USGS
science for a changing world

Using the U.S. Geological Survey National Water Quality Laboratory LT-MDL to Evaluate and Analyze Data

Open-File Report 2008-1227

U.S. Department of the Interior
U.S. Geological Survey

USGS
science for a changing world

Using the U.S. Geological Survey National Water Quality Laboratory LT-MDL to Evaluate and Analyze Data

by Bernadine A. Bonn

Open-File Report 2008-1227

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
DIRK KEMPTHORNE, Secretary

U.S. Geological Survey
Mark D. Myers, Director

U.S. Geological Survey, Reston, Virginia 2008

For product and ordering information:
World Wide Web: http://www.usgs.gov/pubprod
Telephone: 1-888-ASK-USGS

For more information on the USGS—the Federal source for science about the Earth,
its natural and living resources, natural hazards, and the environment:
World Wide Web: http://www.usgs.gov
Telephone: 1-888-ASK-USGS

Suggested citation:
Bonn, B.A., 2008, Using the U.S. Geological Survey National Water Quality Laboratory LT-MDL to evaluate and analyze data: U.S. Geological Survey Open-File Report 2008-1227, 73p.

CONTENTS

Using the U.S. Geological Survey National Water Quality Laboratory LT-MDL to Evaluate and Analyze Data

By Bernadine A. Bonn

ABSTRACT

A long-term method detection level (LT-MDL) and laboratory reporting level (LRL) are used by the U.S. Geological Survey's National Water Quality Laboratory (NWQL) when reporting results from most chemical analyses of water samples. Changing to this method provided data users with additional information about their data and often resulted in more reported values in the low concentration range. Before this method was implemented, many of these values would have been censored.

The use of the LT-MDL and LRL presents some challenges for the data user. Interpreting data in the low concentration range increases the need for adequate quality assurance because even small contamination or recovery problems can be relatively large compared to concentrations near the LT-MDL and LRL. In addition, the definition of the LT-MDL, as well as the inclusion of low values, can result in complex data sets with multiple censoring levels and reported values that are less than a censoring level. Improper interpretation or statistical manipulation of low-range results in these data sets can result in bias and incorrect conclusions.

This document is designed to help data users use and interpret data reported with the LT-MDL/LRL method. The calculation and application of the LT-MDL and LRL are described. This document shows how to extract statistical information from the LT-MDL and LRL and how to use that information in USGS investigations, such as assessing the quality of field data, interpreting field data, and planning data collection for new projects. A set of 19 detailed examples are included in this document to help data users think about their data and properly interpret low-range data without introducing bias. Although this document is not meant to be a comprehensive resource of statistical methods, several useful methods of analyzing censored data are demonstrated, including Regression on Order Statistics and Kaplan-Meier Estimation. These two statistical methods handle complex censored data sets without resorting to substitution, thereby avoiding a common source of bias and inaccuracy.

INTRODUCTION

In FY 2000, the U.S. Geological Survey's National Water Quality Laboratory (NWQL) began routinely applying a new reporting procedure for high-demand water methods (Childress and others, 1999). Use of this reporting procedure has continued and is being employed for most analyses on water samples. The reporting procedure does not alter the actual analytical methods used by the NWQL, but only the way in which the results are communicated. Understanding this procedure provides the data user additional information about their data; it also presents the data user with new opportunities and challenges regarding data interpretation, especially in the low range of the method. Moreover, improperly interpreting low range results can bias summary statistics.

Purpose and Scope

The purpose of this paper is to help the data user understand the reporting procedure used by NWQL and show how to apply that understanding to better interpret data from the NWQL. Specifically, the following will be addressed:

- How the reporting parameters used by NWQL (the LT-MDL and the LRL) are defined and used in reporting results of chemical analyses
- How to extract the statistical information that is embedded in these reporting parameters
- How to use that statistical information, combined with other quality control data from the NWQL, to assess the quality of field data, interpret field data, and plan data collection for new projects
- How to properly interpret low range data to prevent distortion of the data distribution which biases summary statistics
- What factors to consider when representing and analyzing censored data
- An exploration of some methods of analyzing censored data

The techniques that will be discussed are most applicable for concentrations in the low range of the chemical method and include, but are not limited to, censored values. Particular attention will be focussed on to how to best represent censored data in light of the question being investigated. In addition, some methods of analyzing censored data will be explored. This paper, however, is not a comprehensive resource of statistical methods for analyzing censored data.

Structure of This Document

This document is divided into four parts:

1. The Background section includes a discussion of analytical error and its role in reporting the results from chemical analyses. Basic concepts such as confidence and uncertainty are reviewed.
2. The NWQL Reporting Procedure section includes a detailed description of the default reporting values used in the new procedure (the LT-MDL and LRL) and how they are applied to data. A comparison of the new reporting procedure with the one used previously by NWQL is included.
3. The Assessing Project Data Quality section describes how to determine if the LT-MDL and LRL are appropriate for a given dataset. This is particularly important if the data user is interpreting data near the low end of the analytical range or planning to use the statistical basis of the new procedure to estimate uncertainty.
4. The Examples sections contain detailed discussions and calculations for a variety of hypothetical questions and datasets.

Throughout this document, and especially for the examples, it is assumed that the reader has a familiarity with basic statistical concepts and tests (mean, median, standard deviation, t and z tests, F test, Chi square test, and others). Readers will find it helpful to have an introductory statistics text available to use as a reference for statistical tables and to review some procedures.

BACKGROUND

Understanding Chemical Analyses

In order to understand the new reporting procedure and how to use it to advantage, it is important to consider the nature of chemical analyses—in particular—the issues of detection and quantification. *Detection* addresses the question: "Is the analyte of interest present in the sample?" *Quantification* addresses the question: "What is the concentration of the analyte of interest in the sample?" Although quantification seems to be the more difficult question, it is in fact, often more straightforward to answer than detection.

Analytical error and uncertainty—All chemical analyses have errors. Errors can be classified as either systematic or random.

Systematic errors are always in the same direction. For example, some methods produce results that are low because the analyte degrades before it can be detected; other methods produce results that are high because of background contamination. Bias is the result of systematic error. Systematic errors can be discovered by a quality assurance/quality control (QA/QC) program. Often, the problems that lead to these errors can be corrected, thereby eliminating the error in future analyses. Data users need to be aware of possible systematic errors in their data and interpret results accordingly. Systematic error will only be discussed in this paper in the context of methods that are known to have low analyte recovery.

Unlike systematic errors, random errors occur with equal frequency in both directions and are unavoidable because they are caused by fundamental limitations on the ability to make perfect measurements. Random errors translate into analytical uncertainty—a region that surrounds the reported value. Almost all of the issues and examples presented in this paper relate to the analytical uncertainty produced by random error.

Because of analytical uncertainty, the reported result of a chemical analysis can never be assumed to equal the true concentration of an analyte. Rather, it is an estimate of the true concentration. Analytical uncertainty is sometimes explicitly included with the reported result (for example, 4.5 ± 0.2 mg/L), but more commonly it is either implied by significant figure conventions or not reported at all. All analytical results have associated uncertainty, regardless of whether the uncertainty is reported with the result. For a chemical analysis to be accurate or correct, the region of uncertainty around the reported value must include the true concentration. Of course, the true concentration is never known exactly. Confidence is based on the risk that the region of uncertainty surrounding the reported result does not encompass the true concentration.

The size of the uncertainty depends on three things: (1) the nature of the analytical method, (2) the chemical and physical properties of the analyte, and (3) the degree of confidence desired by the data user. Some analytical methods tend to produce large uncertainties (that is, they tend to have high variability); some tend to produce small uncertainties (they tend to have low variability). Methods that tend to produce small uncertainties are considered precise.

The choice of analytical method is partially determined by the chemical and physical properties of the analyte. Not every method will work for every analyte and some compounds are notori-

ously difficult to analyze. Some of the properties that can cause analytical difficulties include chemically reactive compounds that degrade before and during analysis, volatile compounds that evaporate, and compounds that are difficult to isolate from similar compounds. Results for such analytes may never be as precise or accurate as those for analytes that are easier to assay.

The data user cannot change the inherent nature of a method or the properties of an analyte, but they can modulate the size of the uncertainty by requiring more or less confidence. For a given method, accepting lower confidence shrinks the size of the uncertainty and requiring higher confidence expands the size of the uncertainty (fig. 1). This may seem counterintuitive, but remember that high confidence has to do with minimizing the chance of being wrong, not with minimizing the size of the uncertainty. It is a sad irony that the analytes whose methods are associated with large uncertainties often seem to be the same ones for which data users desire high confidence.

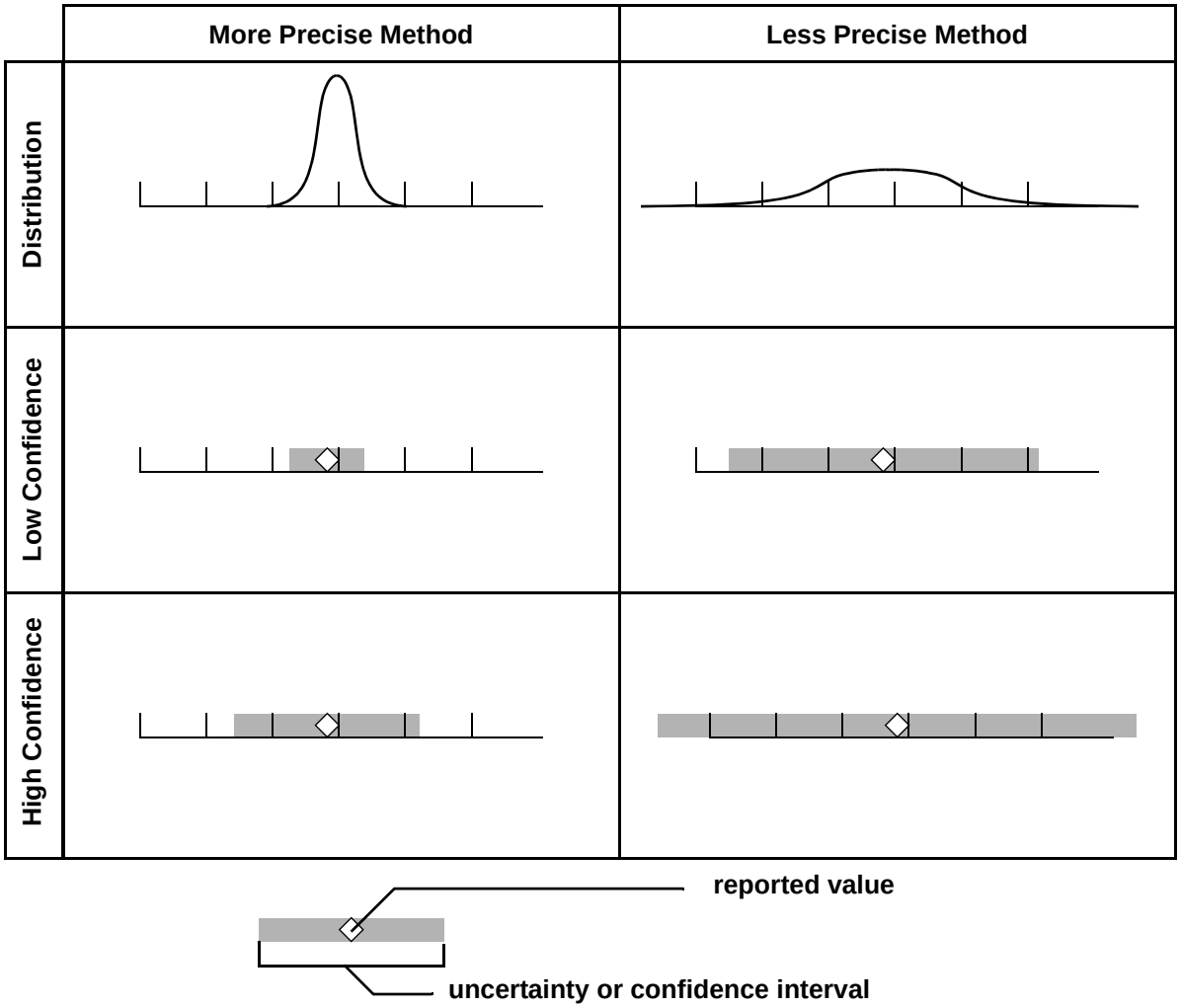


Figure 1. The analytical method and the confidence required determine the size of the uncertainty. For the same degree of confidence, more precise methods tend to have smaller uncertainties than less precise methods. For the same method, the more confidence required, the larger the size of the uncertainty.

Uncertainty can be determined in a variety of ways. Sometimes uncertainty is based solely on instrument resolution. For example, a mass measurement produced by a centigram balance has an inherent uncertainty of ±0.01 g, whereas the same measurement made using an analytical balance has an uncertainty of ±0.0001 g. Sometimes uncertainty incorporates a wide variety of factors and is determined by the professional judgement of the analyst. Uncertainty also can be defined using statistical methods. In this paper, a statistically defined confidence interval will be used to calculate uncertainty. For example, a 90% confidence interval indicates that there is a 10% chance that the region of uncertainty does not include the true result. Of course, the true result of the analysis might not be the true concentration of the analyte if the method is biased (always produces results that are too high or too low).

Developing Intuition about Uncertainty and Confidence

Although it may seem counterintuitive at first, a large region of uncertainty implies greater confidence than a small region of uncertainty (assuming no method or analyte differences). Confidence has to do with how sure the data user is that the region of uncertainty surrounding the measured value actually encompasses the true value. The following analogy illustrates these concepts.

Suppose that I want to specify the location of my sister. It is a weekday; she is a nurse that works at a local clinic. If I want a very high degree of confidence in my assessment of her location, I might say "She is on the planet Earth." With this extremely large region, the chance that I am wrong is virtually zero, but the region of uncertainty is too large to be useful. If I choose a smaller region of uncertainty, "She is in the metropolitan area," I have a greater chance of being wrong—in other words, less confidence. It is possible that she is out of town attending a conference or on vacation, but these situations are not very likely because I know that they are rare occurrences in her routine. If I choose an even smaller region of uncertainty, "She is in the clinic office on Main Street," my chance of being wrong is much greater. She might be home sick, taking a day off, or out at lunch. Every time I choose a smaller region of uncertainty, I also decrease the confidence that the region captures her true location.

This analogy can be extended to consider the nature of different analytes. Suppose I want to specify the location of my cousin who is a sales representative. I am still quite certain that he is on the planet Earth, but if I limit the neighborhood of his location to "metropolitan area," I have a very good chance of being wrong. I know that he travels a great deal. Because of their inherent natures and behaviors, my sister is much easier to locate than my cousin. As chemical analytes, she would be relatively easy to analyze and he would be more difficult.

The analogy can be extended further to consider the analytical method. At this point I have assessed an individual's location based on a knowledge of his or her habits. This is not a very precise method, but it is easy and inexpensive. If I require a small region of uncertainty and high confidence, I would need to use a more precise method for determining location, such as fitting my sister or cousin with a radio transmitter. This would require more work on my part and be more expensive than the less-precise, but easy, method. Similarly, choosing a more precise analytical method decreases the size of the uncertainty for chemical analyses, but it usually costs more. In addition, a more precise method that is biased may not provide an estimate that is closer to the true value than an unbiased but less precise method.

To summarize, the size of the uncertainty for a measured value depends on the nature of the analyte and the method used to analyze it, and on the degree of confidence required by the data user.

Quantification and detection—A quantitative result is essentially an observation in the form of a number. It was obtained when some physical measurement (such as the volume of a titrant or the magnitude of an electronic signal) was mathematically manipulated to obtain an estimate of analyte concentration. The data user must calculate or assume an appropriate amount of uncertainty around the reported result.

In contrast, detection is not an observed number, but a decision. Detection imposes a judgement on an analysis that was not required for quantification: Is this sample different from a sample that does not contain the analyte (a blank)? Making that decision is complicated by the fact that all quantitative results have uncertainty. To be judged a detection, the region of uncertainty surrounding the reported value should not include zero (fig. 2). When an analyte is reported as "detected," the data user should be confident that the analyte is indeed present in the sample. The converse is not true, however. If an analyte is reported as "not detected," it may still be present, but the concentration is so low that a blank sample could have produced the same analytical signal.

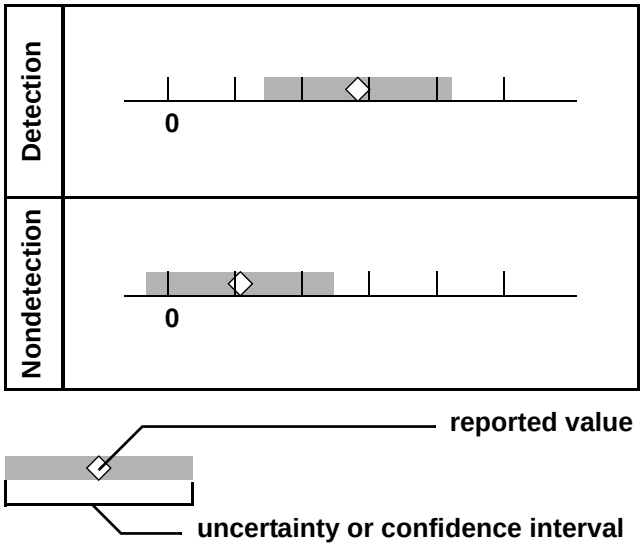


Figure 2. If the region of uncertainty around a result includes zero, then the result is not considered a detection.

Methods of Reporting Analytical Results

An analytical result can be reported in a variety of ways. Regardless of the conventions that are used, all analytical results should be reported with some indication of the associated uncertainty.

Most robust method—In an ideal world, every sample would be analyzed several times so that the uncertainty for that sample would be firmly established. In this utopia, every value would be reported with its unique standard deviation and the number of replicate analyses; the data user could apply statistical theory to calculate the desired confidence interval for every value (see *A Closer Look—Most Robust Method,* on the following page.). In this case, the data user makes decisions about the question of detection and the laboratory simply reports the data. In this situation, that laboratory would use *all* analytical measurements, even the negative values.

Unfortunately, there are many reasons why this approach is impractical. Collecting a large enough sample for the analysis of many replicates may not be possible. Routinely analyzing many replicates for each sample also would greatly increase analytical costs.

A Closer Look— Most Robust Method

PROCEDURE

1. Sample is submitted to laboratory.

2. At the laboratory:

- *Sample is homogenized.*
- *Sample is split into several fractions.*
- *Splits are analyzed independently.*
- *Hypothetical results for five splits: 3.252, 3.283, 3.291, 3.258, 3.264.*
- *Mean ($\bar{x}$) and standard deviation (s) are calculated: $\bar{x}$ =3.269600, s=0.016682.*

3. Laboratory reports result as: 3.270, s=0.017 (n=5), or they report the individual values.

4. Data user interprets value:

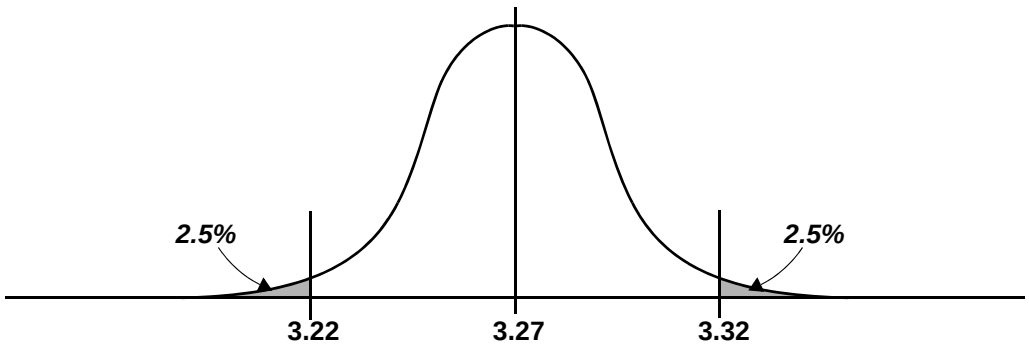
- *Data user decides that 5% error is tolerable.*
- *Student's t for a 2-tail test, 4 degrees of freedom, 95% confidence: t=2.78.*
- *Calculate confidence limits: 2.78 × 0.017 = 0.047.*

RESULT

Data user publishes result as 3.27± 0.05 .

