Title

The Sequential Probability Ratio Test: An efficient alternative to exact binomial testing for Clean Water Act 303(d) evaluation.

Permalink

Authors

Bartroff, Jay
Bay, Steven
Goldstein, Larry
et al.

Publication Date

DOI

The Sequential Probability Ratio Test: An Efficient Alternative to Exact Binomial Testing for Clean Water Act 303(d) Evaluation

Connie Chen[1,‡], **Matthew O. Gribble**[2,‡,*], **Jay Bartroff**[3], **Steven M. Bay**[4], and **Larry Goldstein**[3]

[1]Department of Statistics, Carnegie Mellon University, Pittsburgh, PA

[2]Department of Environmental Health, Emory University Rollins School of Public Health, Atlanta, GA

[3]Department of Mathematics, University of Southern California, Los Angeles, CA

[4]Southern California Coastal Water Research Project, Costa Mesa, CA

Abstract

The United States's Clean Water Act stipulates in section 303(d) that states must identify impaired water bodies for which total maximum daily loads (TMDLs) of pollution inputs into water bodies are developed. Decision-making procedures about how to list, or delist, water bodies as impaired, or not, per Clean Water Act 303(d) differ across states. In states such as California, whether or not a particular monitoring sample suggests that water quality is impaired can be regarded as a binary outcome variable, and California's current regulatory framework invokes a version of the exact binomial test to consolidate evidence across samples and assess whether the overall water body complies with the Clean Water Act. Here, we contrast the performance of California's exact binomial test with one potential alternative, the Sequential Probability Ratio Test (SPRT). The SPRT uses a sequential testing framework, testing samples as they become available and evaluating evidence as it emerges, rather than measuring all the samples and calculating a test statistic at the end of the data collection process. Through simulations and theoretical derivations, we demonstrate that the SPRT on average requires fewer samples to be measured to have comparable Type I and Type II error rates as the current fixed-sample binomial test. Policymakers might consider efficient alternatives such as SPRT to current procedure.
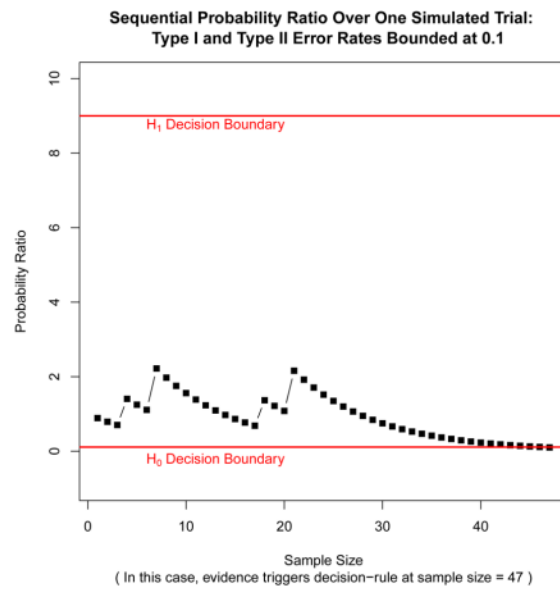
Graphical Abstract

---

[*]Address Correspondence to: Matthew O. Gribble, Ph.D., Department of Environmental Health, Emory University Rollins School of Public Health 1518 Clifton Rd NE, Mailstop 1518-002-2BB, Atlanta, Georgia 30322, Phone: 404-712-8908, Fax: 404-727-8744.
[‡]These authors contributed equally to this work: Connie Chen and Matthew O. Gribble

Conflicts of Interest

The authors have no conflicts of interest to declare.

