BJC/OR-416

Improved Methods for Calculating Concentrations Used in Exposure Assessments

Date Issued—January 2000

Prepared by the
Lockheed Martin Energy Research Corporation

Prepared for the
U.S. Department of Energy
Office of Environmental Management

BECHTEL JACOBS COMPANY LLC
managing the
Environmental Management Activities at the
East Tennessee Technology Park
Oak Ridge Y-12 Plant Oak Ridge National Laboratory
Paducah Gaseous Diffusion Plant Portsmouth Gaseous Diffusion Plant
under contract DE-AC05-98OR22700
for the
U.S. DEPARTMENT OF ENERGY

PREFACE

The purpose of this report is to describe computer programs that provide improved methods of determining the concentrations appropriate for use in ORO risk assessments. According to EPA, because of uncertainty associated with any estimate of exposure concentration, the 95% upper confidence limit of the arithmetic mean will be used for the reasonable maximum exposure in risk assessment. The SAS macros described in the report provide an efficient way of calculating this value, as well as other important summary statistics. It is especially noteworthy that the programs include methods that are appropriate when nondetects are present, when the distribution of the underlying data is unknown, and when both situations apply. This work was performed under Work Breakdown Structure 1.4.12.2.3.04.05.03 (Activity Data Sheet 8304). This Work Breakdown Structure is entitled Risk Assessment: Decision Support. The ultimate objective of this report is to provide and explain new statistical software that should be applied to improve the concentration estimates applied in ORO risk assessments. This document was previously released as a draft with the document number ES/ER/TM-211.

CONTENTS

PREFACE iii

FIGURES vii

TABLES vii

ACRONYMS ix

EXECUTIVE SUMMARY xi

1. INTRODUCTION 1

2. BACKGROUND STATISTICAL INFORMATION 4

3. CALCULATION OF CONFIDENCE LIMITS: UNCENSORED CASE 6
   3.1 CONFIDENCE LIMITS OF THE MEAN OF NORMALLY DISTRIBUTED UNCENSORED DATA 6
   3.2 UPPER CONFIDENCE LIMIT OF THE LOGNORMAL MEAN OF UNCENSORED DATA 7

4. CALCULATION OF THE MEAN, CONFIDENCE LIMITS, AND TOLERANCE BOUNDS THAT APPLY WHEN DATA ARE EITHER UNCENSORED OR LEFT CENSORED 9
   4.1 LOGNORMAL AND NORMAL MODEL ESTIMATES 9
   4.2 PRODUCT LIMIT ESTIMATES 14
      4.2.1 Mathematical Explanation of the PLE and its Application 14
      4.2.2 Simple Example of Calculation of the PLE 17

5. USING THE SAS MACROS 21

6. AN EXAMPLE 23

7. DISCUSSION 23

8. REFERENCES 26

APPENDIX A: THE LNOR MACRO A-1
APPENDIX B: THE PLE MACRO B-1
APPENDIX C: THE LOGCONF MACRO C-1
APPENDIX D: THE SAS PROGRAM USED TO PRINT THE INPUT DATA, TO CALL THE SAS MACROS, AND TO PRODUCE THE OUTPUT FILES D-1

FIGURES

1 The step function which is the cumulative distribution function for the sample used in the example . . 18
2 The step function shown in Fig. 1 with the addition of labeled rectangles 19
3 Lognormal distribution function (DF) estimate and 95% confidence bounds for groundwater lead at Station ST-006 30
4 PLE (black lines) and 95% confidence bounds (gray lines) for groundwater lead at Station ST-006 31

TABLES

1 The steps in the calculation of the product limit estimate (PLE) 18
2 Calculations of areas of rectangles in Figure 2 20
3 Remaining steps involved in the calculation of the SE of the mean of the PLE 21
4 Groundwater lead concentrations (mg/L) from Y-12 Fuel Station 27
5 SAS macro lnor.sas output data set contents for groundwater lead (Y-12 Fuel Station) 32
6 Output of SAS macro lnor.sas for groundwater lead (Y-12 Fuel Station) 34
7 SAS macro ple.sas output data set contents for groundwater lead (Y-12 Fuel Station) 36
8 Output of SAS macro ple.sas for groundwater lead (Y-12 Fuel Station) 38
9 Groundwater barium concentrations (mg/l) from Y-12 Fuel Station 39
10 SAS macro logconf.sas output data set contents for groundwater barium (Y-12 Fuel Station) 40
11 Output of SAS macro logconf.sas for groundwater barium (Y-12 Fuel Station) 40

