	0.0251	0.0026	0.0026	0.0016
FW	Chloride (Dissolved)	402.8244	399.9348	235.6043	148.1553	147.2930	85.6583
FW	Chloride (Total Residual)	8.8996	8.8549	5.3655	7.0857	7.0897	4.2629
MW	Chloride (Total Residual)	6.0892	6.0547	3.6550	4.8312	4.8648	2.9257
FW	Chlorpyrifos	0.0389	0.0387	0.0230	0.0264	0.0266	0.0159
MW	Chlorpyrifos	0.0052	0.0051	0.0031	0.0036	0.0036	0.0022
FW	Chromium (Hex)	7.0260	7.0291	4.2147	6.4415	6.4741	3.9466
MW	Chromium (Hex)	515.2405	517.9707	311.3287	32.2077	32.3522	19.5016
MW	Copper	2.2483	2.2343	1.3675	1.9969	1.9777	1.1653
FW	Cyanide	10.3048	10.3143	6.0867	3.3496	3.3486	2.0324
MW	Cyanide elsewhere	0.4684	0.4664	0.2773	0.6442	0.6508	0.3970
MW	Cyanide Pt Roberts to Pt Wilson	4.2624	4.2830	2.5609	1.8036	1.7794	1.0482
FW	DDT (and metabolites)	0.5152	0.5176	0.3109	0.0006	0.0006	0.0004
MW	DDT (and metabolites)	0.0609	0.0607	0.0370	0.0006	0.0006	0.0004
FW	Endosulfan	0.1030	0.1028	0.0618	0.0361	0.0360	0.0215
MW	Endosulfan	0.0159	0.0607	0.0370	0.0056	0.0006	0.0004
FW	Endrin	0.0843	0.0840	0.0484	0.0015	0.0015	0.0009
MW	Endrin	0.0173	0.0173	0.0105	0.0015	0.0015	0.0009

Fresh-water / Marine Water	Parameter	For LTA based on 95th percentile					
		Acute criteria			Chronic criteria		
		Input LTA	Sample Mean	Sample StDev	Input LTA	Sample Mean	Sample StDev
FW	Heptachlor	0.2436	0.2445	0.1467	0.0024	0.0024	0.0014
MW	Heptachlor	0.0248	0.0250	0.0151	0.0023	0.0023	0.0014
FW	Hexachlorocyclohexane (Lindane)	0.9368	0.9330	0.5495	0.0515	0.0514	0.0307
MW	Hexachlorocyclohexane (Lindane)	0.0749	0.0743	0.0438	NA	NA	NA
MW	Lead	98.3641	98.3316	58.7006	5.2176	5.2064	3.0683
FW	Mercury	0.9836	0.9853	0.5947	0.0077	0.0078	0.0047
MW	Mercury	0.8431	0.8441	0.4895	0.0161	0.0162	0.0098
MW	Nickel	34.6616	34.6592	21.0313	5.2821	5.2749	3.1781
FW	Parathion	0.0304	0.0301	0.0180	0.0084	0.0085	0.0052
MW	Pentachlorophenol (PCP)	6.0892	6.0726	3.6020	5.0888	5.1076	3.0509
FW	Polychlorinated Biphenyls (PCBs)	0.9368	0.9456	0.5776	0.0090	0.0090	0.0055
MW	Polychlorinated Biphenyls (PCBs)	4.6840	4.6603	2.7898	0.0193	0.0193	0.0116
FW	Selenium	9.3680	9.3087	5.5775	3.2208	3.2186	1.9436
MW	Selenium	135.8361	135.3794	80.4981	45.7349	45.5678	27.0581
FW	Silver	168.6242	168.1632	102.2786	122.3892	121.9166	72.1004
MW	Silver	0.8900	0.8874	0.5299	NA	NA	NA
FW	Toxaphene	0.3419	0.3438	0.2019	0.0001	0.0001	0.0001
MW	Toxaphene	0.0984	0.0987	0.0584	0.0001	0.0001	0.0001
MW	Zinc	42.1560	42.5834	26.0208	52.1764	52.5913	31.4419

Table 3A3. Comparison of input Long-Term Averages based on 99th percentile of lognormal distribution (calculated per EPA, 1997) and output sample means of 10,000 randomly generated values.