Sequential Probability Ratio Over One Simulated Trial: Type I and Type II Error Rates Bounded at 0.1

Keywords

sequential probability ratio test; water quality; Clean Water Act; study design; Sediment Quality Objectives

INTRODUCTION

In the United States, the Clean Water Act (CWA) Section 303(d) requires states to identify impaired water bodies and to recommend total maximum daily loads (TMDLs) for contaminants affecting impaired waters, such that water bodies adhering to those TMDLs will eventually comply with water quality standards.[1] After the USEPA Administrator's approval of a state's recommended list of impaired water bodies and implementation of TMDLs, impaired water bodies are then monitored to determine whether they have attained or not yet attained the water quality standards. If a water body meets water quality standards it may be removed from the list of impaired waters (e.g., delisted). There are major regional differences within the United States in how the 303(d) listing criteria are implemented,[2] so we focus on the regulatory framework within California, although our findings from this example may be informative for other settings considering efficient alternative study designs for 303(d) evaluation. In particular, we focus on the decision rule for sediment quality as an indicator of whether a water body is impaired under the Clean Water Act.

The California Water Code section 13191.3(a) requires the state to develop standards for listing and delisting water bodies per the CWA.[3] Listing decisions are based on the frequency of exceedance of water quality standards (binary decision variable), which, for constituents such as bacteria, dissolved oxygen, contaminants, or nutrients, are numeric criteria or objectives.[4] For listing evaluations based on sediment quality in bays and estuaries, California has adopted a sediment quality objective based on a "multiple lines of evidence" approach that considers contaminant levels, sediment toxicity and sediment

macrofaunal community condition.[5–7] These multiple lines of evidence are integrated and assessed to determine whether the sediment quality objective has been attained at a given station[8] which reduces the multiple possible considerations for sediment quality into a binary decision variable suitable for evaluating exceedance frequency and responding to the 303(d) listing and delisting requirements of the CWA.

There have been several methods proposed for the analysis of binary water or sediment quality data. In 2003, Shabman and Smith recommended striking a balance between the desired Type I and Type II error rates for any 303(d) regulatory test.[9] The California EPA in 2004 considered several fixed sample size methods for Section 303(d) analyses[10] and opted for a variation on the exact binomial test with upper limits of 0.2 for both the Type I and Type II error rates as the basis for its listing decisions; delisting decisions are based on maximum error rates of 0.1. California has specified the number of maximum number of exceedances (failures) for a specified number of total samples, leading to a range of Type I and II error rates allowed for different sample sizes (Figure 1, Figure 2, Supplementary Material).

Application of the exact binomial test in California's 303(d) decisions requires a substantial number of samples to attain the specified error rates. For example, a minimum of 28 samples, with no more than 2 exceedances, is required to remove a site (water or sediment segment) from the 303(d) list.[11] For evaluating sediment quality, monitoring costs to obtain the minimum sample size to evaluate delisting could easily exceed $200,000.[10] Use of an alternative method with similar performance (i.e., Type I and II error rates), but reduced sample size requirements, would reduce the cost of compliance monitoring.

An alternative approach that can make the same assumptions as the exact binomial (i.e., independent and identically-distributed Bernoulli observations), called the *sequential probability ratio test* (SPRT), uses the data obtained from previous testing to evaluate whether adequate evidence exists at that time to favor a null or alternative hypothesis.[12,13] This is conceptually similar to the sequential Bayesian updating proposed by Qian and Reckhow for longitudinal environmental monitoring data to determine if a water body following a TMDL has attained water quality standards,[14] but here, in addition to being focused on a binary variable, our inferential goal includes making the decision of whether a water body should be listed as impaired under 303(d) and subjected to TMDL requirements. California's current regulatory testing paradigm has a parallel structure for the listing and the delisting decisions, and in this analysis we are comparing against an alternative test that also facilitates parallelism between the listing and delisting procedures.

The objective of this study is to contrast the performance of the sequential probability ratio test with California's current procedure, a fixed-sample exact binomial test, through theoretical derivations and simulation studies. Our comparison metrics are the expected number and standard deviation of the number of required samples to obtain the same Type I and Type II error rates as for the corresponding fixed-sample binomial tests. Our comparison here focuses on one alternative approach making similar assumptions to current regulatory practice, as this apples-to-apples comparison can best highlight the potential gains from more efficient methods. However, it should be noted that other efficient designs and analysis

approaches, making alternative assumptions, might be even more useful for informing 303(d) listing and delisting decisions.