DISCUSSION

Replicate analyses are performed to allow a statistical determination of the region of uncertainty at a user-chosen confidence level. Normally distributed values are assumed—a reasonable assumption for random analytical variability. Standard statistical tables are used to look up the value of Student's t. A 2-tailed value is used, placing half of the error in one tail and half in the other tail. The distribution for this example is illustrated below (not to scale). Note that this procedure produces a well-defined quantitative result for ***a single sample****. No information about field or sampling variability or bias is included. Furthermore, this result may be different from the "true" concentration if the laboratory method is biased.*



Less robust method: reporting and detection levels—When replicate analyses of each sample are not possible, assumptions and compromises are required and the data user must be mindful of them. Sometimes each sample is analyzed only once. A single analysis provides absolutely no information about the uncertainty, but the data user still needs that information for the analytical result to be meaningful. In such cases, the laboratory adopts procedures for reporting the data that provide the data user with some indication of the random error associated with the analytical method. Reporting procedures vary among laboratories and analytical methods.

MRL. One commonly used option is for a laboratory to establish a minimum reporting level (MRL). An MRL divides the data into two groups. Results that are greater than the MRL are reported. Results that are less than the MRL are censored; that is, they are reported simply as "less than the MRL" (< MRL). Methods for choosing MRLs vary and no single procedure is universally used. In general, the data user can assume that a laboratory is confident that the analyte is present in the sample if its concentration exceeds the laboratory's MRL. No other information about uncertainty is implied. Some laboratories choose MRLs so that the relative standard deviation for values greater than the MRL (reported values) is approximately constant. This is very useful for the data user to know because it provides an estimate of the random error associated with analytical results and can be used to estimate a region of uncertainty. In many cases, however, the MRL has no statistical basis and is based on the professional judgement of the analyst.

Imposing an MRL does not alter the fact that all data have associated analytical error. Because of analytical error, most data users would not assume that two samples, each analyzed only once, with reported concentrations of 0.15 and 0.16 mg/L, were different. However, data users sometimes make the erroneous assumption that two samples with reported concentrations of <0.15 and 0.16 mg/L are different. That is, they assume that if the true analyte concentration of a sample were greater than the MRL, then the results for this sample would not be censored. This assumption is false.

Analysis of a sample having a true concentration equal to the MRL has a 50% chance of producing a result that is less than the MRL and a 50% chance of producing a result that is greater than the MRL. Consequently, based on a single analysis, a data user cannot confidently conclude that a sample with a reported concentration near the MRL is different from a sample with a reported concentration of <MRL (fig. 3).

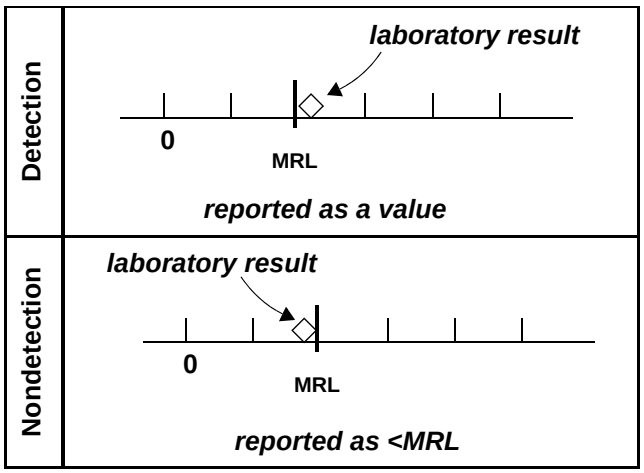


Figure 3. The results shown here would be reported differently, even though they could easily be replicate results for the same sample.This is because the results fall on opposite sides of the MRL.

USEPA-MDL. Another common approach is for a laboratory to use the U.S. Environmental Protection Agency's method detection limit (USEPA-MDL) procedure. Unlike MRLs that can be based on many different factors, the USEPA-MDL is determined using a clearly prescribed procedure (U.S. Environmental Protection Agency, 1998) (see *A Closer Look—USEPA-MDL,* on the following page*)*. The USEPA-MDL is statistically defined as the smallest concentration that can be measured and reported with 99% confidence that the analyte concentration is greater than zero.

When the measured concentration exceeds the USEPA-MDL, the compound is considered "detected." Assuming that the sample is well behaved, this conclusion (detection) could occur due to random noise in the analytical method a maximum of 1% of the time if the value measured were equal to the USEPA-MDL. In other words, the chance of a false detection is no more than 1%. When the analytical result does not exceed the USEPA-MDL, the analyte is considered "not detected with adequate confidence," which is different from "not present."

The USEPA-MDL procedure is based on one set of replicate analyses of an ideal sample (usually analyte-spiked blank water); as such, the USEPA-MDL indicates the best performance of the analytical method. The analytical method may not perform as well on field samples, which probably contain substances in addition to analyte and water. For example, other competing analytes or humic substances can cause matrix interferences. Therefore, the data user should be aware that the USEPA-MDL may not be applicable to environmental samples that are not well behaved or have substantial matrix interferences.

Using the USEPA-MDL as an MRL. Usually, an MRL is set at a concentration that is considerably greater than the USEPA-MDL. This is because the relative analytical uncertainty is large at concentrations near the USEPA-MDL; when the analyte concentration is equal to the USEPA-MDL, the region of uncertainty (defined as the 99% confidence interval) is ±100%. Occasionally, however, a laboratory will use the USEPA-MDL as their MRL—both to censor data and as the default value for reporting the censored data. In this case, when the laboratory measures a value that is less than the USEPA-MDL, it is censored and reported as "< USEPA-MDL." The use of the USEPA-MDL as the default reporting value for censored data is not related to its statistical definition. The USEPA-MDL is designed to limit the chance of erroneously concluding that an analyte is present in a sample when it actually is not. In other words, the USEPA-MDL protects against false positives. If a sample does not contain the analyte, there is only a 1% chance that random error in the analytical method will produce a result greater than or equal to the USEPA-MDL. *The USEPA-MDL does not indicate the minimum concentration that can be detected with confidence (99% of the time).* In other words, it does not protect against false negatives. Analysis of a sample having a true concentration equal to the USEPA-MDL has a 50% chance of producing a result that is less than the USEPA-MDL and a 50% chance of producing a result that is greater than the USEPA-MDL.

A Closer Look— USEPA-MDL

PROCEDURE

1. Laboratory estimates a value for the MDL, for this example 0.05.

2. Laboratory prepares and analyzes a solution of spiked blank water:

- *Analyte concentration in spike is 1–5 times the estimated MDL.*
- *In this example, spike is 0.135.*
- *Seven replicates of spike are analyzed.*
- *Hypothetical results for spikes: 0.1345, 0.1402, 0.1358, 0.1297, 0.1366, 0.1410, 0.1383.*
- *Mean ($\bar{x}$) and standard deviation (s) are calculated: $\bar{x}$ =0.136586, s=0.003827.*

3. Laboratory interprets result:

- *Student's t for a 1-tail test, 6 degrees of freedom, 99% confidence: t=3.14.*
- *Calculate USEPA-MDL: 3.14 × 0.003827 = 0.012.*
- *Value is much smaller than originally estimated, making spike concentration too high.*

4. Laboratory prepares and analyzes a new spike solution:

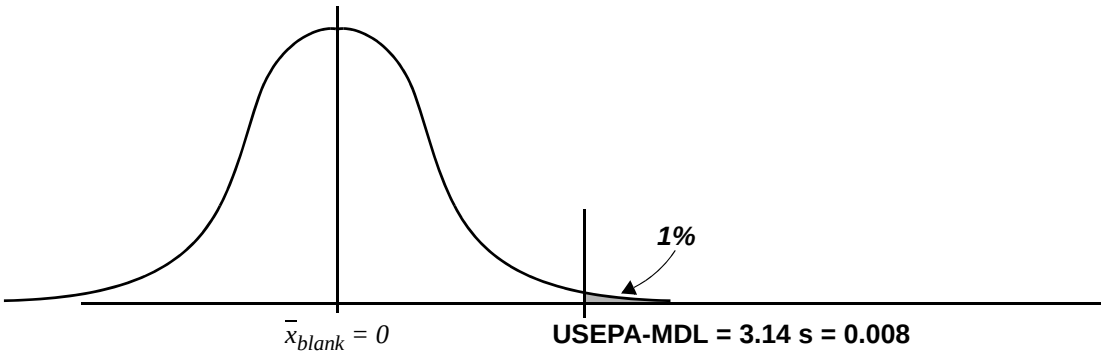
- *New spike concentration in this example is 0.040.*
- *Hypothetical results are: 0.0366, 0.0409, 0.0427, 0.0358, 0.0391, 0.0383, 0.0404.*
- *$\bar{x}$=0.039114, s=0.002438.*

5. Laboratory interprets new result:

- *USEPA-MDL= 3.14 × 0.002438 = 0.008.*
- *Spike concentration is within desired range (1-5 times USEPA-MDL).*
- *Analytical results greater than or equal to 0.008 are considered detections; that is, they are different from zero with 99% confidence.*

DISCUSSION

Analytical variability for a true blank (concentration equals zero) is assumed to be the same as that for a low concentration spike. Analytical variability is assumed to be normally distributed. Standard statistical tables are used to look up the value of Student's t. A 1-tail value is used, because the interest is in limiting the incidence of false positives. The distribution for this example is illustrated below (not to scale).



NWQL REPORTING PROCEDURE

Fundamentals (or the ABCs of the NWQL's LT-MDLs and LRLs)

Historically, the NWQL used MRLs when reporting data. Procedures for determining MRLs were not consistent among methods and sometimes not well documented. More recently, USEPA-MDLs were used as MRLs for some of the newer methods. In 1999, the NWQL established a new reporting procedure that has been implemented for most water matrix methods (Childress and others, 1999). The procedure involves two parts: the use of a statistic called the long-term method detection limit (LT-MDL) and the implementation of a new reporting convention. The new method is outlined as follows (see also *Comparing the Old and New Reporting Procedures,* on the following page).

1. A long-term detection limit (LT-MDL) is determined for each analyte/method combination on an annual basis. The LT-MDL is very similar to the USEPA-MDL. The LT-MDL is statistically defined identically to the USEPA-MDL—the smallest concentration that can be measured and reported with 99% confidence that the analyte concentration is greater than zero. Like the USEPA-MDL, it is obtained from replicate analyses of spiked blank water and, therefore, it applies to well-behaved samples. It differs from the USEPA-MDL in that it incorporates variability due to the different instruments and analysts that are part of a production laboratory like NWQL and in that it is calculated over an extended period of time.

2. All analytical results greater than the LT-MDL are reported. These results can be considered detected (different from zero with 99% confidence), provided that the samples are well behaved.

3. Analytical results less than the LT-MDL are reported if the analytical method includes analyte-specific identification—usually a matching spectral signature. In such methods, the detector (a mass spectrometer or photodiode array spectrometer) confirms the identity of the analyte. NWQL calls these methods "information-rich." Results less than the LT-MDL are reported with a qualifying E-code.

4. Analytical results less than the LT-MDL are censored if the analytical method does not specifically identify the analyte. Examples of nonspecific detection methods include retention time and absorbance at a particular wavelength. These results are reported as "less than the laboratory reporting level" (< LRL). The LRL is calculated as LRL= z x LT-MDL. The value of z depends on the mean recovery of spike samples used to determine the LT-MDL. If the mean recovery is 100%, then z=2; if the mean recovery is less than 100%, then z>2. Provided that the sample is well behaved, the data user can assume with 99% confidence that the concentration in the sample is less than the LRL. In other words, there is no more than a 1% chance that the true concentration of the analyte in the sample exceeds the LRL due to random analytical error.

5. A qualifying E-code accompanies analytical results between the LT-MDL and LRL because relative uncertainty is greater in this region than above the LRL.

Comparing the Old and New Reporting Procedures

The table below shows hypothetical laboratory results and how those results would be reported using the various procedures. Remember that the data user only receives the reported results and not the instrument derived values.

Sample	Instrument Derived Value	Reported Result Using		
		Old Procedure MRL=0.050	New Procedure LT-MDL = 0.030 LRL = 0.060	
			Not Information-Rich Analytical Method	Information-Rich Analytical Method
1	–0.008	< 0.050	< 0.060	< 0.060
2	0.015	< 0.050	< 0.060	E 0.015
3	0.048	< 0.050	E 0.048	E 0.048
4	0.051	0.051	E 0.051	E 0.051
5	0.076	0.076	0.076	0.076

DISCUSSION

The old procedure. *When a value is reported as <0.050, the data user knows that the value measured by the laboratory was less than 0.050. However, that does not mean that the true analyte concentration is less than 0.050 (50% of the time, a sample with a true concentration of 0.050 will be reported as <0.050). Notice that Samples 1, 2, and 3 would be indistinguishable to the data user, when in fact Samples 3 and 4 really had the closest instrument values.*

The new procedure (for most methods)*. When a value is reported as "<0.060" (<LRL), the data user knows that the laboratory obtained a value less than 0.030 (the LT-MDL). The data user can also conclude that the true analyte concentration is less than 0.060 with no more than a 1% chance of it actually being greater than 0.060. (Using the LRL minimizes the risk of a false negative.) The results for Samples 3 and 4 are E-coded because they are between the LT-MDL and the LRL and, therefore, subject to greater relative analytical error than larger values. In particular, false negatives are not uncommon in this region. Based solely on these single measurements, the data user could not conclude with confidence that the true concentrations of Samples 1 through 4 were different; the use of the LRL as the default reporting level and the use of the qualifying E-code are consistent with that fact.*

The new procedure (for information-rich methods). *When a value is reported as "<0.060" (<LRL), the data user knows that the laboratory was unable to positively identify the analyte (assuming laboratory quality control criteria were met). The LRL is used because it is the default value for nondetections. The result for sample 2 is reported with an E-code. The analyte's presence was confirmed by the method, but because it is less than the LT-MDL, this result is subject to greater relative analytical error (especially contamination problems) than larger values.*

NOTES

In this example, the LT-MDL < MRL < LRL. This will not necessarily be the case. Most frequently, the MRL>LT-MDL. If the MRL<LT-MDL, it is an indication that previously reported data less than the LT-MDL have high uncertainty. If such data are considerably less than the LT-MDL and the method is not information-rich, the possibility of false positives should be considered.

In this example, it was assumed that the lowest laboratory standard used in the analysis (LS) is approximately equal to the LRL, which will usually be the case. If not, some additional values may be E-coded. See Childress and others (1999) for an explanation of the role of the LS in coding.

Why the LT-MDL/LRL Reporting Procedure is Good News for Data Users

Provides more information—The LT-MDL/LRL reporting method provides the data user with information about method performance at the NWQL. Because the LT-MDL and LRL are statistically defined to describe routine lab performance on well-behaved samples, the data user can use them to estimate the expected incidence of false positives and false negatives and to estimate uncertainty in the low end of the analytical range. This information can be used to help decide if an existing analytical method is adequate for a proposed project. In addition, this information gives the data user more power in assessing data quality and in interpreting project data. Using information embedded in the LT-MDL to assess data quality is addressed in the next section. Specific examples of using the LT-MDL to help plan a project, assess data quality, and interpret data are given in the examples sections.

Provides more data—The LT-MDL/LRL reporting procedure often results in the data user receiving fewer censored data than before. This is because data having values between the LT-MDL and the LRL are now reported (with an E-code); previously, it is likely that these values would have been censored. In addition, the censoring criterion (the LT-MDL) commonly is smaller than the previously used criterion (the MRL). In rare cases, the LT-MDL may be larger than the old MRL. Such an occurrence might seem to be a disadvantage because it results in more data being censored than before, but those previously uncensored data were not reliably different from a blank. In other words, if the LT-MDL is larger than the old MRL, it is evidence that the old MRL was too small and resulted in an increased incidence of false positives.

How the LT-MDL/LRL Reporting Procedure Poses Challenges for Data Users

Complicated data sets—The LT-MDL is reassessed annually, and it and the LRL can change over time. Consequently, data sets can be complicated by multiple LRLs and LT-MDLs. In addition, reported values less than the LRL and sometimes less than the LT-MDL may be common in some data sets, as well as an increased number of values with an E-code. Data users need to develop strategies for working with such data sets. Several examples of strategies to deal with these data set complications are presented in the examples sections.

Potential for misinterpretation— Low level data produced by LT-MDL/LRL reporting procedure can be incorrectly interpreted in a way that results in distortion of the data distribution (Helsel, 2005). As previously described, an analytical result that is less than the LT-MDL is reported as less than the LRL (for a method that is not information-rich) and an analytical result that is between the LT-MDL and LRL is reported as a value with an E-code. If the data user interprets the censored value as being potentially between 0 and the LRL and interprets the E-coded value as being potentially between the LT-MDL and the LRL, the data distribution becomes distorted. In this erroneous interpretation, the uncertainty of the E-coded value is assumed to be less than that of the censored value, when, in fact, the uncertainties are about the same. The uncertainties are roughly equal because both values are in the low-level range of the method. Depending on the confidence that the data user requires, these two results should be treated either as (A) indistinguishable (high confidence required), or (B) the E-coded value being greater than the censored

value (less confidence required). Methods for avoiding this problem are presented in several examples later in this paper.

Need for careful data quality assessment—For some projects, the LT-MDL/LRL reporting procedure will result in data sets that contain many low-level and E-coded reported values that previously would have been censored. These data have considerable uncertainty. In order to correctly interpret such data, data users will need to carefully examine quality control samples. Similarly, data users who plan to make use of the statistical basis of the LT-MDL and LRL need to ensure that their samples behaved as expected during analysis. In addition to using all available laboratory quality assurance information, additional quality control samples may be needed.

ASSESSING PROJECT DATA QUALITY

As with the old MRL, the LT-MDL and an LRL are not determined individually or uniquely for each sample. Unlike the old method, however, the LT-MDL/LRL reporting procedure is based on some well-defined assumptions about the behavior of samples and the analytical performance of the method. Data users can evaluate various quality control sample results to determine if there is any evidence that the LT-MDL and LRL do not apply to their samples.

Every laboratory implements procedures to ensure that their analytical results are correct. An incorrect result occurs when the true concentration is not within the region of uncertainty of the reported concentration. Standardized procedures help minimize the chances of obvious errors. The Inorganic and Organic Blind Sample Program (BSP) administered by the Branch of Quality Systems helps uncover method bias and systematic errors, such as laboratory contamination or loss of analyte during the analytical process. Statistical results from the BSP also can be used to estimate the analytical random error. The BSP helps identify errors that apply to all samples. In contrast, individual data users need to worry about sample-specific errors—ill-behaved samples. Often this occurs due to a matrix interference; another substance that is present in the sample and interferes with the analyte of interest. Depending on the type of interference, the reported concentration may be always high or always low or highly variable.

It is up to the data user to determine whether an LT-MDL is appropriate for their particular samples. In some cases, it may be necessary to examine chromatograms or talk to the analyst about the performance of specific samples.

Comparing Your Sample Results to Routine Lab Performance

The LT-MDL, LRL and the BSP's evaluations provide information about the results that can be expected when the NWQL analyzes spiked blank water. The data user needs to determine if the performance on their field samples is comparable. Several tests can be done.

Simple checks—Data users can check for several conditions that may indicate problems with analytical performance or the behavior of particular samples or sample types. These include:

- A sample reported as a nondetection with an analyst-raised LRL (an LRL higher than the normal LRL for the method)
- Analyst-raised LRLs that occurred during a particular time period or for particular sample types
- The use of E-codes on values greater than the LRL
- Surrogate recoveries that are outside of control limits

Note that the absence of these conditions does not guarantee that the LT-MDL and LRL apply to your field samples.

Standard deviation—The standard deviation of values from field spikes and replicate samples can be compared to the results obtained by the laboratory. If the results are significantly different from the laboratory results, then the data user has evidence that these samples are not well

behaved. The basic procedure is outlined below; several examples are given in the examples sections.

1. Calculate the standard deviation for replicate samples and replicate field spikes.
2. Pool the standard deviations if appropriate.
3. Obtain laboratory estimates for the standard deviation for the method.These can be found at the following Web sites:

Information	URL
LT-MDL documentation	http://bqs.usgs.gov/ltmdl accessed on 3/20/2007
Blind samples—organic methods	http://bqs.usgs.gov/OBSP accessed on 3/20/2007
Blind samples—inorganic methods	http://bqs.usgs.gov/bsp accessed on 3/20/2007

4. Use the F-test or Levene's test to compare field results with lab estimates.

Incidence of false positives and contamination—Blank samples (field blanks as well as lab and BSP blanks) are presumed to have an analyte concentration of zero. Because of random error, a blank can produce an analytical signal. The reporting procedure is devised so that there is only a 1% probability that a blank will produce a signal equal to or greater than the LT-MDL. Blank samples that routinely are reported as having detectable concentrations are evidence of background contamination or interference from another analyte.