Fresh-water / Marine Water	Parameter	For LTA based on 99th percentile					
		Acute criteria			Chronic criteria		
		Input LTA	Sample Mean	Sample StDev	Input LTA	Sample Mean	Sample StDev
FW	Aldrin/Dieldrin	0.8027	0.8022	0.4822	0.0010	0.0010	0.0006
MW	Aldrin/Dieldrin	0.2280	0.2281	0.1383	0.0010	0.0010	0.0006
MW	Ammonia (un-ionized NH3)	0.0748	0.0749	0.0445	0.0185	0.0185	0.0109
MW	Arsenic	22.1547	22.1947	13.2177	18.9876	18.9623	11.3729
MW	Cadmium	13.4855	13.4978	7.9594	4.9051	4.8762	2.9091
FW	Chlordane	0.7706	0.7754	0.4647	0.0023	0.0023	0.0014
MW	Chlordane	0.0289	0.0290	0.0177	0.0021	0.0021	0.0013
FW	Chloride (Dissolved)	276.1316	274.8881	166.9309	121.3097	120.4512	72.1695
FW	Chloride (Total Residual)	6.1006	6.0839	3.7498	5.8018	5.7494	3.4282
MW	Chloride (Total Residual)	4.1741	4.1892	2.4846	3.9558	3.9728	2.4024
FW	Chlorpyrifos	0.0266	0.0269	0.0163	0.0216	0.0218	0.0131
MW	Chlorpyrifos	0.0035	0.0035	0.0021	0.0030	0.0030	0.0018
FW	Chromium (Hex)	4.8162	4.8198	2.8984	5.2743	5.2799	3.3069
MW	Chromium (Hex)	353.1915	349.4278	205.2844	26.3717	26.3852	15.6677
MW	Copper	1.5412	1.5265	0.9125	1.6350	1.6243	0.9570
FW	Cyanide	7.0638	7.0698	4.1830	2.7427	2.7370	1.6408
MW	Cyanide elsewhere	0.3211	0.3210	0.1937	0.5274	0.5240	0.3199
MW	Cyanide Pt Roberts to Pt Wilson	2.9219	2.8903	1.7135	1.4768	1.4848	0.8969
FW	DDT (and metabolites)	0.3532	0.3518	0.2097	0.0005	0.0005	0.0003
MW	DDT (and metabolites)	0.0417	0.0415	0.0249	0.0005	0.0005	0.0003
FW	Endosulfan	0.0706	0.0706	0.0419	0.0295	0.0294	0.0177
MW	Endosulfan	0.0109	0.0415	0.0249	0.0046	0.0005	0.0003
FW	Endrin	0.0578	0.0579	0.0349	0.0012	0.0012	0.0007
MW	Endrin	0.0119	0.0119	0.0073	0.0012	0.0012	0.0007