EXPERIMENTAL (MATERIALS AND METHODS)

Historical Sediment Quality Data

The Southern California Bight Monitoring Program has, since 1994, coordinated a regional sediment quality monitoring survey approximately every 5 years from up to 400 stations around Southern California.[15] This large longitudinal sediment quality monitoring dataset provides a useful resource for evaluating the performance of methods to assess 303(d) compliance on realistic datasets. We recoded the five sediment quality assessment categories in the public database into the following binary categories, consistent with how these data would be used for regulatory decision-making under current practice: *meets the sediment quality objective* ("unimpacted", "likely unimpacted") or *fails to meet the sediment quality objective* ("possibly impacted", "likely impacted", "clearly impacted"). Data assessments in the database classified as "inconclusive" were excluded from analysis, just as they would also have been excluded from informing regulatory decisions. For this analysis, we focused on regional monitoring data (N=46) from San Pedro Bay in southern California, which includes Los Angeles and Long Beach Harbors. To facilitate replication of our simulations, these data are available in the Supplementary Material.

Simulated Data

We simulated Bernoulli data with the same success probability (i.e., "exceedance rate") as observed in the historical regional monitoring data (code available in Supplementary Material).

Statistical Approach

Idealized Comparisons of Exact Binomial Test vs. SPRT—Theoretical sample sizes for the exact binomial test and for the sequential probability ratio test for specified Type I and Type II error rates are derived and presented in Tables 1 and 2. Details on the theory for sequential probability ratio test are provided in the Supplementary Material. Because the sample size of the sequential probability ratio test is a random variable, we also assessed the performance of this procedure using simulated datasets. We first evaluated the empirical sample size requirements of the sequential probability ratio test to obtain the previously specified Type I and Type II error rates (Table 3). The code for all simulations is provided in the Supplementary Material.

Simulations Comparing Current California Regulatory Test Procedure vs. Truncated SPRT—The simultaneous test employed by California for regulatory purposes is a decision-rule defined by the observed numbers of failures and total numbers of trials, with different combinations detailed in tables in the *Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List.*[11] We converted these decision-rules into corresponding Type I and Type II error rates (Supplementary Material) for comparisons with SPRT. To summarize, California's decision rules currently allow Type I error rates ranging from 0.0009 to 0.16 and Type II error rates ranging from 0.0008 to

0.1885 (Supplementary Material), consistent with never allowing Type I nor Type II error rate to exceed 0.20 for listing decisions, or 0.10 for delisting decisions.

The fact that there is no fixed, *a priori* limit to the number of samples required by the SPRT to arrive at a decision could be an obstacle for environmental decision-makers. Therefore, as a proof-of-feasibility, we modified the SPRT by adding a truncation rule that would declare a water body "impaired" if no decision had yet been reached by the truncation point. We used a conservative truncation cutoff of twice the required number of observations from the corresponding current regulatory test under comparison. For example, in comparison with the state's decision-rule for 50 samples, the truncated SPRT was forced to make a decision by the 100th sample, although it typically terminated in less than 50 samples (see Figures 1, 2). In general, the lower a truncation threshold for declaring impairment, the fewer extremes are included in the sample mean (and standard deviation) number of samples for making a decision. If we apply the Precautionary Principle[16] and interpret ambiguous truncated outcomes as "impaired" when truncated, then for 303(d) listing decisions the Type I error rate increases and Type II error rate decreases, while for 303(d) delisting decisions, the Type I error rate decreases and Type II error rate increases. Therefore, the comparisons provided between the truncated SPRT and the state's current test are approximate matches, as the two tests will have slightly different empirical error rates. How different the error rates are between these two methods depends on how often the stochastic process underlying SPRT invokes truncation.

RESULTS AND DISCUSSION

Theoretical and simulation-estimated sample sizes for the sequential probability ratio test without truncation were lower than the exact binomial test, for the same power and type I error rate (Tables 1–3). For a fixed Type I and Type II error rate, the required sample sizes for non-truncated SPRT in simulations were slightly higher than the theoretical sample sizes, but had standard deviations almost as large as the expected values (Tables 2–3). Thus, the performance of the non-truncated SPRT varies according to the data on which it is used, and truncation might protect against excessive sampling.

