Appendix B
Statistical Method Used for Compliance Determinations for SAR

PIONEER CO-0047767, East Spanish Peaks: Wet, Sarcillo, Burro, Cow, Reilly, Santisteven and other Unnamed Canyons

PIONEER CO-0048003, West Spanish Peaks: Guajatoyah Creek, North Fork of the Purgatoire, Parras Canyon

PIONEER CO-0047776, Lorencito: Lorencito Canyon and its tributaries

XTO CO-0048054, Lorencito: Lorencito Canyon and its tributaries

XTO CO-0048062, Alamocito: Alamocito, Apache, Cherry, Ciruela, Gallegos and Unnamed Canyons

Adopted from:
WQCD 2016 303(d) Listing Methodology
Administrative Action Hearing Draft for Commission Review - March 2015
published Jan 8, 2015

APPENDIX B

Assessing Attainment of Ambient-Based Water Quality Standards in Colorado

Ambient-based water quality standards have been adopted in Colorado in limited circumstances where the table value standard cannot be met as a result of either natural conditions or irreversible, man-induced conditions. Each ambient-based standard is a site-specific characterization of *existing quality*[1] derived from "available representative data". Once an ambient-based standard has been adopted, attainment is assessed using recent, representative data.

The mechanics of setting the ambient-based standard and assessing its attainment are the same, but the characterizations are carried out with different, possibly overlapping, data sets. For dissolved metals, for example, the chronic standard is set equal to the 85th percentile of the available, representative concentration data, and the acute standard is set equal to the 95th percentile[2]. When existing quality is assessed, the 85th percentile of the available, representative concentration data in a subsequent data collection is compared to the chronic standard, and the 95th percentile is compared to the acute standard.

Assessment determines if water quality continues to meet the level of ambient quality originally characterized by the standard. In the current assessment methodology, the same quality is maintained (i.e., the standard is attained) if the assessed value does not exceed the standard. If the assessed value exceeds the standard, water quality is considered to be impaired. McBride (2005) calls this a "face value" test because it does not include consideration of sampling error.

Current assessment methodology for ambient-based standards has proven problematic. Successive assessments may yield opposite conclusions about the maintenance of existing quality. Changed assessment conclusions can have significant practical ramifications when they cause water bodies with ambient-based standards to move on and off of the 303(d) list, as has happened in a number of cases (Table 1).

Table 1. Historical changes to listings based on ambient-based standards for selected stream segments.

Assessed	COARLA01a[3]	COARLA01b	COARLA01c	COARMA04a	COSPBD01
1998	List: Se, SO4, Fe	List: Se, SO4, Fe	List: Se, SO4, Fe	List: Se	

[1] 31.5(20) "EXISTING QUALITY" means the 85th percentile of the data for total ammonia, nitrate, and the dissolved metals, the 50th percentile for total recoverable metals, the 15th percentile of such data for dissolved oxygen, the geometric mean of such data for E. coli, and the range between the 15th and 85th percentiles for pH. For temperature, for the purposes of implementing the chronic standard, "existing quality" means the maximum WAT in a three year period.

[2] Concentrations corresponding to the specific percentiles are estimated from the available data using the PERCENTILE function in EXCEL. In most cases, these concentrations are interpolated values.

[3] COARLA01 was split into segments 1a, 1b, and 1c in 2002.

Assessed	COARLA01a[3]	COARLA01b	COARLA01c	COARMA04a	COSPBD01
2002	Delist: Se, SO4, Fe	Delist: Se, SO4			
2004	List: Se	List: Se; Delist: Fe	List: Se		
2006	List: Fe				List: Se
2007				Set ambient std	Set ambient std
2008	List: SO4, Delist: Fe			Delist: Se	
2010					Set seasonal std
2012				List: Se	Delist one season

In retrospect, it should not be surprising that successive assessments of ambient-based standards could yield different conclusions even in the absence of any water quality change. Successive assessments based on the same percentile (e.g., 85th) are affected by normal variability in the available concentration data. Seasonal patterns, stochastic variation, and sampling or analytical error all contribute to that variability. Consequently, we might expect about half of the assessed values to be larger than the standard and half smaller.

When an assessment shows that an ambient-based standard is exceeded by even a small amount, the water body may be placed on the 303(d) list. Assuming no trend in ambient concentrations, it is equally likely in the next assessment cycle that the assessed value will fall below the standard. When the assessed value for a listed water body falls below the standard, the water body is removed from the 303(d) list. Thus, the examples in Table 1 of water bodies going on and off of the 303(d) list is consistent with statistical expectations for the current assessment methodology.

Having water bodies move on and off the 303(d) list creates two problems. The first problem is that it takes time and effort to develop or revise the 303(d) list. The second problem is that listing has practical ramifications for dischargers. Both problems can be addressed by adding an explicit level of confidence to assessments of ambient-based standards. The addition of a defined level of confidence would not affect the underlying definitions of existing quality or ambient-based standards, but would establish the reliability of conclusions drawn from assessments.

Statistical Approach

Increasing the reliability of conclusions drawn from assessments of ambient-based standards is based on the statistical concept of the *confidence interval*. The confidence interval is often viewed as the region around an estimate (i.e., the assessed concentration) within which the true concentration (i.e., the standard) is thought to be located[4]. If the confidence interval of

[4] McBride (2005; p. 58) explains why this simplistic view is not strictly correct. Nevertheless, it is useful for communicating the approach.

the assessed concentration (e.g., 85th percentile) does not include the standard, then the assessed concentration is significantly different from the standard.

The width of the confidence interval, and thus the range of concentrations it spans, is determined in part by the desired level of confidence. When the level of confidence is set to 95%, for example, it means there is only a 5% probability (a 1-in-20 chance) of mistakenly concluding that the assessed concentration differs from the standard (i.e., a Type 1 error). Setting the risk of a mistake to 5% (a 1-in-20 chance) would improve the reliability of future assessments compared to the current approach.