Fresh-water / Marine Water	Parameter	For LTA based on 99th percentile					
		Acute criteria			Chronic criteria		
		Input LTA	Sample Mean	Sample StDev	Input LTA	Sample Mean	Sample StDev
FW	Heptachlor	0.1670	0.1681	0.1005	0.0020	0.0020	0.0012
MW	Heptachlor	0.0170	0.0168	0.0100	0.0019	0.0019	0.0011
FW	Hexachlorocyclohexane (Lindane)	0.6422	0.6391	0.3800	0.0422	0.0421	0.0250
MW	Hexachlorocyclohexane (Lindane)	0.0514	0.0517	0.0319	NA	NA	NA
MW	Lead	67.4275	67.7079	41.2900	4.2722	4.3171	2.6020
FW	Mercury	0.6743	0.6696	0.3904	0.0063	0.0063	0.0038
MW	Mercury	0.5779	0.5819	0.3561	0.0132	0.0133	0.0080
MW	Nickel	23.7602	23.8078	14.4897	4.3250	4.2987	2.5777
FW	Parathion	0.0209	0.0209	0.0126	0.0069	0.0069	0.0041
MW	Pentachlorophenol (PCP)	4.1741	4.1309	2.4668	4.1667	4.1526	2.4910
FW	Polychlorinated Biphenyls (PCBs)	0.6422	0.6437	0.3844	0.0074	0.0074	0.0044
MW	Polychlorinated Biphenyls (PCBs)	3.2108	3.1957	1.9003	0.0158	0.0158	0.0096
FW	Selenium	6.4217	6.4374	3.8362	2.6372	2.6398	1.5653
MW	Selenium	93.1141	93.1165	55.0964	37.4478	37.2742	21.9396
FW	Silver	115.5900	115.4887	68.6670	100.2124	100.2502	59.4642
MW	Silver	0.6101	0.6144	0.3758	NA	NA	NA
FW	Toxaphene	0.2344	0.2343	0.1368	0.0001	0.0001	0.0001
MW	Toxaphene	0.0674	0.0677	0.0399	0.0001	0.0001	0.0001
MW	Zinc	28.8975	28.6322	17.5779	42.7221	42.5279	24.9355

Table 3A4. Exceedance rates (percent of observations exceeding standards) of single "1-day" observations and running "4-day" averages above chronic TSCs, and ratios of the "1-day" to "4-day average" exceedance rates. Results are based on 10,000 randomly generated concentrations from lognormal distributions defined by Long-Term Averages based on 90th, 95th, and 99th percentiles, for each parameter.

Freshwater /Marine Water	Parameter	90th Percentile			95th Percentile			99th Percentile		
		% Exceedance		Ratio 1-day / 4-day	% Exceedance		Ratio 1-day / 4-day	% Exceedance		Ratio 1-day / 4-day
		1-day	4-day avg		1-day	4-day avg		1-day	4-day avg	
FW	Aldrin/Dieldrin (Dieldrin/Aldrin)	18.84	10.01	1.88	14.45	5.00	2.89	6.89	1.12	6.15
MW	Aldrin/Dieldrin (Dieldrin/Aldrin)	19.46	10.85	1.79	14.17	4.86	2.92	7.83	1.22	6.42
MW	Ammonia (un-ionized NH3)	18.85	10.13	1.86	14.41	5.23	2.76	7.81	1.12	6.97
FW	Arsenic	18.74	9.42	1.99	14.12	5.27	2.68	7.52	1.03	7.30
MW	Arsenic	19.14	10.09	1.90	14.29	4.88	2.93	7.83	1.03	7.60
MW	Cadmium	19.19	11.18	1.72	14.37	4.89	2.94	7.24	0.99	7.31
FW	Chlordane	19.56	10.46	1.87	14.89	5.39	2.76	7.66	1.30	5.89
MW	Chlordane	18.64	9.29	2.01	14.47	5.34	2.71	7.56	1.13	6.69
FW	Chloride (Dissolved)	18.85	9.28	2.03	14.09	4.98	2.83	7.42	1.04	7.13
FW	Chlorine (Total Residual)	18.79	8.79	2.14	14.73	5.03	2.93	7.34	0.99	7.41
MW	Chlorine (Total Residual)	18.69	9.03	2.07	14.39	5.72	2.52	7.78	1.32	5.89
FW	Chlorpyrifos	19.26	9.35	2.06	14.58	5.04	2.89	7.78	1.21	6.43
MW	Chlorpyrifos	19.41	11.08	1.75	14.34	4.99	2.87	7.66	1.12	6.84
FW	Chromium (Hex)	19.14	9.33	2.05	14.29	5.42	2.64	7.83	1.51	5.19
MW	Chromium (Hex)	18.69	9.28	2.01	14.29	5.22	2.74	7.43	1.10	6.75

