VIRGINIA WATER RESOURCES RESEARCH CENTER

Using Single Sample Information to Evaluate Criteria for Waterbody Health Risk

2015 Report of the Academic Advisory Committee
for
Virginia Department of Environmental Quality

SPECIAL REPORT

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
BLACKSBURG, VIRGINIA

SR57-2015
July 2015

This report is available online at http://vwrrc.vt.edu.

Virginia Water Resources Research Center (MC 0444)
210 Cheatham Hall, Virginia Tech
310 West Campus Drive
Blacksburg, VA 24061
(540) 231-5624
FAX: (540) 231-6673
E-mail: water@vt.edu

Stephen Schoenholtz, Director

USING SINGLE SAMPLE INFORMATION TO EVALUATE CRITERIA FOR WATERBODY HEALTH RISK

2015 Report of the Academic Advisory Committee for Virginia Department of Environmental Quality

by:

Eric P. Smith
Statistics Department
Virginia Tech

Edited by:
Jane L. Walker

Members of the 2015 Academic Advisory Committee to the
Virginia Department of Environmental Quality

Stephen H. Schoenholtz, Chair
Virginia Water Resources Research Center /
　　Department of Forest Resources and
　　Environmental Conservation
Virginia Tech

E. Fred Benfield
Department of Biology
Virginia Tech

Paul Bukaveckas
Department of Biology / Center for
　　Environmental Studies / Rice Center for
　　Environmental Life Sciences
Virginia Commonwealth University

Andrew L. Garey
Department of Biology / Rice Center for
　　Environmental Life Sciences
Virginia Commonwealth University

Gregory C. Garman
Department of Biology / Center for
　　Environmental Studies
Virginia Commonwealth University

Carl Hershner
Department of Biology / Center for Coastal
　　Resources Management
Virginia Institute of Marine Science
College of William and Mary

Wu-Seng Lung
Department of Civil and Environmental
　　Engineering
University of Virginia

Kevin J. McGuire
Virginia Water Resources Research Center /
　　Department of Forest Resources and
　　Environmental Conservation
Virginia Tech

Daniel McLaughlin
Virginia Water Resources Research Center /
　　Department of Forest Resources and
　　Environmental Conservation
Virginia Tech

Leonard A. Shabman
Resources for the Future

Eric P. Smith
Department of Statistics
Virginia Tech

Leonard A. Smock
Department of Biology / Rice Center for
　　Environmental Life Sciences
Virginia Commonwealth University

Kurt Stephenson
Department of Agricultural and Applied
　　Economics
Virginia Tech

Jane L. Walker
Virginia Water Resources Research Center
Virginia Tech

Gene Yagow
Department of Biological Systems
　　Engineering
Virginia Tech

Carl E. Zipper
Department of Crop and Soil Environmental
　　Sciences
Virginia Tech

Summary

Recommendations from the U.S. Environmental Protection Agency (EPA) published in 2012 for bacteria water-quality evaluation are based on criteria using a geometric mean (GM) and a statistical threshold value (STV). If the GM calculated from water samples taken at a monitoring site exceeds the recommended GM criterion or if 10% of the samples exceed the recommended STV, then the waterbody is in violation. The recommendations indicate a minimum of four samples be used for calculations.

In this report, evaluation of water quality using a single sample is statistically compared to the EPA approach for waterbodies that are in compliance and for those that are not in compliance. When the waterbody is truly in compliance with the recommended GM, the probability of a false declaration (declaring the waterbody to be in violation) for the single sample approach is below 0.5 (50%) and decreases as the true GM of the waterbody decreases. When the waterbody is truly in violation, the false declaration (saying the waterbody is in compliance), as based on a single sample, decreases from 0.5 at the GM criterion to close to zero for waterbodies with GMs that are just below the STV.

When multiple samples are available, the probability of declaring a waterbody to be in violation increases as a function of sample size regardless of whether or not the waterbody is truly in violation or not. Hence there is an increase in the true declaration of a violation (when the waterbody is truly in violation) as well as in the false declaration of a violation (when the waterbody is truly in compliance). Relative to the GM approach, the single sample approach will almost always have higher error rates.

The EPA approach does not involve a statistical test and error rates for GMs on or near the boundary of the decision rule. For waterbodies near the criterion, false declarations do not decline sharply as a function of sample size. When the GM is equal to the numerical criterion, the probability of declaring the waterbody as a health risk when it is not is 0.5 regardless of the sample size.

EPA Recreational Water Quality Criteria Recommendations

In 2012, the EPA provided nationally recommended criteria for protecting human health from bacteria in coastal and non-coastal waterways designated for recreational use (*2012 Recreational Water Quality Criteria*, EPA-820-F-12-052). The recommendations from EPA are intended to focus on magnitude, frequency, and duration of exposure and are based on the geometric mean (GM) and statistical threshold value (STV). The GM is calculated as the nth root of the product of the *n* observations ($y_1, y_2, ..., y_n$) collected over a specified period of time, (equivalently, as the exponential of the average of the log-transformed observations):

$$\sqrt[n]{y_1 y_2 \dots y_n}$$

The STV is calculated as the 90th percentile of the observations.

The 2012 recommendations by EPA are summarized in Table 1. The guidelines provide two recommendations based on different estimated illness rates. For assessment purposes, the waterbody GM should not be greater than the recommended GM magnitude in any 30-day interval, and there should not be greater than a ten percent excursion frequency of the selected STV magnitude in the same 30-day interval.

Table 1. Criteria recommendations from EPA.

Indicator	Recommendation 1 CFU/100 ml		Recommendation 2 CFU/100 ml	
	GM	STV	GM	STV
Enterococci	35	130	30	110
E. Coli	126	410	100	320

Recommendation 1 = Estimated Illness Rate: 36 per 1,000 primary contact recreators;
Recommendation 2 = Estimated Illness Rate: 32 per 1,000 primary contact recreators;
CFU/100 ml = colony forming units per 100 milliliters of sample; GM = geometric mean;
STV = statistical threshold value

EPA advises states to collect weekly samples, at least, to evaluate the GM and STV during the 30-day assessment period, and it suggests conducting even more frequent sampling for popular swimming beaches. Furthermore, EPA's recommendations state (p. 42), "The number of samples, to be collected by a state in determining if WQS [Water Quality Standards] have been exceeded, is not an approvable element of a WQS package (Florida Public Interest Research Group vs. EPA, 2007). Therefore states should not include a minimum sample size as part of their criteria submission." Thus, any and all available data collected during the assessment period for the monitoring site are to be used to calculate the GM.