A level of confidence other than 95% could be used, but there are tradeoffs. A higher level of confidence, such as 99%, has the advantage of reducing the risk of Type 1 errors to 1% instead of 5%, but it would also result in broader confidence intervals. Having broader confidence intervals makes it less likely that an exceedance will be identified because it is more likely that the standard will fall within the interval. Conversely, a lower level of confidence, such as 90%, yields a narrower confidence interval, but an increased risk (10% probability of a Type 1 error, instead of a 5% probability) of claiming that a segment is impaired when it is not.

For most assessments, regulators are interested only in situations where the assessed concentration is significantly larger than the standard. Thus, the null hypothesis can be defined to assume that the assessed concentration is less than or equal to the standard (i.e., H_0: Assessed concentration $\leq$ Standard); the test is one-sided. Rejection of the null hypothesis for this one-sided test means that the assessed concentration is significantly larger than the standard. In this one-sided case, a 5% probability of a Type 1 error defines the risk of claiming that a water body is impaired when it is not.

Selection of a specific statistical approach is affected by the number of assessments that the Division must undertake on a regular basis. If assessments were required only occasionally for one constituent at one site, there are a number of parametric and non-parametric tests to choose from. However, the Division must contend with about 170 ambient-based standards that have been adopted in 100 water bodies across the state. Running tests separately for each constituent in each water body during each assessment cycle would not be practical for the Division, but may be appropriate and acceptable for parties that may have a narrower focus.

One way to keep the large workload manageable is to define the confidence interval for the assessed value in terms of percentiles rather than concentrations. Defining the confidence interval of a percentile is inherently non-parametric and well-suited to typical assessment data sets for which the distribution usually is not known in advance. Furthermore, it would be difficult, especially with small sample sizes, to validate a distribution. If a distribution cannot be assumed or validated, application of parametric methods becomes questionable.

Confidence intervals for percentiles are a function only of sample size. Tables can be developed to define confidence intervals that would be applicable to any constituent at any site for which the assessed data set consisted of the same number of measured concentrations. Thus, an assessment for zinc in one watershed and one for copper in another

watershed would use the same confidence interval for percentiles as long as both sites had the same number of observations.

Assessment with percentiles is best understood with some graphical examples, and a good place to begin is with the current assessment methodology. The 85th percentile is featured in this example because it is the most common among ambient-based standards, but the concept applies equally well to the other percentiles (95th or 50th). The current methodology locates the concentration that corresponds to the 85th percentile[5] of the assessed data set and compares it to the standard. When the assessed value (85th percentile) is larger than the standard, current assessment methodology registers an exceedance (Figure 1). The magnitude of the exceedance, in terms of concentration, may be large or small, but the outcome is the same. However, the current methodology does not specify the reliability of the conclusion.

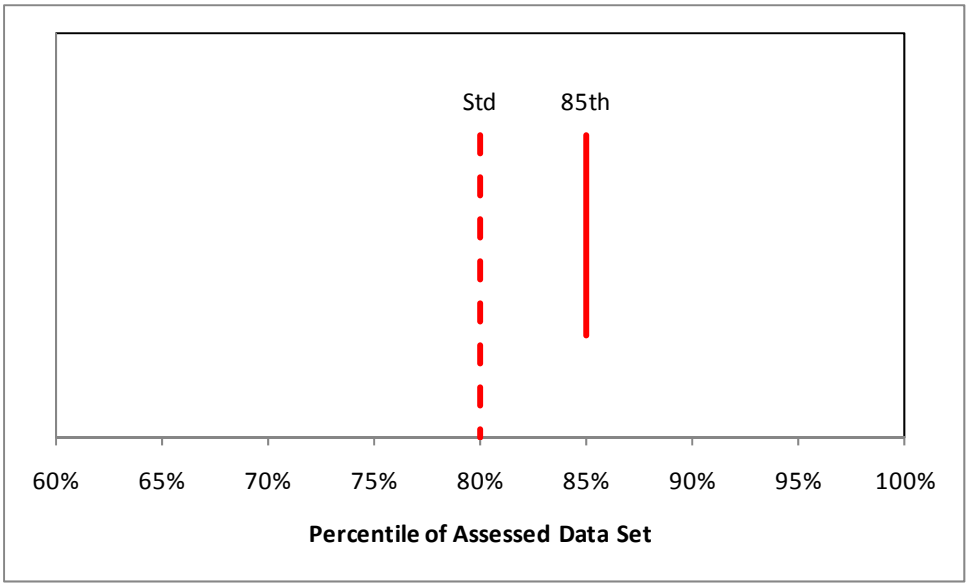

Figure 1. Graphical representation of a scenario where the 85th percentile of the assessed data (solid red line) exceeds the standard (dashed red line). For convenience in presenting the example, the concentration of the standard is represented as a percentile (80th) of the assessed data set. Current assessment methodology would interpret the result as an exceedance of the standard.

Reliability can be specified by defining a confidence interval around the 85th percentile. For example, when a confidence level of 95% is specified, the confidence interval constructed around the estimate (i.e., the 85th percentile) has a 95% probability of containing the true value (i.e., the standard). Making the right call 95% of the time is a very reliable basis for decision-making.

Building on the scenario used for Figure 1, a one-sided confidence interval is constructed for the 85th percentile of the assessed data. A statistical test is formalized with a null hypothesis

[5] The PERCENTILE function in EXCEL is used to determine the concentration that corresponds to the 85th percentile of all concentrations in the assessed data set. High concentrations correspond to high percentiles; thus, the 85th percentile is a high concentration within the assessed data set. For reasons to be explained later, the Excel function has shortcomings that should be considered in future assessments.

(H_0) stating that the 85th percentile of the assessed data set is less than or equal to the standard. The test is one-sided to determine if the assessed value (85th percentile) exceeds the standard because assessment is focused on exceedances. The null hypothesis is rejected when the LCL exceeds the standard.

