September 25, 1990

Branch of Systems Analysis Technical Memorandum 90.1
Subject: STATISTICS -- Second in the Series of BSA Technical Memoranda

Attached is the second in a series of briefing papers on statistical issues of interest in hydrology. The first briefing paper was distributed with BSA Tech. Memo 89.1. Each paper will present issues which are applicable to ongoing work in the District offices, and will represent activities either of the Branch of Systems Analysis or related work in the field of applied statistics. They attempt only to survey these issues; the appropriate references should be sought for more detail.

The second briefing paper, "Less Than Obvious: Statistical Treatment of Data Below the Reporting Limit" is scheduled for publication in the Dec. 90 issue of *Environmental Science and Technology*. It describes procedures for interpreting data sets which include values known only to be less than a threshold value (called "censored data") such as data below a laboratory reporting limit. These procedures are also applicable for analysis of historical annual floods known only to be less than some relatively high stage, for water levels below the bottom of the casing, for well yields which include data reported only as less than some small value, etc. The same techniques can also be adapted for use with data known only to be greater than some threshold, such as bacterial plate counts, or heads of flowing wells on old maps.

The briefing paper describes techniques for censored data concerning three purposes: a) computing summary statistics such as the mean; b) performing hypothesis tests and c) computing regression equations. A table summarizing these techniques follows, along with examples of computations of summary statistics for common situations, and locations of available computer code.

One important conclusion is emphasized here: two commonly-used methods for dealing with values below the reporting limit should be avoided. These are deletion of data below the reporting limit, and substitution of arbitrary values such as zero or one-half the reporting limit for "less-thans" prior to computing summary statistics or performing statistical tests.

At this point in time, citations to the original references listed here are appropriate. All procedures have been published in prior journals and books. The summary paper itself may be cited as "accepted by Environmental Science and Technology" until it is published in December 90, when the full reference may be employed.

Questions on this topic should be entered into the STATI continuum on QVARSA.

Dennis R. Helsel
Chief,
Branch of Systems Analysis

Recommended Techniques for Interpretation of Censored Data

Estimation of Summary Statistics

Mean and Standard Deviation	Percentiles
Robust Probability Plot	Robust Probability Plot or MLE

Hypothesis Tests

	One Reporting Limit	Several Reporting Limits
Compare 2 groups:	rank-sum test	tobit regresion
Compare >2 groups:	Kruskal-Wallis test	tobit regression
Severe Censoring (>50%):	above tests, or contingency tables	--

Regression

Small % censoring	Moderate % censoring	Large % censoring
Kendall's robust line	tobit regression	logistic regression
tobit regression	logistic regression	contingency tables

COMPUTATION OF SUMMARY STATISTICS -- COMMON EXAMPLES

Example 1: Data all below some detection limit, and n is odd:

<1 <1 <1 <10 <10 <10 <50.

The mean and std deviation cannot be estimated, as there are no data above the detection limit. For the median and IQR, a great deal of information is present. To compute a median where all data are below one or more detection limits and the sample size is odd, remove the < sign, compute the sample median, and then restore the < sign. Therefore the median is <10. The IQR must equal the sample 75th percentile, as the 25th percentile could equal zero. Here the IQR is <10.

Example 2: Data all below some detection limit, and n is even:

<1 <1 <1 <10 <20 <20

For this situation, the larger of the two center observations (the [n/2]+1 th observation) is used as the median, rather than the average of the two center observations as for uncensored data. Then restore the < sign. In this example the median is <10. The IQR is computed as in example 1, and here would be <20.

Example 3: Data above and below one detection limit:

<1 <1 <1 5 7 8 12 16 25

This is a simplified case of example 3. The probability plot and maximum likelihood methods must be used to compute the mean and standard deviation. The percentiles may also be estimated using these methods, or sample statistics may be computed. The sample median is known to be 7, as less than 50% of the data are censored. Because more than 25% of the data are censored, the sample IQR must be computed as a range. If all the <1's are actually 0, the IQR = 14 - 0 = 14. If all <1's are very close to 1, the IQR = 13. So the sample IQR could be reported as "13 to 14".

Example 4: Data above and below multiple detection limits:

<1 <1 <1 5 7 8 <10 <10 <10 12 16 25

Without the <10's the data could be ordered from smallest to largest, and a median calculated. However, it is unclear whether the <10's are below or above 1, or 5, etc. The probability plot method is therefore used to compute the mean and standard deviation, and the maximum likelihood method for the median and IQR (Helsel and Cohn, 1988). These give the following:

mean = 7.8	median = 2.8
std dev = 6.9	IQR =7.5.

Location of computer code

Estimation of Summary Statistics

MDL.LIB is a library of routines for the Primes which computes both the robust plotting position and MLE estimates of summary statistics. The subroutine MDL within the library is called from the user's program in order to compute the statistics. AUTOMDL.RUN is an example calling program which reads an external data file and computes the statistics, reporting all values to the terminal. A documentation file called **READ.ME.FIRST.is also available for further explanation. All of these files can be retrieved from the RVARES prime using the FTR command, from the directory **<sysgrp>common>lthans**.

<sysgrp>common>lthans	>mdl.lib.	the library of compiled subroutines
	>mdl.f77	fortran source code with data formats.
	>automdl.run	use with data in file specific format.
	>**read.me.first	documentation for the programs.

An example output from MDL is given below. The program will calculate any percentiles requested given sufficient numbers of data, with the default reporting of the 10th, 25th, 50th, 75th and 90th percentiles. Estimates will not be provided for data which have more than 85% of the observations censored.

```
REFERENCE: HELSEL & COHN (1988), WATER RESOURCES RESEARCH
"ESTIMATION OF DESCRIPTIVE STATISTICS FOR MULTIPLY-CENSORED WATER-QUALITY DATA"

FOR FILE SILVER.DATA

               ESTIMATES USING LOG-PROBABILITY PLOTTING
 N   NLT   % LT   # DL   MAX DL       MEAN      STD DEV         Q10         Q25
56    34   60.7   12      25.00       12.51        75.45      21.002E-03  60.488E-03

MEDIAN        Q75         Q90
0.2358       1.373       3.560


               ESTIMATES USING ADJUSTED LOGNORMAL MAXIMUM LIKELIHOOD
 N   NLT   % LT   # DL   MAX DL       MEAN      STD DEV         Q10         Q25
56    34   60.7   12      25.00       4.827        35.07      16.020E-03  67.915E-03

MEDIAN        Q75         Q90
0.3380       1.682       7.132
```

tobit regression

Software to compute Tobit MLE estimates is available on the RVARES Prime. It was developed by Tim Cohn of the Branch of Systems Analysis. Documentation for the software may be found in **<sysgrp>common>lthans>tobit_90.9.doc** , and the procedure itself is in **<sysgrp>common>lthans>tobit_90.9.run**. The procedure is interactive, asking for the input data file name. The output presents the intercept and slope coefficients, their partial likelihood ratios (the test statistic analogous to a partial t-statistic in regression), associated p-values, as well as the estimated degrees of freedom and standard error of the regression line. The format is quite similar to output from the Minitab statistical package. An example of the output is given below.

```
  ADJUSTED MAXIMUM LIKELIHOOD ESTIMATES         50% censoring

 The regression equation is

 Y = -1.646E-01 +  1.157E+00*X1 +  9.392E-01*X2 +  9.227E-01*X3 + 1.131E+00*X4


 predictor           Coef           Stdev
   Constant    -1.646319E-01   1.939827E-01
   X1           1.156589E+00   1.918564E-01
   X2           9.391609E-01   1.823410E-01
   X3           9.226754E-01   1.620001E-01
   X4           1.130758E+00   1.898502E-01

S =  1.153790E+00

APPROX. DF:     36.4
```

Other procedures

The rank-sum test, the Kruskal-Wallis test and contingency tables are available in most commercial statistics packages. The larger packages (SAS, SPSS, Systat, etc.) will contain logistic regression.