Virginia's Recreational Water Quality Program

Current Virginia regulations at section 9VAC25-260-170 (Bacteria; other recreational waters), state that enterococci bacteria shall not exceed a monthly GM of 35 CFU/100 ml in transition and saltwater, and *E. coli* bacteria shall not exceed a monthly GM of 126 CFU/100 ml in freshwater. Geometric means shall be calculated using all data collected during any calendar

month with a minimum of four weekly samples. The section also deals with the situation where there is insufficient data to calculate the GM:

> "If there are insufficient data to calculate monthly geometric means in transition and saltwater, <u>no more than 10% of the total samples in the assessment period shall exceed enterococci 104 CFU/100 ml.</u>"

and

> "If there are insufficient data to calculate monthly geometric means in freshwater, <u>no more than 10% of the total samples in the assessment period shall exceed 235 *E. coli* CFU/100 ml.</u>"

The majority of statewide bacteria-monitoring data in Virginia are collected on a monthly, or sometimes bimonthly, basis. Only designated swimming beaches within localities receive weekly monitoring and then only during the months of May through September. Therefore, in some months, it may not be possible or feasible to collect sufficient data to make the recommended evaluations proposed by EPA. In some cases there may only be one sample collected for a given location and 30-day assessment period.

Decision Rules

In formulating criteria and recommendations, the state and EPA are creating decision rules. Whereas the decision rules are intended to protect human health, they involve collected data, and hence there are uncertainties about the correctness of a decision. In some cases, there will be incorrect decisions. There are two basic decisions that would lead to an incorrect assessment:

A water sample (or set of samples) indicates a health risk when there is not one (a Type I error or a false positive);

A water sample (or set of samples) does not indicate a health risk when in fact there is one (a Type II error or a false negative).

The basic question to be addressed in this document is:
     <u>If only one sample is used, how does this affect the decision rule and associated error rates?</u>

The decision process implies a hypothesis and test. Specifically, the null hypothesis would be interpreted as the water is safe, whereas the alternative hypothesis is the water is not safe. In terms of the EPA recommended GM and STV the following hypotheses result.

$$H_0 : GM \leq GM_{crit}$$

$$H_1 : GM > GM_{crit}$$

or

$$H_0 : p_{90} \leq p_{90,crit} = STV$$

$$H_1 : p_{90} > p_{90,crit} = STV$$

Here GM_{crit} is the critical geometric mean, and $p_{90,crit}$ is the STV. The critical values used in this document correspond to recommendation one in Table 1 for enterococci (GM = 35 CFU/100 ml; STV = 130 CFU/100 ml).

The decision rule for the first test is simply to reject if the estimated GM is greater than the critical GM. For enterococci, this would imply we would reject the null hypothesis if the sample GM for the 30-day assessment period is greater than 35 CFU/100 ml. The decision rule for the second test is to reject the null hypothesis if the 90th percentile of the data obtained in the same 30-day period exceeds the critical value (*e.g.*, 130 CFU/100 ml for enterococci).

There are statistical issues with using the decision rule associated with the recommended GM approach. In statistics, tests for means use a standard error that is reduced as sample size increases; hence the critical value should change with sample size. Thus, from a statistical perspective, the proposed rule is not consistent with standard statistical practice. In standard statistical testing, the critical value changes as a function of sample size to maintain a constant Type I error rate. In the EPA approach, the critical value is fixed regardless of sample size. Hence, the rule implies that the Type I error rate varies with sample size.

Describing Population Distributions

Decision rules may be compared based on theoretical statistical properties of the rule if there is information about a distribution associated with the data that are available. The criteria suggested in the EPA recommendations in Table 1 may be translated into parameters associated with a distribution of bacteria levels in surface waters.

The population distribution often associated with bacteria counts is the lognormal distribution. There are two parameters associated with the lognormal distribution, the mean and the variance. The mean, also called the average or expected value, refers to the central tendency of the distribution. The GM is viewed as the mean of the lognormal distribution. The variance gives a measure of how the data distributes itself about the mean. If the variance is large, the variability of the data set is great; if the variance is small, the variability of the data set is small.

Given the thresholds from Table 1, the variance for the distribution may be calculated using two assumptions. First, assume that the GM in the table is on the boundary for the null hypothesis for the geometric mean. Hence, use the tabled GM (*e.g.*, 35 CFU/100 ml for enterococci) as the GM for the baseline distribution. Second, assume that the STV represents the 90th percentile of the distribution and from this value calculate the standard deviation. Given the mean and the 90th

percentile of the lognormal distribution, the standard deviation (represented by σ) and hence the variance, which equals the standard deviation squared or σ^2, can be calculated.

In determining the standard deviation, it is important to realize that if the observations follow a lognormal distribution, then the log of the observations follow a normal distribution. The mean of the normal distribution is given by log(GM). The standard deviation can be calculated using the following logic and mathematical equations: Given that the standard normal distribution has mean 0 and variance 1, the value from a standard normal distribution associated with the 90th percentile is given by $\Phi^{-1}(0.90)$. Then,

$$\log(p_{90}) = \log(GM_{crit}) + \Phi^{-1}(0.90)\sigma$$
$$so \ \sigma = \left(\log(p_{90}) - \log(GM_{crit})\right) / \Phi^{-1}(0.90)$$

In the equations, phi (Φ) represents the cumulative standard normal distribution function. If a percentile is substituted into the function, the value from a standard normal associated with that percentile is the result. The upper 10th percentile for a standard normal distribution (which corresponds to the 90th percentile of the distribution) is 1.282. The formula allows calculation of the variance of the normal distribution and can be used to calculate quantities associated with the lognormal distribution.

For example, if the GM and STV recommendations are substituted into the equation, we have the following values for the standard deviation for recommendation one: 1.024 for enterococci and 0.920 for *E. coli* and for recommendation two: 1.013 for enterococci and 0.907 for *E. coli*. The standard deviations are consistent with those obtained from raw data as shown in Table 2 for a data set provided by the Virginia Department of Environmental Quality.

Table 2. Summary data of *E. coli* bacteria from monitoring sites in Virginia.

