Extracting Correct Information From Censored Environmental Data*

Richard B. Shepard, PhD†

1 Introduction

Data are just numbers and cannot be used without conversion to information which then is interpreted to yield knowledge. Statistical models are the tools to convert raw data to information. Expertise and experience convert information to knowledge and insight which can be used by decision-makers. This white paper describes the correct conversion of environmental chemical data to information through the application of statistical models that properly incorporate observations below laboratory method detection limits.

Statistics courses for science and business students tend to use well-behaved, normally distributed data and the parametric models that work with them. Unfortunately, environmental chemistry data are usually quite different from the well-behaved examples taught in academic courses.

Environmental chemistry data are almost never normally distributed; at least, not with the constituents with which regulators and the public are most concerned. Toxic metal and organic compound data sets tend to be skewed with high concentration outliers on the right tail. No chemical concentration can be less than zero while the left tail of a normal distribution with most observations at low values can be less than zero. Toxic constituents (inorganic metals and organic compounds) frequently have concentrations less than the detection limits of the instruments used to measure them. Such "less-than" values are also called "nondetects" or "censored".

These censored concentrations are known only to be somewhere between zero and the laboratory's reporting level (RL). Over the past 20 years the fields of medical survival analysis and industrial reliability analysis have greatly improved methods for analyzing censored data–those observations reported only as being above or below a threshold value. Today, methods for correctly summarizing, testing hypothesis, and determining strengths of association and cause-and-effect using data sets with censored data are available. Methods developed for the "greater-thans" (right-censored data) in medical and industrial studies can also be applied to the "less-thans" (left-censored data) of low-level environmental concentrations. Unfortunately, regulators of the environmental community have generally not incorporated these procedures or required their use when permit holders submit results of compliance or baseline monitoring.

These data, and the statistical analyses to use with them, apply to all environmental chemistry data: air, biological tissue, sediment, soil, and water.

The five topics covered in this white paper are:

1. How to describe data.
2. Definition of censored data.
3. The wrong ways censored data have been handled.
4. How to correctly analyze data sets containing censored data.
5. Why you should care.

†President, Applied Ecosystem Services, Inc.

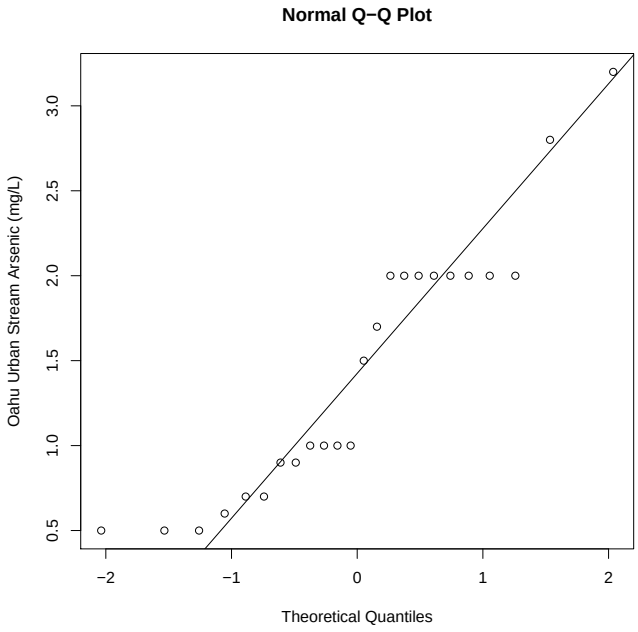

Figure 1: Q-Q plot of arsenic concentrations in streams on Oahu, Hawaii.

2 Information from data

2.1 Exploratory data analysis

The first step to processing any environmental data set is to conduct exploratory data analysis (EDA). Statisticians and environmental data analysts use graphical methods rather than tables of numbers to examine the data set as a whole. Four questions EDA answers are:

1. Are the variables normally distributed?
2. Do we need to transform the data?
3. Are their outliers?
4. What are the moments of the data distribution:
   (a) Where are the data centered?
   (b) How are they spread?
   (c) Are they symmetric, skewed, bimodal?

To determine if the data are normally distributed use a normal Quantile-Quantile (Q-Q) plot. This compares the quantiles of the observed data with those of a normal distribution. Figure 1 is a Q-Q plot of a subset of arsenic concentrations in streams on the Hawaiian island of Oahu. If the data were normally distributed they would closely follow the line from lower left to upper right on the figure. There are observations with the same value at several places that confirm that the data are not normally distributed.