The number of field blanks that are reported with detectable concentrations can be compared to the expected behavior of the method. This is important when interpreting data in the low-range of the method (values less than the LRL and, particularly, values less than the LT-MDL). In this range, identifying any low-concentration blank contamination is necessary. The data user should analyze results from field blanks as well as NWQL results for lab water sets and the BSP's results for blanks. The basic procedure is outlined below; several specific examples are given in the examples sections.

1. Calculate the incidence of presumed false positives as the number of blank samples with reported concentrations greater than the LT-MDL divided by the total number of blank samples.
2. Determine the probability for this incidence of false positives. Table 1 provides such values. The probability also can be calculated using the binomial distribution function: Probability = ${}_{d}^{n}Cp^{d}q^{(n-d)}$, where n is the total number of blank samples, d is the number of detections greater than the LT-MDL, p is the probability of a detection for a blank (p=0.01 for a blank), q is the probability of a nondetection (q=0.99 for a blank) and ${}_{d}^{n}C$ is the binomial coefficient which is calculated as $\frac{n!}{d!(n-d)!}$, where ! indicates factorial.

Table 1: Probability of false positives.
[Calculated using the binomial distribution function as described in step 2 of the preceding section.]

Number of Blank Samples	Total Number of Detections ≥ LT-MDL					
	0	1	2	3	4	5
1	0.99	0.01	—	—	—	—
2	0.98	0.020	0.0001	—	—	—
3	0.97	0.029	0.00030	0.000001	—	—
4	0.96	0.039	0.00059	0.000004	<0.000001	—
5	0.95	0.048	0.00097	0.000010	<0.000001	<0.000001
6	0.94	0.057	0.0014	0.000019	<0.000001	<0.000001
7	0.93	0.066	0.0020	0.000034	<0.000001	<0.000001
8	0.92	0.070	0.0026	0.000053	<0.000001	<0.000001
9	0.91	0.083	0.0034	0.000079	0.000001	<0.000001
10	0.90	0.091	0.0042	0.00011	0.000002	<0.000001
11	0.90	0.099	0.0050	0.00015	0.000003	<0.000001
12	0.89	0.11	0.0060	0.00020	0.000005	<0.000001
13	0.88	0.12	0.0070	0.00026	0.000007	<0.000001
14	0.87	0.12	0.0081	0.00033	0.000009	<0.000001
15	0.86	0.13	0.0092	0.00040	0.000012	<0.000001
16	0.85	0.14	0.010	0.00049	0.000016	<0.000001
17	0.84	0.14	0.012	0.00059	0.000021	0.000001
18	0.83	0.15	0.013	0.00070	0.000027	0.000001
19	0.83	0.16	0.014	0.00083	0.000033	0.000001
20	0.82	0.17	0.016	0.00096	0.000041	0.000001

Notice that the probability of a random detection in blanks increases with the number of blank samples. If 20 blanks were submitted, the probability that 1 (and only 1) of them was reported with a concentration greater than or equal to the LT-MDL is about 17%. This is solely due to analytical variability and would not be considered unusual. However, the chance that analytical variability alone resulted in 2 (and only 2) out of 20 blanks being reported with concentrations greater than or equal to the LT-MDL is only 1.6%. This occurrence is rare enough that it could be considered evidence of background contamination or another problem.

Incidence of false negatives and analyte loss—Spiked samples (field spikes as well as lab and BSP spikes) are presumed to have an analyte concentration greater than zero. Because of random error, a low-level spiked sample can produce an analytical signal less than the LT-MDL. This value would be reported with an E-code for an information-rich method and as a nondetection (<LRL) for a method that is not information-rich. The reporting procedure is devised so that there is a 1% probability that a sample with a concentration equal to the LRL will produce a signal equal to or less than the LT-MDL. Spikes that routinely are reported as nondetections (false negatives) are evidence of analyte loss. False negatives are somewhat more difficult to interpret than false positives. This is because the probability of a false negative depends on an unknown— the true concentration of analyte in a field matrix spike.

The concentrations of field matrix spikes should be chosen to best meet the needs of the data user. The organic spike kits from NWQL often result in analyte concentrations that may be an order of magnitude or more greater than the LRL. At such concentrations, field matrix spikes are not very useful in determining method performance in the low range of the method (less than the LRL) and are not a good test of the LRL value. The data user has some choices.

If the data that are to be interpreted are expected to be significantly greater than the LRL, the data user can prepare field matrix spikes in the usual manner. Such spikes will provide information about method bias, but not about performance in the low concentration range.

If the data user is primarily interested in data that are in the low range of the method, then field matrix spikes with analyte concentrations near the LRL are needed. These could be prepared by using the standard spiking procedure but substituting a larger volume of sample or by serial dilution of the standard solution. Samples that have concentrations equal to the LRL are expected to result in nondetections 1% of the time, assuming no matrix interferences. Performance of spikes at the LRL can be evaluated using the same procedure as for field blanks. If the analyte concentration is not equal to the LRL, then the probability of a nondetection must be calculated using the estimate of the standard deviation. This probability can then be used in the formula for the binomial distribution. Examples are given in the examples sections.

EXAMPLES

The remainder of this document details a number of examples that make use of the information inherent in the LT-MDL. Examples include quality assurance, project planning, and data interpretation. These examples are not intended to be a comprehensive set of all possible situations that a data user might encounter. Rather, it is hoped that they will provide the data user with "food for thought" regarding how to handle their data. No two data sets or projects are alike and approaches must vary accordingly. The data user must use his or her best judgement about a particular situation, make a decision, and then document what was done.

The examples that follow assume that the data user is familiar with basic statistical methods and has access to statistical tables and software. Tables of statistics (t values, F values, etc.) are not included, but generally can be found in the appendices of standard statistical texts. Some statistical references are given at the end of this report.

The examples given here are for illustration purposes. The analytes are hypothetical and concentration units are not specified. NWQL statistics such as LT-MDLs and standard deviations are made up to reflect the process used at the NWQL, but do not correspond to any particular analyte or method.

Advice to the data user

Know the question you are trying to answer.

- *Don't interpret low-level data if you are only interested in values near the upper end of the data distribution.*
- *Decide what level of certainty is required to answer the question.*

Know your data and its limitations.

- *Remember that all analytical results have uncertainty. This is not the lab's fault, it is life.*
- *Some samples are easier to analyze than others. Yours might be easy or difficult.*
- *Performance of instruments, analysts, and data users varies from day to day. You don't get to choose whether your samples are analyzed on a good day or a bad day.*
- *Do enough QC to ensure that you know the quality of your data.*

Tailor your approach to your question and your data.

- *Remember that different questions and data sets require different interpretive methods.*
- *Don't do extensive (and expensive) QC to assess low-level data if you are only interested in values near the upper end of the data distribution or if you do not need a high level of certainty.*
- *Ask yourself if the quality of your data allows the type of analysis you are doing.*

Be mindful of lurking variables.

- *Matrix effects, low-level contamination and analyte losses can be very important, especially at low concentration levels.*
- *Make sure that differences between groups of samples are not because one type of sample has a simpler matrix than another and therefore had fewer nondetections.*

Document your approach and explain the reasons behind it.

Statistical Methods and Censored Data

The application of statistical methods to environmental chemical data is complicated by the presence of censored data. Historically, very few statistical methods accommodated such data sets. The most common approach to this problem was simply substituting a value for the nondetection. The usual choices for the substituted value are zero, the detection limit (DL), ½DL, and $DL/(\sqrt{2})$. At best, substitution is an arbitrary method that minimally affects data interpretation; at worst it can lead to bias and erroneous conclusions. Although substitution is still used, newer statistical methods have been developed that are better for censored data. The only option worse than substitution is to discard the censored data altogether; this approach should never be used. For an unorthodox description of the pitfalls of using substitution, see the cookie company analogy on the following pages.

One alternative to substitution is to use statistical methods that rely on data ranks, rather than actual data values. Calculation of percentiles for summary statistics, Wilcoxen or Mann-Whitney U tests for comparison tests, and Spearman or Kendall coefficients for correlation are such methods. (See Examples 3 and 4 in this document.) Although these methods produce nonbiased results, they have some limitations. For example, they can produce summary statistics that are censored. In addition, they cannot be directly applied to complicated data sets that contain multiply-censored data or reported values less than the censoring level.

Another way to approach censored data is to convert the data values to categorical data and then apply appropriate statistical methods. Calculation of frequencies for summary statistics, Chi-square analysis for comparison tests, and logistic regression are examples of this approach. (See Example 8 in this document.) These methods are not used as commonly as rank tests, but can be especially advantageous when the categories are chosen based on the purpose of the investigation.

A variety of methods have been developed recently that accommodate complicated data sets. Some of these methods are fully or partially parametric, meaning that they require assumptions about the shape of the distribution. Others are fully nonparametric. Two of these methods are used in the examples in this document: Regression on Order Statistics (Examples 4 and 7) and Kaplan-Meier Estimation (Example 7). For more information about these and other such methods see Helsel (2005) and Helsel and Hirsch (1992).

This document is designed to (a) help data users think about their data and (b) illustrate several statistical methods that avoid the problems of simple substitution methods and thereby lead to more accurate interpretations. Data interpretation will always be an issue and results can be biased either by ignorance or by design. New methods to interpret data and calculate statistics for a mixture of semiquantitative, nonquantitative and quantitative data are relatively recent and may not be widely used. Data interpretation can be improved as follows:

- Minimize the amount of semiquantitative and nonquantitative data by using the best available laboratory methods;
- Know which data are semiquantitative or nonquantitative and which data are quantitative;
- Understand the pitfalls of the statistical method used for data interpretation;
- Use new approaches and methods to interpret mixtures of semiquantitative, nonquantitative and quantitative data.

An Analogy for Chemical Analysis and Data Interpretation

The NWQL uses detection levels (LT-MDLs) and reporting levels (LRLs) to provide the data user with NWQL's best estimate of the concentration ranges that are quantitative (above the reporting level value), semiquantitative (between the reporting level and the detection level), and not quantitative (below the detection level). The following analogy illustrates how different data users might use or misuse these reporting and detection levels. It was first printed in the NWQL Newsletter, January 1999, and shown in an updated format here.

A famous cookie company produces one million cookies a day and is concerned about cookie pilfering by its elf employees. They hire a cookie detective to search the worker elves as they leave the factory. The cookie detective thought about the task of detecting pilfered cookies. The cookie detective knew he could reliably count and identify a quantity as small as one cookie if he found it. Ah, but finding just one pilfered cookie was a problem. The detective knew that he couldn't reliably identify every elf that was absconding with just one cookie. A cookie or two could be stashed in an inside pocket or concealed in some other way. How many cookies would an elf have to take to be caught most of the time? After testing various cookie-stashing methods, he concluded that it would be fairly difficult to stash a dozen cookies and walk out without detection. A dozen cookies, then, was the minimum number of cookies that he could routinely and reliably find. Now, what if an elf was found with crumbs or parts of cookies? Well, the elves make cookies all day and everyone knows that "cookies crumble," so a few crumbs should be expected. Parts of cookies, therefore, wouldn't count as evidence of theft.

The detective summarized the search method for management and elves as follows: Finding at least one whole cookie was evidence of a theft. Any elf who tried to pilfer 12 or more cookies would almost surely be caught. Elves who pilfered 1 to 11 cookies would sometimes be caught and sometimes get away with it. Cookies eaten by elves on the job don't count. That is biodegradation. Crumbs don't count. That is a blank contamination problem. More concisely, LT-MDL = 1 cookie and LRL = 12 cookies. (The mathematical relation between the LT-MDL and the LRL is different in this case from what the NWQL uses because NWQL is basing its definition on a different statistical probability than the cookie detective.)

Next, the cookie detective set about detecting cookie theft by elves. He carefully measured and honestly reported the data to cookie company management and to the elf union. He was shocked to find that the use of his data depended on the perspective and personal agenda of who was interpreting the data. Here is what he found.

Case 1—*Company management wants to find out how much money is being lost to cookie-pilfering elves. So, the management elf in charge of loss estimation adds up all of the sure detections of cookie theft and considers that sum to be the minimum loss. But he knows that the cookie detective doesn't catch all of the pilfering elves. He assumes that every elf who wasn't caught (nondetections) had actually pilfered 11 cookies (<LRL) and gotten away with it. He further assumed that all of the elves caught stealing just a few cookies (low-level detections) had really stashed 11 cookies, but all of them hadn't been found. (Those elves are clever rascals.) Then he adds these to the detected losses from thefts of 12 or more cookies to arrive at an estimated maximum cookie loss. He reports the potential range of cookie loss to the chief elf. The chief elf is shocked that the elves could be stealing so many cookies.*

Case 2—*The cookie business hasn't been so good lately and the company needs to renegotiate the contract with its worker elves. The management elf in charge of the negotiation wants to justify that the worker elves need to make some hefty concessions. He thought about using the high estimate of cookie theft that the chief elf had shown him, but he thought that the estimate was biased and too high. After all, not every elf stole 11 or more cookies all the time. So, he assumed that every elf who wasn't caught (nondetections) had actually pilfered six cookies (half the LRL) and added that result to the detected thefts. That number was probably high, too, because some of the employees took nothing and very few probably took 6. So the average was probably much less than six, but it was an estimate and he wanted a bargaining chip. He presents the estimated theft loss to the elf union and says they had better compromise.*

Case 3—*The elves' union representative doesn't like the way management counted cookie theft. He wants to assume that every elf who wasn't caught stealing was innocent. But, he knows that occasionally a cookie is pilfered and he doesn't want to sound as biased as that lousy company negotiator. To show his good faith, he is willing to assume that every nondetection could be counted as 0.5 cookie (half the LT-MDL). He reasons that, sure, a few elves pilfer a cookie now and then, but the problem is small and if the company paid the elves a decent salary, they wouldn't have to steal cookies to feed the hungry little elves at home. He realizes that this assumption might be biased low, but he doesn't want to admit that cookie pilfering by elves is a problem. He also wants a bargaining chip (also called a chocolate chip by the wily elves). He presents his theft estimate and requests binding arbitration.*

Case 4—*An individual elf has been watching the manipulation of data by the company and her union and she is offended. She has never stolen a single cookie in her life—not even when they were making double chocolate chip. Yet everyone seems to have assumed that she pilfered cookies. She doesn't care if they assume 0.5 or 1 or 6 or 11. Her pilfering was zero. She is sick and tired of everyone not realizing what a terrific employee she is. She feels wrongly accused and is considering filing suit for defamation of character. Character is important to elves.*

Case 5—*The company is going to lay off elves. It decided that pilfering elves would be the first to go. A particular elf had never pilfered cookies, but making cookies wasn't lucrative and his twin elflets' 4th birthday was coming up and he didn't have a present. So, one day he pilfered two cookies—one for each of the twins. He wasn't very good at pilfering and got caught. He was fired. He thought this wasn't fair. This was his first offense and it was only two cookies. The elf who worked next to him pilfered six cookies every day and smuggled them out in a cleverly concealed cookie pouch under his elf hat. That guy never gets caught. The fired elf complains to the company, but it reasons that someone who was caught with even one cookie must have pilfered more than someone who hasn't been caught. The elf decides to talk with the union representative.*

All of the aforementioned cases mimic actual practices with real water-quality data. The problem with cookie detection and analytical chemistry is that detection is not perfect—there is a gray area, a range that is semiquantitative. The problem can be minimized by improving detection ability, but neither cookie detectors nor NWQL can achieve detection capability that quantitates zero. The NWQL uses LT-MDLs and LRLs together to communicate ranges of certainty and uncertainty associated with its measurements. Those data will be used by someone. Substitution methods, as used by all of the elves in this analogy, often lead to flawed results.

Quality Assurance Example 1

Is the variability for my samples consistent with the LT-MDL?

Description: As part of a QA program, two sets of triplicates were sent to the lab for analysis. The samples could be from different sites or collected at different times. The results for the analyte were:

	Sample 1	**Sample 2**
Replicate A	0.0096	0.0153
Replicate B	0.0116	0.0142
Replicate C	0.0115	0.0144
$\bar{x}$	0.0109	0.0146
s	0.00113	0.000586
n	3	3

The LT-MDL for this analyte is 0.002, which was based on 21 samples. From the NWQL Web pages, the standard deviation used to calculate the LT-MDL was s=0.00083. Is the variance of these field samples consistent with the LT-MDL with 95% confidence?

Analysis: An F-test is used to determine if the standard deviations of the sample replicates is greater than the standard deviation from the LT-MDL. Because there is no reason to believe that the true variance for field samples actually could be less than that for spiked blank water, a 1-tail test will be used. By similar reasoning, there is no need to compare the standard deviation for sample 2.

The equation for the F test is: $F = \frac{s_1^2}{s_2^2}$,where $s_1 > s_2$.

For sample 1 replicates: $F = \frac{0.00113^2}{0.00083^2} = 1.85$

The critical value for F for a 1-tail test with 95% confidence: $F_{2,20}$=3.49.

The calculated value of F does not exceed the critical value of F; therefore, there is no significant difference. The performance on these field replicates is consistent with lab performance on LT-MDL blank (reagent water) spikes. There is no reason to conclude that the standard deviation measured by the lab does not apply to this data set or that the LT-MDL is not a valid estimate of the detection level.

Quality Assurance Example 2

What does it mean if the variability for my samples is not consistent with the LT-MDL?

Description: As part of a QA program, four sets of triplicates were collected at different locations or on different dates. The results are in the table below, which also includes a variety of statistics that were calculated from the replicates. The LT-MDL for this analyte is 0.002, based on 21 samples with a standard deviation of 0.00083. See Quality Assurance Example 1 to review how to calculate F.

	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**
Replicate A	0.0096	0.0097	0.0132	0.0532
Replicate B	0.0116	0.0078	0.0124	0.0419
Replicate C	0.0115	0.0083	0.0084	0.0473
$\bar{x}$	0.0109	0.0086	0.0113	0.0475
s	0.00113	0.00098	0.00257	0.00565
n	3	3	3	3
$CV = \frac{s}{\bar{x}}$	10%	11%	23%	12%
$\bar{x}/LT\text{-}MDL$	5.5	4.3	5.7	24
F	1.85	1.41	9.60	46.3
$Critical\ F_{2,20}\ (95\%)$	3.49	3.49	3.49	3.49

Analysis: For samples 1 and 2, the F value does not exceed the critical F value, showing that the variability for these analyses is consistent with that obtained by NWQL for LT-MDL blank spikes. The F values for samples 3 and 4 (9.60 and 46.3, respectively) exceed the critical F value, indicating that the variances for these samples are significantly different from (and larger than) the LT-MDL blank spikes. Interpreting what this difference means is not the same for these two samples. In the case of sample 4, the concentration is not in the low range of the method; the ratio of the average concentration to the LT-MDL is 24. The standard deviation is not expected to remain constant as the concentration increases above about 5 times the LT-MDL. The results for sample 4 replicates must be compared to NWQL results for comparable concentrations, rather than results for low-level concentrations. The Blind Sample Program (BSP) evaluates performance in the ultralow, low, medium and high range of a method. Data from the BSP for this analyte were low range spiked at about 10×LT-MDL CV=15.2% and medium range spiked at about 50×LT-MDL CV=12.4%. Comparing these values to the CV for sample 4 (12%) indicates that the behavior of sample 4 replicates is comparable to that

obtained by the BSP and that there is no evidence that sample 4 is ill behaved. This is not the case for sample 3. The sample 3 mean concentration is on the high side of the region that should be described by the LT-MDL standard deviation ($\bar{x}$/LT-MDL=5.7); however, the standard deviation for sample 3 is significantly higher than what is expected for that region. This is an indication that sample 3 may not be well behaved. At this point, the data user needs to use other knowledge about sample 3 to try to understand why this may be the case and to decide if this could apply to other samples in the larger data set. Assumptions that rely on the validity of the LT-MDL, such as estimates of uncertainty or incidence of false positives and false negatives, may not be applicable to sample 3 and other samples like it.

Quality Assurance Example 3

What can I do if the variability for my samples is not consistent with the LT-MDL?

Description: As part of a QA program, three sets of replicates were collected. The results are in the table below, which also includes a variety of statistics. The LT-MDL for this analyte is 0.040, based on 19 samples with a standard deviation of 0.0157. See Quality Assurance Example 1 to review how to calculate F.