LESS THAN OBVIOUS:
Statistical Treatment of Data Below the Reporting Limit

Dennis R. Helsel
U.S. Geological Survey
410 National Center
Reston, Virginia 22092
(703) 648-5713

As trace substances in the world's soils, air and waters are increasingly investigated, concentrations are more frequently being encountered which are less than limits deemed reliable enough to report as numerical values. These "less-than" values -- values stated only as "<rl", where rl is the "reporting limit" or "limit of quantitation"(1) or "determination limit"(2) -- present a serious interpretation problem for data analysts. For example, compliance with wastewater discharge regulations is usually judged by comparing the mean of concentrations observed over some time interval to a legal standard. Yet sample means cannot be computed when less-thans are present. Studies of ground-water quality at waste-disposal sites commonly involve comparisons of two groups of data (upgradient versus downgradient wells). Usually t-tests are employed for this purpose, and yet the t-test requires estimates of means and standard deviations which again are impossible to obtain unless numerical values are fabricated to replace any less-thans present in the data. The results of such tests can vary greatly depending on the values fabricated. Therefore, estimates of summary statistics (such as mean, standard deviation, median, and interquartile range) which best represent the entire distribution of data, both below and above the reporting limit, are necessary to accurately analyze environmental conditions. Also needed are hypothesis test procedures that provide valid conclusions as to whether differences exist among one or more groups of data. These needs must be met using the only information available to the data analyst: concentrations measured above one or more reporting limits, and the observed frequency of data below those limits.

This paper discusses the most appropriate statistical procedures given that data have been reported as less-thans. It does not consider the alternative of reporting numerical values for all data, including those below reporting limits -- see references (3)-(6) for discussion of this alternative.

METHODS FOR ESTIMATING SUMMARY STATISTICS

Methods for estimating summary statistics of data which include less-thans (statisticians call these "censored data") can be divided into the three classes discussed below: simple substitution, distributional, and robust methods. Recent papers have documented the relative performance of these methods (7-11). The first three papers compared the abilities of several estimation methods in detail over thousands of simulated data sets (7-9). In (10) these methods were applied to numerous water-quality data sets, including those which are not similar to the assumed distributions required by the distributional methods. A single case study is reported in (11). Only (9) deals with censoring at multiple reporting limits. Large differences were found in these methods' abilities to estimate summary statistics.

Which summary statistics are appropriate?

Environmental-quality data are usually positively skewed, sometimes very highly skewed (7, 12-14). This is especially true for data close to zero, as are data which include censored values, because the lower bound of zero ensures a positive skew. A typical pattern is one where most data have low values, but a few high "outliers" occur. In such cases the mean and standard deviation are strongly affected by a small number of the highest observations. They may be quite sensitive to the deletion or addition of even one observation, and are therefore poor measures of central value and variability. For positively skewed data, the mean may be exceeded by less than half of the observations, sometimes by as little as 25% or less. Thus the mean is not a good estimate of the central value of those data. Similarly, the standard deviation will be inflated by outliers, implying a variability larger than that shown by the majority of the data set. The mean and standard deviation are useful when concerned with mass loadings of a constituent, such as when computing the average sediment concentration at a river cross section. Large concentrations at one point in the cross section should increase the overall mean value. However, when the strong influence of one large value will distort summaries of data characteristics, such as the "typical" sediment characteristics found over many streams, the mean and standard deviation are usually not appropriate measures.

Alternative measures of central value and variability for skewed data are percentile parameters such as the median and interquartile range (IQR). The median by definition has 50% of the data above it, and 50% below. Unlike the mean, it is not strongly affected by a few low or high "outlier observations." It is a more stable (or "resistant") estimator of typical value for skewed data than is the mean, and will be similar to the mean for symmetric (non-skewed) data. Often the "geometric mean", the mean of logarithms of the data, is computed for the same purpose. The geometric mean is an estimate of the median (in original units) when the logarithms are symmetric. The IQR, like the median, is largely unaffected by the lowest or highest data values. It is the 75th percentile minus the 25th percentile, and thus is the range of the central 50% of the data. It equals 1.35 times the standard deviation for a normal distribution. However for the skewed distributions common to environmental monitoring data, the IQR will often be much smaller than the standard deviation, and a better estimate of variability of the bulk of the data.

The median and IQR have another advantage when applied to censored data: when less than 50% of the data are below the reporting limit, the sample median is known. Similarly, when less than 25% of the data are censored, the sample IQR is known. No "fix-ups" are necessary to obtain sample estimates.

Comparisons of estimation methods

Methods may be compared based on their ability to replicate true population statistics. Departures from true values are measured by root mean squared error (RMSE), which combines both bias and lack of precision. Methods with lower RMSE are considered better.

Class 1: Simple substitution methods (figure 1). These methods substitute a single value such as one-half the reporting limit for each less-than value. Summary statistics are calculated using both these fabricated numbers along with the values above the reporting limit. These methods are widely used, but have no theoretical basis. As figure 1 shows, the distributions resulting from simple substitution methods have large gaps, and do not appear realistic.

All of the studies cited above determined that simple substitution methods performed poorly in comparison to other procedures (7-11). Substitution of zero produced estimates of mean and median which were biased low, while substituting the reporting limit resulted in estimates above the true value. Results for the standard deviation and IQR, and for substituting one-half the reporting limit, were also far less desirable than alternative methods. With the advent of convenient software (11) for other procedures there appears to be no reason to use simple substitutions for such computations. As large differences may occur in the resulting estimates, and as the choice of value for substitution is essentially arbitrary without some knowledge of instrument readings below the reporting limit, estimates resulting from simple substitution are not defensible.