Number of samples	Sample mean	Mean of log values	Geometric mean	Standard deviation	Standard deviation log values	Location
13	110.769	4.24401	69.6870	134.720	0.94996	Beaver
23	69.000	3.63989	38.0878	120.493	0.88360	Calfpasture
12	118.750	4.51882	91.7276	99.024	0.76601	Chickahominy
22	122.500	4.00970	55.1305	223.360	1.08588	Mechunk
50	72.241	3.83825	46.4443	94.487	0.83570	Ni River
50	72.241	3.83825	46.4443	94.487	0.83570	Po River
22	57.500	3.64741	38.3750	72.847	0.79244	Ramseys
22	57.500	3.64741	38.3750	72.847	0.79244	Rivanna
24	83.333	4.05818	57.8687	85.550	0.82705	Taylor Creek

Probabilities and Declarations

Once the mean and standard deviation (or variance) are calculated, the non-compliance probabilities may be calculated for single or multiple samples. The probability a site is declared

5

as a health risk may be calculated as the P(*Y*>criteria) for a single sample measurement, *Y*. If the samples are independent, then the probability of having at least one sample exceeding the criteria is given by $1-(1-P(Y>\text{criteria}))^n)$.

For the criterion based on the GM, the probability that a geometric mean exceeds the criteria is calculated as

$$P(\hat{GM} > GM_{crit} \mid GM, \sigma)$$

If the log(GM) is considered to be normally distributed, these values can be computed for different sample sizes, geometric means, and standard deviations.

Figure 1 displays a graph of lognormal distributions for enterococci and *E. coli* for hypothetical distributions with a GM and STV that are within compliance (taller, black curves) and distributions at the border of compliance (shorter, red curves). Vertical lines are drawn to represent the state criteria and the EPA guidelines. Compliance means were selected for illustrative purposes.

For enterococci, the line at the 104 CFU/100 ml represents the state criterion for sample sizes less than four, and the line at 130 CFU/100 ml is the EPA recommended STV. The enterococci GMs are 20 CFU/100 ml for the distribution representing compliance (taller, black curve) and 35 CFU/100 ml for the distribution at the border of compliance (shorter, red curve). Non-compliance probabilities for an individual observation exceeding the GM criterion are 0.144 (taller, black curve) and 0.250 (shorter, red curve) for these distributions.

For *E. coli*, the line at 235 CFU/100 ml indicates the state criterion for sample sizes below four, and the line at 410 CFU/100 ml equals the EPA recommended STV. The GMs for *E. coli* are 53 CFU/100 ml for the distribution representing compliance (taller, black curve) and 126 CFU/100 ml for the distribution at the border of compliance (shorter, red curve). Non-compliance probabilities for an individual observation exceeding the GM criterion are 0.031 (taller, black curve) and 0.068 (shorter, red curve) for these *E. coli* distributions.

Note that the figures have a different scale on both the *x*- and *y*-axes so it appears that the probabilities are actually higher for the second set of figures (those associated with *E. coli*).
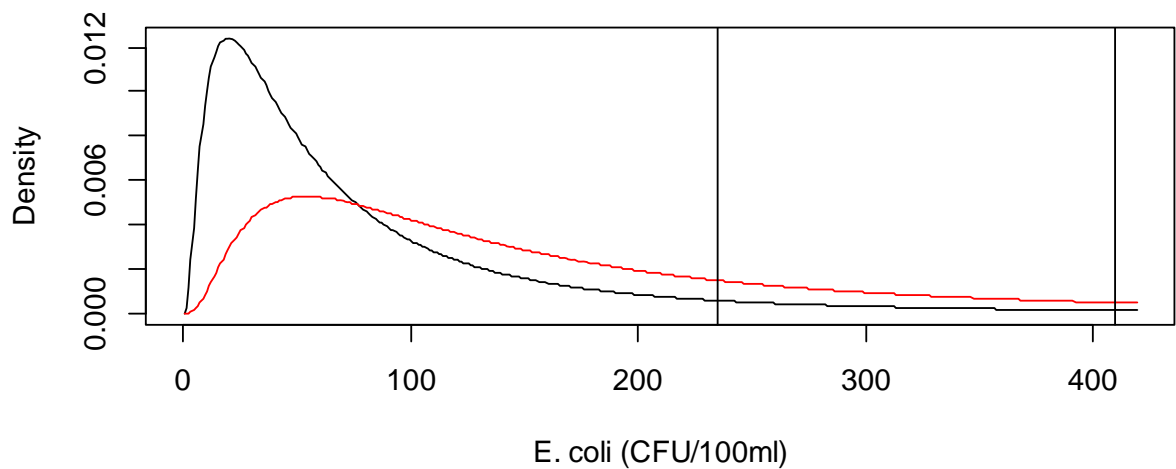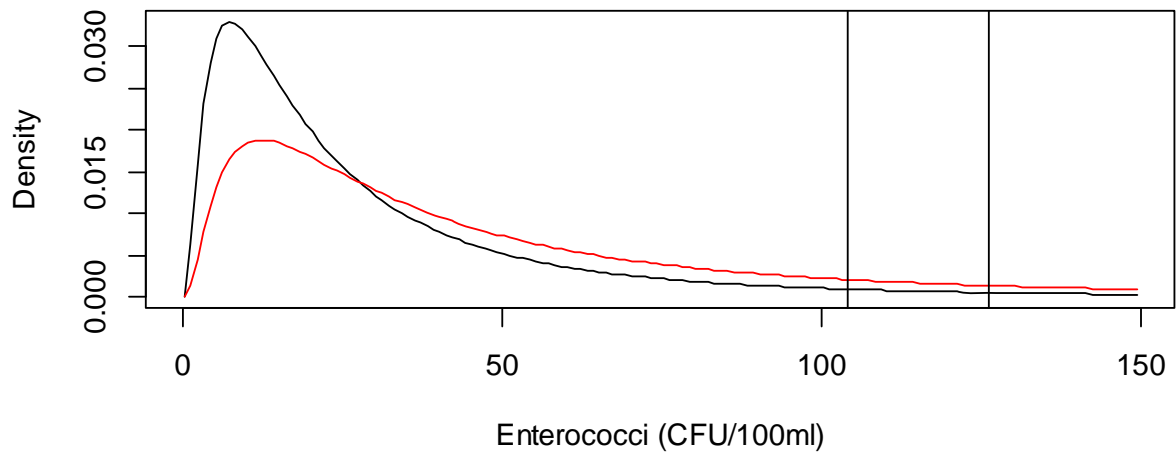
Figure 1. Plot of lognormal densities for enterococci and *E. coli* using values from recommendation one in Table 1. For enterococci, the taller, black curve was developed using a geometric mean (GM)=20 CFU/100 ml, σ =1; the shorter, red curve was made assuming GM=35 CFU/100 ml, σ=1.02. For *E. coli*, the taller, black curve was created using GM=53 CFU/100 ml, σ=1; the shorter, red curve has GM=126 CFU/100 ml, σ=0.92. Vertical lines are drawn at the Virginia criterion for sample sizes less than four (left) and the EPA recommended statistical threshold value (right).