ACRONYMS

DQA	data quality assessment
EPA	United States Environmental Protection Agency
LCL	lower confidence limit
MLE	maximum likelihood estimate
ORO	Oak Ridge Operations
PLE	product limit estimate
RME	reasonable maximum exposure
SAS	Statistical Analysis System
SE	standard error
UCL	upper confidence limit

EXECUTIVE SUMMARY

In order to estimate the potential health or environmental effects of a particular substance in a given medium at a particular location, an estimate is needed of the concentration of the substance that is present. Under current Environmental Protection Agency (EPA) guidance for risk assessment (EPA 1992, 1995a), the average concentration is the value of the exposure to be used in such estimation.

Because only a finite number of samples can be taken, the average concentration cannot be determined precisely. For this reason, EPA requires that a 95% upper confidence limit (95% UCL) on the arithmetic average concentration be calculated to estimate exposure concentration used in risk assessments. The 95% UCL of the average concentration is the value that, when calculated for an infinitely large number of randomly drawn subsets of site data, will equal or exceed the true average 95% of the time.

Commonly used methods for calculating exact confidence limits on the mean require assumptions about the underlying distribution of values. For example, it is commonly assumed that the data are either normally or lognormally distributed. Lognormal data are data that are normally distributed after the logarithms of the data (to any base) are taken. Before this transformation has been made, the distribution of the data is skewed with a long right tail. Frequently samples of concentrations of contaminants in the environment appear to have a lognormal distribution.

The problem of nondetects (also called left censoring) occurs commonly for environmental data. A "nondetect" is an observation that is below the level of detection of the analytical method. The limit of detection is generally defined as the lowest concentration that can be determined to be statistically different from a blank specimen. The limit of detection is an imprecise quantity that can vary from sample to sample and laboratory to laboratory. Several methods of low reliability are commonly used when analyzing left-censored data. These include substituting 0 for nondetects, substituting the detection limit divided by 2 for nondetects, or procedures which involve graphing the data and replacing the nondetects with values that fit the assumed underlying distributions.

In either the normal or lognormal case, it is possible to estimate exact confidence limits on the mean using an uncensored random sample from the distribution. When there is censoring, approximations are needed.

This report describes macros developed for use with SAS software[1]. These macros simplify calculation of 95% UCLs and of other environmental summary statistics based on the normal and lognormal models, as well as on a nonparametric method. Two of the macros account for nondetects in ways that are more sophisticated than commonly used methods. This is important because environmental data are often left censored. The summary statistics provided by the macros are needed for Baseline Risk Assessment Reports and in other applications in environmental restoration.

The SAS macros described in this report should provide the basis for development of exposure concentrations for all ORO risk assessments in which the sample selection procedure emulates simple random sampling. While the methods provide a more reliable way of analyzing data that are left censored, it is important to realize that these methods are also applicable to normal and lognormal data that are not censored. (See report BJC/OR-271 for an overview of the data evaluation process.) It is recommended that much weight

[1]SAS® and SYSTAT® are registered trademarks used to identify products or services of SAS Institute, Inc

be given to the Product Limit Estimate (PLE) approach because it is free of assumptions about the underlying distributions, and it is particularly well adapted for handling nondetects. It will often be necessary to report the confidence limits and other summary statistics for the lognormal and normal distributions, which are also easily calculated using these macros. However, it should be realized that the lognormal approach, in particular, can sometimes lead to large errors in estimating confidence limits. When the 95% UCL of the lognormal mean is many times larger than the 95% UCL of the PLE mean, there would seem to be good reason to apply the latter in risk estimation.

1. INTRODUCTION

In order to estimate the potential health or environmental effects of a particular substance in a given medium at a particular location, an estimate is needed of the concentration of the substance that is present. Under current Environmental Protection Agency (EPA) guidance for risk assessment (EPA 1992, 1995), the average concentration is the desired value of the exposure to be used in such estimation. The average is suggested for use based on the following:

(1) carcinogenic and chronic noncarcinogenic toxicity criteria are based on estimated lifetime average exposures to low levels of such substances;

(2) the average concentration is most representative of the concentration to which individuals would be exposed over time at a site.