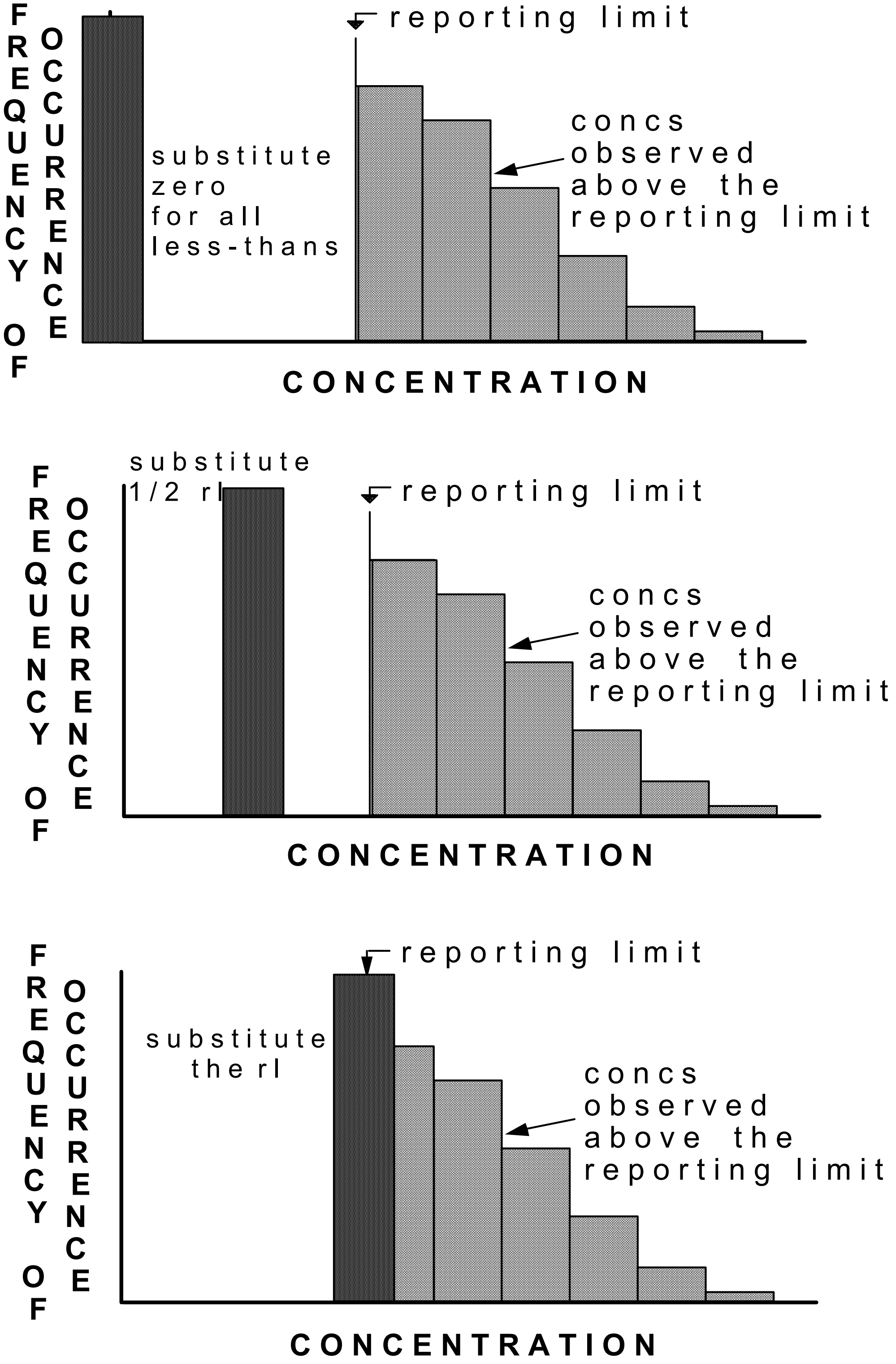


Figure 1. Histograms for simple substitution methods.

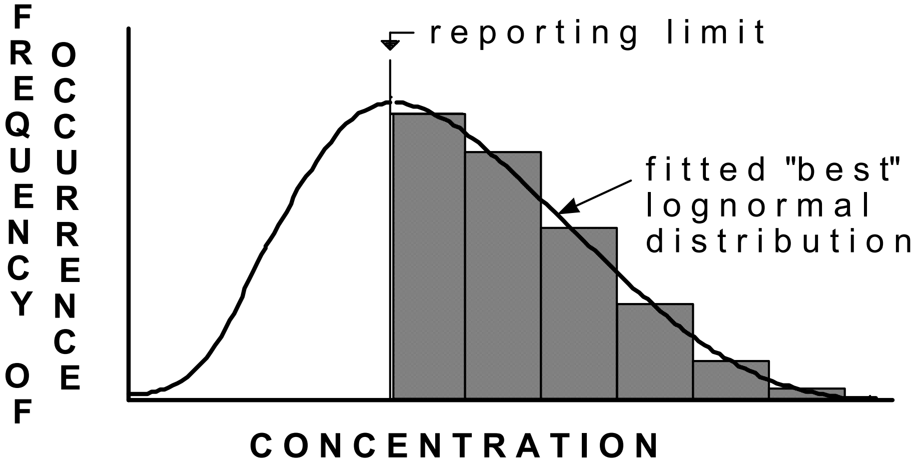
Class 2: Distributional methods (figure 2). Distributional methods use the characteristics of an assumed distribution to estimate summary statistics. Data both below and above the reporting limit are assumed to follow a distribution such as the lognormal. Given a distribution, estimates of summary statistics are computed which best match the observed concentrations above the reporting limit and the percentage of data below the limit. Estimation methods include maximum-likelihood estimation or MLE (15), and probability plotting procedures (16). MLE estimates are more precise than probability plotting, and both methods are unbiased, when observations fit the assumed distribution exactly and the sample size is large. However, this is rarely the case. When data do not match the observed distribution, both methods may produce biased and imprecise estimates (7, 9). Thus the most crucial consideration when using distributional methods is how well the data can be expected to fit the assumed distribution. Even when distributional assumptions are correct, MLEs have been shown to produce estimates with large bias and poor precision for the small (n=5, 10, and 15) sample sizes considered common for environmental data (8). MLE methods are commonly used in environmental disciplines such as air quality (17) and geochemistry (12).

Assuming a lognormal distribution for concentrations, MLEs for larger (n=25, 50) data sets have provided excellent estimates of percentiles (median and IQR) for a variety of data distributions realistic for environmental studies, even those which are not lognormal. However, they have not worked as well for estimating the mean and standard deviation (7, 10). There are two reasons why this is so.

First, the lognormal distribution is flexible in shape, providing reasonable approximations to data which are nearly symmetric and to some positively-skewed distributions which are not lognormal. Thus the lognormal can mimic the actual shape of the data over much of the distribution, adequately reproducing percentile statistics even though the data were not truly lognormal in shape. However, the moment statistics (mean and standard deviation) are very sensitive to values of the largest observations. Failure of the assumed distribution to fit these observations will result in poor estimates of moments.

Second, there is a transformation bias (see box) inherent in computing estimates of the mean and standard deviation for any transformation -- including logarithms -- and then retransforming back to original units (18) (19). Percentiles, however, can be directly transformed between measurement scales without bias. Estimates of mean and standard deviation computed in transformed units by MLEs or other methods are biased upon retransformation. Several studies have included methods which attempt to correct for this bias (9, 11, 12).

Maximum Likelihood (MLE) -- fits 'best' lognormal distribution to the data, and



then determines summary statistics of the fitted distribution to represent the data.

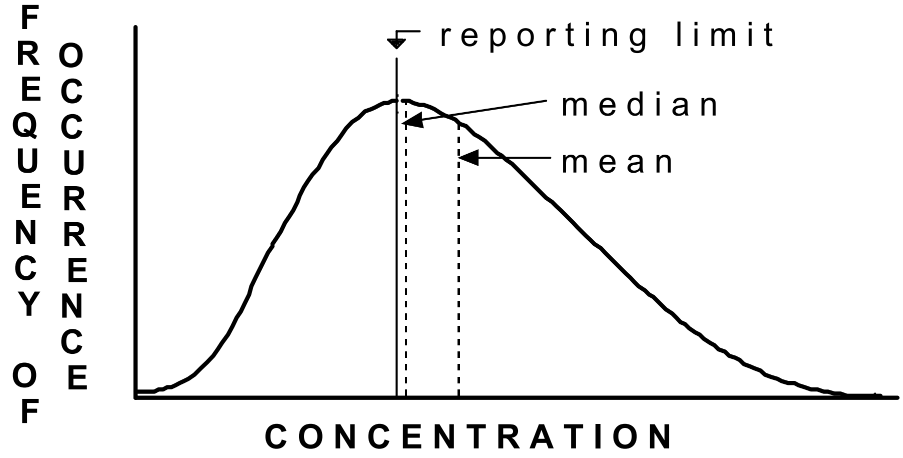


Figure 2. Distributional (MLE) method for computing summary statistics.

Transformation Bias -- a simple example

Original Data		Base 10 logarithms	
1		0	
10		1	
100		2	
1000		3	Mean of logs retransformed
10000		4	= 10^2 = 100
2222.2	**Mean**	**2**	≠ 2222.2

Two less-frequently used distributional methods are a "fill-in with expected values" MLE technique (8) and a probability plot method which estimates the mean and standard deviation by the intercept and slope, respectively, of a line fit to data above the reporting limit (16). Probability plot methods are easy to compute with standard statistics software, an advantage for practitioners. Both methods suffer from transformation bias when estimates are computed in one scale and then retransformed back into original units. Thus the probability plot was recommended for estimating the geometric mean (16), but would not work well for estimating the mean in original units because of transformation bias. Both methods should be slightly less precise than MLEs.