The main focus of the report is on the probability of a false declaration. Five graphs and sets of calculations are used below to evaluate the quality of a method based on a single observation approach versus the geometric mean approach.

1. What is the probability a single observation will lead to a false risk declaration when the true geometric mean would not indicate a risk?

This calculation estimates the Type I error associated with using a single sample. The error is calculated as the probability that a single observation exceeds the recommended enterococci GM criterion (35 CFU/100 ml) when the true GM is less than or equal to 35 CFU/100 ml, *i.e.*, P(single observation>35 given true GM ≤ 35). Different values for the true GM, ranging from 3 CFU/100 ml to 35 CFU/100 ml, and standard deviation, ranging from 0.85 to 1.05, were used.

A graph of the probability that a single sample, *Y*, exceeds the GM criterion for enterococci (35 CFU/100 ml) is displayed in Figure 2 and shows several features. First, the probability of a false risk signal is small if the true GM is small (*i.e.*, there is little chance that a sample will declare a violation when the true GM is small). The risk probability increases as the true GM increases as expected (*i.e.*, as the true GM for a population in compliance gets closer to the violation criterion, the probability increases that a sample will declare the waterbody to be in violation). The probability reaches a maximum of 0.5 (50%) when the true GM equals 35 CFU/100 ml. There appears to be little effect due to changes in the standard deviations. The single sample approach will have a moderate (>30%) chance of giving a false risk declaration when the true GM is between 20 CFU/100 ml and 35 CFU/100 ml.

The bottom curve in Figure 2 represents the error associated with the GM approach using a sample size of *n*=4. It is clear that unless the GM is small or close to the critical GM, the probability is lower for the GM approach.
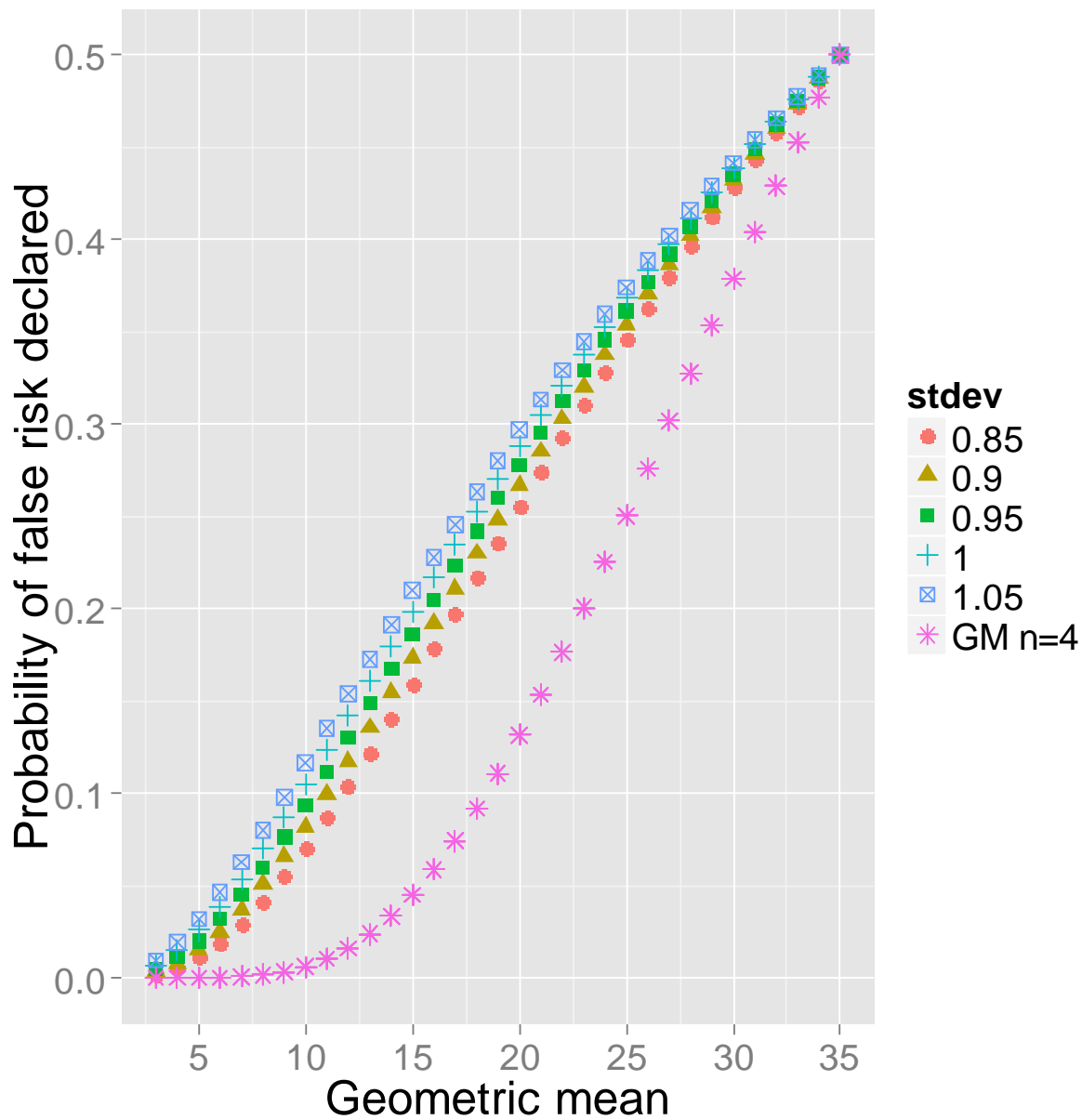
Figure 2. Plot of the probability that a single observation exceeds the criterion of 35 CFU/100 ml for distributions of enterococci with different true geometric means and standard deviations (stdev). The bottom set of points is based on the sample GM test using a sample size of four.