In this example, the standard falls within the confidence interval for the 85th percentile (Figure 2). Thus, the concentration corresponding to the 85th percentile of the assessed data set is not larger than the standard. The null hypothesis is not rejected, and there would be no justification for listing the water body.

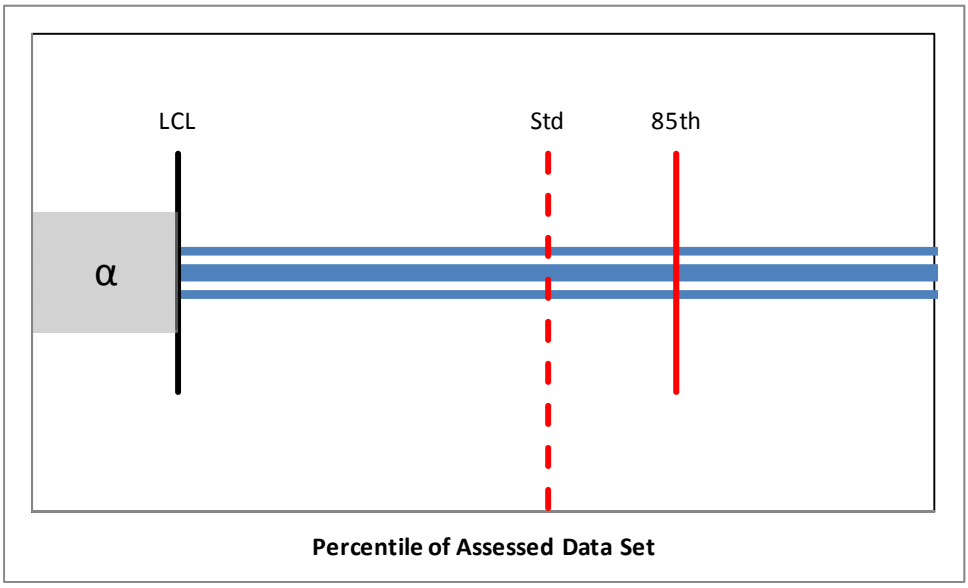

Figure 2. One-tailed confidence interval with the lower (LCL) confidence limit added to the scenario shown in Figure 1. Although the 85th percentile of the assessed data exceeds the standard, the difference is not statistically significant because the LCL of the 85th percentile does not exceed the standard. The critical region (gray region marked with alpha) extends to the left (lower percentiles) of the LCL. The confidence level for the interval is 1-α.

The scenario in Figure 2 is now changed so that the assessed value is significantly larger than the standard (Figure 3). The standard now corresponds to a low percentile of the assessed data distribution. With this scenario, the lower confidence limit (LCL) of the assessed value (85th percentile) exceeds the standard, and the null hypothesis is rejected. The assessed value is significantly greater than the standard, and the outcome would support a listing decision.

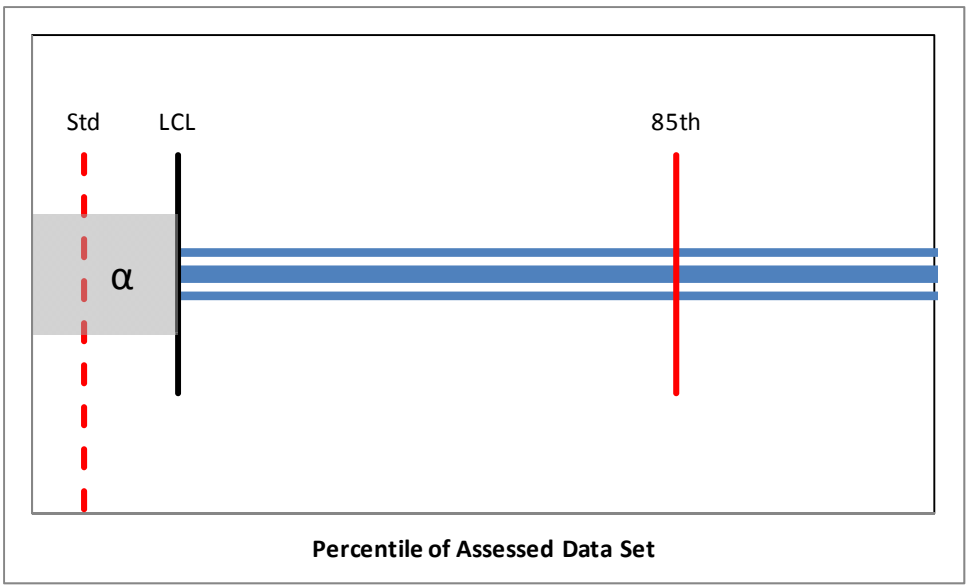

Figure 3. The scenario shown in Figure 2 is modified so that ambient concentrations are high relative to the standard. The LCL is greater than the standard, meaning that the 85th percentile of the assessed data is significantly greater than the standard.

Statistical Implementation

Defining the LCL is the key to a defensible statistical assessment of Colorado's ambient-based standards. There are a number of possible approaches (see Helsel and Hirsch 2002), but there are compelling reasons for the Division to focus on non-parametric methods, as explained below. A brief overview of common approaches provides a useful introduction to the concepts before tailoring an approach to our needs.

Overview of Methods

Variance and sample size are required for locating the LCL because the confidence interval is a statement about uncertainty. Since most environmental data sets are not normally distributed (or are too small to test for normality) a non-parametric test is preferred in most cases. A non-parametric method for locating the LCL makes no assumptions about the underlying statistical distribution of the data. Non-parametric methods for defining confidence intervals rely on the binomial distribution for defining variance. Exact and approximate non-parametric methods are available.