The average of interest is actually the time-averaged concentration. In some cases, this is approximated using the spatial average. For example, if the medium is contaminated soil, then the spatially-averaged concentration can be used to approximate the time-averaged concentration if one assumes that the exposed individual moves randomly across the exposure area.

Because it is impractical to characterize sites completely regarding the exposure concentration, it is necessary to address the uncertainty when estimating the average using a finite number of samples. Indeed, EPA (1989, pp. 6-19 and 22) requires that an estimate of the upper 95 percent confidence limit (95% UCL) on the arithmetic average concentration be calculated, and that the smaller of the maximum detected concentration and the 95% UCL be used to estimate the exposure concentration used in risk assessments.

In risk assessment, the reasonable maximum exposure (RME) is the maximum exposure that is reasonably expected to occur at a site. It is important to realize that according to EPA (1989, p. 6-19) the statistic that is used for the RME is the 95% UCL on the arithmetic average. The emphasis in this report is on more reliable methods for calculating the 95% UCL. The 95% UCL of the average concentration is the value that, when calculated for an infinitely large number of randomly drawn subsets of site data, will equal or exceed the true average 95% of the time.

The validity of any estimate of the 95% UCL is dependant on the quality of the data used. Much of the theory behind summary statistics such as means and UCLs is based on the assumption that the data values used were obtained by random sampling. EPA (1995b) provides a discussion of the importance of random sampling and how it can be accomplished.

Commonly used methods for calculating exact confidence limits on the mean require assumptions about the underlying distribution of values. For example, it is commonly assumed that the data are either normally or lognormally distributed. In both of these cases, it is possible to estimate exact confidence limits on the mean using an uncensored random sample from the distribution. Lognormal data are data that are normally distributed after the logarithms of the data (to any base) are taken. Before this transformation has been made, the distribution of the data is skewed with a long right tail. Frequently samples of concentrations of contaminants in the environment appear to have a lognormal distribution. This is the main reason why much of this report deals with methods for calculating the mean for the lognormal distribution, together with

estimates of the uncertainty about it. Despite the seemingly simple connection between the normal and lognormal distributions, the methods of calculating means, and their degree of uncertainty, differ substantially.

EPA (1995b) discusses methods for testing whether data fit the normal or lognormal distributions. (For the lognormal case, usually the natural logarithms of the data are taken and the transformed data are tested to see if they follow the normal distribution.) Often the number of samples is far too sparse to provide any means of reliably determining the shape of the underlying distribution. On the other hand, large sample sizes can lead to the rejection of a particular distribution even though, in actuality, the distribution may be adequate. In some instances, neither the normal nor the lognormal distribution is appropriate. The above difficulties are one of the reasons why a nonparametric estimate, together with estimates of the uncertainty about it, is particularly attractive. Nonparametric approaches are not dependent upon one's guessing correctly which type of underlying distribution is appropriate. For very small sample sizes, EPA (1995b) recommends that nonparametric hypothesis tests "be selected during Step 3 of the DQA Process in order to avoid incorrectly assuming that the data are normally distributed when, instead, there is simply not enough information to make this determination." (Step 3 of the DQA process is the one that involves selection of the "most appropriate procedure for summarizing and analyzing the data".) The nonparametric approach that will be considered is based on the product limit estimate (PLE ; Kaplan and Meier, 1958).

The problem of left censoring occurs commonly for environmental data. "Left censoring" means that some of the observations (often denoted "U" or "*") are below the level of detection of the analytical method. (Often a sizeable fraction of the observations are nondetects.) For example, assume that a given instrument under particular sampling conditions cannot detect a concentration of substance R equal to, or lower than, 0.30 units. Perhaps 20 samples are tested and only 12 of them yield detectable concentrations, perhaps ranging from 0.32 to 3.40 units/gram of soil. The remaining samples are reported as nondetects of value 0.30U. These 8 samples are important because they show that 8 of 20 samples were $\leq$ 0.30 units per gram of soil. It is possible that all of them were zero units per gram of soil..