Class 3: Robust methods (figure 3). These methods combine observed data above the reporting limit with below-limit values extrapolated assuming a distributional shape, in order to compute estimates of summary statistics. A distribution is fit to the data above the reporting limit by either MLE or probability plot procedures (7, 9), but the fitted distribution is used only to extrapolate a collection of values below the reporting limit. These extrapolated values are not considered as estimates for specific samples, but only used collectively to estimate summary statistics. The robustness of these methods result primarily from their use of observed data rather than a fitted distribution above the reporting limit. They also avoid transformation bias by performing all computations of summary statistics in original units.

Robust methods have produced consistently small errors for all four summary statistics in simulation studies (7, 9), as well as when applied to actual data (10). Robust methods have at least two advantages over distributional methods for computation of means and standard deviations. First, they are not as sensitive to the fit of a distribution for the largest observations because actual observed data are used rather than a fitted distribution above the reporting limit. Second, estimates of extrapolated values can be directly retransformed and summary statistics computed in the original units, avoiding transformation bias.

Probability plot: regression of log of concentration versus normal score used to extrapolate "fill-in" values below the reporting limit.

These "fill-ins" are retransformed back to original units, and combined with data above the reporting limit to compute estimates of summary statistics.

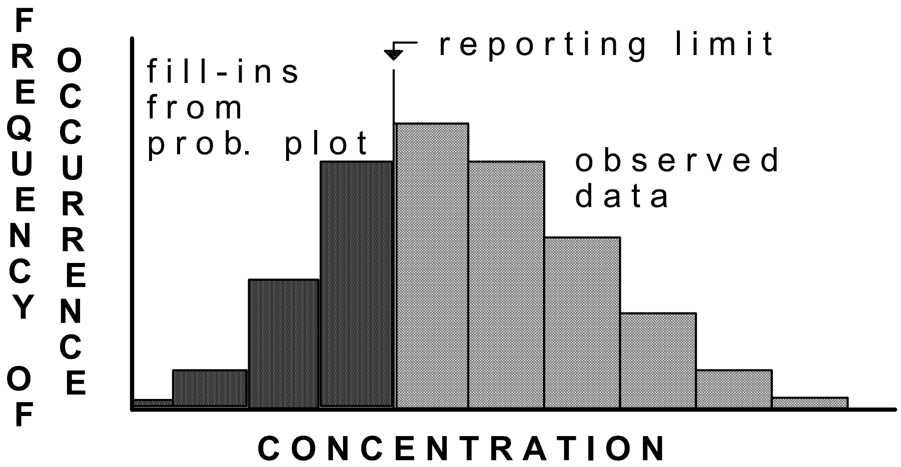


Figure 3. Robust (probability plot) method of estimating summary statistics

Recommendations

Robust procedures have substantial advantages over distributional methods when concentrations cannot be assumed to follow a defined distribution. In practice, the distribution of environmental data is rarely if ever known, and may vary between constituents, time periods, and locations. It is therefore not surprising that robust methods have been recommended for estimating the mean and standard deviation (7, 9). Either robust probability plot or distributional MLE procedures perform well for estimating the median and IQR (7-9). Use of these methods rather than simple substitution methods for environmental data should substantially lower estimation errors for summary statistics.

Multiple Reporting Limits

Data sets may contain values censored at more than one reporting limit. This commonly occurs as limits are lowered over time at a single lab, or when data having different reporting limits are combined from multiple laboratories. Estimation methods belonging to the above three classes are available for this situation. A comparison of these methods (9) again concluded that robust methods provide the best estimates of mean and standard deviation, and MLEs for percentiles. For example, in figure 4 the error rates for six estimation methods are compared to the error that would occur had all data been above the reporting limit (shown as the 100% line). Figure 5 shows the same information when the data differ markedly from a lognormal distribution (9). The simple substitution methods (ZE, HA and DL: substitution of zero, one-half and one times the reporting limit, respectively) have more error in most cases than does the robust probability plot method MR. Where the substitution methods have lower

RMSE, it is an artifact of constant, strongly biased estimates, also not a desirable result. The maximum likelihood procedure MM and the MLE adjusted for transformation bias AM show themselves to be excellent estimation methods for percentiles, but suffer from large errors when estimating the mean and standard deviation.

In summary, use of MLE for estimation of percentiles, and the robust probability plot method for estimating the mean and standard deviation, should greatly decrease errors as compared to simple substitution methods for data with multiple reporting limits.

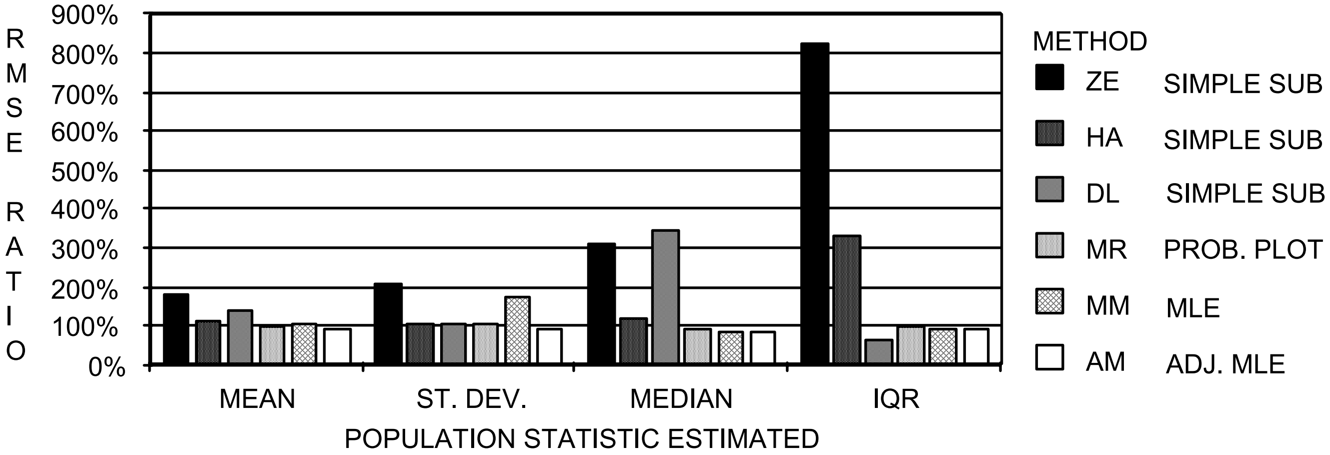


FIGURE 4 -- Error rates (RMSE -- root mean square error) of six multiple-detection methods divided by error rates for uncensored data estimates, in percent, for data similar to a lognormal distribution (9)