The Clopper-Pearson equations are used to determine exact confidence intervals for percentiles (Equation 1), but the computation is tedious, especially for large sample size. Less well-known, and much less tedious, is a direct calculation (Equation 2) using the F distribution (Leemis and Trivedi 1996), which can be evaluated in EXCEL with the FINV function.

APPENDIX B

$$\sum_{k=y}^{n} \binom{n}{k} p_L^k (1-p_L)^{n-k} = \alpha$$

Equation 1. Clopper-Pearson equation for the lower confidence interval. Probabilities are evaluated for successive proportions (p_L) until the sum reaches the desired exceedance level (α). Each proportion is calculated from the number of successes (k) out of the number of trials (n).

$$p_L = \frac{1}{1 + \frac{n-k+1}{kF_{2k,2(n-k+1),1-\alpha}}}$$

Equation 2. Exact lower confidence limit defined in terms of the F distribution; k, n, p_L, and α are defined in Equation 1. The result is identical to that obtained by iteration with Equation 1.

An exact solution for the confidence interval may sound like the ideal approach, but it is not well-aligned with Colorado's assessment needs, which would benefit from a target percentile that applies to data sets of any size. Because the binomial is a discrete distribution, assessments would logically be based on the [integer] number of samples that exceed the standard. However, the discreteness of the distribution precludes locating the LCL exactly for most sample sizes since the 85th percentile corresponds to an integer value of k only when sample size, n, is a multiple of 20[6]. When sample size is not a multiple of 20 and the confidence interval must be calculated with an integer, the resulting confidence interval would be larger than 95%. In other words, it becomes less likely that an exceedance will be identified.

Approximate methods also exist for defining confidence intervals for percentiles. In fact, many statisticians (e.g., Agresti and Coull 1998) recommend approximate methods because the exact method yields confidence intervals that tend to be too large (i.e., *exact* is something of a misnomer). A number of approximate methods have been developed for estimating confidence intervals. Historically, the Wald confidence interval has been recommended, especially when sample size is large[7]. It is also the easiest to understand.

The Wald test could be used to evaluate the null hypothesis that the estimated 85th percentile (i.e., the assessed value of a recent data set) is equal to the true 85th percentile (i.e., the standard, which characterizes existing quality). The difference between the estimate ($\hat{p}$) and the standard (p_0) is assumed to be approximately normally distributed at larger sample size. When the difference is divided by the standard error of the estimate (Equation 3), the result can be compared to standard normal deviates (z). Inverting the test yields a two-sided, 100(1-α)% confidence interval for p_0 (Equation 4).

[6] Given that the percentile is calculated as p=k/n, the first integer combination that delivers p=0.85 is when k=17 and n=20.

[7] A sample size of 30 or more is often regarded as large, but this rule of thumb may only be helpful where the central limit approximation is applicable. See Brown et al (2001) for a brief review of common sample size recommendations.

$$z = \frac{(\hat{p} - p_0)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

Equation 3. Wald test statistic for the difference between estimated and true values of a percentile.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Equation 4. Inversion of the Wald test statistic to yield a confidence interval for the estimated percentile.

For many years, the Wald interval was the recommended approach for large sample size. A perusal of the literature shows that there is not much agreement now on what constitutes "small", especially where percentiles are extreme (e.g., close to zero or to 100%). At small sample size, which is common for assessments, the Wald interval tends to be too small to accurately define the 95% confidence interval. However, concerns about the performance of the Wald interval extend beyond the issue of sample size. Brown et al (2001) have shown that the Wald interval also exhibits "erratic behavior" even when sample size is large.

An alternative approximate method that seems to have broad support in the statistical literature is the Wilson, or Score Test, interval (Brown et al 2001 & 2002, Agresti and Coull 1998). It is an inversion of the score test, and the development of the equation is reviewed in Agresti and Coull (1998). The Wilson interval is proposed in preference to the Wald interval for improving assessment of ambient-based standards. The equation for the Wilson interval is somewhat intimidating (Equation 5), but it is manageable on a spreadsheet.

$$\left\{\hat{p} + z_{\alpha/2}^2/2n \pm z_{\alpha/2}\sqrt{\left[\hat{p}(1-\hat{p}) + z_{\alpha/2}^2/4n\right]/n}\right\}/(1 + z_{\alpha/2}^2/n)$$

Equation 5. The Wilson confidence interval for an estimated percentile.

Adaptation of Wilson Interval Method

The Wilson interval method is an improvement over the exact method and the Wald interval, but the discreteness issue remains. In exact and approximate methods, the target proportion ($\hat{p}$ = k/n) is formally defined by the [integer] number of "successes" (k) relative to sample size (n). As mentioned previously, most sample sizes do not match exactly the proportion (e.g., 85%) used for assessment.

The idiosyncrasies of Colorado's assessment method prompt consideration of a departure from the usual statistical approach of scoring "successes" on the basis of individual samples. In current methodology, assessment is based on locating the concentration that corresponds to the 85th percentile of the assessed data set. Usually, this step is carried out with the PERCENTILE function in EXCEL, which interpolates between measured concentrations. In other words, the assessed value rarely matches a measured concentration from the assessed data set.

APPENDIX B

The equation for the Wilson interval requires values for $\hat{p}$, n, and z. There is no computational impediment to finding the LCL for any percentile in any data set. Hence, the equation is used to define the LCL for the three conventionally-assessed percentiles (50th, 85th, and 95th) over a wide range of possible sample sizes (n = 5 through 100) with a one-sided, 95% confidence level (i.e., $\alpha = 0.05$). The values are summarized in **Supporting Table 1**. However, due to a quirk in the way EXCEL calculates percentiles, the values in **Supporting Table 1** must be adjusted, as explained later, when assessments are undertaken with EXCEL spreadsheets.

Translation of the LCL into an Excel-compatible Percentile

Percentiles in Excel are calculated by a method that does not match the calculations used to establish the LCLs. The computational differences are not major, but become increasingly important for small sample sizes. Insofar as the Excel function is widely used, we have incorporated an adjustment such that **Supporting Table 2** contains Excel-compatible LCLs. The basis for the adjustment and the interpretation of **Supporting Table 2** are given below.