Freshwater /Marine Water	Parameter	90th Percentile			95th Percentile			99th Percentile		
		% Exceedance		Ratio 1-day / 4-day	% Exceedance		Ratio 1-day / 4-day	% Exceedance		Ratio 1-day / 4-day
		1-day	4-day avg		1-day	4-day avg		1-day	4-day avg	
MW	Copper	18.55	8.38	2.21	13.56	4.19	3.24	7.18	0.99	7.25
FW	Cyanide	18.68	8.99	2.08	14.17	5.03	2.82	7.60	1.14	6.67
MW	Cyanide Pt Roberts to Pt Wilson	19.25	10.29	1.87	13.34	4.66	2.86	7.78	1.03	7.55
MW	Cyanide elsewhere	18.58	9.77	1.90	14.76	5.39	2.74	7.41	1.06	6.99
FW	DDT (and metabolites)	18.99	9.58	1.98	14.64	5.12	2.86	7.73	1.12	6.90
MW	DDT (and metabolites)	18.74	9.13	2.05	14.06	4.87	2.89	7.58	1.16	6.53
FW	Endosulfan	18.10	9.09	1.99	14.23	5.03	2.83	7.45	1.09	6.83
MW	Endosulfan	18.74	9.13	2.05	14.06	4.87	2.89	7.58	1.16	6.53
FW	Endrin	18.95	9.40	2.02	13.79	5.21	2.65	7.79	1.00	7.79
MW	Endrin	19.31	10.26	1.88	14.58	4.95	2.95	7.43	1.12	6.63
FW	Heptachlor	19.47	10.12	1.92	13.83	4.64	2.98	7.55	1.07	7.06
MW	Heptachlor	19.55	10.22	1.91	14.20	5.07	2.80	7.18	0.99	7.25
FW	Hexachlorocyclohexane (Lindane)	17.81	8.29	2.15	14.13	5.26	2.69	7.52	1.00	7.52
MW	Lead	18.55	9.96	1.86	14.07	4.64	3.03	7.82	0.93	8.41
FW	Mercury	18.88	10.21	1.85	14.99	5.20	2.88	7.69	1.11	6.93
MW	Mercury	18.81	9.97	1.89	14.29	5.57	2.57	8.00	1.14	7.02
MW	Nickel	18.97	10.17	1.87	14.15	5.42	2.61	7.42	0.96	7.73
FW	Parathion	18.82	9.65	1.95	14.52	5.69	2.55	7.52	1.04	7.23
MW	Pentachlorophenol (PCP)	18.40	9.66	1.90	14.82	5.09	2.91	7.69	0.88	8.74

Freshwater /Marine Water	Parameter	90th Percentile			95th Percentile			99th Percentile		
		% Exceedance		Ratio 1-day / 4-day	% Exceedance		Ratio 1-day / 4-day	% Exceedance		Ratio 1-day / 4-day
		1-day	4-day avg		1-day	4-day avg		1-day	4-day avg	
FW	Polychlorinated Biphenyls (PCBs)	18.55	10.30	1.80	14.85	5.21	2.85	7.53	0.94	8.01
MW	Polychlorinated Biphenyls (PCBs)	19.55	9.53	2.05	13.69	4.85	2.82	7.37	1.11	6.64
FW	Selenium	18.70	9.20	2.03	14.17	5.13	2.76	7.48	0.84	8.90
MW	Selenium	19.43	9.41	2.06	13.72	4.63	2.96	7.62	0.92	8.28
FW	Toxaphene	19.23	9.95	1.93	14.01	4.69	2.99	7.56	1.09	6.94
MW	Toxaphene	19.28	10.56	1.83	14.27	5.09	2.80	7.58	1.02	7.43
MW	Zinc	18.84	9.53	1.98	14.85	5.32	2.79	7.55	0.99	7.63