	Sample 1	**Sample 2**	**Sample 3**
Replicate A	0.110	0.249	0.304
Replicate B	0.153	0.197	0.330
Replicate C	—	0.170	0.374
Replicate D	—	0.241	—
$\bar{x}$	0.132	0.214	0.336
s	0.030	0.0373	0.0354
n	2	4	3
$CV = \frac{s}{\bar{x}}$	23%	17%	11%
$\bar{x}$/*LT-MDL*	3.3	5.4	8.4
F	3.65	5.64	5.08
Critical F (95%)	$F_{1,18}$=4.41	$F_{3,18}$=3.16	$F_{2,18}$=3.55

Analysis: In this example, the F values for samples 2 and 3 exceed the critical F values, showing that the variability for these samples exceeds the variability from LT-MDL blank spikes. The F value for sample 1 does not exceed the critical F value, but it is fairly close. Based on the data here, it seems likely that the standard deviation for this data set in general might exceed that for LT-MDL spikes. This analysis does not indicate why this is the case—difficult matrix, variable decomposition during shipping, sloppy collection or sample preparation are all possibilities. Regardless of the reason for the greater standard deviation, the data user decides that this entire data set is not consistent with the LT-MDL as published and decides to recalculate a detection limit based on the performance of these data. First, the data user uses an F-test to compare the standard deviation of each of these samples to the other. The reader can verify that there are no significant differences among these samples. Then, the data user pools the standard deviations using the following formula:

$$s_{pool}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2}{n_1 + n_2 + \ldots + n_k - k}$$

Substituting the appropriate values from this example into the above equation yields:

$$s^2_{pool} = \frac{1(0.030)^2 + 3(0.0373)^2 + 2(0.0354)^2}{2 + 4 + 3 - 3} = \frac{0.007580}{6} = 0.001263$$

Taking the square root gives s=0.0355. There are 6 degrees of freedom for this calculation (denominator from s_{pool} calculation). For a 1-tail test with 99% confidence and 6 degrees of freedom, t=3.14. This results in an estimated MDL of (3.14)(0.0355)=0.111 for these samples. To obtain a new estimate for the reporting level, the data user needs to know the mean recovery. From the NWQL Web site, the data user finds that the mean recovery for the LT-MDL spikes was 85%. A spike sample sent in by the data user had a similar recovery. A new estimate of the reporting level is then calculated as

$$RL = \frac{2}{\text{percent recovery}} \times MDL = \frac{2}{0.85} \times 0.111 = 0.261 .$$

The data user chooses to use these alternative values (MDL=0.111, RL=0.261) in data interpretation rather than the NWQL LT-MDL and LRL of 0.040 and 0.080. (Note that the NWQL did not adjust the LRL for percent recovery in this example.) Based on the field sample replicates, the data user decides that the laboratory LT-MDL and LRL are too low for these particular sample types and could lead to erroneous conclusions for this data set.

One note of caution for the data user: the analysis described here tests only whether the sample variability exceeds the LT-MDL variability (1-tail test). It does not check the reverse because the reverse should not occur except by chance. The data user should not use this method to calculate an MDL and RL that are less than the LT-MDL and LRL.

Quality Assurance Example 4

The results for my blanks were reported as less than a concentration that was greater than some of my field samples—do I have a contamination problem?

Description: As part of a QA program, five field reagent water blanks were sent to NWQL for analysis, randomly, over the course of project sample collection. The LT-MDL for the analyte is 0.015. All five of the samples were reported as “0.030” with a qualifying code of “<.” Is there any evidence of contamination in the low-end of the analytical range? Two field samples (non blanks) were reported as “0.020” with a qualifying code of “E.” The analytical method was not considered information-rich

Analysis: All of the blank samples were reported as nondetections. This means that the actual result measured by the lab was less than 0.015 (the LT-MDL) for each of these samples. The fact that the default reporting value is greater than the LT-MDL is to be expected. The LRL, which is 2 times the LT-MDL for an analysis with 100% recovery, is used as the default reporting value for nondetections. Therefore, there is no indication of low-level contamination. The data user should also check the lab set blank and the BSP blank results to verify this conclusion.

The two field samples had concentrations reported because the actual results measured by the lab were greater than the LT-MDL. If the standard deviations are consistent with those of the LT-MDL blanks (see Quality Assurance Examples 1–3), then there is no more than a 1% chance that the true concentration in either of these two field samples was actually zero.

Quality Assurance Example 5

The results for my blanks were not all "less thans"—do I have a contamination problem? The method was not information-rich.

Description: As part of a QA program, five field reagent water blanks were sent to NWQL for analysis, randomly, over the course of project sample collection. The LT-MDL for the analyte is 0.015. The results were < 0.030, <0.030, E0.018, <0.030, and <0.030. The analytical method used was not considered information-rich. Is there any evidence of contamination in the low-end of the analytical range?

Analysis: In this example, four of the five blanks were reported as nondetections. That means that four of the five samples produced results that were less than the LT-MDL (0.015). The data user needs to determine the chance of obtaining one detection out of five blank samples simply due to random noise. The chance of any individual blank being reported as a value greater than the LT-MDL is 0.01, provided that the method is working properly and there are no contamination problems. This is based on the definition of the LT-MDL. To calculate the probability of one detection out of five samples, the binomial distribution function is used:

$$P = {}^{n}_{d}C p^{d} q^{(n-d)} = \frac{n!}{(n-d)!(d)!} p^{d} q^{(n-d)}$$

In this equation, n is the number of samples, d is the number of detections, p is the probability of a detection, $q = 1-p$ (the probability of nondetection) and C is the mathematical combinatorial function. Results of this calculation for p= 0.01 can be found in Table 1. For completeness, the calculation for this example is shown here and yields:

$$P = \frac{5!}{4!1!}(0.01)^{1} 0.99^{4} = 0.048$$

The probability of one in five blanks having a detection solely by chance is about 0.05 (5%, or 1 in 20). A check of the lab set blank and the BSP blank results for this method shows no evidence of contamination either for the lab set or on a routine basis. Now the data user must make a decision. Depending on the project's needs, the data user may consider 5% not particularly rare and conclude that this was an unfortunate chance occurrence and not an indication of problems with contamination or ill-behaved samples. A different data user might apply a more stringent requirement and be concerned about possible low-level contamination. The second data user might choose to have more blanks analyzed. Neither approach is better than the other. What is important is that the approach is based on the certainty requirements of the project.

Quality Assurance Example 6

The results for my blanks were not all "less thans"—do I have a contamination problem? The method was information-rich.

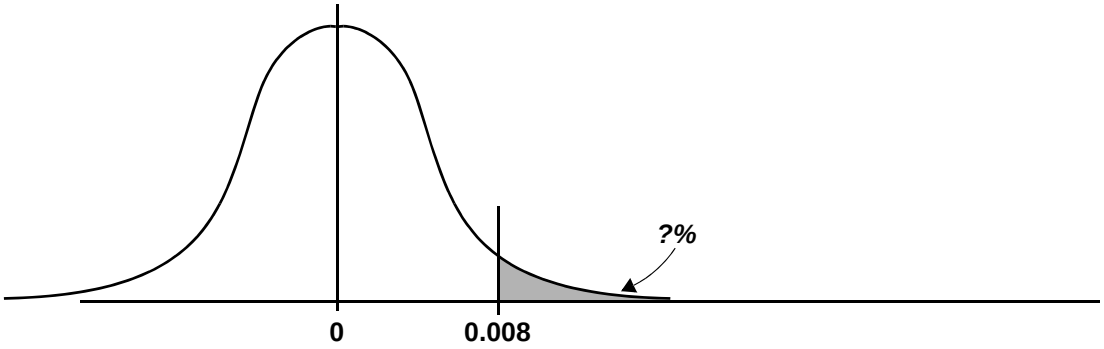
Description: As part of a QA program, five field reagent water blanks were sent to NWQL for analysis, randomly, over the course of project sample collection. The LT-MDL for the analyte is 0.015. The results were < 0.030, E0.010, E0.018, E0.009, and E0.005. The analytical method used was considered information-rich What do the results for the blanks indicate?

Analysis: This example is very similar to the previous one. Like that example, out of five blanks, there is one detection greater than the LT-MDL. What differs is that for this information-rich method, three of the other four blanks were reported with estimated values, albeit values less than the LT-MDL. The analyst had additional qualifying information that included positive identification of the analyte of interest. Having a high incidence of low-level detections, as in this example, is a good indication of background contamination. This is a common occurrence for some analytes (phthalates, for example). The source could be from the field, from the lab or from both. The set blanks and BSP results would indicate the level of contamination, if any, that is typical for this analysis.

A reasonable action for the data user in such a case is to censor their data at a minimum reporting level that is based upon the reported values for the blanks. If enough blank samples were analyzed, the data user could censor all data at the 99th or 95th percentile of the blank values. When few blanks were analyzed, the data user could censor the data at the maximum blank value. In this example, the data user might decide to censor all values less than or equal to 0.018 and consider such values not reliably different from a blank.

Alternatively, the data user can calculate the probability of detections at a level smaller than the LT-MDL and then make a decision about how to interpret low-level data. To do this, the binomial distribution function is used in the same way as in the previous example except that the probability of a detection is no longer 0.01 (because of blank contamination) and therefore table 1 is not applicable. In this case, it is necessary to calculate the probability of a detection at the level of interest.

If the data user is interested in detections as low as 0.008, then the probability that a blank gives a signal of 0.008 or greater must be determined. This involves calculating the probability that a blank produces a signal that falls within the shaded upper tail of the normal distribution curve as shown in the illustration on the following page.



A standardized t value is calculated using the formula: $t = \frac{x - \bar{x}}{s}$.

In this case, x equals 0.008 (the value of interest) and $\bar{x}$ equals zero (because the sample is a blank). The standard deviation can be obtained from NWQL, but the data user can also approximate it from the LT-MDL. Generally, the NWQL uses about 20 samples to determine an LT-MDL. Assuming 20 samples (19 degrees of freedom) and 99% confidence, the value of t is 2.54. The standard deviation then is 0.015/2.54=0.0059. If the number of samples were known exactly, then this would not be an approximation. Substituting into the formula yields

$$t = \frac{0.008 - 0}{0.0059} = 1.36$$

The probability that a blank yields a result greater than 0.008 is obtained by looking up the value 1.36 in a table of t values, specifying 19 degrees of freedom and a 1-tail test. From such a table, the standardized area under the unshaded (larger) part of the distribution curve is 0.905, making the probability of a value falling in the shaded area equal to 9.5%.

The calculation is now just like the previous example. Out of the five blanks, two were either not detected or reported at less than 0.008 (probability=90.5%) and three were reported as greater than or equal to 0.008 (probability =9.5%) Using the binomial distribution function yields:

$$P = \frac{5!}{3!2!}(0.905)^2(0.095)^3 = 0.007$$

The probability of three in five blanks having a value of 0.008 or more solely by chance is about 0.007 (0.7%). The data user considers this event too unlikely and concludes that this was probably not a chance occurrence and that the blanks indicate low-level contamination. By examining the other lab QC data, the data user may be able to determine if this is a general laboratory problem or a field-derived contamination.

Quality Assurance Example 7

What does it mean if one of my spikes is a nondetect?

Description: A project manager anticipates that a number of samples will contain concentrations of analyte near the LRL and therefore submits samples spiked at concentrations near the LRL value of 0.030. Two field samples were collected and each was split into two fractions, one of which was spiked with analyte to produce a concentration increase of 0.030 above background. All four samples were analyzed. The results are shown in the table below.

	Sample 1	**Sample 2**
Background (unspiked)	0.047	<0.030
Spiked (0.030 additional)	0.063	<0.030
Δspike	0.016	0 (highly uncertain)

Analysis: The true concentration is unknown for all four samples. However, the data user is sure that the concentration in the two spiked samples should be at least 0.030. Because this value is the LRL, the probability of nondetection is no more than 1%. The binomial distribution function can be used as before. Out of two spikes, one was a detection and one was not, giving:

$$P = \frac{2!}{1!1!}(0.99)^1(0.01)^1 = 0.02$$

If neither sample contained the analyte before spiking, then the probability of getting one detection and one nondetection simply by chance is 2%. Because these samples might have contained analyte before spiking, the probability of one nondetection is less than 2%. At this point, the data user must make a judgement call. It is reasonable to conclude that sample 1 contained analyte before spiking. These data suggest that some analyte may have been lost or that the method may have a low bias for these samples. The addition of spike to sample 1 increased the concentration by 0.016, or about half of what would be expected. The addition of spike to sample 2 was not discernible. With only two samples, the data are inconclusive and it is difficult to know if this behavior was due to random chance, if something was wrong with the spike solution, or if matrix effects compromised the analyses of these samples. The data user has just enough data to be suspicious and possibly worried, but not enough data to take any definitive action. The next example shows an approach that is less likely to leave the data user in such a quandary.

Quality Assurance Example 8

What can I do if my spikes show a poor recovery?

Description: As in the previous example, the project manager anticipates that a number of samples will contain concentrations of analyte near the LRL. The LRL is 0.030 as before, and the samples were spiked so the expected concentration increase would be 0.030. This time, two sets of ***triplicate*** field samples and spikes were submitted. That is, each raw sample was split into two fractions, one of which was spiked. The spiked and unspiked fractions were each split into three fractions which were sent in for analysis. Although such a sampling scheme requires a large volume of the original field sample, having the data from triplicate analyses is advantageous because it allows the calculation of standard deviations. The results are shown in the table below. Note that this analytical method must have been information-rich because one reported result is smaller than the LT-MDL of 0.015. (Note: the concepts in this example also apply to methods that are not information-rich.)

	Triplicate Results	$\bar{x}$	**s**	**∆spike**
Sample 1				
Background (unspiked)	<0.030, <0.030, <0.030	<0.030	?	~0.0147
Spiked (0.030 additional)	E0.010, E0.019, E0.015	0.0147	0.0045	
Sample 2				
Background (unspiked)	0.053, 0.042, 0.050	0.0483	0.0057	0.0157
Spiked (0.030 additional)	0.060, 0.065, 0.067	0.0640	0.0036	

Analysis: The results for sample 1 indicate that the analyte was probably not present in the unspiked sample. Spikes of sample 1 clearly show that analyte is present, although one of the spikes is less than the LT-MDL. Given that the spikes were done at a level of 0.030, the same as the LRL, the chance of detecting each spike is 99% when added to a sample containing no analyte. The chance of obtaining two values above and one value below the LT-MDL can be calculated from the binomial distribution function:

$$P = \frac{3!}{1!2!}(0.99)^2(0.01)^1 = 0.03 = 3\%$$

This result is somewhat rare and probably indicates analyte losses or low bias. Reproducibility is fine. The standard deviation for the sample 1 replicates (0.0045) is not statistically different from the s for the LT-MDL blank spikes (0.006) (see Quality Assurance Example 1 for how to check this). The values for the sample 1 spikes, however, appear to be less than they should be, which supports the concern about analyte loss or low bias in the analytical method. Analyte losses probably mean that the LRL is too low for these samples.

Sample 2 clearly contained analyte before spiking. The addition of analyte to the spiked samples is also clearly discernible, but the values again appear to be smaller than they should be. The reproducibility is comparable to what is expected for the method (0.006).

In general, both of these results indicate that some of the spike is being “lost.” Possible causes of this include inefficient extraction, analyte degradation, and volatilization. Some methods routinely produce low results for such reasons. The data user decides to compare the study results to those from the BSP and the lab set spike. Those data do not show a similar loss of analyte, indicating that the problem is not part of routine NWQL procedures. Alternatively, the apparent loss of analyte might be due to a matrix problem. A matrix issue may or may not apply to all samples in this project. At this point, the data user might contact NWQL personnel who can look more carefully at the raw results for these analyses to see if there is any indication as to why they are apparently low. The data user may also have additional knowledge about the sample matrix.

The data user may decide that samples for this particular project have a matrix interference that causes a diminished signal. For example, samples with high amounts of dissolved organic matter may not be extracted efficiently. In such a case, the data user should probably raise the reporting level (RL) for these samples. First, a recovery factor is calculated by dividing the actual recovery by the expected recovery. For samples 1 and 2, respectively, the recovery factors are:

$$\frac{0.0147}{0.030} = 0.490 \text{ and } \frac{0.0157}{0.030} = 0.523$$

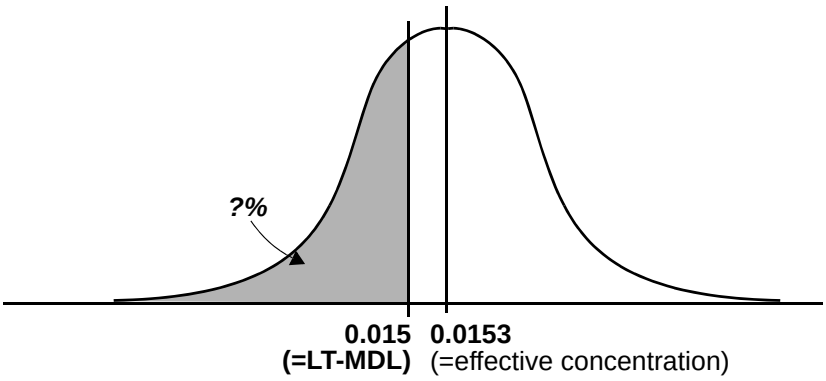
The average recovery factor is 0.51, making the new RL for these samples equal to:

$$\text{RL} = \frac{\text{LRL}}{\text{recovery factor}} = \frac{0.030}{0.51} = 0.059$$

This means that a sample with a true concentration of 0.059 has only a 1% chance of being reported as less than the LT-MDL (0.015). Note that in this case, the RL is no longer equal to twice the LT-MDL.

The results for sample 1 can be reevaluated, knowing that the chance of having a value reported as less than the LT-MDL in this instance is greater than 1%. First, determine the probability that a sample with a true concentration of 0.030 is reported with a concentration of less than the LT-MDL. In this example, a sample with a true concentration of 0.030 has an “effective concentration” of $0.030 \times 0.51 = 0.0153$ (concentration $\times$ recovery factor). The distribution is illustrated in the figure on the following page. The shaded area is the fraction of the distribution that is less than the LT-MDL. To obtain the probability associated with the shaded area, calculate a standardized t value and look up the probability in a table of t-values. To calculate the standardized t value, subtract the effective concentration from the LT-MDL and then normalize by assuming a standard deviation of 0.006 (based on the LT-MDL).

$$t = \frac{0.015 - 0.0153}{0.006} = -0.05$$



The value -0.05 is looked up in a table of *t* values, assuming 19 degrees of freedom and a 1-tail test. From such a table, the probability of a value falling in the shaded area shown above is 0.48 or 48%. Therefore, the probability of a detection (a value of 0.015 or greater) is 52% and the probability of a nondetection is 48%. For sample one, out of three spikes, two were greater than the LT-MDL and one was less than the LT-MDL. Using the binomial distribution function yields:

$$P = \frac{3!}{1!2!}(0.52)^2(0.48)^1 = 0.39$$

The probability that one in three spikes with a concentration of 0.030 is reported as less than the LT-MDL is 39%. This is obviously likely and supports the idea that raising the RL is an appropriate action.

Project Planning Example 1

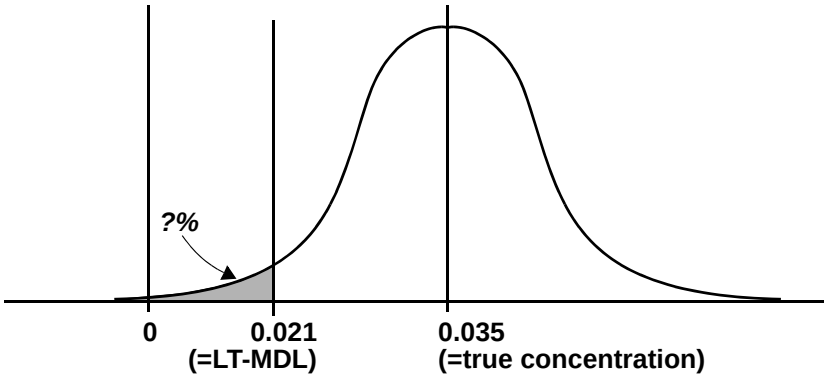
I want to compare values to a criterion at the low end of the analytical range. Is that possible with this analytical method?

Description: The LT-MDL for a particular analyte is 0.021, based on a standard deviation of 0.008 and 16 samples. The LRL is 0.042. As part of a proposed project, ambient values will be compared to a criterion which is 0.035. How feasible is this? What sorts of errors are involved? How can errors be minimized?

Analysis—Part A: The main purpose of this project is identifying values that are greater than or equal to 0.035. Assume that the project manager has decided that it is very important that the actual concentrations in samples reported as nondetections (<LRL, <0.042) are really less than 0.035. The project will be designed with this objective in mind.