2. What is the probability a single observation will lead to a false declaration of no risk when in fact there is a risk?

This calculation estimates the Type II error associated with using a single sample. For enterococci, the probability that a true risk is not declared can be illustrated as the probability that an observation is less than or equal to 35 CFU/100 ml when the true GM is above 35 CFU/100 ml, $i.e.$ P(single observation $\leq$ 35 given true GM>35). The probability was calculated

for GMs from 35 CFU/100 ml to 120 CFU/100 ml and for standard deviations from 0.85 to 1.05. The GM of 120 CFU/100 ml was chosen as the upper limit as it is a value considerably above the critical GM yet below the STV (130 CFU/100 ml). Figure 3 plots this probability as a function of the true GM. As expected, the probability that a true risk is not declared declines as the GM increases. When the true mean is above 95 CFU/100 ml, the probability is quite low that a sample would be less than 35 CFU/100 ml. This finding suggests that a single sample will have a high probability of leading to a correct decision when the true GM is high. However, when the true GM is between 35 CFU/100 ml and 55 CFU/100 ml, the single sample approach will have a moderate chance (>30%) of a false signal (not declaring a risk when one is actually present).

The bottom curve in Figure 3 represents the probability associated with the GM approach using a sample size of n=4. The GM approach is clearly superior in that the error rate will be lower. When the GM is close to the critical value, the error rate for the single sample approach will be close to the GM approach.
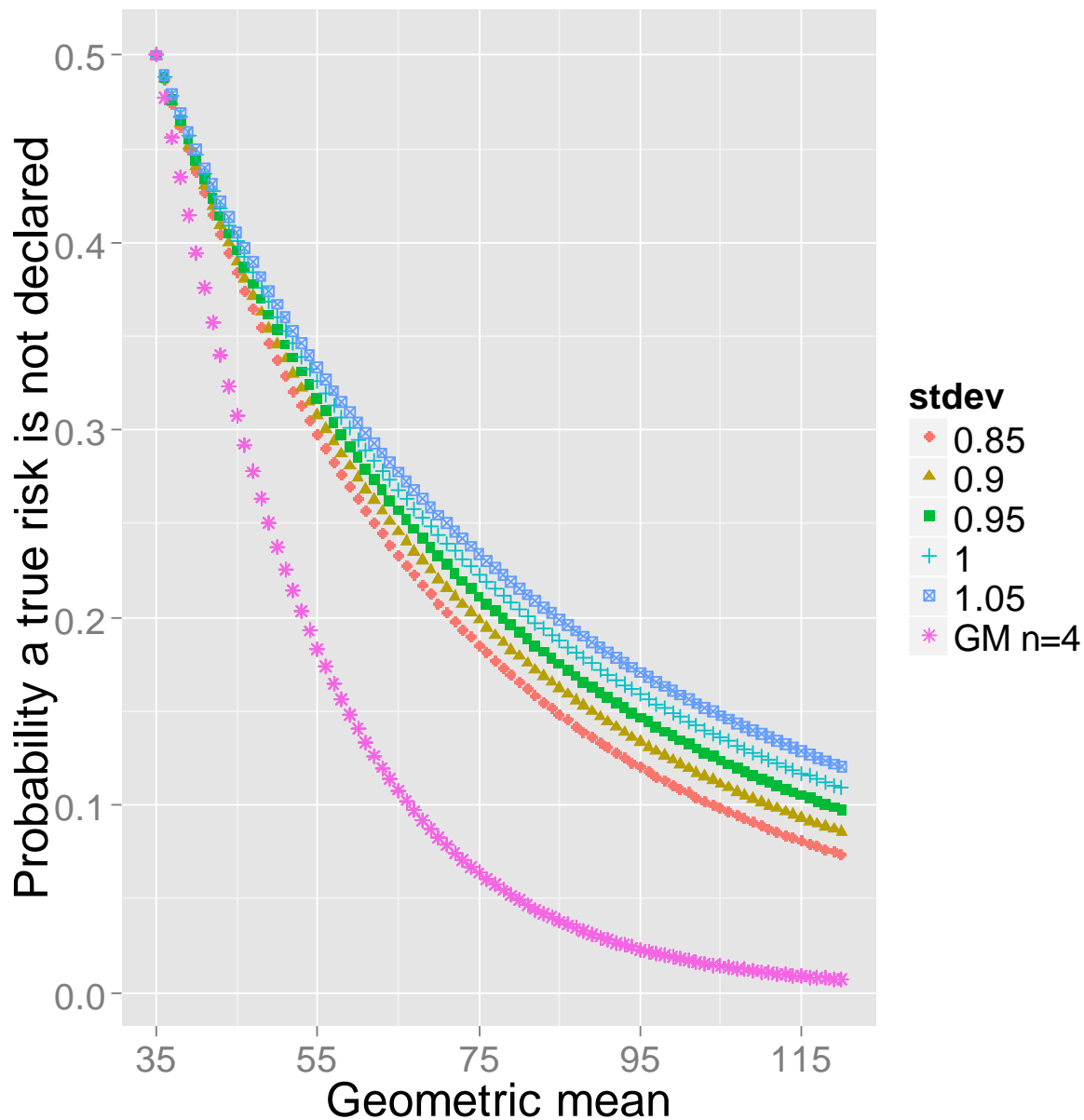
Figure 3. Plot of the probability that a true risk is not declared based on a single sample for different GMs and standard deviations (stdev). The bottom set of points is based on the sample GM test using a sample size of four.