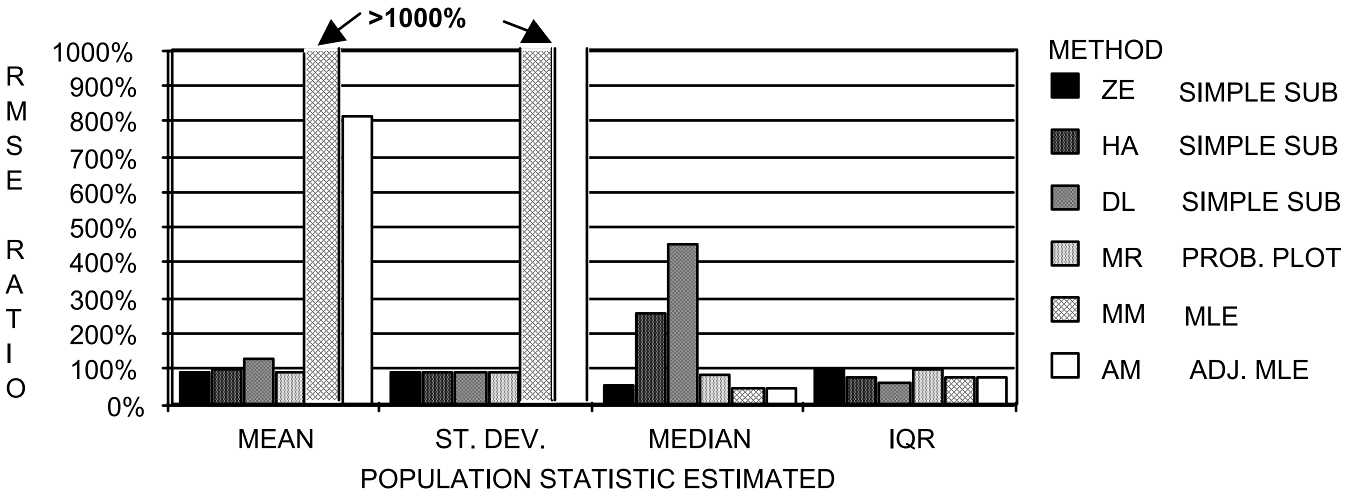


FIGURE 5 -- Error rates (RMSE -- root mean square error) of six multiple-detection methods divided by error rates for uncensored data estimates, in percent, for data not similar to a lognormal distribution. (9)

Software for Computations

MLE methods require advanced computational software. These and other distributional methods for single reporting limits, including the distributional (slope-intercept) probability plot estimator, were recently made available (11) to the scientific community. The robust probability plotting method for a single reporting limit can easily be computed by most commercially available statistics software. Normal scores ("NSCORES" of Minitab, or "PROC RANK" within SAS, etc.) are first computed with all less-thans set to slightly different values all below the reporting limit. Second, a linear regression equation is developed using only the above-limit observations, where log of concentration is the y variable and normal scores the x variable. Estimates for the below-limit data are then extrapolated using this regression equation from normal scores for the below-limit data. Finally, extrapolated estimates are retransformed into units of concentration, combined with above-limit concentration data, and summary statistics computed.

Fortran code for multiple reporting limit techniques is available by sending a self-addressed, stamped envelope and a formatted 3 1/2 inch disk (IBM or Macintosh format) to the author.

METHODS FOR HYPOTHESIS TESTING

Methods for hypothesis testing of censored data can also be classified into the three types of procedures: simple substitution (class 1), distributional or parametric (class 2), and robust or nonparametric (class 3). Parametric statistical tests are frequently used in environmental assessments. They assume that data follow some distributional shape, usually the normal distribution. Parameters (true population statistics) such as the mean and standard deviation are estimated in order to perform the test. When censoring is present, values are often fabricated in order to estimate these parameters (class 1). Problems caused by fabrication are illustrated below. Parametric tests are also available which do not require substitutions for less-thans (class 2). Where the distributional assumptions are appropriate, these relatively unknown tests have great utility. Investigators have also deleted censored data prior to hypothesis testing. This latter approach is the worst procedure, as it causes a large and variable bias in the parameter estimates for each group. After deletion, comparisons made are between the upper X% of one group versus the upper Y% of another, where X and Y may be very different. Such tests have little or no meaning.

Alternatively, nonparametric tests can be performed (20). These tests simply rank the data, and judge whether the ordering of the data points indicate that differences occur, trends exist, etc. No fabrication of data values is required, as all censored data are represented by ranks which are tied at values lower than the lowest number above the reporting limit. These tests generally have greater power than parametric tests when the data do not conform to a normal distribution (20, 21).

As an example of the differences between hypothesis test methods for censored data, tests were performed which determine whether or not means or medians significantly differ between two groups. Two data sets were generated from lognormal distributions having the same variance, but with differing mean values. Sample statistics for the two data sets before and after censoring are given in Table 1.

Prior to any censoring, group means are shown to be significantly different by a t-test (p=0.04, Table 2), and by a t-test for regression slope equal to zero. The latter is performed by designating the data set each observation belongs to as either a zero or one. This binary variable is then used as the explanatory (independent) variable in a linear regression. Though identical to the t-test prior to censoring, a variation of the regression approach will become the distributional (class 2) method for censored data used later. The equivalent nonparametric test, the rank-sum test, produces a much lower p-value (p=0.003). This lower p-value is consistent with the proven greater power of the nonparametric test to detect differences between groups of skewed data (21, 22), as compared to the t-test.

Suppose that these data represent dissolved arsenic concentrations. A typical reporting limit for dissolved arsenic is 1 μg/L, and therefore all data below 1.0 would be recorded as <1. Censoring these data sets at 1 produces 14 less-than values (70%) in group A, and 5 less-than values (23%) in group B.

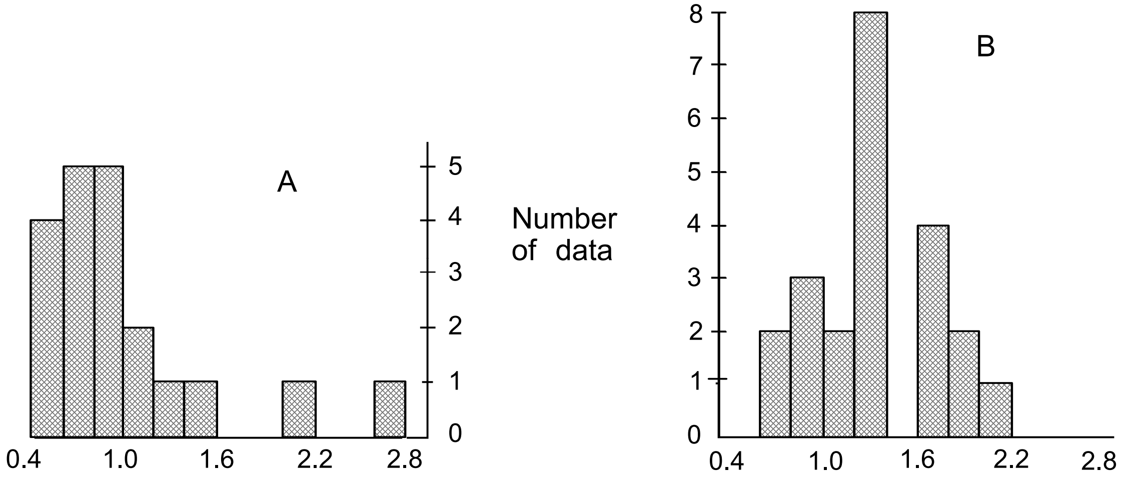
The class 1 method for comparing two groups of censored data is to fabricate data for all less-than values, and include these "data" with detected observations when performing a t-test. No *a priori* arguments for fabrication of any particular value between 0 and the reporting limit can be made. Substituting zero for all less-than values, the means are declared significantly different (p = 0.01). Yet when the reporting limit of 1.0 is substituted, the means are not found to be different (p = 0.19). The conclusion is thus strongly dependent on the value substituted! This example shows that fabrication of data followed by a t-test must be considered too arbitrary for use, especially for legal or management decision purposes, and should be avoided.

The distributional (class 2) method for hypothesis testing also requires an assumption of normality, but does not involve substitution of values for censored data. Instead, a t-test is performed using a regression procedure for censored data known as tobit regression (23, 24). Tobit regression uses both the data values above the reporting limit, and the proportion of data below the reporting limit, to compute a slope coefficient by maximum likelihood. For a two-group test, the explanatory variable in the regression equation is the binary variable of group number, so that data in one group have a value of 0, and in the other group a value of 1. The regression slope then equals the difference between the two group means, and the t-test for whether this slope differs from zero is also a test of whether the group means differ. Tobit regression is also discussed later in the section on regression. One advantage to Tobit regression for hypothesis testing is that multiple reporting limits may easily be incorporated. The caution for its use is that proper application does require the data in both groups to be normally distributed around their group mean, and for the variance in each group to be equal. For large amounts of censoring these restrictions are difficult to verify.