First, calculate the probability of not detecting the analyte in a sample that has a true concentration of 0.035. (A sample with a true concentration of 0.035 is the most difficult case of the objective.) This example assumes that the recovery factor is 100% (no analytical bias). Detection is defined as an analytical result that is greater than or equal to the LT-MDL (0.021).

$$t = \frac{0.021 - 0.035}{0.008} = -1.75$$



Looking up the probability in a table of t values, with 15 degrees of freedom, a 1-tail test yields a probability of 0.048. This means there is about a 1 in 20 chance (0.05) of a sample with a true concentration of 0.035 being reported as a nondetection. The project planner decides that 1 in 20 is too many exceedences to miss.

One way to reduce the chance of erroneously not detecting the analyte is to analyze replicates for every sample. Consider analyzing duplicates. The possible results for a sample with a true concentration of 0.035 are detailed in the table for duplicate samples on the following page.

.

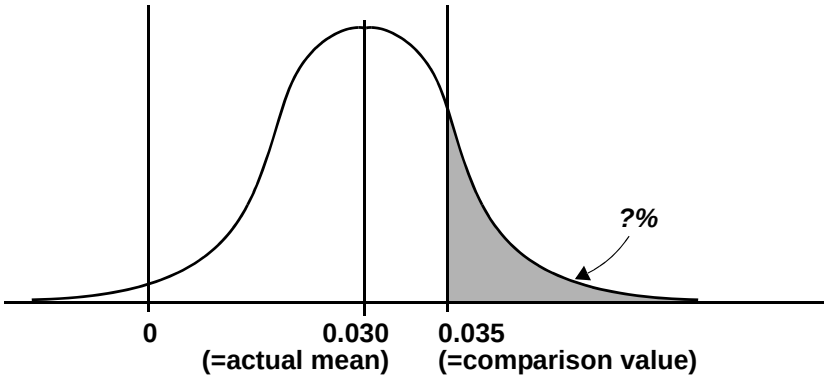
Potential Outcome	Probability Calculation	Probability
Detected in both duplicates	$P = \frac{2!}{0!2!}(0.95)^2(0.05)^0 = 0.903$	90.3%
Detected in only one duplicate	$P = \frac{2!}{1!1!}(0.95)^1(0.05)^1 = 0.095$	9.5%
Not detected in either duplicate	$P = \frac{2!}{2!0!}(0.95)^0(0.05)^2 = 0.0025$	0.3%

If every sample is analyzed in duplicate, the project can effectively lower the default "less than" reporting value. If the analyte is not detected in either duplicate, then the data user can conclude with 99.7% certainty that the true concentration is less than 0.035, despite the fact that the reported result was <LRL or <0.042. Of course, this comes with a price: the cost of duplicate sampling.

Analysis—Part B: In this case, assume that the project manager has a different objective—minimizing the chances of concluding that the concentration in a sample exceeded 0.035 when it actually did not. Recognizing that all measurements have error, the project manager decided to determine the probability that a sample exceeded 0.035, when it actually was 0.030 or less. (The choice of 0.030 by the project manager is arbitrary; choosing 0.030 allows for a 14% margin of safety.)

$$t = \frac{0.035 - 0.030}{0.008} = 0.625$$

A table of t values shows that the probability associated with a t of 0.625, 15 degrees of freedom and a 1-tail test is 0.27 or 27%, which the project manager considers too high.The project manager considers having every sample analyzed in duplicate.



Potential Outcome	Probability Calculation	Probability
Reported as ≥0.035 in both duplicates	$P = \frac{2!}{0!2!}(0.27)^2(0.73)^0 = 0.073$	7.3%
Reported as ≥0.035 in one duplicate and as <0.035 in one duplicate	$P = \frac{2!}{1!1!}(0.27)^1(0.73)^1 = 0.394$	39.4%
Reported as <0.035 in both duplicates	$P = \frac{2!}{2!0!}(0.27)^0(0.73)^2 = 0.533$	53.3%

There is still a 7% chance that a sample with a concentration of 0.030 would be reported as greater than or equal to 0.035 for both duplicates. The project manager considers this error too high, although it is better than the 27% chance for one sample. The manager considers analyzing triplicates.

Potential Outcome	Probability Calculation	Probability
Reported as ≥0.035 in all three replicates	$P = \frac{3!}{0!3!}(0.27)^3(0.73)^0 = 0.02$	2%
Reported as ≥0.035 in two replicates and as <0.035 in one replicate	$P = \frac{3!}{1!2!}(0.27)^2(0.73)^1 = 0.16$	16%
Reported as ≥0.035 in one replicate and as <0.035 in two replicates	$P = \frac{3!}{2!1!}(0.27)^1(0.73)^2 = 0.43$	43%
<0.035 in all three replicates	$P = \frac{3!}{3!0!}(0.27)^0(0.73)^3 = 0.39$	39%

The project manager considers 2% an acceptable error rate and plans to submit a triplicate split for every sample. If every one of the three triplicates produces a result greater than or equal to 0.035, the data user will conclude that the true concentration of the sample could have been 0.030 or greater.

Analysis—Comparing A and B: Parts A and B of this example show the two types of error encountered in determining detection or deciding if a value exceeds a criterion. One type of error is concluding detection or exceedence when it actually did not occur; the other type of error is concluding nondetection or nonexceedence when it did actually occur. These two types of error are always a trade off. The less likely the first type of error, the more likely the second type, and vice versa. Analyzing replicate samples can allow the data user to constrain both types of error simultaneously, but introduces a third category: “too close to call.” For example, suppose that a sample was analyzed in

duplicate and the result was one nondetection and one 0.035. It would be difficult to conclude with any certainty whether or not this sample actually exceeded 0.035. If the concentration actually was 0.035, there would be a 50% chance of obtaining one reported value greater than 0.035 and one reported value less than 0.035. If a project manager requires greater precision, the only options are to increase the number of replicates, thereby decreasing the standard error of the mean, or to ask the analytical laboratory to use a different method of chemical analysis which has less variability.

Project Planning Example 2

I want to compare two groups. Is the precision of this analytical method adequate?

Description: Some projects involve comparing two or more groups of data to each other. When this is the case, the analytical uncertainty must be substantially smaller than the difference that the project is attempting to discern. For example, a project is proposed to determine if there is an observable chemical difference between typical urban streams and similar streams which have undergone streambank restoration. Will the analytical method be precise enough to distinguish between the two?

Analysis: As a rule of thumb, the means of two groups must differ by at least three standard deviations (more for small data sets) for a discernible difference to be statistically observable. The standard deviation here measures the total variability—spatial, temporal, sampling, and analytical. Typically, the analytical variability will account for only a tiny fraction of the total variability. In any event, it is important to make sure that the analytical variability is acceptable for the proposed project.

The LT-MDL is calculated as the product of the standard deviation for the method and a t-multiplier. The t-multiplier will never be less than 2.3*. Therefore, the routine analytical standard deviation of the method will not be larger than the LT-MDL divided by 2.3.

Suppose that the LT-MDL for a particular analyte is 0.0067. The maximum routine analytical standard deviation is 0.003 (0.0067/2.3), for values that are in the low range of the method ($<5\times$LRL or $<10\times$LT-MDL). The minimum difference that can be observed between group means therefore would be about 0.009, assuming that all other forms of variability are negligible. If this value is greater than or even near the value of the difference that is expected, the project manager must use a different analytical method with lower variability or alter the project in some other way.

*[Note: the t-multiplier comes from the Student's t distribution. The t-multiplier in the LT-MDL will never be less than 2.3 because that is the t-value for an infinite number of degrees of freedom at a probability of 0.99.]

Data Interpretation Example 1

I've heard that it is easy to misinterpret nondetections in a way that introduces bias in the summary statistics. How do I avoid that? The analytical method is not information-rich.

Description: The LT-MDL for a particular analyte is 0.003 and the LRL is 0.006. The data are E0.004, E0.005, <0.006, <0.006, 0.008, and 0.010. What are some options for handling these data? What should be avoided?

Analysis: Bias is introduced when the data are interpreted in a way that is inconsistent with the rank order of the *actual lab measurements*. It is important to remember that the LRL is a default value rather than a measured one. The lab measurement leading to a reported value of "<LRL" was less than the LT-MDL. Analytical results for samples having true concentrations between the LT-MDL and the LRL are associated with a larger relative error (the reason for the E-code) and a greater incidence of false negatives than those for samples having true concentrations greater than the LRL. The following table shows several methods of acceptable interpretation which could be used in calculating summary statistics or comparing groups. The choice of method depends upon the needs of the data user.

Data as Reported from NWQL	Actual Lab Measurement	Used in Data Analysis	
		Value	Rank
Most conservative approach—*Values that have high uncertainty are not compared to one another, but rather, are ranked equally. This is appropriate when great certainty is required or when the QA needed to interpret results near the LT-MDL is absent.*			
<0.006	0 ≤ measurement <0.003, high uncertainty	<0.006	2.5
<0.006	0 ≤ measurement <0.003, high uncertainty	<0.006	2.5
E0.004	0.004, high uncertainty	<0.006	2.5
E0.005	0.005, high uncertainty	<0.006	2.5
0.008	0.008	0.008	5
0.010	0.010	0.010	6
Least conservative approach—*This approach is useful when characterizing a distribution and when the data user can tolerate greater uncertainty. There should be adequate QA (blanks and low-level spikes) to demonstrate that the LT-MDL is appropriate for the data set.*			
<0.006	0 ≤ measurement <0.003, high uncertainty	<0.003	1.5
<0.006	0 ≤ measurement <0.003, high uncertainty	<0.003	1.5
E0.004	0.004, high uncertainty	0.004	3
E0.005	0.005, high uncertainty	0.005	4
0.008	0.008	0.008	5
0.010	0.010	0.010	6

Data as Reported from NWQL	Actual Lab Measurement	Used in Data Analysis	
		Value	Rank
Intermediate approach—*The distinction between detection and nondetection is preserved, but low-level detections (considered nonquantitative detections) are not compared to one another. Results from field blanks should show that the incidence of false positives is not inconsistent with the LT-MDL. If adequate low-level spike data are absent, the data user must be tolerant of the possibility of an increased chance of false negatives in the nondetection group.*			
<0.006	0 ≤ measurement <0.003, high uncertainty	<0.003	1.5
<0.006	0 ≤ measurement <0.003, high uncertainty	<0.003	1.5
E0.004	0.004, high uncertainty	0.003 – 0.006	3.5
E0.005	0.005, high uncertainty	0.003 – 0.006	3.5
0.008	0.008	0.008	5
0.010	0.010	0.010	6
INCORRECT approach—*The nondetections encompass values that exceed an E-coded detection. This is inconsistent with lab measurements and leads to bias, regardless of the method used. Using the LRL as the censoring level while including individual values between the LT-MDL and LRL will bias results from methods such as regression on order statistics (ROS or probability plot), Kaplan-Meier estimation, and maximum likelihood estimation (MLE).*			
<0.006	0 ≤ measurement <0.003, high uncertainty	<0.006	3.5
<0.006	0 ≤ measurement <0.003, high uncertainty	<0.006	3.5
E0.004	0.004, high uncertainty	0.004	1
E0.005	0.005, high uncertainty	0.005	2
0.008	0.008	0.008	5
0.010	0.010	0.010	6

Data Interpretation Example 2

I've heard that it is easy to misinterpret nondetections in a way that introduces bias. How do I avoid that? The analytical method is information-rich.

Description (Information-rich methods): The LT-MDL for a particular analyte is 0.003 and the LRL is 0.006. The data are E0.002, E0.004, E0.005, <0.006, 0.008, and 0.010. What are some options for handling these data? What sort of interpretation should be avoided because it introduces bias?

Analysis: The essential requirement for treating data from information-rich methods is the same as that for methods that are not information-rich: make sure that the data are interpreted in a way that is consistent with the rank order of the *actual lab measurements*. Data sets for information-rich methods are sometimes complicated by reported values less than the LT-MDL which are extremely uncertain. The following table shows several methods of acceptable interpretation which could be used in calculating summary statistics or comparing groups. The choice of method depends upon the needs of the data user.

Data as Reported from NWQL	Actual Lab Measurement	Used in Data Analysis	
		Value	Rank
Most conservative approach—*Values that have high uncertainty are not compared to one another, but rather, are ranked equally. This is appropriate when great certainty is required or when the QA needed to interpret low levels is absent.*			
<0.006	analyte confirmation insufficient	<0.006	2.5
E0.002	0.002 very high uncertainty	<0.006	2.5
E0.004	0.004, high uncertainty	<0.006	2.5
E0.005	0.005, high uncertainty	<0.006	2.5
0.008	0.008	0.008	5
0.010	0.010	0.010	6
Least conservative approach—*This approach is useful when characterizing a distribution and when the data user can tolerate considerable uncertainty in individual values. There should be adequate QA (blanks and low-level spikes) to demonstrate that the LT-MDL is appropriate for the data set. Because information-rich methods require evidence that identifies the analyte, using zero for nondetection is not equivalent to a substitution method. Zero is the value that is most faithful to the analytical result—sufficient evidence of presence was not observed.*			
<0.006	analyte confirmation insufficient	0.000	1
E0.002	0.002 very high uncertainty	0.002	2
E0.004	0.004, high uncertainty	0.004	3
E0.005	0.005, high uncertainty	0.005	4
0.008	0.008	0.008	5
0.010	0.010	0.010	6

Data as Reported from NWQL	Actual Lab Measurement	Used in Data Analysis	
		Value	Rank
Intermediate approach 1—*Detections that are less than the LT-MDL are grouped with nondetections. This is especially useful to minimize the possibility of false positives (usually due to carry over between analyses). Values between the LT-MDL and LRL are grouped together because of their increased uncertainty relative to values greater than the LRL.*			
<0.006	analyte confirmation insufficient	<0.003	1.5
E0.002	0.002 very high uncertainty	<0.003	1.5
E0.004	0.004, high uncertainty	0.003 – 0.006	3.5
E0.005	0.005, high uncertainty	0.003 – 0.006	3.5
0.008	0.008	0.008	5
0.010	0.010	0.010	6
Intermediate approach 2—*nondetections, detections less than the LT-MDL, and detections between the LT-MDL and LRL are treated as three separate groups. This approach maximizes the ability to observe detection. Adequate QA (blanks and low-level spikes) is important. In particular, users of this approach should check lab blanks as well as field blanks for incidence of low-level false positives that may indicate carry over between analyses.*			
<0.006	analyte confirmation insufficient	0.000	1
E0.002	0.002 very high uncertainty	0.000 – 0.003	2
E0.004	0.004, high uncertainty	0.003 – 0.006	3.5
E0.005	0.005, high uncertainty	0.003 – 0.006	3.5
0.008	0.008	0.008	5
0.010	0.010	0.010	6
INCORRECT approach—*The nondetections encompass values that exceed several E-coded detections. This is inconsistent with lab measurements and leads to bias, regardless of the method used. Using the LRL as the censoring level while including individual values less than the LRL will bias results from methods such as regression on order statistics (ROS or probability plot), Kaplan-Meier estimation, and maximum likelihood estimation (MLE).*			
<0.006	analyte confirmation insufficient	<0.006	4
E0.002	0.002 very high uncertainty	0.002	1
E0.004	0.004, high uncertainty	0.004	2
E0.005	0.005, high uncertainty	0.005	3
0.008	0.008	0.008	5
0.010	0.010	0.010	6

Data Interpretation Example 3

How do I interpret data and calculate statistics when I have nondetections? The low-level values are not important to my study.

Description: The LT-MDL and LRL for a particular analyte are 0.004 and 0.008, respectively. The data user is interested in reporting some basic descriptive statistics such as medians and percentiles. The data user also would like to perform some statistical tests such as comparing groups and correlations. Most of the data exceed the LRL value. The data user is not particularly concerned with values less than the LRL. The criterion for aquatic health for this analyte is 0.040, a value that is 10 times the LT-MDL. Data for one group in the data set are 0.015, 0.024, 0.019, 0.031, 0.010, <0.008, 0.023, E0.006, 0.046, 0.018, and 0.022. How should nondetections be handled?

Analysis: The data user decides to use a simple, robust method to handle this data set. The user censors the data at the LRL (0.008) and the data are ranked as shown in the table below. The censored data are shaded. (Note that the sample reported as E0.006 produced a larger analytical signal than the sample reported as <0.008, but both are ranked equally because the data are being censored at the LRL).

Data as reported from NWQL	Data as used by data user	Rank
<0.008	<0.008	1.5
E0.006	<0.008	1.5
0.010	0.010	3
0.015	0.015	4
0.018	0.018	5
0.019	0.019	6
0.022	0.022	7
0.023	0.023	8
0.024	0.024	9
0.031	0.031	10
0.046	0.046	11

The data user decides to report nonparametric summary statistics such as percentiles as shown in the table below. Detailed calculations for this example are shown in the box on page 45.

90th percentile	0.043
75th percentile	0.024
median	0.019
25th percentile	0.010
10th percentile	<0.008

The 10th percentile here includes, but is not limited to, nondetections. It should be interpreted literally—values less than 0.008. Depending on the data set, the censoring point may fall much higher in the distribution; in other words, a greater proportion of the data set could be censored. This is not a problem because the data user made the decision that the high end of the distribution is what is important. Low level detections are not pertinent to data interpretation for this project.

The data user also can perform nonparametric statistical tests on the ranks of censored values (for example, Spearman correlation, Wilcoxon group comparisons, etc.). For large data sets, parametric methods such as *t*-tests and Analysis of Variance can be performed on the ranks because parametric tests applied to ranks approximate the nonparametric tests for large data sets. For small data sets such as this one, however, the exact nonparametric tests are preferred.

This method of handling low-level values is advantageous because it is straightforward and requires no assumptions. Although data below the LRL are not interpreted, that should not be considered a disadvantage. Not interpreting data below the LRL is often the most appropriate approach, especially for projects that are primarily concerned with identifying areas where high concentrations are a threat to aquatic or human health. The data user needs to decide what is appropriate based on the intent of the study.

Calculating percentiles

Percentiles can be calculated many ways; for example, SAS (Proc Univariate) offers 5 different methods. There is no general agreement about which methods are best for which applications. The statistics produced by all methods are very similar for large data sets and for medians, but diverge for small data sets and for large or small percentiles. To avoid highly uncertain estimates, the range of percentiles calculated should be limited to between $\frac{100}{n}$ and 100– $\frac{100}{n}$, where *n* is the number of data values. In this document, the method described by Helsel and Hirsch (1992) is used. It is identical to SAS Proc Univariate Definition 4 (SAS Institute, 1990). This method produces estimates of upper percentiles that are slightly greater than and lower percentiles that are slightly less than those of the other methods.

Procedure:

The general formula is: $(100 \times p)^{th} percentile = x_i + frac(x_{i+1} - x_i)$

where *i* and *frac* are the integer part and fractional parts of (*n*+1)*p*, respectively, *n* is the number of values in the data set, and *p* is the quantile (the percentile represented as a fraction).

1. Calculate the nearest data (x_i and x_{i+1}) and the weighting factor (*frac*) for each percentile.

- For this dataset, *n*=11. For the 90th percentile, *p*=0.90
- Calculate *i* and *frac*.
  (n+1)p = (11+1)(0.90) = 10.8, therefore i=10 and frac=0.80
- Choose nearest data using the index, *i*, after ordering the data from smallest to largest.
  $x_i = x_{10} = 0.031$ *and* $x_{i+1} = x_{11} = 0.046$

2. Calculate the percentile value.
90th percentile = 0.031 + 0.80(0.046 – 0.031) = 0.043

A spreadsheet showing the results for this data set is shown at the right. For this dataset, the calculation should be limited to the 9th to 91stpercentile range.

Quanitile (p)	*(n+1)p=*	*i =*	*frac =*	*Value*
0.90	10.80	10	0.80	0.043
0.75	9.00	9	0.00	0.024
0.50	6.00	6	0.00	0.019
0.25	3.00	3	0.00	0.010
0.10	1.20	1	0.20	<0.008

About Excel percentiles. Because of its ubiquity and ease of use, Microsoft Excel is often used to calculate percentiles. The method used by Excel differs from Helsel and Hirsch and all of the SAS methods. Because of how Excel ranks data, the Excel Percentile Function may not be the best method for datasets that contain many tied data values.

Data Interpretation Example 4

How do I calculate summary statistics when I have nondetections? I am not concerned with individual values, but want to characterize the distribution. The analytical method used was not information-rich.