3. How does the probability of a decision that a site is not in compliance change as a function of sample size?

To address this question, we assume multiple observations are available. The probability of non-compliance is based on the product of probabilities that at least one observation in a group of *n* observations exceeds a STV (as the sample size is 10 or less). Figure 4 is a plot of the probability of non-compliance as a function of sample size for different GMs. The figure is

based on the probability that at least one observation exceeds the state recommended enterococci STV (104 CFU/100 ml). Note that the probability of identifying a waterbody as not in compliance when it is truly not in compliance is above 0.75 when the sample size is around ten. However the false signal is also high. For example, if the true GM is 30 CFU/100 ml, then the probability that one or more samples will exceed the STV is more than 60% with ten samples. This finding suggests a high false positive rate (Type I error) for the STV approach when 104 CFU/100 ml is used as the decision criterion and sample sizes are around ten.
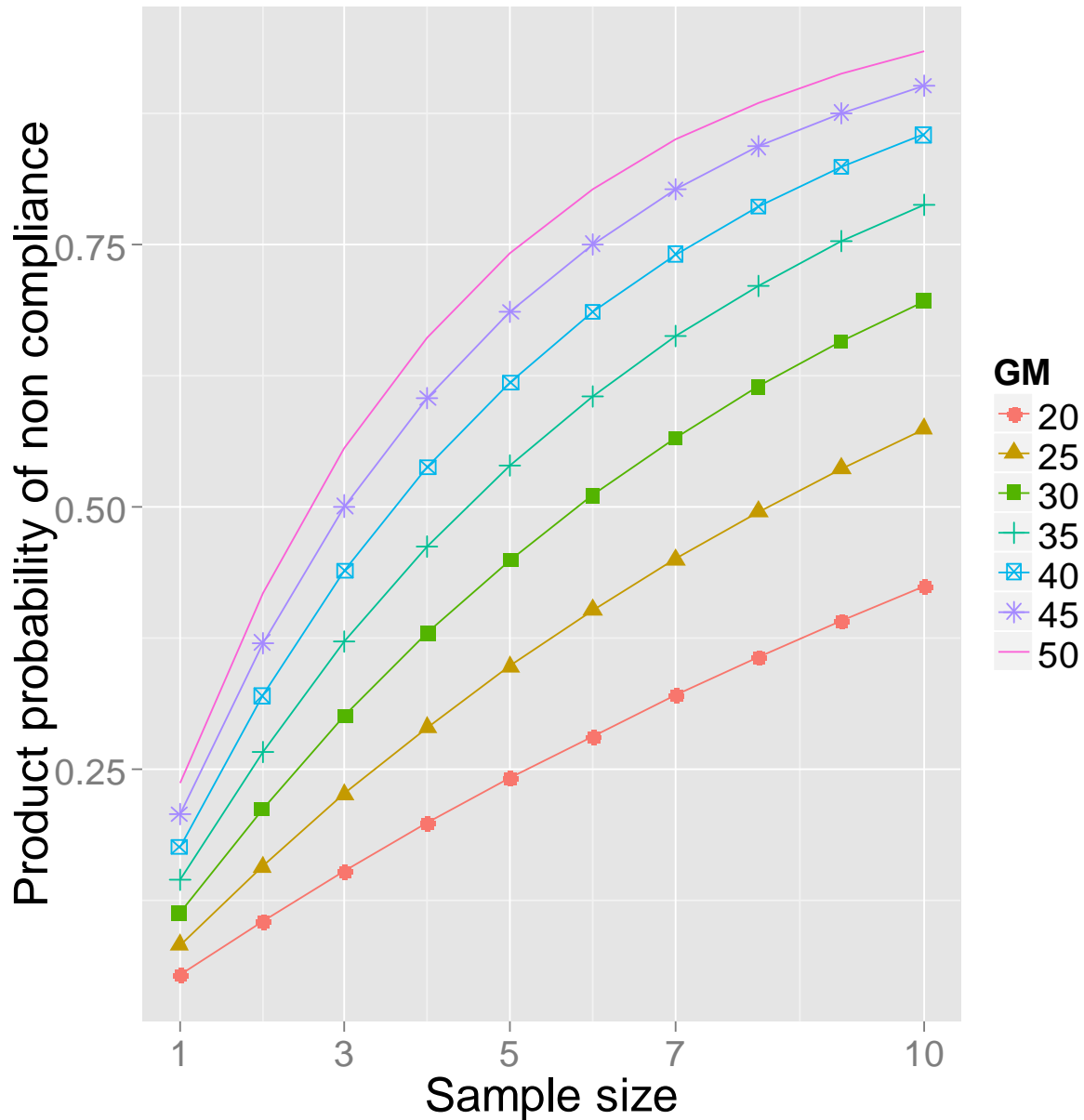

Figure 4. Plot of probability that a site is identified as not in compliance with the criterion for different sample sizes for enterococci using the statistical threshold value (STV)=104 CFU/100 ml as the criterion. The standard deviation is based on the 90th percentile from recommendation one in Table 1.

4. How much better is a decision rule based on the geometric mean relative to a rule based on a single observation as sample size increases?

One way to compare the decision process that is based on a single observation with the decision process that is based on multiple observations is to compare the probability of a correct decision for the two approaches. Figure 5 is a plot that shows the improvement in decision for the rule based on a sample of size n relative to that of a rule based on one sample. In other words, this graph represents the ratio of the probability of a correct decision for the different methods. The improvement in the GM approach relative to the single observation approach is calculated as follows:

- If true GM<35 then the improvement is given by $P(\hat{GM} < 35)/P(Y < 35)$.
- If true GM>35 then the improvement is calculated as $P(\hat{GM} > 35)/P(Y > 35)$.
- For the case where the true GM=35, the ratio is 1.0.