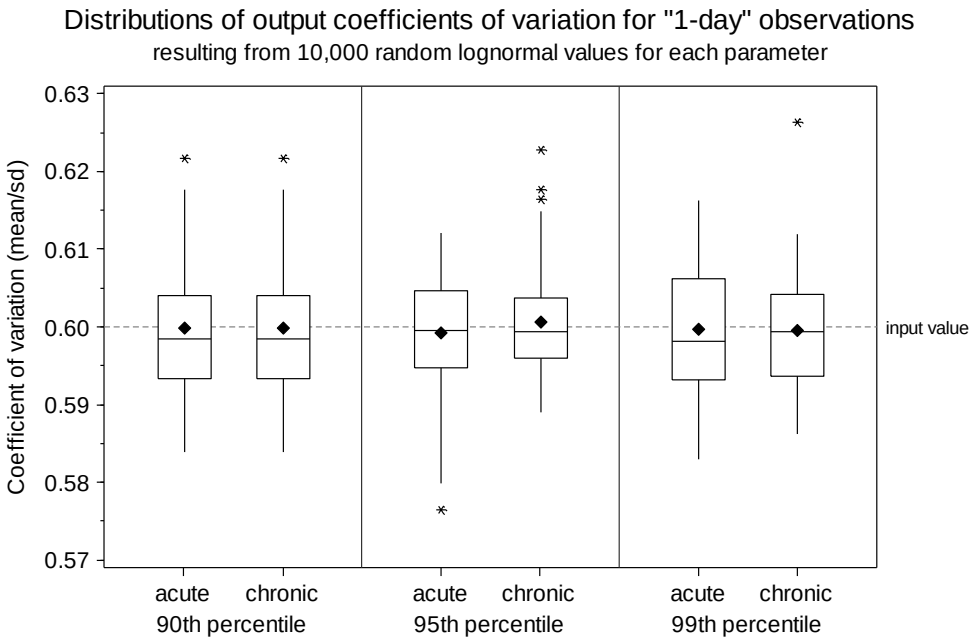


Figure 3A1. Sample coefficients of variation resulting from Monte Carlo simulations of 1-day observations, for input Long-Term Averages based on lognormal percentiles. Means are indicated by solid diamonds.

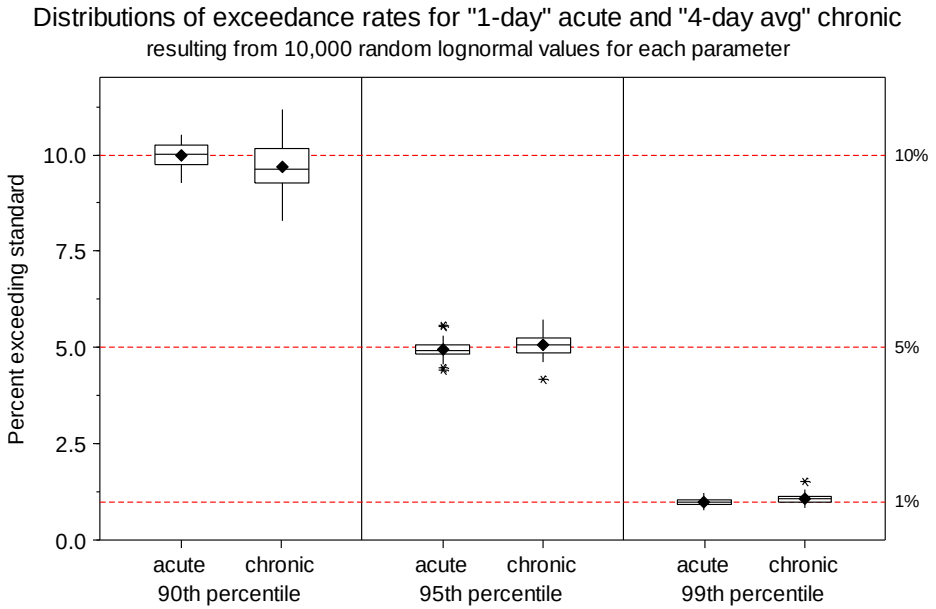


Figure 3A2. Exceedance rates for Monte Carlo simulations of 1-day observations compared to acute criteria and 4-day running averages compared to chronic criteria for each parameter, for input Long-Term Averages based on lognormal percentiles. Means are indicated by solid diamonds.