Description: An ongoing project has a moderately sized set of ambient chemical data. The purpose of the project is to statistically describe the distribution of analyte concentrations in the environment. Because the data will be viewed collectively, an individual value can have large uncertainty without compromising the interpretation. Much of the data falls in the low-level range, including values between the LT-MDL (0.050) and LRL (0.100) and many nondetections. Because the analytical method was not information-rich, the laboratory censored values below the LT-MDL and reported them as <LRL, the same as nondetections. How should this data set be handled?

Analysis: Because most of the distribution falls in the low-level range, the data user would prefer not to censor the data at the LRL—too much information would be lost. However, before proceeding, the data user must examine the QA data to make sure that the samples were well behaved during analysis. If this is not the case for some samples, the data user may choose to omit those samples from interpretation and explain the reasons for doing so. For example, the data user noticed that blanks collected in June of a particular year had evidence of contamination, but that no other blanks did. The amount of contamination was similar in magnitude to low-level environmental concentrations measured in this project. Based on this information, the data user concluded that the June data were seriously compromised and decided to omit samples collected during that time period from the final analysis. Note that this decision is somewhat arbitrary. Another data user might choose to retain the questionable data. The decision to omit or to include data that are known to be compromised should not be made lightly. The data user should examine the effect of either omitting or retaining these data and use that information to inform their decision.

Once the data user has assessed the quality of the analytical results, nondetections must be handled. Nondetections are reported as <LRL, but the LRL is simply the default "less than" reporting value that was selected to minimize the incidence of false negatives (reporting an analyte as not present when it actually is present). When an analyte is reported as not detected, the signal observed by the analyst was in the level of noise for a non-information-rich method (as in this case) and less than the signal equivalent to the LT-MDL. Consequently, *the value that most faithfully reports the observed analytical signal is <LT-MDL*. The reason the NWQL does not report it this way is that a sample which had a true concentration just greater than the LT-MDL has a high probability of being reported as <LT-MDL (almost 50%; more if an analyte has a recovery of less than 100%). Reporting the value as <LRL limits the chance of a false negative error to 1%. For the purpose of describing a distribution, however, high certainty is not needed for each value; rather, the most faithful representation is needed.

The data user reassigns "<LT-MDL" to the results that were originally reported as "<LRL." This is not altering the data. The data user has chosen to accept a higher level of uncertainty for nondetections than that used by NWQL. In this situation, the LT-MDL is a more appropriate default "less than" value for nondetections than is the LRL. A sample data set is shown below.

Data as reported from NWQL	Data as used by data user	Rank
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
< 0.100	<0.050	7
E 0.057	E 0.057	14
E 0.061	E 0.061	15
E 0.081	E 0.081	16
E 0.090	E 0.090	17
E 0.091	E 0.091	18
E 0.093	E 0.093	19
0.103	0.103	20
0.119	0.119	21
0.133	0.133	22
0.134	0.134	23
0.137	0.137	24
0.184	0.184	25
0.248	0.248	26
0.537	0.537	27
0.542	0.542	28
0.544	0.544	29
1.17	1.17	30

The data user then decides what method to use to determine summary statistics. Two good methods are the rank method and the robust "regression on order statistics" (robust ROS) method. Both methods are appropriate for most data sets. A detailed dis-

cussion of these two methods follows, and the calculated summary statistics are shown at the end of the section.

Other methods such as Kaplan-Meier and Maximum Likelihood Estimation (MLE) may also be used, but are not shown here. When only one censoring value is present, the Kaplan-Meier method does not yield much more information than the rank method and the estimated mean will be biased if the lowest value is censored as it is in this example. The Kaplan-Meier method is shown in Data Interpretation Example 7. The Maximum Likelihood Estimation method requires a relatively large data set (at least 50 values) and an assumption about the shape of the distribution. An example of the use of MLE is not included in this document.

The rank method is simple and involves no assumptions about the underlying distribution. It requires a single censoring point below which all data are censored. Therefore, data sets with multiple detection limits or reported data less than the censoring level (such as is possible with information-rich methods) must be recensored, causing the loss of some information. Because the percentiles are determined by using <LT-MDL as the lowest rank, some of the statistics produced by the rank method will be censored values and some statistics may be ties (as are the 10th and 25th percentiles in this case, both of which are <0.050). The rank method does not yield parametric statistics such as the mean or standard deviation.

The robust ROS method uses a probability plot procedure to “fill in” censored values (Helsel and Hirsch, 1992; Helsel, 2005). Although an underlying data distribution is an inherent assumption of this method, the overall method is not fully parametric, because summary statistics are calculated using the fill-in values (assumed data distribution) combined with the noncensored data (no assumed distribution). The robust ROS method can accommodate multiple detection limits and detected values that are less than the censoring limit without recensoring (see Data Interpretation Example 7). This method also can produce parametric summary statistics. Detailed calculations for this example are shown in the boxes on the following pages.

Summary statistics obtained from each method are shown below.

	Rank Method	Robust ROS Method
10th percentile	<0.050	0.008
25th percentile	<0.050	0.020
median	0.071	0.071
75th percentile	0.135	0.135
90th percentile	0.542	0.542
mean	—	0.153
standard deviation	—	0.247

ROS Method —Data Interpretation Example 4 ***(page 1 of 2)***

A spreadsheet was used for the robust ROS method in this example. For a detailed discussion of the theory involved in this method see Helsel (2005).

Part 1—The noncensored values are used to obtain a regression equation. A log-normal distribution is assumed.

1. Compute normal scores for the detected values.

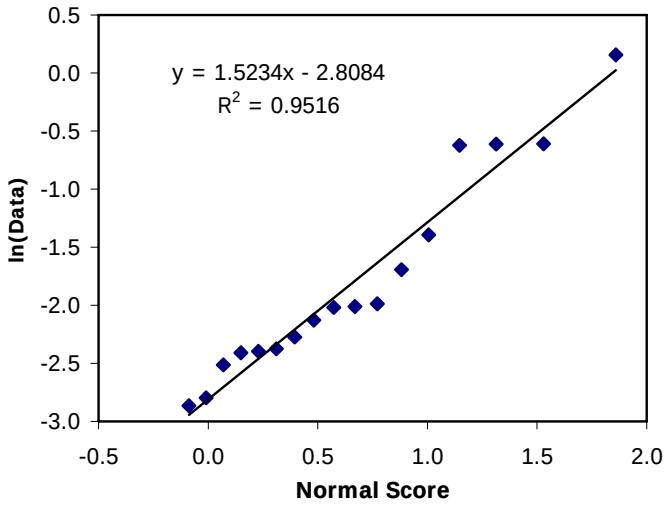
- List the noncensored values in order from largest to smallest. (*See* ***Detected Data*** *column*.)
- Calculate the probability of detection as: number of detections/total number.
  P(detection) = 17/30 = 0.5667
- Calculate the probability level increment as: P(detection)/(number of detections + 1).
  Increment = 0.5667/18 = 0.0315
- Assign probability levels to the detected values. (*See* ***Prob Level*** *column*.)
  For the highest data value: Prob Level = 1 – Increment
  For subsequent values: Prob Level = Previous Prob Level – Increment
  Check: Prob Level for the smallest detection should equal: 1 – P(detection) + Increment
- Calculate the normal score (z-score) associated with each probability level by using the inverse of the normal distribution function.(*See* ***Normal Score*** *column.*)

2. Take the natural logarithm of the detected values. (*See* ***ln(Data)*** *column*.)

3. Graph the ln(Data) versus the Normal Score and calculate the equation of the regression line.

The portion of the spreadsheet resulting from these steps is shown below.

Detected Data	Prob Level	Normal Score	ln(Data)
1.17	0.969	1.859	0.16
0.544	0.937	1.530	-0.61
0.542	0.906	1.314	-0.61
0.537	0.874	1.146	-0.62
0.248	0.843	1.005	-1.39
0.184	0.811	0.882	-1.69
0.137	0.780	0.771	-1.99
0.134	0.748	0.669	-2.01
0.133	0.717	0.573	-2.02
0.119	0.685	0.482	-2.13
0.103	0.654	0.395	-2.27
0.093	0.622	0.311	-2.38
0.091	0.591	0.229	-2.40
0.090	0.559	0.149	-2.41
0.081	0.528	0.070	-2.51
0.061	0.496	-0.009	-2.80
0.057	0.465	-0.088	-2.86



ROS Method —Data Interpretation Example 4 (continued) ***(page 2 of 2)***

Part 2—The regression equation is used calculate fill-in values to represent the censored values in the calculation of summary statistics.

1. Compute normal scores for the censored values.

- Calculate the probability of nondetection as: 1 – P(detection)
  P(nondetection) = 1 – 0.5667 = 0.4333
- Calculate probability level increment as: P(nondetection)/(number of nondetections + 1)
  Increment = 0.4333/14 = 0.03095
- Create a list of probability levels for censored values. (*See* ***Censored Prob*** *column*, below)
  Beginning with: Censored Prob = P(nondetection) – Increment
  For subsequent values: Censored Prob = Previous Censored Prob – Increment
  The number of Censored Probability Levels should equal the number of censored values and the least Censored Probability Level should equal the Increment (within round-off error).
- Calculate the normal score (z-score) associated each probability level by using the inverse of the normal distribution function. (*See* ***Censored Norm*** *column.*)

2. Use the regression equation from Part 1 to calculate the natural logarithms of fill-in values. (*See* ***ln(Fill-in)*** *column.*)

3. Calculate fill-in data values from their natural logarithms. *(See* ***Fill-in Values*** *column.)*

The portion of the spreadsheet resulting from these steps is shown at the right.

Censored Prob	Censored Norm	ln (Fill-in)	Fill-in Values
0.402	-0.247	-3.185	0.041
0.371	-0.328	-3.308	0.037
0.340	-0.411	-3.435	0.032
0.310	-0.497	-3.566	0.028
0.279	-0.587	-3.703	0.025
0.248	-0.682	-3.847	0.021
0.217	-0.784	-4.002	0.018
0.186	-0.894	-4.170	0.015
0.155	-1.016	-4.357	0.013
0.124	-1.156	-4.570	0.010
0.093	-1.323	-4.824	0.008
0.062	-1.539	-5.153	0.006
0.031	-1.867	-5.653	0.004

Part 3—The detected values (all 17 values in the ***Detected Data*** column on the previous page) and the fill-in values (all 13 values in the ***Fill-in Values*** column at the right) are combined and used to calculate summary statistics. The reader can verify the summary statistics calculated from this hybrid data set. Note that fill-in values do not correspond to any specific samples and should never be reported as data.

Data Interpretation Example 5

How do I calculate summary statistics when most of my data are nondetections or low-level values? The analytical method was information-rich and many of the reported values are less than the LT-MDL.

Description: An ongoing project has large set of ambient chemical data. The data are to be used to calibrate and verify a water quality model. No individual value is of particular importance and the data user can tolerate considerable uncertainty regarding each value. Most of the data are below the LRL (0.10) and about 20 percent of the data are nondetections. An information-rich method was used, so some values below the LT-MDL (0.050) are reported. How should these data be handled?

Analysis: The first priority for the data user is to carefully examine the QA data to determine if the low-level samples seem to be well behaved. The data user must be satisfied that the behavior of different samples are comparable to each other and to the routine laboratory performance. If this is not the case, then the data user may choose to adjust values so that they are comparable. Here are two examples. Suppose that the results of low level spikes frequently were reported as nondetections. In such a case, the data user may choose to censor data based on the performance observed. In another case, supposed that one tributary routinely had higher concentrations of humic substances than the others, and that these samples also had matrix problems that led to poor recovery. In this case, the data user may decide to apply sample-specific or site-specific recovery factors to adjust the data in this project. In both of these cases, the data user would need adequate data to calculate censoring levels or recovery factors and must document whatever action was taken. See Quality Assurance Example 8 for more information related to recalculating censoring levels.

Assuming that any data problems have been addressed by the data user, the next task is to decide how to handle nondetections. Censoring the data at the LRL is undesirable because too much valuable information will be lost. Recall that although nondetections are reported as <0.10 (the LRL value), they were not uniquely measured as less than 0.10. The LRL is the default reporting value and is based on probability. Furthermore, for an information-rich method, when an analyte is reported as not detected it usually means that not only was the observed signal in the range of noise, but no positive identification of the analyte was made. In other words, evidence of the analyte's presence was insufficient. For such samples, *zero is the value that is most faithful to the analytical result*. The data user assigns a concentration of zero to all nondetections. This results in no censoring of the data set whatsoever, the data are amenable for use in a model, and no special techniques are needed to calculate summary statistics. Note that each individual value is quite uncertain—reported values below the LT-MDL have a considerable risk of being false positives and zeros have a considerable risk of being false negatives. For this application, however, certainty of individual values is not necessary. Assigning a value of zero to nondetections in this case is different than simply using zero

in a substitution method. Because this analytical method was information-rich, the data user knows that the analyst did not observe sufficient evidence of the analyte to report its presence.

Data as reported from NWQL	Data as used by data user		Rank	
< 0.100	0.000		4	
< 0.100	0.000	*High risk of false negative.*	4	
< 0.100	0.000		4	10^{th} percentile=0.000
< 0.100	0.000	*Evaluate low-level spike*	4	
< 0.100	0.000		4	
< 0.100	0.000		4	
< 0.100	0.000		4	
E 0.006	0.006		8	25^{th} percentile=0.005
E 0.007	0.007		9	
E 0.008	0.008	*High risk of false positive, which decreases as values increase*	10	
E 0.012	0.012		11	
E 0.013	0.013		12	
E 0.016	0.016		13	
E 0.017	0.017		14	
E 0.023	0.023	*Evaluate blank*	15	median=0.030
E 0.037	0.037		16	
E 0.039	0.039		17	
E 0.045	0.045		18	
E 0.048	0.048		19	
E 0.052	0.052	◄*LT-MDL = 0.050*	20	
E 0.058	0.058		21	
E 0.060	0.060		22	
E 0.071	0.071		23	75^{th} percentile=0.074
E 0.082	0.082		24	
E 0.090	0.090		25	
0.110	0.110	◄*LRL = 0.100*	26	
0.133	0.133		27	
0.138	0.138		28	90^{th} percentile=0.138
0.175	0.175		29	
0.182	0.182		30	

Data Interpretation Example 6

My data were collected over several years, during which the LT-MDL changed. What can I do to simplify my data set?

Description: A project collected data over 3 years. During that time, the LT-MDL and LRL for two of the analytes changed. The reported data are shown in the table below.

Analyte	Year 1	Year 2	Year 3
A (information-rich)	<0.006 E0.002 E0.004 E0.005 0.006 0.010	<0.006 <0.006 E0.003 0.006 0.008	<0.008 E0.003 E0.005 E0.005 E0.006 0.009
B (not information-rich)	<0.006 <0.006 E0.004 E0.005 0.006 0.010	<0.006 <0.006 E0.003 0.006 0.008	<0.008 <0.008 E0.005 E0.005 E0.006 0.009

Note that these data demonstrate how the nature of the method affects the options for the data user. The “laboratory signal” in this data set was identical for both methods. The method for analyte A was information-rich; the method for analyte B was not. Data for both analytes have multiple detection limits, which complicates data analysis. Is there a way to simplify this data set to eliminate the problems associated with multiple censoring levels.

Analysis: The laboratory spike samples that are used to determine the LT-MDL and LRL are submitted on an ongoing basis, but the results are evaluated annually. If the LT-MDL calculated for new data is significantly different (on a statistical basis) from the LT-MDL for the previous year, the LT-MDL (and the LRL) will change. Some change is to be expected, because the standard deviation that is used to calculate the LT-MDL is only an estimate of the true standard deviation of the analytical process, which can never be known with absolute certainty. Generally, the changes will be relatively small. If the LT-MDL or LRL change by a large amount, it is indicative of a major change such as a significant equipment upgrade. If some of the data are from a time before the LT-MDL/LRL procedure was implemented, and if the LRL is considerably higher than the older MDL but the method itself has not changed, then it is indicative that the old MDL was too small. In such a case, it would be reasonable to recensor the data at the new LRL value.

In this example, the LRL changes by a small amount (0.006 to 0.008) over the 3-year span of the study. A pooled LT-MDL and LRL can be calculated that span the entire study

period. The data user looks up the LT-MDL values and the number of lab spikes that they were based upon on the NWQL Web pages. These are shown in the second and third columns in the table below. A statistical table is used to obtain the value of Student's t at n-1 degrees of freedom and 99% confidence (1-tail). The standard deviation, s, is calculated as LT-MDL/t.

Year	LT-MDL	n	Student's t	s
Year 1	0.0030	21	2.53	0.00119
Year 2	0.0026	24	2.50	0.00104
Year 3	0.0038	19	2.55	0.00149

The pooled standard deviation is given by:

$$s_{pool}^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2}{n_1 + n_2 + \ldots + n_k - k}$$

Substituting the appropriate values from this example into the above equation yields:

$$s_{pool}^2 = \frac{20(0.00119)^2 + 23(0.00104)^2 + 18(0.00149)^2}{(21+24+19-3)} = \frac{9.316 \times 10^{-5}}{61} = 1.527 \times 10^{-6}$$

Taking the square root gives s=0.00124. This is the appropriate statistic to compare with field estimates of standard deviation (see Quality Assurance Example 1).

The pooled standard deviation also can be used to calculate a 3-year MDL. There are 61 degrees of freedom for this calculation (denominator from s_{pool} calculation). For a 1-tail test with 99% confidence and 61 degrees of freedom, t=2.39. This results in a 3-year MDL of (2.39)(0.00124)=0.00296, or 0.003 for these samples.

The 3-year MDL is useful for the analyte A data (information-rich method) if the data user wants to account for highly uncertain values. The table below shows the application of the 3-year MDL and Intermediate Approach 1 from Data Interpretation Example 2.

Year 1		Year 2		Year 3	
Reported by NWQL	Interpretation	Reported by NWQL	Interpretation	Reported by NWQL	Interpretation
<0.006	0.000 – 0.003	<0.006	0.000 – 0.003	<0.008	0.000 – 0.003
E0.002	0.000 – 0.003	<0.006	0.000 – 0.003	E0.003	0.003 – 0.006
E0.004	0.003 – 0.006	E0.003	0.003 – 0.006	E0.005	0.003 – 0.006
E0.005	0.003 – 0.006	0.006	0.006	E0.005	0.003 – 0.006
0.006	0.006	0.008	0.008	E0.006	0.006
0.010	0.010			0.009	0.009

Notice that the nondetected values that were reported as <0.006 or <0.008 (using the 1-year LRL values) by the NWQL are interpreted as <0.003 (using the 3-year LT-MDL). Because the LRL was a default value, substituting a different, but statistically defined default value is valid. Recall that it is important to use the LT-MDL as the censoring value in data analysis when values less than the LRL are included (see Data Interpretation Example 2). Applying the 3-year LT-MDL to data from an information-rich method eliminates multiple detection limits.

Unfortunately, a multiple-year MDL does not work as well when the method is not information-rich. This is because values less than the annual LT-MDL are reported as "less thans" by NWQL. The multiple-year MDL will be less than the annual LT-MDLs for at least one of the individual years (because it is a type of average). For that year, the data user has no way of knowing if the laboratory value associated with <LT-MDL was greater than or less than the multiple-year MDL. This is the case for the result reported as <0.008 for analyte B in year 3 for this example. The data user knows that the laboratory observed a signal that was less than 0.004 (LT-MDL for year 3), but does not know how that observed signal compared to 0.003 (the 3-year MDL). Consequently, a multiple-year MDL cannot be applied to all nondetections when the method is not information-rich. The multiple-year LRL can only be used when it is greater than the annual-LRL. This is not particularly useful. When the LT-MDL and LRL do not change very much (as in this example), the best approach is to censor all the data using the highest LT-MDL (0.004). This is shown in the table below; the intermediate approach was used (see Data Interpretation Example 1).

Year 1		Year 2		Year 3	
Reported by NWQL	Interpretation	Reported by NWQL	Interpretation	Reported by NWQL	Interpretation
<0.006	<0.004	<0.006	<0.004	<0.008	<0.004
<0.006	<0.004	<0.006	<0.004	<0.008	<0.004
E0.004	0.004 – 0.008	E0.003	<0.004	E0.005	0.004 – 0.008
E0.005	0.004 – 0.008	0.006	0.004 – 0.008	E0.005	0.004 – 0.008
0.006	0.004 – 0.008	0.008	0.008	E0.006	0.004 – 0.008
0.010	0.010			0.009	0.009

Data Interpretation Example 7

How do I calculate summary statistics when I have several detection levels or when I have reported values that are less than the censoring level?