2.2 Summary statistics

The description of data summarizes the set by locating its center (mean or median), dispersion around that center (the standard deviation), and any asymmetry around the center (skewness) based on a defined distribution (normal, lognormal, bimodal, etc.) Statisticians and environmental data analysts frequently use plots (box-and-whisker diagram, cumulative frequency distribution) to visually present the data set as well as a table of the summary statistics.

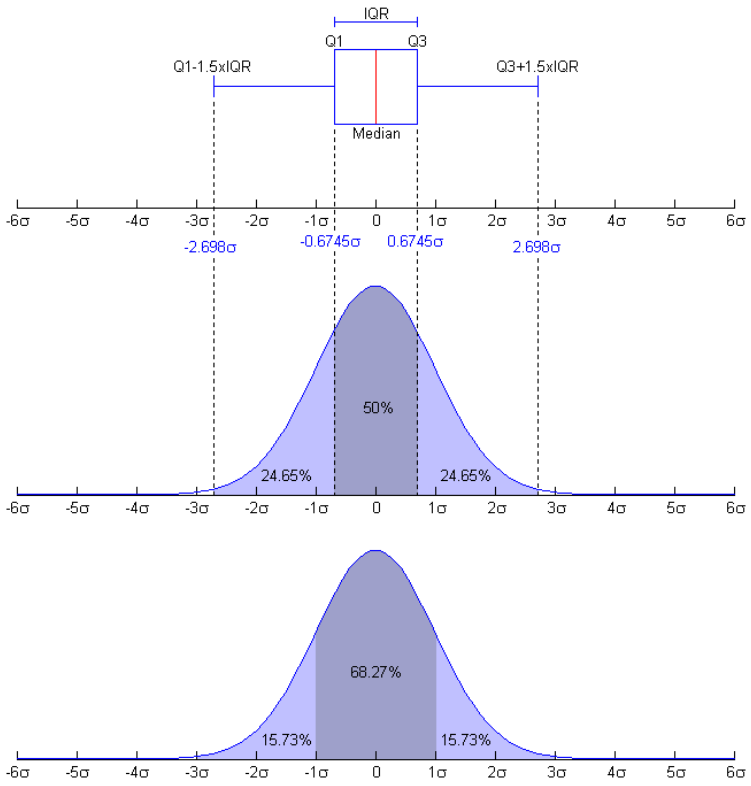

Figure 2: Boxplot and its relationship to a normal curve. Environmental data extremes (such as water quality threshold exceedences) are at the right tail of the distribution where they represent approximately 0.35% of all observed values.

A boxplot, or box-and-whisker plot (Figure 2) displays the median and variance of a single variable. While the midpoint of the box (shown as a line or a dot) is normally the median it can also be the mean. The 25% and 75% quartiles (Q_{25} and Q_{75}, indicated as Q1 and Q3 in the figure) define the hinges (the ends of the box) and the difference between the hinges is called the spread, or inter-quartile range (IQR). Lines (whiskers) are drawn from each hinge to 1.5 times the IQR or to the most extreme value of the spread, whichever is the smaller. Points outside these values, usually drawn as circles, are usually considered as outliers.

The IQR (width of the box) represents 50% of all plotted values and is roughly equivalent to 1 standard deviation (1σ) on either side of the mean in a normal distribution. The ends of the whiskers define the range of 99.3% of all plotted values and is roughly equivalent to 3 standard deviations (3σ) either side of the mean.

Boxplots not only make it easy to quickly understand the important parameters of the data but they are outstanding in displaying data distributions of multiple variables relative to each other (Figure 4). These figures are useful not only in examining environmental data but in presenting data in a format that is easily understood by non-technical decision-makers and others interested in looking closely.

Two examples illustrate the communication capabilities of boxplots. The first (Figure 3) describes the Oahu urban stream arsenic concentrations. The reporting limit (RL) is the horizontal line at 2 mg/L. Most of the reported values are below that RL which supports the interpretation that there is little (or no) health threat from these waters.

The second example (Figure 4) is the distribution of cadmium in streams in the Colorado plateau and the southern Rocky Mountains. The RL is 0.6 mg/L and differences in the concentration distributions is easily seen in the two boxplots.