Percentiles in Excel are set such that the smallest value is defined as 0% and the largest value is defined as 100%. The formula is p' = (k-1)/(N-1), where k is the rank of the observation and N is the number of observations in the data set (see Schoonjans et al 2011). Development of the LCL is based on the binomial distribution, which defines percentiles for each ranked observation as k/N. The largest concentration is still set to 100%, but the smallest observed concentration is 1/N rather than 0%. Thus, when the LCL is converted to a concentration, it could be smaller than the smallest observed value (and thus represent a percentile between zero and 1/N).

The difference between the two formulas affects conversion of the LCL to a concentration using EXCEL functions. Direct conversion of the LCL from **Supporting Table 1** with an Excel function would not yield the correct value. The error is very small when sample size is large, but cannot be ignored at small sample sizes. Therefore, an adjustment should be made.

The adjustment relies on simple algebra to translate the LCL from **Supporting Table 1** into LCL' that is compatible with Excel functions. In the Excel formula, where LCL' = (k-1)/(N-1), the binomial formula, LCL = k/N, is rearranged to enable substitution for k (k=LCL*N). After substituting for k, the equation for the translation is LCL' = (LCL*N-1)/(N-1). This translation is used to produce an Excel-compatible set of LCLs in **Supporting Table 2**. The threshold concentration can be derived directly with the Excel function as follows, PERCENTILE(*concentration_data*,LCL') where the assessed concentration data are in a defined range of cells.

Comparison to Exact Method

Exact and approximate methods can be compared only in those few cases where the integer values of k yield the target percentiles because the proposed method for locating the LCL is not restricted to discrete values of k (p=k/n). For the 50th percentile, the comparison can be made for all even values of n; it shows that the exact method is more conservative, as expected, especially at small sample size. The exact and approximate methods agree within

10% when n exceeds about 20 (Figure 4.). Agreement between the two methods is generally even better for the 85th and 95th percentiles than the 50th percentile.

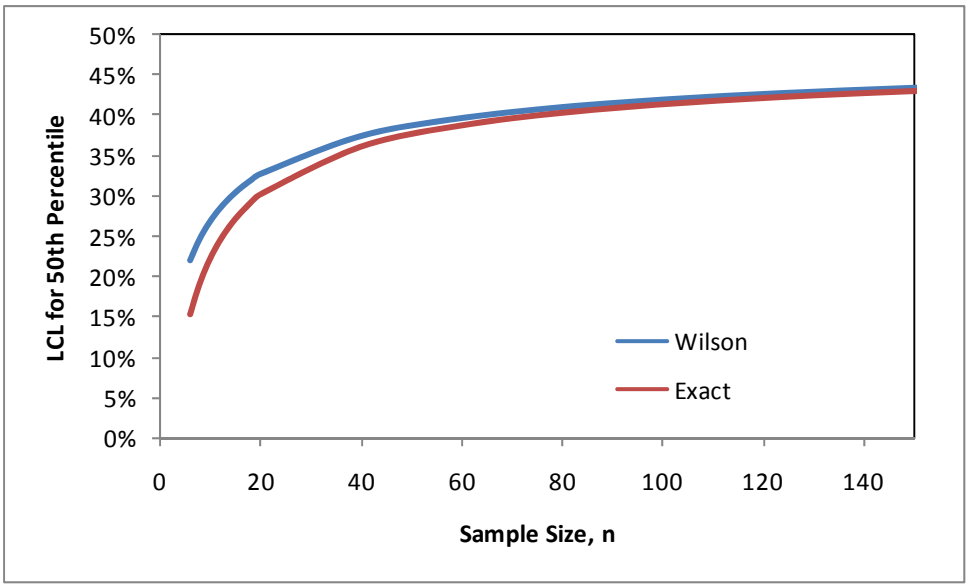

Figure 4. Comparison of LCLs defined by Wilson and exact methods for the 50th percentile. Comparison is limited to sample sizes with even numbers because the exact method can be evaluated only with integer values.

Sample Size

Sample sizes for water quality assessments tend to be small, often less than 10 samples for a segment. As is evident from **Supporting Table 1**, confidence intervals for percentiles are broader when sample sizes are small. Although LCL values can be calculated with the Wilson interval equation for virtually any sample size, there is a practical reason for avoiding very small sample sizes. When the percentile of the LCL is smaller than that of the lowest measured concentration (i.e., <1/N), it would correspond to a concentration smaller than any that were measured. For example, when only four samples are available, the LCL for the 50th percentile would be 0.182, which is smaller than 1/N (=0.25=1/4). When five samples are available, the LCL for the 50th percentile is 0.204, and this is larger than 1/N (=0.200). For the purpose of making listing decisions with ambient-based standards, at least five samples are required.

Setting a minimum of five samples for assessment of ambient-based standards differs from current assessment practice. With the latest version of Colorado's assessment methodology, a firm decision to place a segment on the 303(d) list requires more than ten samples (assuming the decision is based solely on concentration data for one constituent). With 4 to 10 samples, a listing decision must be backed up with evidence in addition to measured concentrations. For three or fewer samples, high concentrations would at most trigger further sampling (M&E list).

Adoption of the LCL table could simplify decision-making for the listing methodology by having only two pathways related to sample size. In order to make a listing decision, there

must be at least five representative[8] samples. When there are at least five samples, no additional supporting information is required because conclusions are equally reliable whether sample size is five or ten or fifty. When there are less than five representative samples, no action should be taken. If there is insistence on having a pathway to the M&E list, it should be on the basis of a table with a smaller confidence interval (e.g., 90%), which would also include a lower threshold for sample size.