Description: A data user has a data set that is complicated because some of the chemical analyses were done by a local laboratory and some by NWQL. The local laboratory censors all data below their quantitation limit of 0.50. Data from NWQL dating from early in study was reported using an LRL of 0.40. Partway through the study, however, a new method became available at NWQL and the project manager switched to that method because it performed better than the old one. The LRL for the new method was 0.16. Neither the old nor the new NWQL methods were information-rich. To further complicate matters, matrix problems with some samples resulted in NWQL reporting “less thans” with raised reporting levels. The data set is shown in the table below.

Source	Data
Local lab	<0.5, <0.5, <0.5, <0.5, 0.6, 0.9, 1.0, 1.3, 1.9, 2.8
NWQL — old method	<0.40, <1.8, E0.24, E0.38, 0.73
NWQL — new method	E0.12, 0.29, 0.68, 0.89, 1.5

The data user is interested in reporting some basic descriptive statistics such as medians and percentiles, and would like to obtain an estimate of the mean. How is this done with such a complicated data set?

Analysis: Before combining data from different sources, the data user must determine if the different methods are yielding comparable results. Split spike samples should have been sent to both the local lab and NWQL. For the purpose of this example, it will be assumed that adequate QA has shown that no significant differences exist between results from the different laboratories and methods.

The next task is to make sure that the "less than" values in the data set are censored appropriately (see Data Interpretation Example 1). The data from the local lab is censored at only one level and no values are reported less than the censoring level, so no modifications are necessary. Nondetections from NWQL are reported using an LRL to reduce the chance of false negatives to 1%. The actual signal observed by the lab for nondetections, however, was less than the LT-MDL and that should be used as the censoring value here. Therefore, the reported value of <0.40 will be used as <0.20 and <1.8 will be used as <0.9 (assuming 100% recovery).

This results in a combined data set with three censoring levels (0.2, 0.5 and 0.9) and containing reported values less than all three of these censoring levels. The simplest approach for a data set such as this is to recensor it at the highest "less than" value, in this case 0.9. That is not a good option here, however, because too much information would be lost.

Calculating summary statistics for data sets that contain different censoring points requires special methods. Simple ranking methods are not applicable because order is ambiguous. Two methods that can handle multiple censoring levels with or without reported values less than the censoring levels are Kaplan-Meier estimation and regres-

sion on order statistics (ROS or probability plotting). Both will be shown for this example. Maximum likelihood estimation (MLE) also is acceptable, but an assumed distribution is required and it is best used when the data set contains at least 50 points.

The Kaplan-Meier method involves no assumptions about the underlying distribution and handles multiple censoring points. It produces percentiles (and their standard deviations) as well as estimates of the distribution's mean and standard error. It can be expanded to methods that allow comparison of two or more groups (Helsel, 2005; SAS Institute, 1990).

The major drawback of the Kaplan-Meier method occurs when the smallest value in the data set is censored. In this case, the estimate of the mean will be biased because the value of the smallest censoring level is substituted for the lowest nondetects in the calculation of the mean. (When there is only one censoring level and all values less than the censoring level are reported as "less thans," the Kaplan-Meier method is equivalent to using simple substitution.) The bias increases as the size of the data set decreases and as the number of censored values at the lowest censoring level increases.

The Kaplan-Meier method was developed for survival analysis involving data that are right censored ("greater than" rather than "less than" censoring). Although equations can be developed for left-censored data, texts and software using the Kaplan-Meier method routinely are limited to right-censored data. Fortunately, it is easy to transform left-censored data into right-censored data (and back), and that transformation will be shown in this example. Detailed calculations for this example are shown in the boxes on the following pages.

The robust ROS method requires the assumption of an underlying data distribution (usually log-normal), but is not fully parametric, because summary statistics are calculated using fill-in values combined with the noncensored data. ROS was shown for a data set with a single censoring point in Data Interpretation Example 4. The current example shows its application to data sets with multiple censoring points. Detailed calculations for this example are shown in the boxes on the following pages.

Summary statistics obtained from each method are shown in the table below.

	Kaplan Meier Method	Robust ROS Method
10th percentile	0.12	0.13
25th percentile	0.24	0.25
median	0.60	0.49
75th percentile	1.00	0.98
90th percentile	1.90	1.86
mean	0.74	0.74
standard deviation	0.71	0.69

Kaplan-Meier Method —Data Interpretation Example 7 ***(page 1 of 3)***

A spreadsheet was used for the Kaplan-Meier method in this example. For a detailed discussion of the theory involved in this method, see Helsel (2005) and SAS (1990).

1. Transform the data by "flipping" each value.

- List all values (including the censored ones) in order from largest to smallest. Include a column for the code for censored values. Do not list any value more than once. Include a column for the number of observed values which will account for multiple incidences of the same value. (*See* ***Value***, ***Code***, *and* ***#Obs*** *columns*.) An index column (***i***) also is included for clarity in the example spreadsheet, but is not required for the calculations.
- Pick an arbitrary value greater than the largest data value. In this case, a value of 3 was used. "Flip" the data by subtracting each data point from this arbitrary value. (*See* ***Flip*** *column*.) This column is sublabelled "t_i" to clarify later calculations. An additional column (***Flip Code***) was included here to show that operation of flipping reverses the direction of the comparison operator; this column is not needed for the calculations. (The flipped data, t_i, are equivalent to the time periods in survival analysis.)
- From this point onward in the calculation, the flipped values are used.

2. Calculate the Incremental Probability— the probability of surviving to the next index level.

- Determine the number of values in the cohort at level i (number at risk). This is the number of flipped values greater than or equal to the flip value at index i. Be sure to account for values with multiple observations. (*See* ***n_i*** *column*.)
- Determine the number of values in the cohort at level i that will not survive to the level i+1 (number of deaths). This is the change in the number of flipped values ($n_i - n_{i+1}$), not counting right-censored values (>), which represent subjects that leave the survival analysis without dying. Be sure to account for values with multiple observations. (*See* ***d_i*** *column*.)
- Calculate the number of values in the cohort that survive to level i+1 as $n_i - d_i$. (*See* ***s_i*** *column*.)
- Calculate the incremental probability as s_i/n_i for each level i. (*See* ***Incremental Probability*** *column, which shows the fraction and the decimal result for each row.*)

3. Calculate the Cumulative Probability— the probability of surviving through the flipped value i—which also is known as the Survival Function (S).

- The Cumulative Probability at level i is the product all the incremental probabilities, up to and including level i as shown by the equation below. (*See* ***CumP*** *or* ***S_i*** *column*.) This is based on the principles of conditional probability.

$$S_i = \prod_{i=1}^{\kappa} \frac{s_i}{n_i}$$

4. The standard error of the Cumulative Probability can be calculated by the formula

$$\text{Standard Error of } S_i = S_i \cdot \sqrt{\sum_{j=1}^{i} \frac{d_j}{n_j \cdot s_j}}$$

(See the three columns with the spanning head of ***Calculate Standard Error PctIls****.)*

5. Plot the survival function and untransform the data to obtain percentiles of the original data.

- The Cumulative Probability is plotted as a step function of Flip and shown at the left below the spreadsheet.
- Cumulative Probability is plotted as a step function of the Original Values and shown at the right below the spreadsheet. Percentiles can be obtained directly from the graph as shown and are shaded on the spreadsheet. Note that this is a different method of calculating percentiles than described in Data Interpretation Example 3.

6. Calculate the mean.

- The mean of Flip is the area under the Cumulative Probability curve. This is most easily done by calculating the areas of individual rectangles and then adding them. The formula is

$$\text{mean Flip} = \sum_{i=0}^{k-1} S_i \cdot (t_{i+1} - t_i)$$

- The mean of the original data is calculated by subtracting the mean of Flip from the arbitrary value used to flip the data in step 1. In this case:

  mean original = 3 – 2.262 = 0.738

7. Calculate the variance and standard error of the mean and the standard deviation of the distribution (which are the same for Flip as for the original data).

- The formula for the variance of the mean is:

$$\text{Variance Mean} = \frac{m}{m-1} \sum_{i=1}^{k-1} A_i^{\,2} \cdot \frac{d_i}{n_i \cdot s_i} \quad \text{where} \quad A_i = \sum_{j=i}^{k} S_j \cdot (t_{j+1} - t_j)$$

  where *k* is the number of unique values (maximum of the ***i*** column), and *m* is the number of detections (sum of the d_i column).

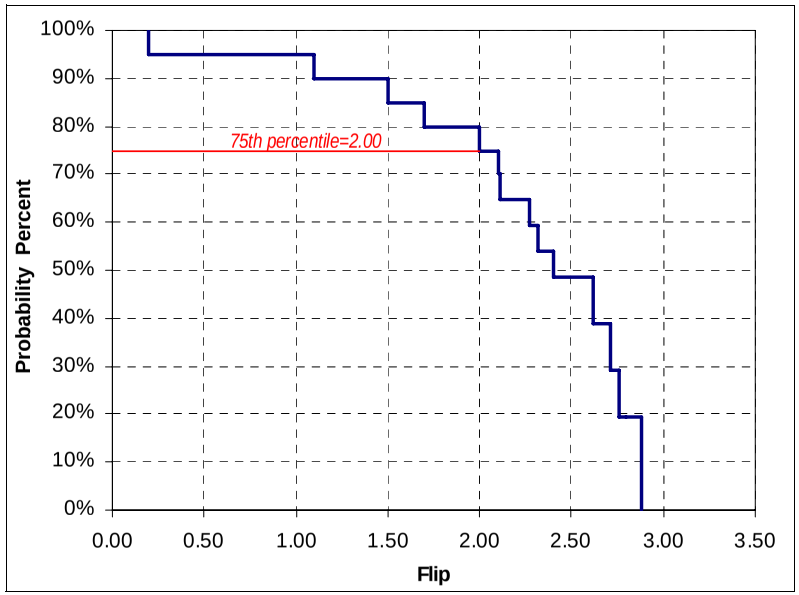
  (See the two columns with the spanning head of ***Calculate SE Mean****.)*
- The standard error of the mean is the square root of the variance of the mean.
- The standard deviation of the distribution is the standard error of the mean times the square root of *n*, the number of total observations (sum of the ***# Obs*** column).

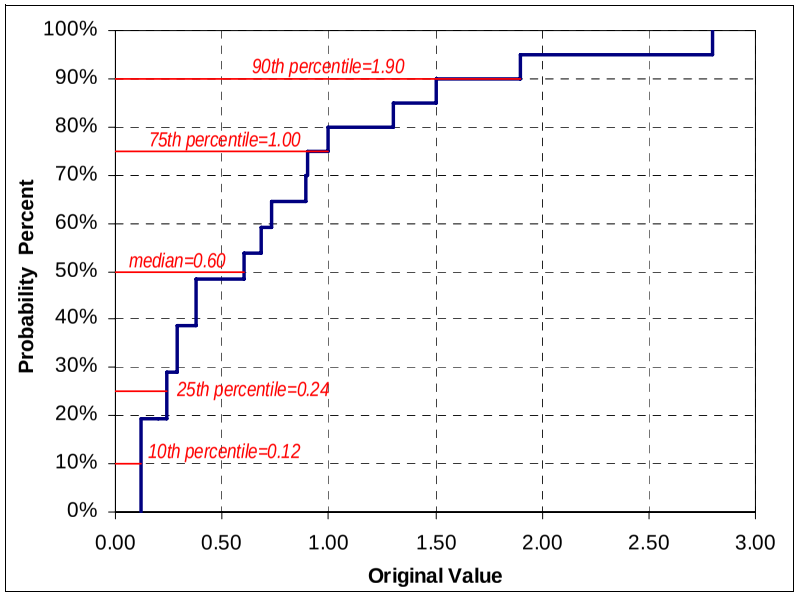
  Standard deviation = 0.259 √20 = 0.717

i	Value	Code	# Obs	Flip t_i	Flip Code	# >= Flip n_i	# > Flip d_i	# Survive s_i	Incremental Probability fraction= decimal	Cum P(>Flip) $S(t_i)$	------ Calculate Standard Error Pctls ------ factor $d_i / (n_i * s_i)$	sumfactor Σ 0 to i	Std Error $S(t_i) * SQRT(sumfact_i)$	Calculate Mean rect area $S(t_i) * (t_{i+1} - t_i)$	--- Calculate SE Mean --- A_i Σ i to k-1	Sq Sum Term $A_i^2 * d_i / (n_i * s_i)$
0				0.00						100.0%				0.200		
1	2.80		1	0.20		20	1	19	19/20= 0.9500	95.0%	0.00263	0.00263	4.9%	0.855	2.062	1.12E-02
2	1.90		1	1.10		19	1	18	18/19= 0.9474	90.0%	0.00292	0.00556	6.7%	0.360	1.207	4.26E-03
3	1.50		1	1.50		18	1	17	17/18= 0.9444	85.0%	0.00327	0.00882	8.0%	0.170	0.847	2.35E-03
4	1.30		1	1.70		17	1	16	16/17= 0.9412	80.0%	0.00368	0.01250	8.9%	0.240	0.677	1.69E-03
5	1.00		1	2.00		16	1	15	15/16= 0.9375	75.0%	0.00417	0.01667	9.7%	0.075	0.437	7.97E-04
6	0.90		1	2.10		15	1	14	14/15= 0.9333	70.0%	0.00476	0.02143	10.2%	0.000	0.362	6.25E-04
7	0.90	<	1	2.10	>	14	0	14	14/14= 1.0000	70.0%	0.00000	0.02143	10.2%	0.007	0.362	0.00E+00
8	0.89		1	2.11		13	1	12	12/13= 0.9231	64.6%	0.00641	0.02784	10.8%	0.103	0.355	8.10E-04
9	0.73		1	2.27		12	1	11	11/12= 0.9167	59.2%	0.00758	0.03541	11.1%	0.030	0.252	4.81E-04
10	0.68		1	2.32		11	1	10	10/11= 0.9091	53.8%	0.00909	0.04451	11.4%	0.043	0.222	4.50E-04
11	0.60		1	2.40		10	1	9	9/10= 0.9000	48.5%	0.01111	0.05562	11.4%	0.048	0.179	3.57E-04
12	0.50	<	4	2.50	>	9	0	9	9/9= 1.0000	48.5%	0.00000	0.05562	11.4%	0.058	0.131	0.00E+00
13	0.38		1	2.62		5	1	4	4/5= 0.8000	38.8%	0.05000	0.10562	12.6%	0.035	0.073	2.64E-04
14	0.29		1	2.71		4	1	3	3/4= 0.7500	29.1%	0.08333	0.18895	12.6%	0.015	0.038	1.19E-04
15	0.24		1	2.76		3	1	2	2/3= 0.6667	19.4%	0.16667	0.35562	11.6%	0.008	0.023	9.02E-05
16	0.20	<	1	2.80	>	2	0	2	2/2= 1.0000	19.4%	0.00000	0.35562	11.6%	0.016	0.016	0.00E+00
17	0.12		1	2.88		1	1	0	0/1= 0.0000	0.0%						

k= 17 n= 20 m= 14

Mean (flip)= 2.262 Var Mean= 0.0253
Mean (original)= 0.738 Std Error= 0.159
Std Dev Dist= 0.711





ROS Method —Data Interpretation Example 7

A spreadsheet was used for the robust ROS method in this example. For a detailed discussion of the theory involved in this method, see Helsel (2005).

Part 1—The noncensored values are used to obtain a regression equation. A log-normal distribution is assumed.

1. Compute normal scores for the detected values.

- List all noncensored values in order from largest to smallest. Subdivide the data into intervals bounded by the censoring values. In this example there are 4 intervals. (*See* ***Detected Data*** *column on next page*.)
- Calculate the probability of detection in each interval.

  For the interval whose lower bound is the highest censoring level (0.9 in this example), this is the number of detections greater than the highest censoring level /total number:

  $P(\geq 0.9) = 6/20 = 0.300$

  The presence of nondetections complicates the calculation of the detection probability in the other intervals. For each of these, the data in the intervals greater than the one of interest and the nondetections at the upper bound of the interval of interest are removed. This yields a conditional probability of detection. In this example, for the second interval it is the probability that a detection is greater than or equal to 0.5, <u>*given that*</u> it is less than 0.9 (P(≥0.5| <0.9)). When this conditional probability is multiplied by the probability of the given condition (P(<0.9) here), the result is the probability of detection in the interval (P(≥0.5 and <0.9)).

  $P(\geq 0.5 \text{ and } < 0.9) = P(\geq 0.5 \mid < 0.9) \cdot P(< 0.9) = 4/13 \cdot 0.700 = 0.215$

  where $P(< 0.9) = 1 - P(\geq 0.9) = 1 - 0.300 = 0.700$

  The calculations for the other sections are shown below.

  $P(\geq 0.2 \text{ and } < 0.5) = P(\geq 0.2 \mid < 0.5) \cdot P(< 0.5) = 3/5 \cdot 0.485 = 0.291$

  where $P(< 0.5) = 1 - [P(\geq 0.9) + P(\geq 0.5 \text{ and } < 0.9)] = 1 - 0.300 - 0.215 = 0.485$

  $P(\geq 0 \text{ and } < 0.2) = P(\geq 0 \mid < 0.2) \cdot P(< 0.2) = 1/1 \cdot 0.194 = 0.194$

  where $P(< 0.2) = 1 - 0.300 - 0.215 - 0.291 = 0.194$
- Calculate the probability level increment for each interval as:

  P(detection in the interval)/(number of detections in the interval + 1).

  Increment for $0.9 - \infty = 0.300/7 = 0.0427$
  Increment for $0.5 - 0.9 = 0.215/5 = 0.0431$
  Increment for $0.2 - 0.5 = 0.291/4 = 0.0727$
  Increment for $0 - 0.2 = 0.194/2 = 0.0969$

- Assign probability levels to the detected and the censoring values. (*See* ***Prob Level*** *column.*)

  For the highest data value: Prob Level = 1 – Increment for top interval

  For subsequent values in top interval: Prob Level = Previous Prob Level – Increment

  For the first censoring value: Prob Level = Previous Prob Level – Increment

  For the highest data value in the next interval:

  Prob Level = Prob Level of Censoring Value – Increment for new interval

  Continue in this fashion, changing the increment at each interval bounded by the censoring boundaries.

  Check: The Prob Level for the censoring values should differ by the probabilities of detection in each interval. In this example, the probability at 0.9 is 0.700 and the probability at 0.5 is 0.485. The difference of these is 0.215 which is the detection probability of the interval ≥0.5 and <0.9.

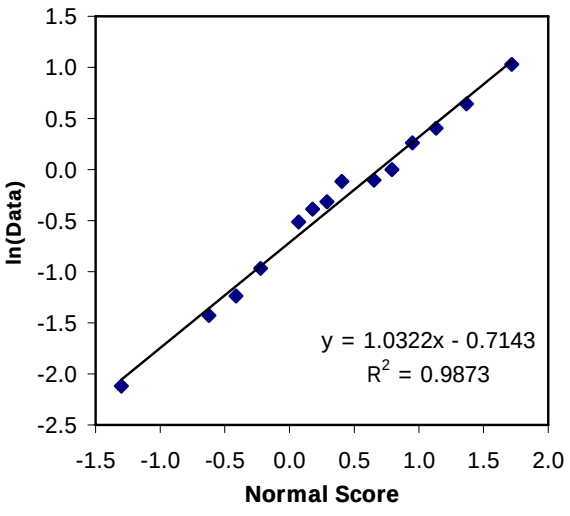
- Calculate the normal score (z-score) associated with each probability level by using the inverse of the normal distribution function. (*See* ***Normal Score*** *column.*)

2. Take the natural logarithm of the detected values. (*See* ***ln(Data)*** *column.*)

3. Graph the ln(Data) versus the Normal Score and calculate the equation of the regression line.

The portion of the spreadsheet resulting from these steps is shown below.

Censor Value	# Cnsrd	Detected Data	Prob Level	Normal Score	ln(Data)
		2.8	0.957	1.718	1.03
		1.900	0.914	1.368	0.64
		1.500	0.871	1.133	0.41
		1.300	0.829	0.949	0.26
		1.000	0.786	0.792	0.00
		0.900	0.743	0.652	-0.11
0.9	1		0.700		
		0.890	0.657	0.404	-0.12
		0.730	0.614	0.289	-0.31
		0.680	0.571	0.178	-0.39
		0.600	0.528	0.069	-0.51
0.5	4		0.485		
		0.380	0.412	-0.223	-0.97
		0.290	0.339	-0.415	-1.24
		0.240	0.267	-0.623	-1.43
0.2	1		0.194		
		0.120	0.097	-1.299	-2.12



Part 2—The regression equation is used to calculate fill-in values to represent the censored values in the calculation of summary statistics.