The nonparametric (class 3) equivalent is the rank-sum test. It considers the 19 less-than values tied at the lowest value, with each assigned a rank of 10 (the mean of ranks 1-19). The resulting p-value is 0.002, essentially the same as for the original data, and the two groups are easily declared different. Thus in this example the nonparametric method makes very efficient use of the information contained in the less-than values, avoids arbitrary assignment of fabricated values, and accurately represents the lack of knowledge below the reporting limit. Results do not depend an a distributional assumption (25).

When severe censoring (near 50% or more) occurs, all of the above tests will have little power to detect differences in central values. The investigator will be inhibited in stating conclusions about the relative magnitudes of central values. Other characteristics must be compared. Instead, contingency tables (class 3) can test for a difference in the proportion of data above the reporting limit in each group (20). The test can be used when the data are reported only as detected or not detected. It may also be used when response data can be categorized into three or more groups, such as: below detection, detected but below some health standard, and exceeding standards. The test determines whether the proportion of data falling into each response category differs as a function of different explanatory groups, such as different sites, land use categories, etc.

TABLE 1--Characteristics of Two Lognormal Data Groups (A and B)



A		B
20	no. observations	22
1.00	mean	1.32
0.57	std. deviation	0.39
0.65	25th percentile	1.07
0.85	median	1.25
1.07	75th percentile	1.66
14	no. <rl	5

TABLE 2-- Example of Significance Tests Between Groups A and B

Hypothesis test used	test statistic	p
Uncensored data		
t-test (Satterthwaite approx.)	-2.13	0.04
regression with binary variable	-2.17	0.04
rank-sum test	-2.92	0.003
After imposing artificial reporting limit		
t-test		
less-thans = 0.0	-2.68	0.01
less-thans = 0.5	-2.28	0.03
less-thans = 1.0	-1.34	0.19
tobit regression with binary variable	-2.28	0.03
rank-sum test	-3.07	0.002

Hypothesis testing with multiple reporting limits

More than one reporting limit is often present in environmental data. When this occurs, hypothesis tests such as comparisons between data groups are greatly complicated. It can be safely said that fabrication of data followed by computation of t-tests or similar parametric procedures is at least as arbitrary with multiple reporting limits as with one reporting limit, and should be avoided. The deletion of data below all reporting limits prior to testing should also be completely avoided.

Tobit regression (class 2) can be utilized with multiple reporting limits. Data should have a normal distribution around all group means and equal group variances to use the test. These assumptions are difficult to verify with censored data, especially for small data sets.

One robust method which can always be performed is to censor all data at the highest reporting limit, and then perform the appropriate nonparametric test. Thus the data set
<1 <1 <1 5 7 8 <10 <10 <10 12 16 25
would become <10 <10 <10 <10 <10 <10 <10 <10 <10 12 16 25.
and a rank-sum test performed to compare this with another data set. Clearly this produces a loss of information which may be severe enough to obscure actual differences between groups (a loss of power). However, for some situations this is the best that can be done.

Alternatively, nonparametric score tests common in the medical "survival analysis" literature can sometimes be applied to the case of multiple reporting limits (26). These tests modify uncensored rank test statistics to compare groups of data. The modifications allow for the presence of multiple reporting limits. In the most comprehensive review of these score tests (27), most of them were found inappropriate for the case of unequal sample sizes. Another crucial assumption of score tests is that the censoring mechanism must be independent of the effect under investigation (see box). Unfortunately, this is often not the case with environmental data. The Peto-Prentice test with asymptotic variance estimate was found to be the least sensitive to unequal sample sizes and to differing censoring mechanisms (27).

In summary, robust hypothesis tests have several advantages over their distributional counterparts when applied to censored data. These advantages include: (1) no need to document adherence to a normal distribution. This is difficult to do with censored data; (2) greater power is achieved for the skewed distributions common to environmental data; (3) comparisons are made between central values such as the median, rather than the mean; and (4) data below the reporting limit are incorporated without fabrication of values or bias. Information contained in less-than values is accurately used, not misrepresenting the state of that information.

Examples when a score test would be inappropriate.

Score tests are inappropriate when the censoring mechanism differs for the two groups. That is, the probability of obtaining a value below a given reporting limit differs for the two groups when the null hypothesis that the groups are identical is true.

1. Suppose a trend over time was being investigated. The first five years of data were produced with a method having a reporting limit of 10 mg/L; the second five years used an improved method with 1 mg/L as its reporting limit. A score test of the first half of the data versus the second would not be valid, as the censoring mechanism itself varied as a direct function of time.
2. Two groups of data are compared as in a rank-sum test, but most of the data from group A were measured with a chemical method having 1 as its reporting limit, while most of group B were measured with a method having 10 as its reporting limit. A score test would not yield valid results, as the censoring mechanism varies as a function of what is being investigated (the two groups).

Examples when a score test would be appropriate.

A score test yields valid results when the change in censoring mechanism is not related to the effect being measured. Stated another way, the probability of obtaining data below each reporting limit is the same for all groups, assuming the null hypothesis of no trend or no difference is true. Here a score test provides much greater power than artificially censoring all data below the highest reporting limit before using the rank-sum test.

1. Comparisons were made between two groups of data collected at roughly the same times, and analyzed by the same methods, even though those methods and reporting limits changed over time. Score tests are valid here.
2. Differing reporting limits resulted from analyses at different laboratories, but the labs were assigned at random to each sample. Censoring is thus not a function of what is being tested, but is a random effect, and score tests would be valid.

When adherence to a normal distribution can be documented, Tobit regression (class 2) offers the ability to incorporate multiple reporting limits regardless of a change in censoring mechanism. Score tests (class 3) require consistency in censoring mechanism with respect to the effect being tested.

METHODS FOR REGRESSION

With censored data the use of ordinary least squares (OLS) for regression is prohibited. Coefficients for slopes and intercept cannot be computed without values for the censored observations, and substituting fabricated values may produce coefficients strongly dependent on the values substituted. Four alternative methods capable of incorporating censored observations are described below. The first and last approaches, Kendall's robust fit (28) and contingency tables (20), are nonparametric (class 3) methods requiring no distributional assumptions. Robust correlation coefficients are also mentioned (20). Tobit and logistic regression (24, 29), the second and third methods, fit lines to data using maximum likelihood (class 2). Both methods assume normality of the residuals, though with logistic regression the assumption is after a logit transformation. As before, assumptions are sometimes hard to check with censored data.

The choice of method depends on the amount of censoring present, as well as on the purpose of the analysis. For small amounts of censoring (below 20%), either Kendall's line or the tobit line may be used. Kendall's would be preferred if the residuals were not normally distributed, or when outliers are present. For moderate censoring (20-50%), Tobit or logistic regression must be used. With large amounts of censoring, inferences about concentrations themselves must be abandoned, and logistic regression employed. When both the explanatory and response variables are censored, tobit regression is applicable for small amounts of censoring. For larger amounts of censoring, contingency tables or rank correlation coefficients are the only option.

1. Kendall's robust line fit . When one censoring level is present, Kendall's rank-based procedure for fitting a straight line to data can test the significance of the relationship between a response and explanatory variable (28). An equation for the line, including an estimate of the slope, is usually also desirable. This can be computed when the amount of censoring is small. Kendall's estimate of slope is the median of all possible pairwise slopes of the data. To compute the slope with censoring, twice compute the median of all possible slopes, once with zero substituted for all less-thans, and once with the reporting limit substituted. For small amounts of censoring the resulting slope will change very little, or not at all, and can be reported as a range if necessary. If the slope value change is of an unacceptable magnitude, tobit or logistic regression must be performed.