Examples with Assessment Data

The current methodology for assessing attainment of ambient-based standards can be improved substantially by adding a defined level of confidence for the attainment decision. The statistical justification for the change is strong. Working through examples with real data is a good way to show that the improved approach is practical and efficient. The following examples incorporate data from the historical record for illustrative purposes, and they are not intended for reaching conclusions in the formal assessment process. Standards or segment descriptions may be changed through Commission actions (as happened with segment COARMA04a), and more recent data may be available for assessment. Nevertheless, these examples retain value for comparing old and new assessment methods for ambient-based standards.

One example is taken from Wildhorse Creek above Highway 50, which has ambient-based standards for selenium. Recent measurements (N=36; 2005-2011) of selenium concentrations serve as the assessed data set (Table 2). Most of the observed concentrations exceed table value standards (ch=4.6 ug/L; ac=18.4 ug/L) by a wide margin.

Date	Se, ug/L	Date	Se, ug/L	Date	Se, ug/L
10/13/2005	420	4/3/2007	500	7/1/2010	441
12/19/2005	496	6/5/2007	429	8/1/2010	564
2/15/2006	593	6/13/2007	95	9/1/2010	487
3/2/2006	535	9/20/2007	754	10/1/2010	479
4/26/2006	480	12/18/2007	691	11/1/2010	539
5/31/2006	362	12/1/2009	556	11/16/2010	1900
7/6/2006	9	1/1/2010	355	12/1/2010	618
9/5/2006	410	2/1/2010	646	2/1/2011	554
10/11/2006	361	3/1/2010	641	2/7/2011	1800
12/5/2006	224	4/1/2010	728	3/1/2011	607
2/20/2007	340	5/1/2010	605	4/1/2011	581
3/9/2007	531	6/1/2010	536	5/24/2011	1500

Table 2. Selenium concentrations measured in Wildhorse Creek above Highway 50.

The first step in assessing ambient-based standards by current methodology is to determine the 85th and 95th percentiles[9] of concentrations in the assessed data set. The 85th percentile

[8] The most recent listing methodology describes the factors to be considered when judging if data are "representative". Factors typically include spatial and temporal distribution of sampling effort, as well as analytical considerations and atypical events in the watershed.

[9] Consistent with current methodology, threshold concentrations are determined with EXCEL's PERCENTILE function.

concentration of the assessed data set is 680 ug/L, which exceeds the ambient-based chronic standard of 597 ug/L (Table 3). The 95th percentile concentration is 1575 ug/L, which exceeds the ambient-based acute standard of 708 ug/L. The assessed values exceed the ambient-based standards, which, by current assessment methodology, would trigger a listing.

Metric	Chronic	Acute
Ambient-based Standard	597	708
Assessed 85th Percentile	680	---
Assessed 95th Percentile	---	1575
Current outcome	Exceeded	Exceeded
Sample Size	36	36
LCL percentile (Appendix A)	0.728	0.853
LCL, Excel-compatible (App B)	0.720	0.849
LCL concentration	605	678
New outcome	Exceeded	OK

Table 3. Assessment of data for attainment of ambient-based selenium standards in Wildhorse Creek. LCL percentiles are taken from Appendices with N=36. The ambient-based standards were appropriate for segment COARMA04a at the time the samples were taken, but changes were adopted subsequently by the Commission.

The selenium data from Wildhorse Creek also are assessed with the improved methodology. Based on the sample size of 36, percentiles for chronic and acute LCLs are taken from the columns corresponding to the 85th and 95th percentiles in **Supporting Table 2**. The LCL for the 85th percentile is 0.720 (Excel-compatible value), which corresponds to a selenium concentration of 605 ug/L in the assessed data set. Therefore, the 85th percentile of the assessed data set is significantly larger than the chronic standard, which would trigger a listing by the improved methodology.

The conclusion about attainment can be properly reached based on assessment of the chronic standard alone, but the data also are assessed for attainment of the acute standard for illustrative purposes. In this case, the LCL corresponds to a selenium concentration of 678 ug/L, which is less than the acute standard. If assessment had been based solely on the acute standard, the conclusion from the improved methodology would have been that the assessed value was not significantly larger than the acute standard, which would not trigger a listing.

A second example is taken from Big Dry Creek, which has a seasonal, ambient-based chronic standard for selenium; the acute standard remains equal to the table value standard (TVS). The example deals only with data from the Apr-Oct "season". Recent measurements (N=34; 2006-2010) of selenium concentrations serve as the assessed data set (Table 4). Some of the observed concentrations exceed the chronic TVS (4.6 ug/L), but none exceeds the acute TVS (18.4 ug/L).

APPENDIX B

Date	Se, ug/L	Date	Se, ug/L	Date	Se, ug/L
4/13/2006	8.2	6/19/2008	5.7	8/12/2010	2.3
4/12/2007	9.1	6/18/2009	2.9	9/14/2006	2.8
4/17/2008	6.4	6/10/2010	3.0	9/13/2007	2.9
4/9/2009	7.0	7/13/2006	5.5	9/11/2008	7.6
4/8/2010	5.7	7/12/2007	4.9	9/10/2009	3.7
5/11/2006	2.4	7/17/2008	2.3	9/9/2010	3.5
5/10/2007	3.4	7/9/2009	2.6	10/19/2006	5.6
5/8/2008	4.0	7/8/2010	2.3	10/11/2007	10.3
5/14/2009	7.5	8/10/2006	4.4	10/9/2008	8.1
5/13/2010	3.3	8/9/2007	7.1	10/15/2009	9.5
6/15/2006	2.9	8/14/2008	8.1		
6/14/2007	2.5	8/13/2009	5.3		

Table 4. Selenium concentrations measured in Big Dry Creek above the USGS gage (COSPBD01). Data are shown for Apr through Oct, 2006-2010.