The limit of detection is a statistical concept not a chemical concept. The limit of detection is generally defined as the lowest concentration that can be determined to be statistically different from a blank specimen. The limit of detection is an imprecise quantity that can vary from sample to sample because of variations in matrix interference, calibrations, dilutions, etc. Detection limits are especially likely to vary when samples combine data collected by different laboratories.

With this report, three macros for use with SAS software are being made available. This report provides background, explanation, and discussion of the three macros and their output. The macros, together with brief descriptions, are as follows:

1. Macro "**logconf**" provides summary statistics for lognormal data that are uncensored. Two methods are applied depending primarily on sample size. When any more than a slight proportion of the data is left censored, it is inappropriate to use this macro. Instead, use the two macros described below.

2. Macro "**lnor**"provides summary statistics for the lognormal distribution, as well as for the normal distribution, that take into consideration left-censoring of data. When there is no censoring, the normal-based approach reduces to computing ordinary sample means, standard deviations, confidence limits, etc. The lognormal-based approach reduces to a method called Cox's direct method (see section 3).

3. Macro "**ple**" is a nonparametric alternative that provides summary statistics for the product limit estimate (PLE). It is well adapted for application when the data are left censored. When there is no censoring, it provides the same results as the ordinary mean and its confidence limits.

Only a univariate approach will be taken in this report. That is, concentrations for only one analyte (a chemical or radionuclide) will be considered at a time. This contrasts with a multivariate approach, where concentrations of several analytes are considered simultaneously. When conducting Baseline Risk Assessments, it is also important to report information on the percentiles of the exposure concentrations. The 50th percentile (called the 50%-ile in the computer output of the macros) is the median, and in certain situations the median provides a better measure of the average than the arithmetic mean. (In the normal distribution, the mean, the median, and the mode are identical.) Confidence bounds (or limits) of percentiles are termed tolerance bounds (or limits). In this report confidence limits of estimates of the mean will always be referred to as confidence limits, and confidence limits of percentiles will always be referred to as tolerance bounds. The **lnor** and **ple** macros report whatever percentiles are requested, together with their tolerance bounds. (For normal or lognormal-based tolerance bounds, use the LNOR macro even in the all-detects case.)

It is important to realize that these three macros do not take into consideration the statistical complications caused by right-censored data or by the clustering of data. It is the responsibility of those who apply these macros to realize these limitations, so as to avoid reporting inappropriate measures of concentrations. Right censoring could result if some of the sampled concentrations were so high that they exceeded the upper limit of the measurement device. It would seem that technical adjustments such as further dilution of samples or the use of a less sensitive scale (for radiation) could eliminate the reporting of environmental data that are right censored. Clustered data, in which certain regions of an area of interest are oversampled, violate the assumption that the data constitute a random sample. As stressed earlier, valid application of the macros assumes that the data are collected by random sampling. If it is known that this is not the case, appropriate caution must be made in the presentation and the interpretation of the data.

In addition to the summary statistics mentioned above, the **lnor.sas** and **ple.sas** macros also provide the following basic summary statistics and characteristics of the concentration data that must be reported in a Baseline Risk Assessment:

- total number of samples
- number of samples that are detects
- frequency of detection
- minimum value (including nondetects)
- minimum detected concentration
- maximum detected concentration
- minimum nondetected concentration
- maximum nondetected concentration
- maximum value
- ordinary mean (i.e., the usual arithmetic mean or sample average)
- ordinary standard error of the mean
- 95% LCL of the ordinary mean
- 95% UCL of the ordinary mean

2. BACKGROUND STATISTICAL INFORMATION

Some basic statistical concepts will be reviewed briefly to provide background for the mathematical explanations that will be provided for the procedures implemented by the macros. After the concepts have been introduced in terms of the population of all possible values, explanations will be provided for random samples drawn from those populations.

The set of all possible values for a particular attribute is called the *sample space*, and a *random variable* is any function from a sample space *S* to the real numbers. In most instances in this report, the sample space of interest is that of all possible concentration values, and the random variable is the concentration in a given exposure area.