Research is currently underway on methods based on scores similar to those for hypothesis tests with multiply-censored data that may allow robust regression fits to data with multiple reporting limits (30).

2. Tobit Regression. Censored response data can be incorporated along with uncensored observations into a procedure called tobit regression (23, 24). It is similar to OLS except that the coefficients are fit by maximum-likelihood estimation. MLE estimates of slope and intercept are based on the assumption that the residuals are normally distributed around the tobit line, with constant variance across the range of predicted values. Again, it is difficult to check these assumptions with censored data. Should the data include outliers, these can have a strong influence on the location of the line and on significance tests (Figure 6), as is true with uncensored OLS. Residuals for uncensored data should be plotted versus predicted values, so that linearity and constant variance assumptions can be verified for at least small amounts of censoring. For larger percentages of less-thans, decisions whether to transform the response variable must often be made based on previous knowledge ("metals always need to be log-transformed", etc.). Tobit regression is also applicable when both the response and explanatory variables are censored, such as a regression relationship between two chemical constituents. However, the amount of censoring must be sufficiently small that the linearity, constant variance, and normality assumptions of the procedure can be checked. Finally, Cohn (18) as well as others have proven that the tobit estimates are slightly biased, and have derived bias corrections for the method.

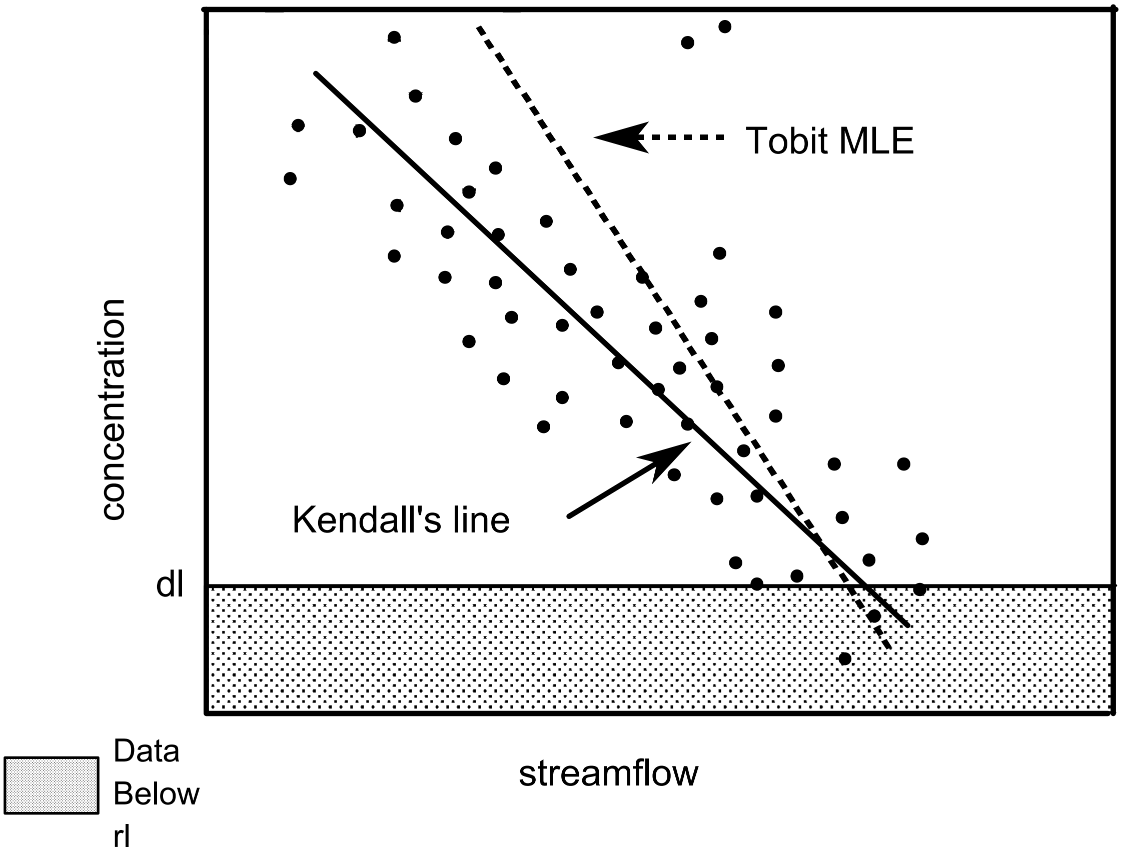


Figure 6. Kendall's and tobit MLE lines for censored data with outliers. Note the tobit line is strongly influenced by outliers.

3. Logistic Regression (29)**.** Here the response variable is categorical. No longer is a concentration being predicted, but a probability of being in discrete binary categories such as above or below the reporting limit. One response (above, for example) is assigned a value of 1, and the second response a 0. The probability of being in one category versus the second is tested to see if it differs as a function of continuous explanatory variable(s). Examples include predicting the probability of detecting concentrations of some organic contaminant from continuous variables such as nitrate concentrations, population density, percent of some appropriate land use variable, or of irrigation intensity. Predictions from this regression-type relationship will fall between 0 and 1, and are interpreted as the probability [p] of observing a response of 1. Therefore [1–p] is the probability of a 0 response.

Logistic regression may be used to predict the probabilities of more than 2 response categories. When there are m>2 ordinal (may be placed in an order) responses possible, (m–1) equations must be derived from the data. For example, if 3 responses are possible (concentrations below rl =0, above rl but below health standards =1, and above health standards =2), two logistic regressions must be computed. First, an equation must be written for the probability of being nonzero (the probability of being above the rl). Next the probability of a 2 (probability of being above the health standard) is also modelled. Together, these two equations completely define the three probabilities p(y=0), p(y=1), and p(y=2) as a function of the explanatory variables.

4. Contingency Tables (20)**.** Contingency tables are useful in the regression context if both explanatory and response variables contain censoring. For example, suppose the relationship between two trace metals in soils (such as arsenic and aluminum) is to be described. The worst procedure would again be to throw away the data below the reporting limits, and perform a regression. Figure 7 shows that a true linear relationship with negative slope could be completely obscured if censored data were ignored, and only data in the upper right quadrant investigated. Contingency tables provide a measure of the strength of the relationship between censored variables -- the phi statistic ϕ (20), a type of correlation coefficient. An equation which describes this relationship, as per regression, is not available. Instead, the probability of y being in one category can be stated as a function of the category of x. For the figure 7 data, the probability of arsenic being above the reporting limit is 21/36 = 0.58 when aluminum is above reporting limit, and 17/18 = 0.94 when aluminum is below the reporting limit.

5. Rank correlation coefficients. The robust correlation coefficients Kendall's tau or Spearman's rho (20) could also be computed when both variables are censored. All values below the reporting limit for a single variable are assigned tied ranks. Rank correlations do not provide estimates of the probability of exceeding the reporting limit as does a contingency table. So they are not applicable in a regression context, but would be more applicable than contingency tables in a correlation context. One such context would be in "chemometrics" (32): the computation of correlation coefficients for censored data as inputs to a principal components or factor analysis.