When the selenium standards for Big Dry Creek are assessed by current methodology, the chronic standard is exceeded, but the acute standard is not (Table 5). The assessment would likely trigger a listing by current methodology. The data from Big Dry Creek also are assessed for the chronic standard with the improved methodology. The LCL for the 85th percentile is 0.716, which corresponds to a selenium concentration of 6.8 ug/L in the assessed data set. Therefore, the 85th percentile of the assessed data set is not significantly larger than the standard and would not trigger a listing. Application of the improved methodology highlights the value of using a defined level of confidence to support listing decisions.

Metric	Chronic	Acute
Ambient-based Standard	7.4	18.4
Assessed 85th Percentile	8.1	---
Assessed 95th Percentile	---	9.3
Current outcome	Exceeded	OK
Sample Size	34	
LCL percentile, Excel compatible	0.716	
LCL concentration	6.8	
New outcome	OK	

Table 5. Assessment of data for attainment of selenium standards in Big Dry Creek. The chronic standard is ambient-based, and the acute standard is TVS. Consequently, only the chronic standard is assessed with the improved methodology. The LCL percentile is taken from Appendix B with N=34.

General Comments

The proposed addition to assessment methodology for ambient-based standards offers an important statistical improvement by establishing a defined level of confidence to support impairment decisions. Obtaining the benefit of a defined confidence level comes at little additional cost because assessment requires almost no additional effort. In addition, the improved methodology retains the practical advantages of the current methodology in that it works for any constituent and requires no assumptions about the underlying statistical distribution.

APPENDIX B

The approach developed by the Division is efficient and effective for routine application, and the Division plans to apply it in the next listing cycle. Nevertheless, it is not the only approach that can provide a defined level of confidence. For example, a bootstrap approach could provide a non-parametric basis for assessment. Where a distribution can be identified, parametric options could be employed. These alternatives may be suitable where the resources are available to invest in developing a statistically-defensible approach for a particular water body.

Improving confidence in assessment decisions by any statistically-defensible method addresses important concerns about future commitment of resources where impairment is identified. By incorporating a defined level of confidence in assessments of ambient-based standards, the Division can be more certain that resources committed to TMDL development, for example, will not be wasted. A more reliable basis for listing decisions also should be well-received by stakeholders, who are affected by listing decisions (or reversals).

Increasing the statistical rigor of assessment also creates more incentive to set minimum requirements for development of ambient-based standards. Sample size and representativeness merit discussion that is beyond the scope of this assessment methodology.

Development of an improved approach for ambient-based standards invites the question of why a similar approach is not also proposed for TVS assessments. In the Division's view, there are important differences between the two kinds of standards. A TVS generally represents a physiological threshold above which concentrations threaten aquatic life. In contrast, assessment of ambient-based standards hinges on detecting degradation of water quality: Have concentrations increased significantly over "existing quality"? Thus, the Division does not recommend a change in current assessment practice for TVS.

References

Agresti, A and BA Coull. 1998. Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician 52(2): 119-126.

Brown, LD, TT Cai, and A DasGupta. 2001. Interval estimation for a binomial proportion. Statistical Science 16(2): 101-133 (includes appended commentary by several authors).

Brown, LD, TT Cai, and A DasGupta. 2002. Confidence intervals for a binomial proportion and asymptotic expansions. The Annals of Statistics 30(1): 160-201.

Helsel, DR, and RM Hirsch. 2002. Statistical methods in water resources. Chapter A3, Book 4 Hydrologic Analysis and Interpretation. Techniques of Water-Resources Investigations of the United States Geological Survey. http://water.usgs.gov/pubs/twri/twri4a3/

Leemis, LM and KS Trivedi. 1996. A comparison of approximate interval estimators for the Bernoulli parameter. The American Statistician 50: 63-68.

McBride, GB. 2005. Using statistical methods for water quality methods. Wiley-Interscience. 313 p.

APPENDIX B

Schoonjans, F, D De Bacquer, and P Schmid. 2011. Estimation of population percentiles. Epidemiology 22(5): 750-751.

Zar, JH. 1999. Biostatistical Analysis, 4th Edition. Prentice-Hall. 663 p. + Appendices.

APPENDIX B

Supporting Table 1: Lower confidence limits (LCLs) for three percentiles of regulatory interest (50th, 85th, and 95th) at assessed sample sizes of 5 through 100. These percentiles should not be used with the PERCENTILE function in EXCEL; the EXCEL-compatible values are presented in Appendix B.