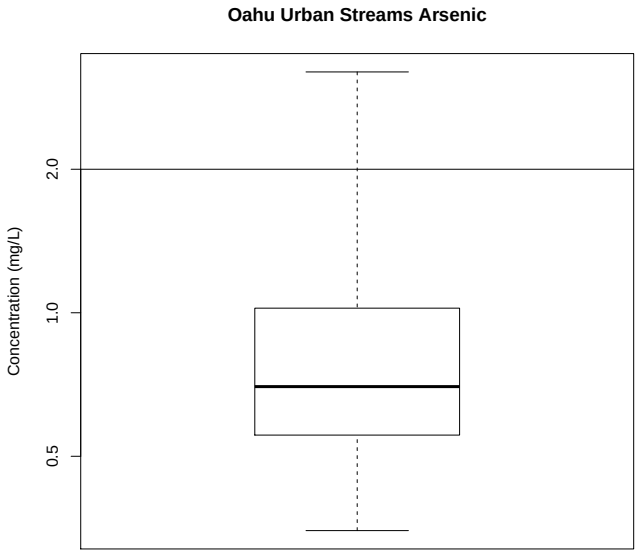

Figure 3: Boxplot of arsenic concentrations in urban streams on Oahu, Hawaii.

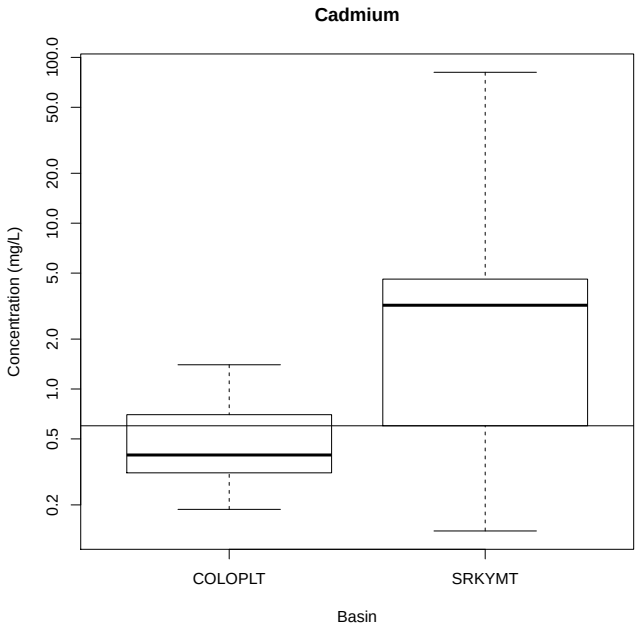

Figure 4: Boxplot of cadmium concentrations in streams of the Colorado Plateau and the southern Rocky Mountains .

2.3 Hypothesis testing

Hypothesis tests compare two or more data sets to determine if they represent samples from the same population. That is, if two data sets have means that are not significantly different from each other or, for multiple groups, if the variability among the data sets is greater than the variability within each data set. Such tests include the *t*-test (for two groups) and analysis of variance (ANOVA) for multiple groups. When the data sets include censored data a maximum likelihood estimation fits the data to a known distribution so samples can be compared with each other.

2.4 Correlation coefficients

Two variables may be associated; that is, changes in one is reflected in changes in the other. The degree of association is measured by a correlation coefficient. The correlation coefficient may range from -1.0 (changes in the two variables are in opposite directions and of the same magnitude) to +1.0 (changes in the two variables are in the same direction and of the same magnitude). When the correlation coefficient is 0.0 the two variables are un-related to each other. Being associated (correlated) is not a measure of cause and effect (see next sub-section).

When chemical constituents have censored (nondetected, less-than) values care must be taken to not substitute an arbitrary value for the censored observations or an incorrect result will be produced. This can have expensive consequences if decisions are based on flawed analyses. The two most commonly applied correlation coefficients are Pearson's ρ (rho) and Kendall's τ (tau).

2.5 Regression equations

Quite often a change in one variable causes a change in a second variable. The first is called the independent variable and the second is the dependent variable. The relationship between the two variables is expressed in a regression equation; the common linear regression equation is $y = mx + b$ where y is the dependent variable, x is the independent variable, m is the slope of the line relating the two variables, and b is the intercept (the value of y when x equals zero). Regression equations, therefore, represent cause-and-effect relationships between two (or more) variables.