For small sample sizes, there is a small increase in the correct decision rate when using the GM compared to a single sample observation (*e.g.*, for four samples, the improvement is around 10-20%). When sample sizes are comparatively large (around ten), the improvement may be more than 30% in favor of the GM method. Furthermore, as the true GM moves away from the criterion (*e.g.*, 20 CFU/100 ml, 50 CFU/100 ml), there is an increase in the ratio, as expected (the GM approach is better than the single sample approach); smaller improvements are evident when the true GM is close to the criterion (*e.g.*, 30 CFU/100 ml, 40 CFU/100 ml). When the true GM is 35 CFU/100 ml, there is no difference between the decision rates.
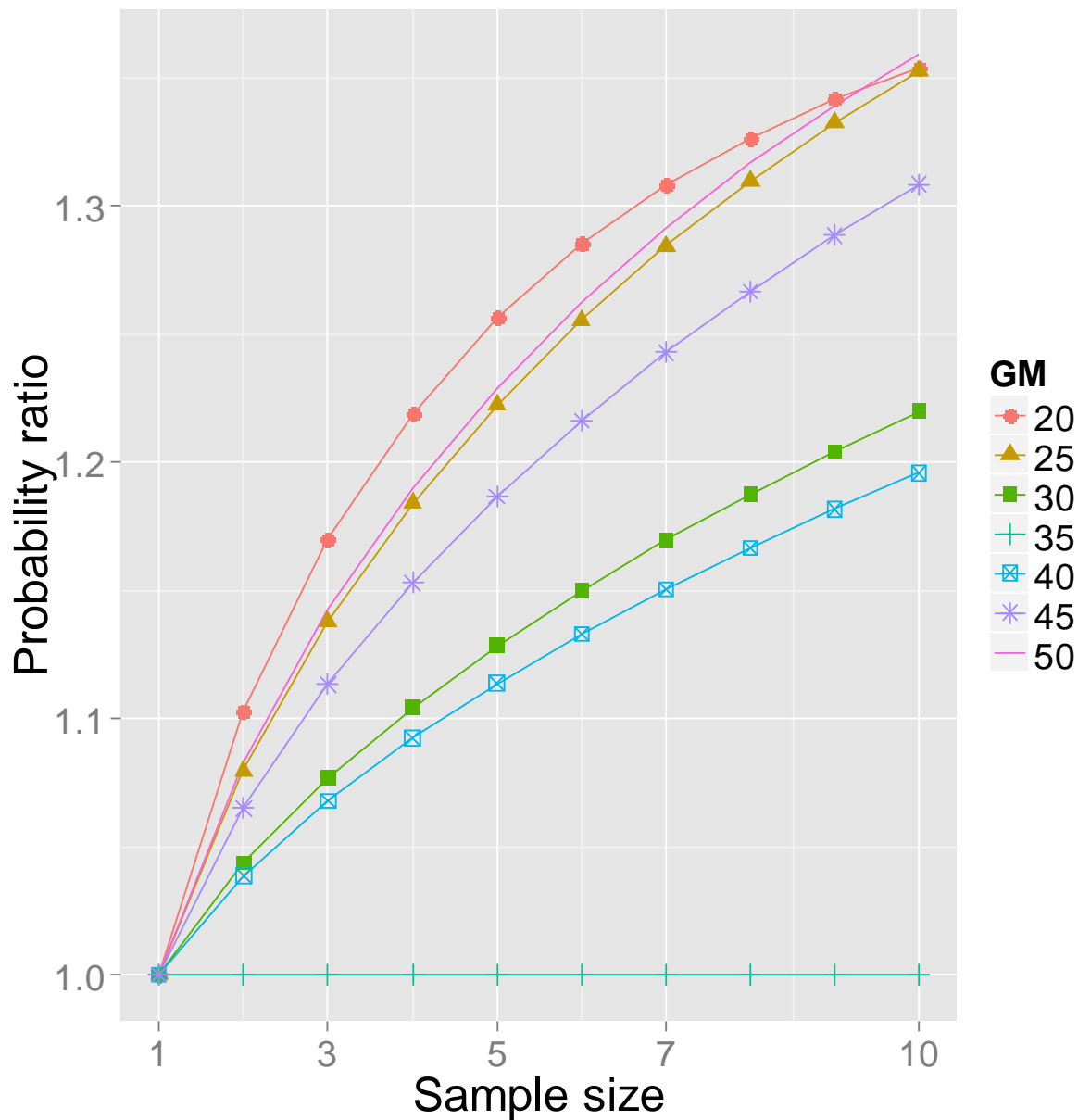
Figure 5. Plot of the improvement in the probability of a correct decision from collecting multiple samples.

5. How much better is a decision to retain or reject the null hypothesis based on the geometric mean rule relative to the single observation rule when multiple observations are used in both rules?

To address this question, assume sample sizes range from one to ten. We can then compare the EPA recommended GM approach with what is referred to as the percentile approach, whereby a site is declared at risk if 10% or more of the observed bacteria counts exceed the Virginia enterococci criterion of 104 CFU/100 ml. To compare the two approaches, we can evaluate the ratio of their probabilities for a correct decision. Note that for small sample sizes ($n \leq 10$),

Virginia's 10% rule would imply that all the observations are below the criterion for a decision of compliance.

The improvement is calculated for enterococci as follows:

- If true GM≤35 then the improvement is given by $P(\hat{GM} \leq 35) / P(Y \leq 104)^n$.
- If true GM>35 then the improvement is calculated as $P(\hat{GM} > 35) / P(Y > 104)^n$.

Figure 6 displays the ratio of the probability of a correct decision for the geometric mean test relative to the percentile test. When comparing the approaches, a ratio greater than one indicates that the geometric mean approach has a better error rate, whereas a ratio of less than one indicates that the percentile method has a better error rate.

When the sample size is small and the true GM is below or at 35 CFU/100 ml, the probability that the estimated GM is less than 35 CFU/100 ml is less than the probability that all measurements are less than 104 CFU/100 ml, and therefore the ratio is less than one (bottom left side of the graph). This result is intuitive as the probability of a sample being less than 35 CFU/100 ml is smaller than the probability of a sample being less than 104 CFU/100 ml: $P(Y<35)$ is smaller than $P(Y<104)$. The graph indicates that the percentile approach (a single sample approach) using 104 CFU/100 ml as a criterion is less likely to have a false signal under these conditions. However, as sample size increases past four, the single sample approach becomes more likely to generate a false signal. Therefore, when the sample size ranges from five to ten and the true GM is below or at 35 CFU/100 ml, the GM approach has a better error rate.

Note that when the GM equals 35 CFU/100 ml, the probability that the sample mean is less than 35 CFU/100 ml is 0.5, and the probability the sample mean is above 35 CFU/100 ml is 0.5. As sample size increases, this probability does not change. In Figure 6, however, the ratio for samples with a true GM of 35 CFU/100 ml increases as a function of sample size. This pattern occurs because although the numerator does not change, the denominator (which reflects the probability that all samples will be below 104 CFU/100 ml) decreases with increasing sample size, resulting in a ratio that increases with sample size.