Let *X* denote a random variable. The *cumulative distribution function*, or cdf, denoted by $F_X(x)$ is given by

$$F_X(x) = \textit{Probability that } X \le x = P[X \le x]$$

Associated with a (continuous) random variable *X* is the *probability density function* (pdf), denoted by $f_X(x)$, which gives probabilities that X is in an interval, as follows:

$$P[a \le X \le b] = \int_a^b f_X(x)\, dx$$

This expression can intuitively be interpreted as "adding up" the continuum of the "probabilities" $f_X(x)$ for $a<X<b$.

There is a simple relationship between the cdf and pdf for a given random variable (in fact, this is sometimes given as the definition of the pdf):

$$F_X(b) = \int_{-\infty}^b f_X(x)\, dx$$

The average, or expected value, of a random variable is, intuitively, the weighted sum of all possible values of the random variable; i.e., each possible outcome is multiplied by its probability of occurring and these products are summed. The precise definition of the expected value or the *mean* (μ) of a random variable *X*, denoted by $\mu = E[X]$, is given by

$$E[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

It is necessary to make statements about the concentration of interest in order to perform a risk assessment. An example of such a statement would be a 95% UCL for the mean. These statements must be made with only a small subset of all possible concentrations. The process of making such statements is called *statistical inference*.

Only a small subset of the entire population can be known, and without knowing the underlying distribution completely, it is impossible to make statements about it with complete (i.e., 100%) confidence. However, if the subset is collected appropriately, statements can be made with a specified level of confidence. The most widely used method of collecting data appropriately is *random sampling*. In essence, the idea in random sampling is that every member of a population has an equal chance of being selected. Methods for sampling data randomly are discussed by EPA (1995b). *Biased sampling* is a broad term applied to any method of collecting a subset such that one does not obtain a representative picture of the population. For valid application of the statistical methods discussed in this report, it is essential that data be collected using methods that protect against bias.

Thus only a finite number of samples from a given distribution can be used to estimate its expected value. Consider n observations $x_1, x_2,...,x_n$. If preferential sampling does not occur spatially or over time, then the samples are considered random and independent. If the values are correlated, they are referred to as clustered. Different statistical methods must be used with data that are correlated, clustered or otherwise interdependent. The methods discussed in the report are not appropriate for such data.

Let x denote a random variable in a sample drawn from a population. The sample *cumulative distribution function* denoted by $\hat{F}_n(x)$, is given by

$$\hat{F}_n(x) = \frac{No.\ of\ sample\ values \le x}{n}$$

This function provides a convenient and familiar way to summarize and display data. A plot of $\hat{F}_n(x)$ versus x makes it easy to visualize the sample, and it provides information on the percentiles and the dispersion of the data. It is also useful for ascertaining the distributional shape of the population from which the sample was taken.

To estimate the true mean of the distribution that is being sampled, the arithmetic mean (or sample mean) is calculated by summing the values of the samples and dividing by the number of samples as shown below. In this report the sample mean is usually referred to as the "ordinary mean".

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

In the common situation of random sampling with no censoring, $\bar{x}$ is an unbiased estimate of the mean. ("Unbiased" means that if a large number of sample means is calculated, the average of these sample means will approach the true mean.)

The *variance* of a random variable provides a measure of the spread in a probability distribution. When the variance is small, it is more likely that the sampled values will be close to the mean of the distribution. The variance is defined mathematically as follows:

$$Var(X) = E(X^2) - \mu^2$$

An unbiased estimate of the sample variance is given by

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

The sample standard deviation is the square root of the estimate of the sample variance and is more commonly reported than the sample variance.

3. CALCULATION OF CONFIDENCE LIMITS: UNCENSORED CASE

3.1 CONFIDENCE LIMITS OF THE MEAN OF NORMALLY DISTRIBUTED UNCENSORED DATA

In this section the concept of the upper and lower 95% confidence limits of the mean (95% UCL and LCL) will be developed for the case where an analyte's concentration distribution is assumed to be normally distributed and the data are both uncensored and unclustered. In the previous section, the basis for the estimation of the sample mean and sample variance were described. The summary statistics are to be calculated for **all** analytes of the data set, independently and one analyte at a time. All three macros provide the 95% UCL, and the **lnor** and **ple** macros also provide the 95% LCL. As noted in the introduction, for risk analysis the 95% UCL is more important.