3 What are censored data

Every analytical process has a minimum concentration below which the presence of a chemical constituent (the signal) cannot be distinguished from meaningless noise. The chemical signal on the measuring instrument is small in relation to the process noise. Measurements are too imprecise to be reported as a fixed number so they are reported as being less than the analytical threshold; for example, "<2 mg/L". This method detection limit is a function of the instrument, analyst, chemical, intererring constituents, degree of dilution, and other factors. All the analyst can report with confidence is that if the constituent is present its concentration is somewhere between zero and the detection/reporting limit. These concentrations are called nondetects, less-thans, and censored data. For consistency this white paper will refer to the threshold between censored and quantified values as the reporting limit (RL).

These censored observations complicate the familiar calculations of descriptive statistics, of testing differences among groups, and of correlation coefficients and regression equations. When done incorrectly the results can be badly distorted and lead to unwarranted fines, penalties, remedial actions, or lawsuits. For the past couple of decades we no longer have the excuse that correctly analyzing data sets containing censored data cannot be done with readily available computing hardware and software. We need to understand the wrong ways of addressing censored data in statistical data before we can appreciate the correct ways of including them.

4 How *not* to analyze censored data sets

Within environmental sciences, the most common procedure continues to be substitution of some fraction of the RL for nondetects even though more than 20 years ago this was known to be wrong. Results when substituting an arbitrary value for censored data are inaccurate statistics, poor and misleading regression models, and incorrect decisions about whether regulatory enforcement or remediation actions are justified. Where current regulatory

agency guidance exists for environmental data analysis the analyst is told to apply one of several equally incorrect methods for computing descriptive statistics:

1. Ignore (or drop) them.
2. Use the reporting limit (RL).
3. Set them equal to zero.
4. Substitute an arbitrary value (for example, ½ the RL).
5. Use a change in percent censored when reporting limits change.

4.1 Ignore (drop) them

Excluding or deleting censored environmental chemistry data is the worst practice done with them. The results have a strong bias in all measures of location (mean and median) and hypothesis tests comparing groups of data. After excluding the 80% of observations that are left-censored nondetects, for example, the mean of the remaining 20% of observations is reported. The report has lost the information contained in 80% of the original data; the proportion of observations in each group that is too low to be quantified. Don't do this.

While it is easy for analysts and regulators to understand that completely ignoring left-censored environmental data is a bad thing the naïve alternative is to substitute some other value for these observations.

4.2 Use the RL

A common approach is to remove the less-than symbol "<" from the data and use the numeric reporting limit as if it was actually measured and quantified. There are two problems with this approach. As with ignoring or deleting nondetects the mean and median values are biased too high, and it appears that all censored concentrations have the same value: the RL. The high bias might suggest there is reason for concern in the results that is not justified and it is intuitively understood that it is unlikely that all nondetected values are actually at the laboratory's reporting limit.

4.3 Set them equal to zero

This is the opposite extreme of setting the censored observations to the RL. The consequences of setting all censored observations to zero is a bias to a low mean/median because the sum is that of the quantified observations but the divisor includes the number of zero values. Not only is valuable information in the raw data lost but the assumption that all censored observations indicate the chemical is not present cannot be supported.

4.4 Substitute an arbitrary value between zero and the RL

Numerous studies have found that substituting one-half of the RL is inferior to other methods. Environmental data analysts have shown that the method represents a significant loss in information compared to other, better methods; that it produces a biased estimate of mean with the highest variability; and that it results in substantial bias unless the proportion of missing data is small, 10 percent or less. The Resource Conservation and Recovery Act (RCRA) guidance recommend substitution only when data sets contain <15% nondetects, in which case the method is "satisfactory". However, that judgment appears based only on opinion rather than on peer-reviewed science. The USEPA's 2004 *Local Limits Development Guidance Appendices* break from this pattern by not recommending substitution methods. Instead, this guidance recognizes that substitution results in a high bias when the mean or standard deviation is calculated and that performance worsens as the proportion of nondetects increases. Substitution introduces more problems today than in the past, because most data today have multiple RLs. Several factors cause multiple RLs, including levels that change over time, samples with different dilutions, interferences from other constituents, different data interpretations for samples sent to multiple laboratories, or variations in RLs because methods for setting them have changed. Regardless of the cause, substituting a fraction of these changing limits for nondetects introduces a signal unrelated to the concentrations present in the samples. Instead, the signal represents the pattern of RLs used. In the end, false trends may be introduced–or real ones canceled out.

Table 1: Summary statistics of arsenic concentrations in urban streams on Oahu, Hawaii.