1. Compute normal scores for the censored values.

- Recall the probabilities associated with each censoring level.
  P(<0.9) = 0.700 *P(<0.5) = 0.485* *P(<0.2) = 0.194*
- Calculate probability level increments as: P(nondetection)/(number of nondetections + 1)
  @ 0.9: Increment = 0.700/2 = 0.350
  @ 0.5: Increment = 0.485/5 = 0.097
  @ 0.2: Increment = 0.194/2 = 0.097
- Create a list of probability levels for censored values. (*See* ***Censored Prob*** *column,* below.)
  Beginning with: Censored Prob = P(nondetection) – Increment
  For subsequent values: Censored Prob = Previous Censored Prob – Increment
- Calculate the normal score (z-score) associated with each probability level by using the inverse of the normal distribution function. (*See* ***Censored Norm*** *column.*)

2. Use the regression equation from Part 1 to calculate the natural logarithms of the fill-in values. (*See* ***ln(Fill-in)*** *column*).

3. Calculate the fill-in data values from their natural logarithms. *(See* ***Fill-in Values*** *column.)*.

The portion of the spreadsheet resulting from these steps is shown at the right.

Censor Level	ND Index	Censored Prob	Censored Norm	ln (Fill-in)	Fill-in Values
0.9	1	0.350	-0.385	-1.112	0.329
0.5	1	0.388	-0.285	-1.009	0.365
0.5	2	0.291	-0.551	-1.283	0.277
0.5	3	0.194	-0.864	-1.606	0.201
0.5	4	0.097	-1.299	-2.055	0.128
0.2	1	0.097	-1.299	-2.055	0.128

Part 3—The noncensored values (all 14 values in the ***Detected Data*** column on the previous page) and the fill-in values (all 6 values in the ***Fill-in Values*** column at the right) are combined and used to calculate summary statistics. The reader can verify the summary statistics calculated from this hybrid data set. Note that fill-in values do not correspond to any specific samples and should never be reported as data.

Data Interpretation Example 8

How do I group my data using one or more cutoff or benchmark values?

Description: A data user would like to compare groups within a data set or develop a model that relates constituent concentration to ancillary factors such as population density, geological substrate, or land use. A large proportion of the data set, however, is comprised of nondetections and detections below the LRL. How does the data user handle this situation?

Analysis: The data user has a variety of options that are all based on the same general approach—the data user picks a "cutoff" value or values and transforms the data into a categorical response. Then, an appropriate statistical method such as chi-square analysis or logit regression is used to analyze the categorical data. Several options for the cutoff value are described below. In all cases, it is important that the data user document what was used as the cutoff value.

Option 1: The data user picks the LT-MDL as the "cutoff" value. This is the common method of using detection to categorize data. For a method that is not information-rich, this method is simple. All values that are greater than or equal to the LT-MDL are considered detections, and results reported as <LRL are considered nondetections.

Data from information-rich methods have the added challenge that detections less than the LT-MDL are reported. These values could be grouped with the nondetections using the LT-MDL as the cutoff value. Alternatively, they could be grouped with the detections if the data user had good results on low-level QC spikes. In this concentration range, there is a high risk of false negatives and false positives, and results are likely to be influenced greatly by matrix effects, sample handling and instrument performance on a given day.

The main advantage of this method is its simplicity. A disadvantage of this method is that grouping all detections together (low concentrations with high concentrations) may obscure too much information. Perhaps a more important disadvantage is that using the LT-MDL as the cutoff value allows a characteristic of a particular analytical method to determine the interpretation of environmental data. Had the samples been analyzed using a different chemical assay, the categorical grouping probably would be different.

Option 2: The data user divides the data into three groups: not-detected, low-level detection, and quantifiable. A likely candidate for the upper cutoff value is the LRL. Values greater than or equal to this value are considered quantifiable. Depending on the needs of the data user, a higher value might be used for the upper cutoff. Some groups, such as the American Chemical Society (Keith and others, 1983) advocate using 10 times the standard deviation (about four times the LT-MDL) as a cutoff limit for quantification. The lower cutoff could be the LT-MDL. In this case, the low-level detection category would include all values reported between the LT-MDL and LRL. Data users might also choose to include detections that are less than the LT-MDL (information-rich methods) in

the category of low-level detection. This method has the advantage of placing detections with greater relative uncertainty into a separate category. It still has the disadvantage of allowing the analytical method to determine the data interpretation.

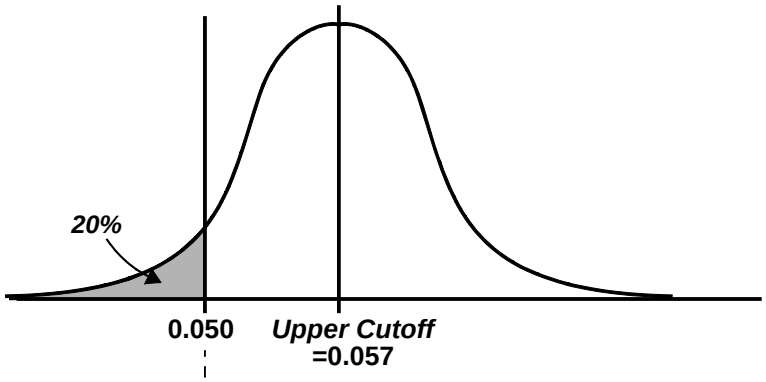
Option 3: The cutoff value chosen by the data user is a benchmark that has meaning relative to the purpose of the study. Examples are as varied as study objectives, but some possible benchmarks include a criterion for the protection of aquatic life, a maximum contaminant level for drinking water, or a reported average background value. Data then would be classified relative to the benchmark. Choosing a cutoff value in this way decouples the interpretation of data from the chemical analytical method. Applying this method is straightforward, provided that the benchmark is greater than or equal to the LRL. If the benchmark is less than the LRL, see Data Interpretation Example 9.

Option 4: The data user divides the data into three groups based on a benchmark and the confidence levels required for exceedence or nonexceedence. The categories would be (1) greater than the benchmark with "*x*"% certainty, (2) less than the benchmark with "y"% certainty, and (3) too close to call. In this case, the data user must specify the certainties required (which could be the same) and calculate cutoff values based on the benchmark and the performance of the chemical method. This approach requires the most work on the part of the data user, but also provides the most meaningful results.

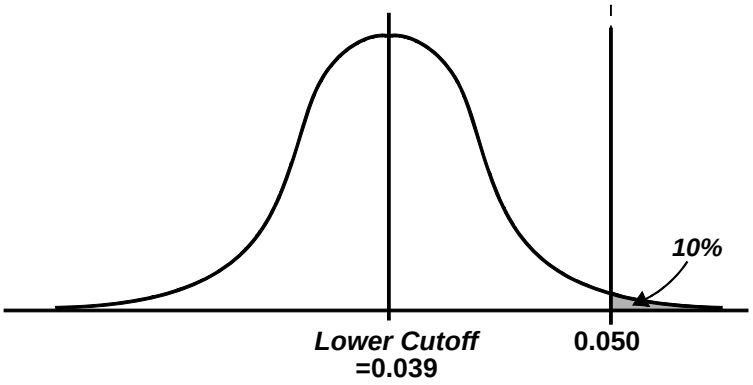
To determine the cutoff values, the data user must have an estimate of the standard deviation of the analytical method in the vicinity of the criterion of interest. The standard deviation could be obtained three different ways depending on the amount of data available and on where the criterion of interest falls within the range of the analytical method. If all samples were analyzed as replicates, the data user could calculate a standard deviation from the pooled replicates. Typically, this will not be the case and the data user will have to rely on statistics from the LT-MDL or the Blind Sample Program. In this case, the data user first must ensure that the analytical method seems to be behaving properly (see Quality Assurance Examples 1–8). If the criterion of interest is in the low range of the analytical method (near the LRL), the standard deviation from the LT-MDL can be used. If the criterion of interest is outside of the low range of the method, then the standard deviation for the appropriate range can be obtained from the Blind Sample Program's data.

For this example, suppose that the LT-MDL for the chemical analysis is 0.02 (based on a standard deviation of 0.008 with 19 degrees of freedom). The data user wants to compare values to an aquatic life criterion of 0.050 with the following certainties specified: the upper cutoff identifies values that have no more than a 20% chance of being reported as less than 0.050, and the lower cutoff identifies values that have no more than a 10% chance of being reported as greater than 0.050. Calculation of each cutoff value is shown on the following page with the appropriate illustration. (Note that this example assumes no bias in the analysis and 100% recovery.)

The value of t for a 1-tail test with α=20% and 19 degrees of freedom is -0.861. Substituting the t-value, criterion, and standard deviation into the equation for t yields:



$$-0.861 = \frac{0.050 - UpperCutoff}{0.008}$$



The value of t for a 1-tail test with α=10% and 19 degrees of freedom is 1.328. Substituting the t-value, criterion, and standard deviation into the equation for t yields:

$$1.328 = \frac{0.050 - LowerCutoff}{0.008}$$

Option 5: One or more cutoff values are chosen based on observable breaks in the data. In this case, the data user plots the data as a cumulative distribution function and examines the plot for natural breaks. Such breaks may represent observed conditions or an underlying explanatory variable and may point the way to future work. Once the break points are determined, they can be applied in the same way that cutoff values are described in options 3 or 4.

Summary: An example data set is shown in the table below with categorizations based on the different options just described. An information-rich analytical method was used. The LT-MDL value is 0.020 and the LRL is 0.040.

Data as reported from NWQL	**Option 1** (cutoff = LT-MDL)	**Option 2** (cutoff = LRL)	**Option 3** (cutoff = benchmark)	**Option 4** (cutoff = benchmark)
<0.040	Below LT-MDL	Not Detected	Below 0.050	Below 0.050, ≤10% error
<0.040	Below LT-MDL	Not Detected	Below 0.050	Below 0.050, ≤10% error
E0.006	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.006	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.007	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.008	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.012	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.012	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.013	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.015	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.016	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.017	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.017	Below LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.023	Exceeds LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.037	Exceeds LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.037	Exceeds LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.038	Exceeds LT-MDL	Low-Level Detection	Below 0.050	Below 0.050, ≤10% error
E0.039	Exceeds LT-MDL	Low-Level Detection	Below 0.050	Too close to call
0.045	Exceeds LT-MDL	Exceeds LRL	Below 0.050	Too close to call
0.048	Exceeds LT-MDL	Exceeds LRL	Below 0.050	Too close to call
0.052	Exceeds LT-MDL	Exceeds LRL	Exceeds 0.050	Too close to call
0.058	Exceeds LT-MDL	Exceeds LRL	Exceeds 0.050	Exceeds 0.050, ≤20% error
0.060	Exceeds LT-MDL	Exceeds LRL	Exceeds 0.050	Exceeds 0.050, ≤20% error
0.071	Exceeds LT-MDL	Exceeds LRL	Exceeds 0.050	Exceeds 0.050, ≤20% error
0.082	Exceeds LT-MDL	Exceeds LRL	Exceeds 0.050	Exceeds 0.050, ≤20% error
0.090	Exceeds LT-MDL	Exceeds LRL	Exceeds 0.050	Exceeds 0.050, ≤20% error
0.110	Exceeds LT-MDL	Exceeds LRL	Exceeds 0.050	Exceeds 0.050, ≤20% error
0.134	Exceeds LT-MDL	Exceeds LRL	Exceeds 0.050	Exceeds 0.050, ≤20% error
0.138	Exceeds LT-MDL	Exceeds LRL	Exceeds 0.050	Exceeds 0.050, ≤20% error
0.175	Exceeds LT-MDL	Exceeds LRL	Exceeds 0.050	Exceeds 0.050, ≤20% error

Data Interpretation Example 9

How do I compare my data to a value that is less than the LRL?

Description: In some cases the aquatic criterion, health guideline, or comparison value of interest is less than the LRL. This becomes a problem in data analysis when the data user tries to determine if a value that is reported as <LRL does or does not exceed this comparison value. For example, suppose that the data user wants to compare data to a criterion of 0.030. The LT-MDL for the method is 0.025 and the LRL is 0.050.

Analysis: A result that is reported as "<LRL" does not mean that the laboratory quantitatively measured the concentration and found it to be less than the LRL. The LRL is simply the default reporting value that the laboratory uses for samples that don't meet minimum detection criteria. For a typical method, < LRL will be reported for samples in which the laboratory found a concentration less than the LT-MDL. For an information-rich method, it means that the laboratory did not have sufficient identification information to report that the analyte was present. When the laboratory reports nondetections as "<LRL," they assumed a maximum false negative error rate of 1%, meaning that the chance of reporting a nondetection is only 1% when the sample's true concentration equals the LRL. That level of certainty may or may not meet your needs as a data user.

In general, values near or less than the LRL will have considerable uncertainty. This means that the data user must be aware that comparing values in this range will have a relatively large chance of error. The table below gives approximate probabilities and relative standard deviations for values in this range. Values in this range may be too uncertain to meet the needs of some studies. The probabilities given in the table were calculated assuming that the method is behaving properly; they do not apply if this assumption is invalid. It is very important to have adequate QA, including blanks and low-level spikes (Quality Assurance Examples 4–8), before trying to interpret values in this range of a method. To reduce uncertainty, the data user can change to a more sensitive analytical method or submit replicate samples for analysis (Project Planning Example 1).

Actual Concentration	**Chance of Being Reported As**			**Approximate Relative Standard Deviation**
	<LRL (typical method) **Value < LT-MDL** (info-rich method)	**≥LT-MDL and <LRL**	**≥LRL**	
0	99%	1%	0.0002%	—
$^1/_3$ LT-MDL	94%	6%	0.005%	130%
$^2/_3$ LT-MDL	78%	22%	0.1%	64%
LT-MDL	50%	49%	1%	43%
$^4/_3$ LT-MDL	22%	72%	6%	32%
$^5/_3$ LT-MDL	6%	72%	22%	26%
2 LT-MDL= LRL	1%	49%	50%	22%

Case 1—Typical analytical method: The reported data values were <0.050, <0.050, <0.050, E0.040, and 0.080. Based solely on these individual measurements, two values are greater than the comparison criterion of 0.030 (0.040 and 0.080) and three values are less than 0.030 (the <0.050 values). The values reported as <0.050 corresponded to an analytical signal that was less than the LT-MDL (0.025), and therefore could be considered less than 0.030, but with considerable uncertainty. The data user can get a rough estimate of the potential for error by using the table above. A sample with a true value of 0.030 (a value slightly greater than the LT-MDL) has a chance between 22 and 50% of being reported as <0.050. Similarly, the data user can estimate a confidence interval around the reported concentration of 0.040 using the approximate relative standard deviation. This estimated value (0.040) is 1.6 times the LT-MDL and the approximate relative standard deviation is about 26%. Given that level of uncertainty, it is possible that the true concentration of this sample could be less than the comparison criterion of 0.030. If the data user needs to be confident that a concentration truly does exceed 0.030 (for example, a regulatory statute), a value of 0.040 with this uncertainty may not be adequate.

Case 2—Information-rich analytical method: The reported data values were <0.050, <0.050, E0.010, E0.040.and 0.080. Based solely on these individual measurements, two values are greater than 0.030 (0.040 and 0.080) and three values are less than 0.030 (E0.010 and the <0.050 values). Because this is an information-rich method, the <0.050 values indicate that the NWQL saw insufficient evidence of the analyte's presence in those samples. This could occur because the analyte was actually not present in the sample or if a matrix effect decreased or masked the signal. Information from blanks and low-level spikes is vital here. The table above does not provide information about the probability of nondetection for information-rich methods. The table is based on the analytical method's variability as measured by the standard deviation of low level results and does not take into account the confirming detection used in information-rich methods. The additional evidence from an information-rich method should decrease the incidence of false negatives and false positives, but that decrease has not been statistically defined. For the E0.010 value, the presence of analyte was confirmed; however, the concentration has considerable uncertainty—a relative standard deviation on the order of 100% (from the table). In addition, the risk of a false positive result is high at this concentration (due to potential carry over from a previous analysis). Replicate values are needed to limit the uncertainty in this region. The data user also could consider using the method of standard additions (described in analytical chemistry texts, for example, Miller and Miller, 1988) to increase precision in this range.

Comparisons to values below the LT-MDL generally should be avoided. The uncertainty in this range is too great. Even using information-rich methods and applying techniques such as the method of standard additions or replicate samples, it is unlikely that the data user would obtain adequate certainty for most comparison purposes.

ACKNOWLEDGMENTS

The author thanks the following people for their contributions to this work: Bill Foreman for his insight into LT-MDL calculations and his thoughtful review; the Office of Water Quality for lab-related performance data; and Chauncey Anderson, Jim Eychaner, Jeff Martin, Jennifer Morace, Callie Oblinger, and Stewart Rounds for their comments and reviews of the manuscript.

ANNOTATED BIBLIOGRAPHY

This bibliography includes the citations found in this document and some additional references. All include a a brief description of the resource.

Childress, C.J.O., Foreman, W.T., Connor, B.F., Maloney, T.J., 1999, New reporting procedures based on long-term method detection levels and some considerations for interpretations of water-quality data provided by the U.S. Geological Survey National Water Quality Laboratory: U.S. Geological Survey Open-File Report 99-193, 19 p.

— *This report describes the rationale and implementation of the USGS LT-MDL procedure.*

Helsel, D.R., 2005, Nondetects and data analysis: Wiley-Interscience, 250 p.

— *This text describes statistical methods that can be applied to data sets containing multiply-censored data. It particularly addresses the sort of problems that can be encountered when interpreting the LT-MDL and LRL. The application of the Robust Regression on Order Statistics Method and the Kaplan-Meier method are shown.*

Helsel, D.R., and Hirsch, R.M., 1992, Statistical methods in water resources: Elsevier, 522 p. Also available online as USGS Techniques of Water-Resources Investigations, Book 4, Chapter A3 (2002): http://pubs.usgs.gov/twri/twri4a3/, accessed April 4, 2008.

— *This text that describes nonparametric and exploratory data analysis methods using environmental examples.*

Kimbrough, D.E., and Wakakuwa, J., 1994, Quality control level—An alternative to detection levels: *Environmental Science and Technology,* v. 28, p. 338-345.

— *This paper describes of a method to determine a laboratory censoring level (the lowest reported concentration) that is based on both accuracy and precision.*

Keith, L.H., Crummett, W., Deegan, J.Jr., Libby, R.A., Taylor, J.K., and Wentler, G., 1983, Principles of environmental analysis: *Analytical Chemistry,* v. 55, p. 2210–2218.

— *This paper describes the American Chemical Society guidelines for the evaluation of environmental chemical data.*

Miller, J.C, and Miller J.N.,1988, Statistics for analytical chemistry, 2nd ed.: John Wiley and Sons, 227 p.

— *This text describes basic statistical methods applied to chemical results. It includes a discussion of detection limits and the method of standard additions.*

Phillips, D.S., 1978, Basic statistics for health science students, W.H. Freeman and Company, 185 p.
— *This very basic text includes discussion and worked examples of many nonparametric tests (including Chi square, medians, signs, and Wilcoxen tests, among others) as well as the typical parametric tests. It also includes exact tables for many of the nonparametric tests for small datasets.*

SAS Institute, 1990, SAS Institute Inc.
SAS procedures guide, Version 6, 3rd ed., 705 p. *and*
SAS/STAT user's guide, Version 6, 4th ed. Volume 1, 943 p. *and*
SAS/STAT user's guide, Version 6, 4th ed., Volume 2, 846 p.
— *These manuals include the applicable equations and descriptions of many techniques, including several percentile methods and Kaplan-Meier Estimation in survival analysis. The SAS manuals are particularly useful because they contain the equations that are used for most techniques.*

She, N., 1997, Analyzing censored water quality data using a non-parametric approach: Journal of the American Water Resources Association, v. 33, p. 615-624.
— *This paper shows the application of the Kaplan-Meier method to the calculation of summary statistics for censored water quality, including how to "flip data" and a comparison of the results of this method to those from other methods.*

U.S. Environmental Protection Agency, 1998, Guidance for data quality assessment, Practical methods for data analysis EPA QA/G-9: EPA/600/R-96/084.
— *This report describes procedures used by the USEPA for performing statistical analyses of dataset that contain censored data. It was designed as a "tool box" of techniques for reviewing data, selecting appropriate statistical tests and verifying statistical assumptions. Both parametric and nonparametric tests are included.*