In the applied comparison against the current listing tests used by California, the truncated SPRT also required on average fewer samples both for listing (illustrated in Figure 1) and delisting (illustrated in Figure 2) decisions.

To make our simulations as comparable to the California current practice as possible, in this analysis we applied the special case of SPRT where the observations in sequence are Bernoulli random variables, but the idea of SPRT is more general and can be used for other sets of independent and identically distributed objects, for example potentially modeling independent and identically distributed, vector-valued *batches of data* as sequential objects, rather than single Bernoulli variables. The alternative flavor of SPRT for batches of data is called group sequential testing.[17] It could be an interesting extension of this work on 303(d) evaluation methods to assess how varying the number of samples per stage in a sequential collection of batches could help optimize the procedure for real-world practicality and statistical efficiency.

Although it appears from this analysis that there could be major efficiency gains in shifting from a simultaneous to a sequential testing framework, in particular with a sequential sampling design and truncated SPRT for delisting decisions, we do not recommend that the SPRT (or its truncated version) be adopted in its current form for regulatory purposes, because there are still major limitations (e.g., failure to account for dependence between samples) to both SPRT and conventional approaches. Rather, we hope our analysis will advance a conversation leading to development of even more efficient and appropriate methods. Our analysis has focused on more efficient designs for future assessments (e.g., for future regulatory testing programs assessing regulatory compliance using data collected over a fixed period such as new data collected within a 2-year regulatory window, comparable to how current Clean Water Act determinations are made using data from the most recent assessments). Neither the SPRT nor the current approach explicitly provides for how to make use of previous observations collected prior to the current regulatory observation window. Several methodological papers focused on improving efficiency of CWA regulatory testing have noted there is often prior information available on the historical water quality and information on data from neighboring sites. The Bayesian power prior method advocated by Duan, Ye and Smith incorporates historical and adjacent site data into a binomial-model water quality assessment via a power prior, but treats "current" data as a batch to be tested simultaneously to provide a likelihood of the parameter value of interest.[18] Similarly, the Bayesian approach encouraged by McBride and Ellis uses a simultaneous sample binomial likelihood but with beta priors.[19] The TMDL compliance method recommended by Qian and Reckhow allows for sequential updating of the likelihood, but was only developed for continuous variables (in the context of attainment of TMDL goals),[14] whereas California determines CWA compliance via a binary decision variable per sample. Combining these ideas, for example by basing a prior for a "failure" parameter on historical information and updating sequentially with each sample using a group sequential test, could make better use of the complete monitoring record available. Another extension that could further strengthen the real-world appropriateness of this method is to formulate a model that would explicitly account for the spatial and temporal autocorrelation between regulatory samples collected across different stations and years; none of the regulatory decision rules we have encountered yet take this spatio-temporal autocorrelation into account.

One concern that might be raised for sequential testing in an environmental monitoring context is that the data points could be "cherry-picked" by an unscrupulous assessor to prioritize the cleanest sampling sites within a water body to be the first sampling stations evaluated, biasing CWA 303(d) decisions toward false "attainment". However, similar concerns about using unreasonable environmental sampling sites also apply to simultaneous study designs and the current "exact binomial test" as well. Any study used for regulatory decisions must include strict requirements for informative sampling.

In conclusion, the SPRT offers an efficient alternative to the current regulatory framework for CWA 303(d) listing and delisting decisions in California. In particular, for CWA 303(d) delisting decisions, adoption of a truncated SPRT that parses any inconclusive (i.e., truncated) results as "impaired" could reduce both the Type I error rate, thus better protecting public health and the environment, and the average required sample sizes, thus

reducing the cost of monitoring for adherence, relative to the status quo. Further work is needed to develop and make accessible to stakeholders (e.g., through easy to use software) new methods for CWA 303(d) evaluation that incorporate historical data, account for the autocorrelation of samples, and update evidence sequentially in order to efficiently make appropriate decisions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. United States Environmental Protection Agency. [Accessed May 26, 2014] Clean Water Act Section 303. Updated March 6, 2012http://water.epa.gov/lawsregs/guidance/303.cfm