Observations	24
No. Censored	13
Median	0.777
Mean	0.945
Standard Deviation	0.656

4.5 Use a change in percent censored

Sometimes looking only at the percentage of the data set that is censored, and comparing that percentage among data sets seems like a reasonable way to avoid the issues caused by substitution. Unfortunately, the percentage of censored observations is not solely related to the number of nondetected observations in the data set. It can also reflect different data set sizes, changes in RL, and all the other factors mentioned in the section above. This approach adds information that does not exist in the data and can result in erroneous decisions.

5 How to correctly analyze censored data sets

Statistical methods to correctly analyze censored data were developed decades ago for medical and industrial studies and are adaptations of statistical models used for uncensored data. Only within the past 15-20 years have they been adapted for environmental data analyses. These methods are not widely used to the detriment of the regulated industries, regulators, and policy makers.

The three most common include one parametric and two nonparametric approaches:

1. Maximum likelihood estimation (MLE; parametric)
2. Regression on order statistics (ROS; nonparametric)
3. Kaplan-Meier survival analysis (K-M; nonparametric)

5.1 Maximum likelihood estimation

Maximum likelihood estimation (MLE) is one class of statistical models (the other two being frequentist and Bayesian). Unlike frequentist models (hypothesis testing, for example) that fit the data to a specific model, MLE fits the model to the specific data set being analyzed. Because the model uses the parameters of a probability distribution MLE is a parametric approach to analyzing censored environmental chemistry data. MLE uses a likelihood function to fit the uncensored (that is, quantified) observations to a probability distribution and identifies the distribution with the maximum likelihood (maximum probability) of describing the data set. Environmental chemistry data are normally distributed very rarely and lognormally distributed most of the time. Infrequently a square root distribution will best fit the data set. Assuming a lognormal distribution is well supported by many analyses of censored data.

Because the shape and distribution of observations are defined for a probability distribution such as the lognormal the censored data can be assigned to assumed concentrations along the left side, below the reporting limit and the quantified (uncensored) observations. The mean, median, and standard deviation are then calculated for the specific data set by applying the known parameters of the distribution. Table 1 shows the results of applying MLE to a small (24 observations) data set of arsenic concentrations in urban streams on Oahu, Hawaii.

MLE is the basis for hypothesis testing of censored environmental chemistry data (as well as the process of estimating water chemistry based on collections of aquatic macroinvertebrates or fish). The means of two groups of data can be tested (the *t*-test) to determine whether both data sets come from the same population of data (the null hypothesis, H_0, is that they do); the variability among different data sets can be tested to determine whether it is greater than the variability within each data set (e.g., ANOVA and the *F*-test). Other multivariate tests are also based on MLE determinations of summary statistics.

Of the three approaches, MLE is the easiest to understand by nontechnical decision-makers. However, MLE does not produce accurate or reliable answers when there are fewer than approximately 50 observations in the data set. When this is the case, the following approach should be used.

Table 2: ROS summary statistics of arsenic concentrations in Oahu urban streams.

Number	24
No. Censored	13
Median	0.700
Mean	0.972
Standard Deviation	0.718

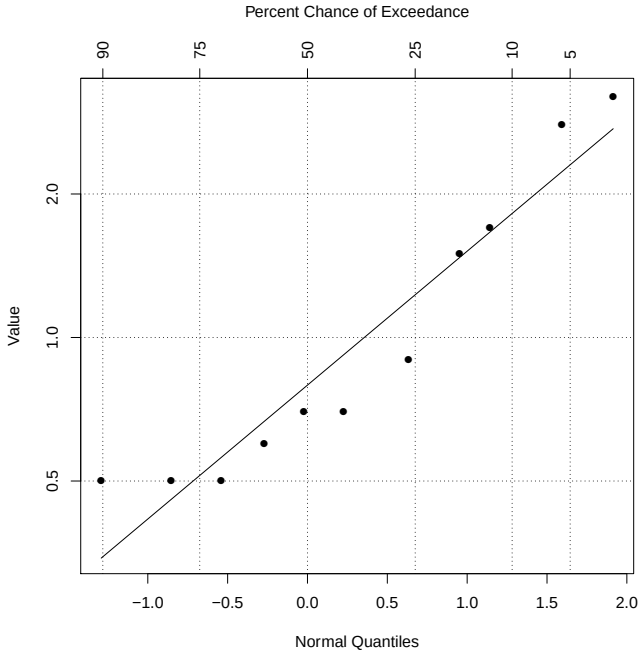

Figure 5: Plot of Oahu urban stream arsenic concentrations applying ROS method.