When the sample sizes are small and the true GM is greater than 35 CFU/100 ml, the GM method is more likely to be correct (upper left part of the graph). However, when sample sizes are larger (five or more), the test based on the percentile approach is superior. This result for larger sample sizes occurs because, when the true GM is say 50 CFU/100 ml, we are more likely to get at least one observation to exceed 104 CFU/100 ml than for a sample GM to be greater than 35 CFU/100 ml.
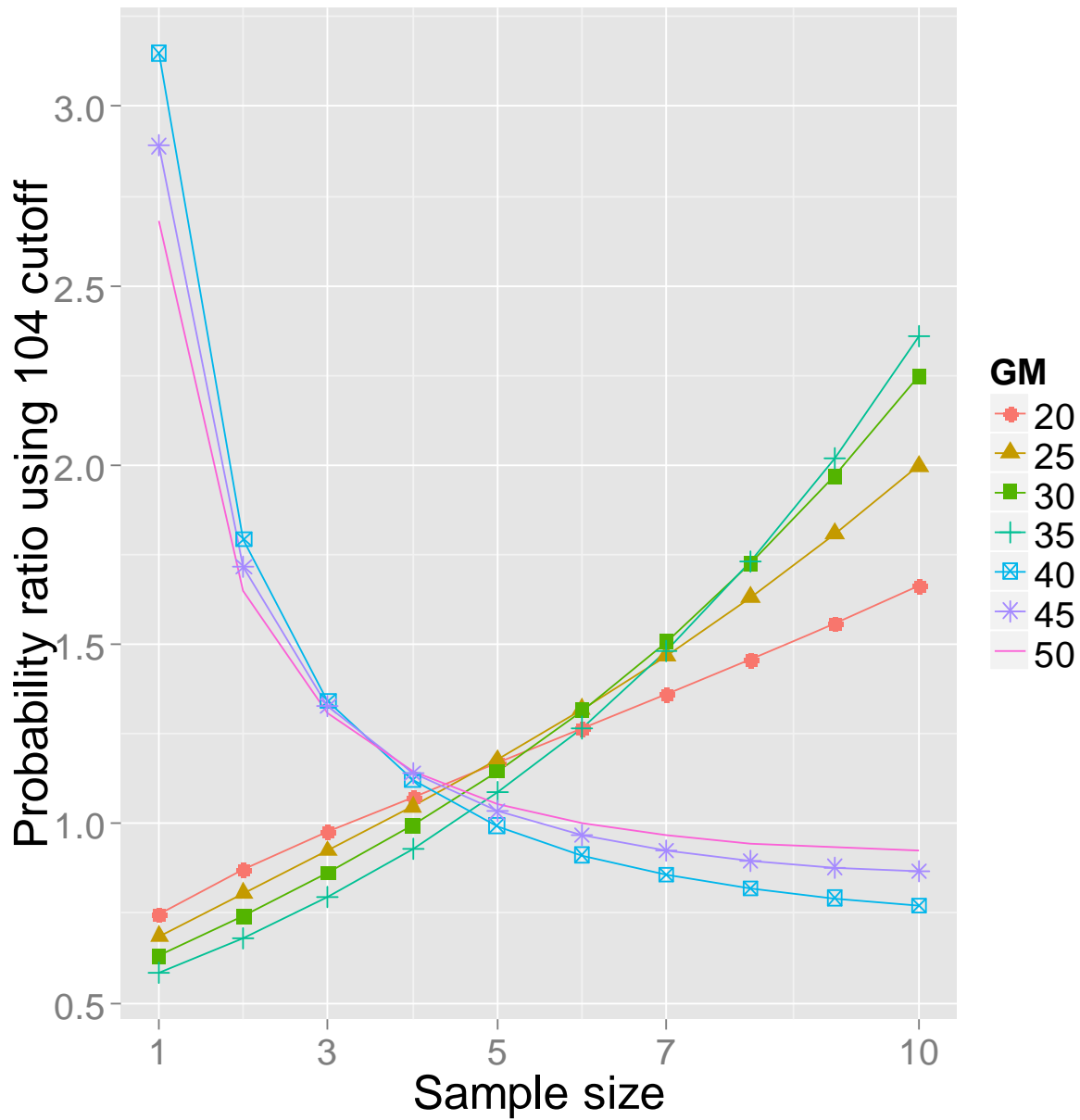
Figure 6. Plot of the ratio of the probabilities of correct decisions using the geometric mean test (with a critical mean of 35 CFU/100 ml) and the percentile test (using a critical value of 104 CFU/100 ml). A value greater than one indicates that the geometric mean approach is superior, whereas a value less than one suggests that the percentile approach is better.

Conclusions and Comments

The use of a health-risk evaluation process based on a GM criterion without taking sample size into account leads to a decision process with error rates that vary with sample size. For small sample sizes, the probability of declaring a waterbody as a health risk when it is not (Type I error) is low to moderate (<0.50) when the true GM is below the criterion (*e.g.*, 35 CFU/100 ml). This false-positive risk decreases as the true GM is reduced and decreases as sample size increases. When the true GM is at the criterion, the probability of declaring the waterbody as a health risk when it is not is 0.5 regardless of the sample size. When the true GM increases above the GM criterion, the false-negative risk (Type II error) decreases as the true GM rises and decreases with increasing sample size.

A single measurement may be used to determine if the water at a site poses a health risk. However, relative to methods using multiple samples, the single sample approach is likely to result in more false signals. This finding is especially true when sample sizes exceed five. With smaller sample sizes, the GM approach outlined in the EPA guidance has error rates that are better than those of the single sample approach; however the difference in error rates, especially when the true GM is close to the critical mean, is small.

One approach that might be considered is to implement a decision approach based on a process that uses two cutoffs rather than one. For example, with enterococci, if one sample is taken and the value is less than 35 CFU/100 ml, declare the site to be not at risk. If the sample is greater than 104 CFU/100 ml, declare the site to be at risk. If the sample is between the two values, the agency would require additional samples before a decision is made. This method may lead to a better balance of error rates.

A similar but more conservative approach would be to set the lower value to below the EPA recommended GM criterion. Specifically, calculate the mean value that results in a probability for a single value to be below some value, say 0.05, when compared with the critical mean. If the critical GM is 35 CFU/100 ml and the standard deviation is 1 (log scale), then this would result in a lognormal distribution with mean of around 7 CFU/100 ml (*i.e.*, 6.75 = exp(log(35)-1.645σ). So if the observed value is below 7 CFU/100 ml, declare the site as safe; if the observed sample is above 35 CFU/100 ml, declare that the site is a risk, and if the value is in between the two criteria, recommend collecting more data.

Although the 2012 EPA nationally recommended criteria are supposed to relate to frequency, magnitude, and duration, the sample programs do not focus on frequency and duration. Magnitude is well measured by the GM, especially for the small sample sizes that often accompany bacteria studies. For frequency and magnitude to be effective, data should be of a higher sampling frequency.