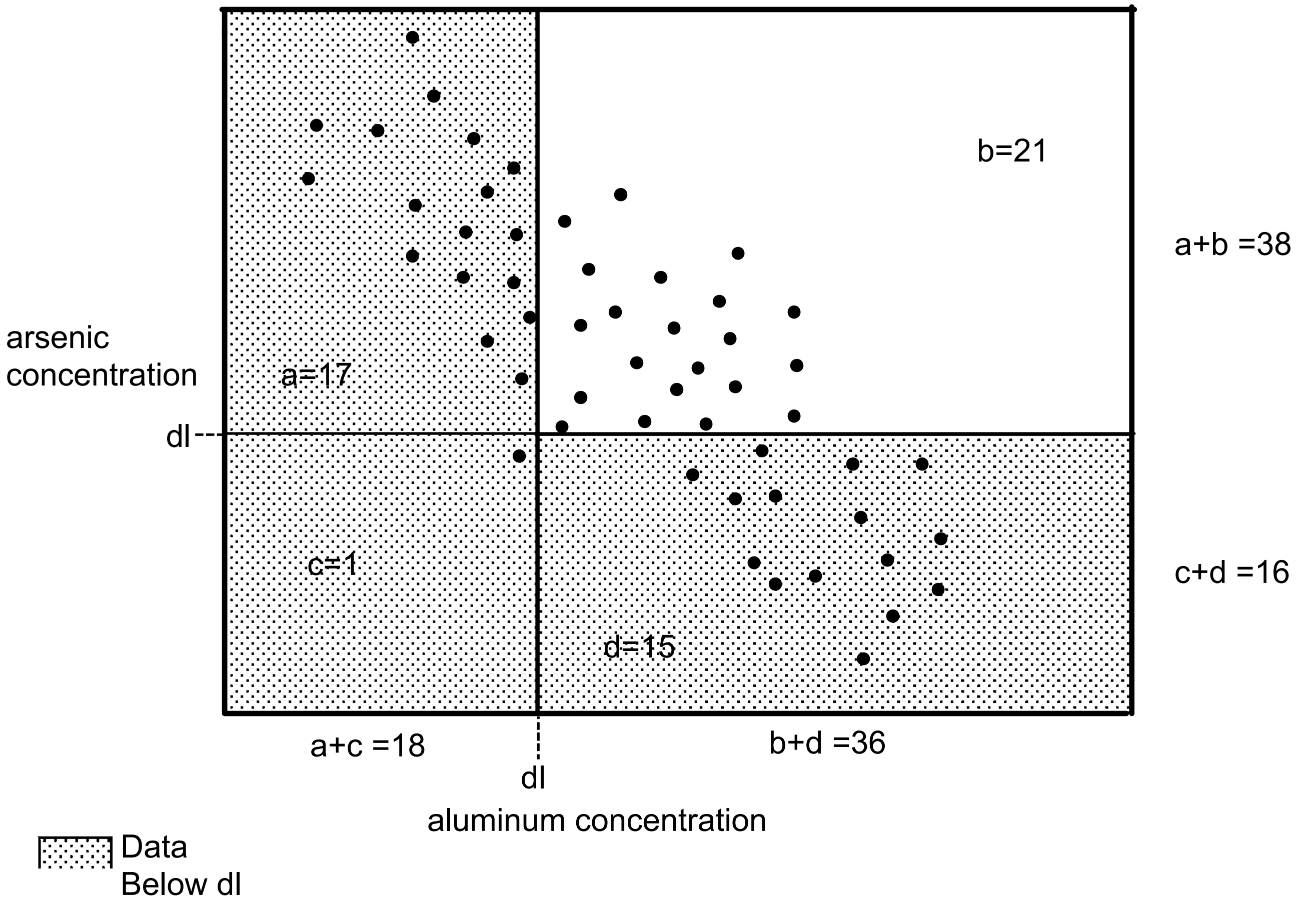


Figure 7. Contingency table relationship between two censored variables.
(Ignoring censored data would produce the misleading conclusion that no relationship exists between the two variables)

In summary, relationships between variables with data below reporting limits can be investigated in a manner similar to regression. Values should not be fabricated for less-thans prior to regression. Instead, for small amounts of censoring and one reporting limit, Kendall's robust line can be fit to the data. For moderate censoring and/or multiple reporting limits, tobit regression can be performed. For more severe censoring of the dependent variable, logistic regression is appropriate. When both response and explanatory variables contain severe censoring, contingency tables can be performed.

CONCLUSIONS

Methods are available which appropriately incorporate data below the reporting limit for purposes of estimation, hypothesis testing, and regression. Deletion of censored data, or fabrication of values for less-thans, leads to undesirable and unnecessary errors.

References

1. Keith, L. H.; Crummett, W.; Deegan, J.; Libby, R. A.; Taylor, J. K.; Wentler, G.; *Anal. Chem.* **1983**, *55*, 2210-18.

2. Currie, L. A. *Anal. Chem.* **1968**, *40*, 586-93.

3. ASTM Subcommittee D19.02. *Annual Book ASTM Standards* **1983**, *11.01*, Chapter D, 4210-4283..

4. Porter, P. S.; Ward, R. C.; Bell, H. F; *Environ. Sci. Technol.* **1988**, *22*, 856-61.

5. Gilliom, R.J.; Hirsch, R. M.; Gilroy, E. J.; *Environ. Sci. Technol.* **1984**, *18,* 530-35.

6. Porter, P. S. in *Monitoring to Detect Changes in Water Quality Series,*; Lerner, D. ed.; IAHS Publication no. 157; International Assoc. of Hydrological Sciences; Wallingford, England, **1986,** pp. 305-15.

7. Gilliom, R.; Helsel, D. *Water Resour. Res.* **1986**, *22*, 135-46.

8. Gleit, A. *Environ. Sci. Technol.* **1985**, *19*, 1201-6.

9. Helsel, D.; Cohn, T. *Water Resour. Res.* **1988**, *24*, 1997-2004.

10. Helsel, D.; Gilliom, R. *Water Resour. Res.* **1986**, *22*, 147-55.

11. Newman, M. C.; Dixon, P. M. *American Environmental Laboratory*, **1990**, pp. 26-30.

12. Miesch, A. *U.S. Geological Survey Professional Paper 574-B*, USGS; Reston, VA, **1967**.

13. Davis, G. D. *Ground Water* **1966**, *4 (4)*, 5-12.

14. Luna, R. E.; Church, H. W. *Journal Applied Meteorology* **1974**, *13*, 910-6.

15. Cohen, A. C. *Technometrics* **1959**, *1*, 217-213.

16. Travis, C.C.; Land, M. L. *Environ. Sci. Technol.* **1990**, *24*, 961-2.

17. Owen, W.; DeRouen, T. *Biometrics* **1980,** *36,* 707-19.

18 Cohn, T. *U. S. Geological Survey Open-File Report 88-350*, USGS; Reston, VA., **1988**.

19. Miller, D. M. *American Statistician* **1984**, *38*, 124-6.

20. Conover, W. *Practical Nonparametric Statistics, 2nd Edition*; John Wiley; New York, **1980**.

21. Blair, R. C.; Higgins, J. J. *Journal of Educational Statistics* **1980**, *5*, 309-35.

22. Hodges, J. L.; Lehmann, E. L. *Annals of Mathematical Statistics* **1956**, *27*, 324-35.

23. Powell, J. L. *Econometrica* **1986**, *54*, 1435-60.

24. Judge, G. G.; Griffiths, W. E.; Hill, R. C.; Lee, T. C. *The Theory and Practice of Econometrics*, Chap. 14, John Wiley: New York, **1980.**

25. Helsel, D.; Hirsch, R. *Water Resour. Bull;* **1987**, *24*, 201-4.

26. Millard, S.; Deverel, S. *Water Resour. Res.* **1988**, *24*, 2087-98.

27. Latta, R. *Jour. American Statistical Association* **1981**, *76*, 713-9.

28. Lehmann, E. *Nonparametrics: Statistical Methods Based on Ranks*; Holden-Day Publishers: Oakland, 1975.

29. Amemiya, T. *Journal of Economic Literature* **1981**, *19*, 1483-1536.

30. McKean, J.; Sievers, G. *Technometrics* **1989**, *31*, 207-18.

31. Breen, J. J.; Robinson, P. E., Eds.; *Environmental Applications of Chemometrics*, ACS Symposium Series 292, American Chemical Society., Washington, D. C. **1985**, 280 p.