5.2 Regression on order statistics

Regression on order statistics (ROS) is a nonparametric approach that calculates summary statistics by least squares regression of probability plots of data ranks, not the actual concentrations themselves. To apply ROS the data values are ordered from lowest to highest; tied values (such as censored observations) are assigned the median rank to each tied value. The ranked data are then analyzed using the Mann-Whitney test (also known as the Wilcoxon or rank-sum test) to produce the mean, median, and standard deviation. It is important to remember that the mean and standard deviation are those of the *logarithms* of the ranked data, not the original data themselves. Care must be taken when transforming these log value to original units to avoid potentially very large errors.

ROS is the preferred approach with small data sets, as is nonparametric models such as the Mann-Whitney and Kruskal-Wallis tests using uncensored (fully quantified) small data sets. Applying ROS to the Oahu urban stream arsenic data produces summary statistics shown in Table 2. With this small (24 observations) data set, these results are more reliable than are those from the MLE example above. The plot of observed values, quantiles, and percent exceedance shows the distribution of these data is adequately close to longnormal so the results can be used with confidence (Figure 5).

5.3 Kaplan-Meier survival analysis

The Kaplan-Meier survival analysis (K-M) approach has been used primarily for medical and industrial data having "greater-than" values, such as the time until a disease recurs or for a product to fail. For this method to be applied to "less-thans", such as low-level chemical concentrations, data values must be individually subtracted from a constant greater than the largest observation value, or "flipped", before applying the Kaplan-Meier survival model. One caution is that estimates of the mean, but not percentiles, will be biased high with this method when

Table 3: Kaplan-Meier (K-M) survival analysis of Oahu urban stream arsenic concentrations.

Mean	0.949
Standard Deviation	0.807
Standard Error	0.165
25% Quartile	0.500
Median	0.700
75% Quartile	0.900

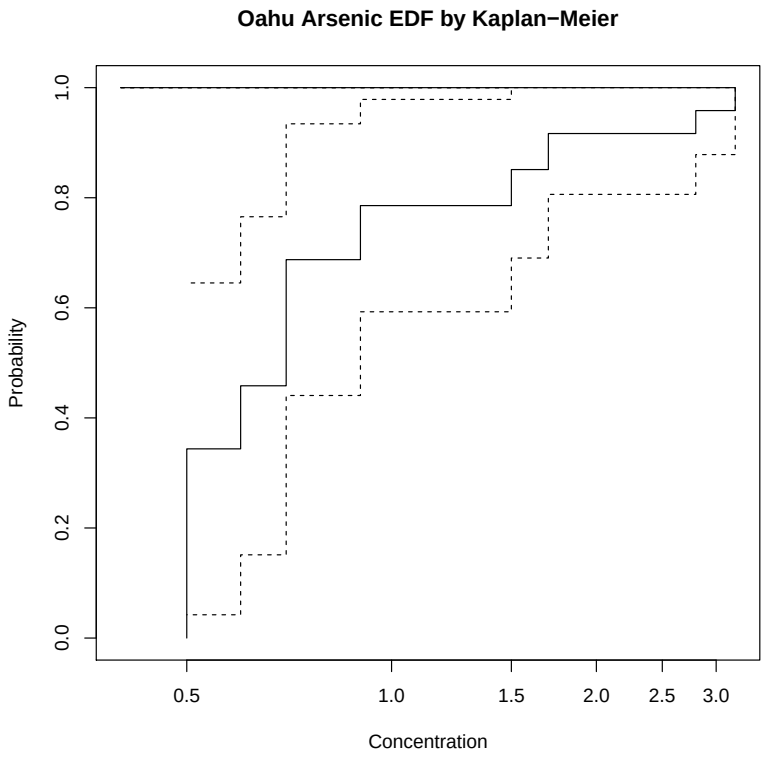

Figure 6: K-M survival analysis of Oahu urban stream arsenic concentrations.

the smallest value in the data set is a nondetect.

Applying K-M to the Oahu urban stream arsenic data yields the summary statistics in Table 3. The analytical results plot (Figure) shows the probability of occurrence of any concentration (the solid line) and the upper and lower 95% confidence limits (the two dotted lines). When you look at this plot you can understand why it is the most difficult of the three approaches to be easily understood by nontechnical decision-makers.