2. Keller AA, Cavallaro L. Assessing the US Clean Water Act 303(d) listing process for determining impairment of a waterbody. Journal of Environmental Management. 2008; 86:699–711. [PubMed: 17270339]

3. [Accessed June 7, 2014] California Water Code, Section 13191.3. Updated March 17, 2014. <http://law.onecle.com/california/water/13191.3.html>

4. Gibbons RD. A Statistical Approach for Performing Water Quality Impairment Assessments. Journal of the American Water Resources Association. 2003; 39(4):841–849.

5. Bay SM, Weisberg SB. Framework for interpreting sediment quality triad data. Integrated Environmental Assessment and Management. 2012; 8(4):589–596. [PubMed: 20886640]

6. Beegan C, Bay SM. Transitioning sediment quality assessment into regulations: Challenges and solutions in implementing California's sediment quality objectives. Integrated Environmental Assessment and Management. 2012; 8(4):586–588. [PubMed: 22987517]

7. Bay SM, Ritter KJ, Vidal-Dorsch DE, Field LJ. Comparison of national and regional sediment quality guidelines for classifying sediment toxicity in California. Integrated Environmental Assessment and Management. 2012; 8(4):597–609. [PubMed: 22730345]

8. State Water Resources Control Board (SWRCB). Water Quality Control Plan for Enclosed Bays and Estuaries; Part I: Sediment quality. Division of Water Quality, State Water Resources Control Board; Sacramento, CA: 2008.

9. Shabman L, Smith E. Implications of Applying Statistically Based Procedures for Water Quality Assessment. Journal of Water Resources Planning and Management. 2003; 129(4):330–336.

10. State Water Resources Control Board (SWRCB). Final Functional Equivalent Document. Division of Water Quality, State Water Resources Control Board; Sacramento, CA: 2004. Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List.

11. State Water Resources Control Board. Water Quality Control Policy for Developing California's Clean Water Act Section 303(d) List. Division of Water Quality, State Water Resources Control Board; Sacramento, CA: 2015.

12. Bartroff, J., Lai, TL., Shih, MC. Sequential Experimentation in Clinical Trials: Design and Analysis, Springer Series in Statistics. Springer; New York (NY), USA: 2013. p. 240

13. Wald, A. Sequential Analysis. John Wiley and Sons, Inc; New York (NY), USA: 1947. p. 212

14. Qian SS, Reckhow KH. Combining model results and monitoring data for water quality assessment. Environmental Science & Technology. 2007; 41(14):5008–13. [PubMed: 17711216]

15. Schiff K, Greenstein D, Dodder N, Gillett DJ. Southern California Bight regional monitoring. Regional Studies in Marine Science. 2016; 4:34–46.

16. Jordan, Andrew, O'Riordan, Timothy. Chapter 3, The precautionary principle: a legal and policy history. In: Martuzzi, Marco, Tickner, Joel A., editors. The precautionary principle: protecting public health, the environment and the future of our children. World Health Organization; 2004.

17. Jennison, C., Turnbull, BW. Group Sequential Methods with Applications to Clinical Trials (Chapman & Hall/CRC Interdisciplinary Statistics). 1. Chapman & Hall; 1999. p. 416

18. Duan Y, Ye K, Smith EP. Evaluating water quality using power priors to incorporate historical information. Environmetrics. 2006; 17(1):95–106.

19. McBride GB, Ellis JC. Confidence of compliance: a Bayesian approach for percentile standards. Water Research. 2001; 35(5):1117–1124. [PubMed: 11268831]

Highlights

- Sequential Probability Ratio Test (SPRT) can mimic assumptions of California's test.

- SPRT needs fewer samples, on average, to obtain similar Type I and Type II errors.