N	$\hat{p}$ = 0.50	$\hat{p}$ = 0.85	$\hat{p}$ = 0.95		N	$\hat{p}$ = 0.50	$\hat{p}$ = 0.85	$\hat{p}$ = 0.95
5	0.204	0.482	0.588		53	0.390	0.752	0.875
6	0.221	0.514	0.625		54	0.391	0.754	0.876
7	0.236	0.540	0.654		55	0.392	0.755	0.877
8	0.249	0.561	0.678		56	0.393	0.756	0.878
9	0.260	0.579	0.698		57	0.394	0.756	0.879
10	0.269	0.595	0.715		58	0.394	0.757	0.880
11	0.278	0.608	0.730		59	0.395	0.758	0.881
12	0.286	0.619	0.742		60	0.396	0.759	0.881
13	0.292	0.630	0.753		61	0.397	0.760	0.882
14	0.299	0.639	0.763		62	0.398	0.761	0.883
15	0.305	0.647	0.772		63	0.399	0.762	0.884
16	0.310	0.654	0.780		64	0.399	0.763	0.884
17	0.315	0.661	0.787		65	0.400	0.763	0.885
18	0.319	0.667	0.793		66	0.401	0.764	0.886
19	0.323	0.673	0.799		67	0.401	0.765	0.886
20	0.327	0.678	0.804		68	0.402	0.765	0.887
21	0.331	0.683	0.809		69	0.403	0.766	0.887
22	0.335	0.687	0.813		70	0.404	0.767	0.888
23	0.338	0.692	0.818		71	0.404	0.768	0.889
24	0.341	0.695	0.821		72	0.405	0.768	0.889
25	0.344	0.699	0.825		73	0.405	0.769	0.890
26	0.346	0.703	0.828		74	0.406	0.769	0.890
27	0.349	0.706	0.832		75	0.407	0.770	0.891
28	0.352	0.709	0.834		76	0.407	0.771	0.891
29	0.354	0.712	0.837		77	0.408	0.771	0.892
30	0.356	0.714	0.840		78	0.408	0.772	0.892
31	0.358	0.717	0.842		79	0.409	0.772	0.893
32	0.360	0.719	0.845		80	0.410	0.773	0.893
33	0.362	0.722	0.847		81	0.410	0.774	0.894
34	0.364	0.724	0.849		82	0.411	0.774	0.894
35	0.366	0.726	0.851		83	0.411	0.775	0.895
36	0.368	0.728	0.853		84	0.412	0.775	0.895
37	0.369	0.730	0.855		85	0.412	0.776	0.895
38	0.371	0.732	0.856		86	0.413	0.776	0.896
39	0.373	0.734	0.858		87	0.413	0.777	0.896
40	0.374	0.735	0.860		88	0.414	0.777	0.897
41	0.376	0.737	0.861		89	0.414	0.777	0.897
42	0.377	0.738	0.863		90	0.415	0.778	0.897
43	0.378	0.740	0.864		91	0.415	0.778	0.898
44	0.380	0.741	0.865		92	0.415	0.779	0.898
45	0.381	0.743	0.867		93	0.416	0.779	0.898
46	0.382	0.744	0.868		94	0.416	0.780	0.899
47	0.383	0.745	0.869		95	0.417	0.780	0.899
48	0.385	0.747	0.870		96	0.417	0.781	0.900
49	0.386	0.748	0.871		97	0.418	0.781	0.900
50	0.387	0.749	0.872		98	0.418	0.781	0.900
51	0.388	0.750	0.873		99	0.418	0.782	0.901
52	0.389	0.751	0.874		100	0.419	0.782	0.901

APPENDIX B

Supporting Table 2: Lower confidence limits (LCLs) for three percentiles of regulatory interest (50th, 85th, and 95th) at assessed sample sizes of 5 through 100. These percentiles are compatible with the PERCENTILE function in EXCEL; see text for explanation.

N	$\hat{p}$ = 0.50	$\hat{p}$ = 0.85	$\hat{p}$ = 0.95		N	$\hat{p}$ = 0.50	$\hat{p}$ = 0.85	$\hat{p}$ = 0.95
5	0.005	0.353	0.485		53	0.378	0.748	0.873
6	0.066	0.417	0.550		54	0.379	0.749	0.874
7	0.109	0.464	0.597		55	0.380	0.750	0.875
8	0.141	0.499	0.632		56	0.382	0.751	0.876
9	0.167	0.527	0.661		57	0.383	0.752	0.877
10	0.188	0.550	0.684		58	0.384	0.753	0.878
11	0.206	0.569	0.703		59	0.385	0.754	0.878
12	0.221	0.585	0.719		60	0.386	0.755	0.879
13	0.234	0.599	0.733		61	0.387	0.756	0.880
14	0.245	0.611	0.745		62	0.388	0.757	0.881
15	0.255	0.622	0.755		63	0.389	0.758	0.882
16	0.264	0.631	0.765		64	0.390	0.759	0.882
17	0.272	0.640	0.773		65	0.391	0.760	0.883
18	0.279	0.648	0.781		66	0.392	0.760	0.884
19	0.286	0.655	0.788		67	0.392	0.761	0.884
20	0.292	0.661	0.794		68	0.393	0.762	0.885
21	0.298	0.667	0.799		69	0.394	0.763	0.886
22	0.303	0.673	0.804		70	0.395	0.763	0.886
23	0.308	0.678	0.809		71	0.396	0.764	0.887
24	0.312	0.682	0.814		72	0.396	0.765	0.888
25	0.316	0.687	0.818		73	0.397	0.766	0.888
26	0.320	0.691	0.822		74	0.398	0.766	0.889
27	0.324	0.694	0.825		75	0.399	0.767	0.889
28	0.328	0.698	0.828		76	0.399	0.768	0.890
29	0.331	0.701	0.831		77	0.400	0.768	0.890
30	0.334	0.704	0.834		78	0.401	0.769	0.891
31	0.337	0.707	0.837		79	0.401	0.769	0.891
32	0.340	0.710	0.840		80	0.402	0.770	0.892
33	0.342	0.713	0.842		81	0.403	0.771	0.892
34	0.345	0.716	0.844		82	0.403	0.771	0.893
35	0.347	0.718	0.847		83	0.404	0.772	0.893
36	0.350	0.720	0.849		84	0.405	0.772	0.894
37	0.352	0.722	0.851		85	0.405	0.773	0.894
38	0.354	0.725	0.853		86	0.406	0.773	0.895
39	0.356	0.727	0.854		87	0.406	0.774	0.895
40	0.358	0.728	0.856		88	0.407	0.774	0.895
41	0.360	0.730	0.858		89	0.407	0.775	0.896
42	0.362	0.732	0.859		90	0.408	0.775	0.896
43	0.364	0.734	0.861		91	0.409	0.776	0.897
44	0.365	0.735	0.862		92	0.409	0.776	0.897
45	0.367	0.737	0.864		93	0.410	0.777	0.897
46	0.368	0.738	0.865		94	0.410	0.777	0.898
47	0.370	0.740	0.866		95	0.411	0.778	0.898
48	0.371	0.741	0.867		96	0.411	0.778	0.898
49	0.373	0.743	0.869		97	0.412	0.779	0.899
50	0.374	0.744	0.870		98	0.412	0.779	0.899
51	0.376	0.745	0.871		99	0.413	0.779	0.900
52	0.377	0.747	0.872		100	0.413	0.780	0.900