5.4 Approach comparison

The summary statistics of the three approaches are similar which demonstrates that they are all valid for use with censored environmental chemistry data (Table 4). Because of the small number of observations (24), the MLE results are less accurate than are the ROS and K-M results.

5.5 Additional statistical analyses

The calculation of means and standard deviations of data sets using one of the approaches appropriate for censored data (MLE, ROS, K-M) allows further statistical analyses to be correctly applied to these data. Among these further analyses are:

Table 4: Summary statistics from three methods.

	Median	Mean	Standard Deviation
ROS	0.700	0.970	0.718
K-M	0.700	0.949	0.807
MLE	0.777	0.945	0.656

- Calculation of confidence intervals for censored data.
- Group tests for censored data (Wilcoxon and Kruskal-Wallis tests).
- Correlation and regression of censored data (Pearson's ρ and Kendall's τ for correlation coefficients, MLE regression equations).
- Multivariate methods for censored data (non-metric multidimensional scaling [NMDS], distance scaling, cluster analysis).

5.6 Summing censored data

Regulatory agencies require permit holders to submit reports of the total quantities of regulated chemicals released to the environment over defined time periods; e.g., the total amount of mercury released from a power plant or mill roaster exhaust stack or metals contained in overburden and other non-ore rocks in a rock disposal area). When the concentrations of these chemical constituents are below laboratory reporting limits, properly incorporating them in the reported totals faces the same constraints as when calculating summary statistics. They cannot be dropped, assumed to all be at the RL or zero or some arbitrary value between those extremes, but must be correctly included.

Summing censored values is the reverse of calculating the mean of a set of numbers. Since the mean is the sum of all values divided by the number of values summed, the sum of data containing censored values is calculated by calculating the mean by one of the appropriate methods and multiplying that number by the number of observations.

Ecological risk assessment (ERA) involves analyzing toxic chemicals, often with concentrations below reporting limits, calculating the total quantities each with a different weight of importance, and estimating toxicities relative to a baseline chemical (2,3,7,8-Tetrachlorodibenzo -p- dioxin, TCDD, commonly called dioxin). The purpose is to calculate an overall effect of complex chemicals such as dioxins, furans, and PCBs on organisms. Each chemical is a class of individual compounds (congeners), and each congener has a different degree of toxicity to different organisms. Toxicity equivalent concentrations (TECs) summarize the toxicity of all congeners in the class by assuming individual congener toxicities are additive. Fish consumption guidelines are critically dependent on TECs so their correct calculation has important consequences for human health and local economies. Because chemical congeners each have different toxicities to different organisms they are "normalized" to the TCDD toxicity using a toxicity equivalency weighting factor (TEF). TEFs are based on consensus of a scientific panels for each organism class.[1] TCDD has a TEF of 1.0 while less toxic congeners have TEFs closer to 0. Measured concentrations are multiplied by the TEF to obtain the TEC for that congener. Total TEC is the sum of individual congener TECs in the sediments, soils, or other media. When congener concentrations are below analytical laboratory reporting limits the sums can accurately be obtained by calculating the means using one of the above approaches and multiplying that by the number of observations to obtain the sum. No substitutions or other guesses are necessary.

6 Why you should care

Collecting and measuring environmental chemistry data requires a lot of time, effort, and money. Correctly extracting all information contained in the data maximizes the return on your investment in acquiring those data. Such data are collected to establish environmental baselines for NEPA impact assessments, NPDES point source water discharge permits, TMDL nonpoint source criteria, new source air discharge permits, and permit compliance monitoring for all of these permits. Environmental chemistry data may also be required to release reclamation or remediation bonds, demonstrate compliance with consent decrees, and litigation support.

[1] Va den Berg, M., et al. 1998. Toxic equivalency factors (TEFs) for PCBs, PCDDs, PCDFs for humans and wildlife. *Environmental Health Perspectives* **160**:775–792.

Avoiding imputation (substitution, guesses) in summarizing these data and applying hypothesis tests of differences among groups demonstrates to regulators and courts that you have taken a "hard look" at the data, that your results are based on all the information in the data, and that no assumptions or use of arbitrary values dilute the robustness of the results you present.

Environmental data analyses have advanced a long way since the end of the last century and there is no reason to not take advantage of the improvements in support of your projects.