- Sequential testing procedures may offer an efficient alternative to current methods.
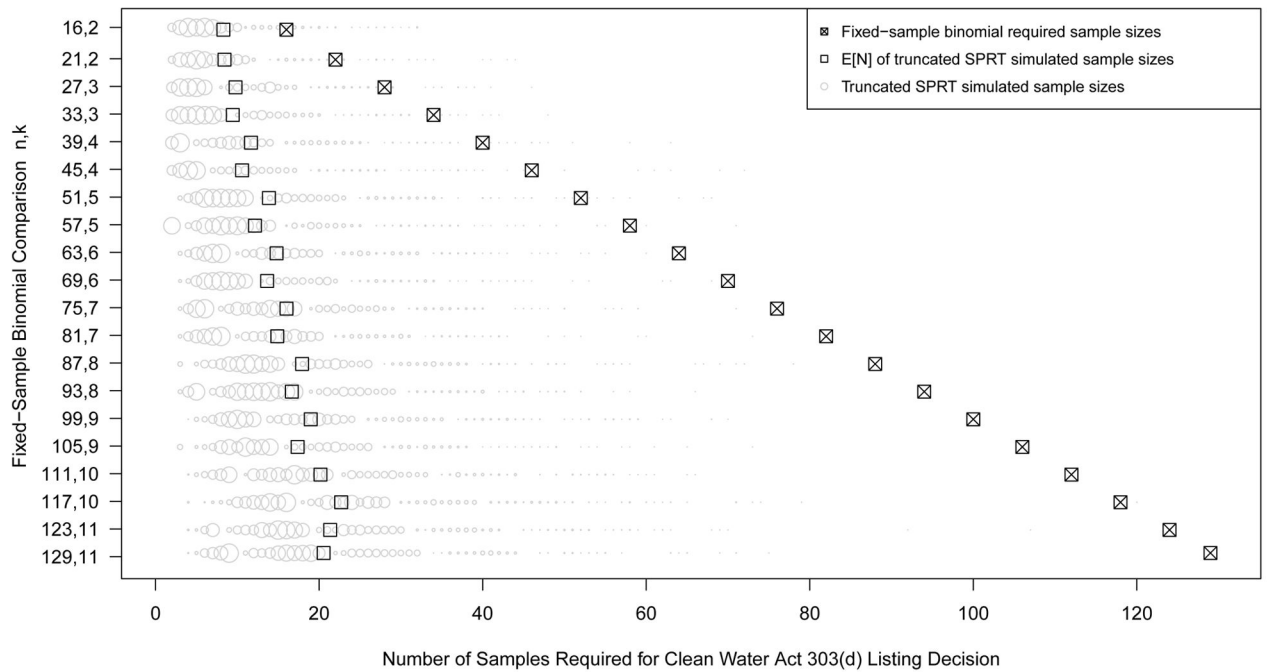
Figure 1. Simulations Comparing Sample Sizes for 303(d) Listing: Sample Sizes under Truncated SPRT vs. Simultaneous Testing

The grey circles have proportionate area to frequency of each sample size observed across simulations (e.g., these are top-down views of histograms). The black squares represent the mean expected sample sizes, per row. The black squares with X through them represent the required sample size from the corresponding fixed-sample test per the state of California's current requirements.

Figure 2. Simulations Comparing Sample Sizes for 303(d) Delisting: Sample Sizes under Truncated SPRT vs. Simultaneous Testing

The grey circles have proportionate area to frequency of each sample size observed across simulations (e.g., these are top-down views of histograms). The black squares represent the mean expected sample sizes, per row. The black squares with X through them represent the required sample size from the corresponding fixed-sample test per the state of California's current requirements.
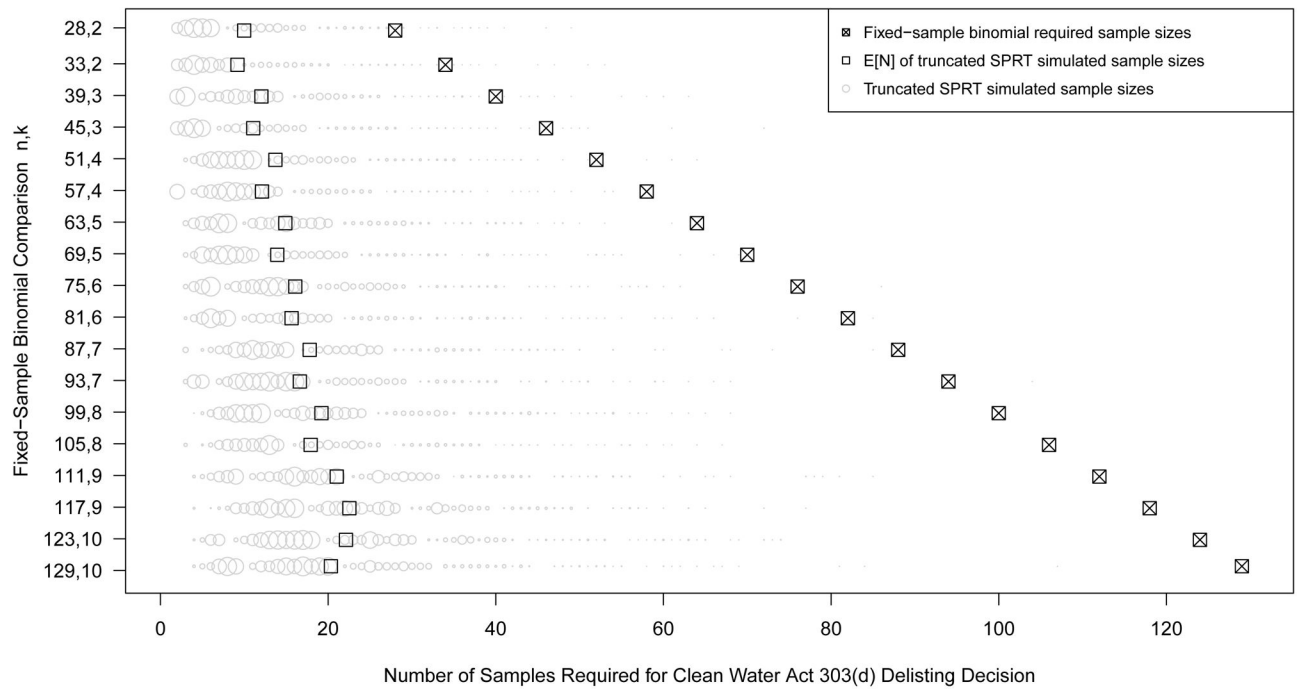
Table 1

Fixed-sample exact binomial test: sample sizes needed to achieve type I error and power.

Type I error	Power = 0.8	Power = 0.9	Power = 0.95
$\alpha = 0.05$	78	109	135
$\alpha = 0.1$	61	86	112
$\alpha = 0.2$	39	63	82

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Sequential probability ratio test: theoretical sample size needed for given type I error and power.

Type I error	Under the Null Hypothesis			Under the Alternative Hypothesis		
	Power = 0.8	Power = 0.9	Power = 0.95	Power = 0.8	Power = 0.9	Power = 0.95
α = 0.05	36.6	54.4	72.2	52.0	64.8	72.2
α = 0.1	31.2	47.9	64.8	37.1	47.9	54.4
α = 0.2	22.7	37.1	52.0	22.7	31.2	36.6

Table 3

Sequential probability ratio test: expected sample size needed for given type I error and power, and standard errors (in parentheses) from 1,000 simulated trials.

Type I error	Under the Null Hypothesis			Under the Alternative Hypothesis		
	Power = 0.8	Power = 0.9	Power = 0.95	Power = 0.8	Power = 0.9	Power = 0.95
α = 0.05	39.3 (31.6)	58.2 (43.9)	72.5 (47.1)	49.0 (36.9)	57.8 (40.4)	65.0 (48.4)
α = 0.1	34.9 (26.4)	50.4 (33.4)	67.6 (43.8)	36.3 (26.9)	45.3 (34.2)	50.6 (42.0)
α = 0.2	27.4 (19.5)	42.6 (27.7)	60.0 (39.3)	22.6 (18.7)	31.3 (27.5)	34.8 (32.9)