$$\frac{\bar{x} - \mu}{s/\sqrt{n}}$$

As was the case for calculation of the sample mean, let $x_1, x_2, \ldots, x_n$ denote a random sample of n values from a normal distribution with unknown mean and variance of μ and σ^2, respectively. If $\bar{x}$ denotes the sample mean and s denotes the sample standard deviation, then it is known that (see, e.g., Casella and Berger 1990, p. 226) the distribution of the ratio depends only on the number of samples n. This distribution is called Student's t-distribution with $n-1$ degrees of freedom, and it is denoted by t_{n-1}.

The 95% UCL is then given by

$$UCL = \bar{x} + t_{n-1}(0.95)\frac{s}{\sqrt{n}}$$

and the 95% LCL by

$$LCL = \bar{x} - t_{n-1}(0.95)\frac{s}{\sqrt{n}}$$

where $t_{n-1}(0.95)$ is the upper 95th percentile of the *t*-distribution with *n-1* degrees of freedom.

Another common summary statistic, which has importance in calculation of confidence limits for the sample mean, is the standard error of the mean (SE, or, as it is referred to in the output of the macros, "Ordinary mean std. err."). It is defined as follows:

$$SE = \frac{s}{\sqrt{n}}$$

It is obvious that the SE is part of the equation for the confidence limits given above. The proper number to use for multiplication by the SE can be read directly from the *t*-table, which is found, for example, in EPA (1995b, p. A-11). In that table the number to use is found in the column headed by ".95" and in the row headed by "n". (It should be noted that "n" in the table means degrees of freedom, which in this case is *n* - 1.)

3.2 UPPER CONFIDENCE LIMIT OF THE LOGNORMAL MEAN OF UNCENSORED DATA

The objective here is to demonstrate how to estimate the mean and its 95% UCL for data that are lognormal, and for which there is no censoring. The approaches described provide the basis for the calculations in the macro named **logconf**. Again, the statistics are to be calculated for **all** analytes of the data set that are lognormally distributed, and they must be calculated independently, one analyte at a time. Unlike the other macros, the **logconf** macro does not calculate lower confidence limits or any tolerance bounds. Also, unlike the other macros, the user cannot select other confidence limits (or tolerance bounds) besides those at the 95% level. These capabilities may be incorporated into later versions of the macros.

Intuitively it might seem possible that one could calculate the mean of a lognormal distribution simply by calculating the logarithms of the data points, finding the mean of those logarithms using the methods described above, and then simply finding the antilogarithm of the answer. However, that approach does not work. To see why, let Z be a random variable having a standard normal distribution (i.e., a mean of 0 and a variance of 1). The quantity $e^{\sigma Z + \mu}$ then has a lognormal distribution with a logscale mean of μ and a logscale variance of σ^2. The expectation of $e^{\sigma Z + \mu}$, that is $E[e^{\sigma Z + \mu}]$, is thus

$$\int_{-\infty}^{+\infty} e^{\sigma z+\mu}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz = e^{\mu}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-(z^2-2\sigma z+\sigma^2)/2}e^{\sigma^2/2}dz = e^{\mu+\sigma^2/2}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{(z-\sigma)^2/2}dz,$$

which is $e^{\mu+\sigma^2/2}$ because the last integral (i.e., $\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{(z-\sigma)^2/2}dz$) is 1. The point here is that the expectation of $e^{\sigma Z+\mu}$ is $e^{\mu+\sigma^2/2}$, not e^{μ}.

Let $x_1, x_2,..., x_n$ denote a random sample from a lognormal distribution (i.e., μ is the arithmetic mean of $\log(x_1)$, $\log(x_2)$, …, $\log(x_n)$), with unknown mean μ and variance σ^2. If $\hat{\mu}$ denotes the logscale sample mean and $\hat{\sigma}$ denotes the log-scale sample standard deviation, then Land (1971) showed that a UCL (optimal in a sense) can be calculated by

$$UCL = e^{\left(\hat{\mu}+0.5\hat{\sigma}^2+\frac{\hat{\sigma}H_{1-\alpha}}{\sqrt{n-1}}\right)}$$

where $H_{1-\alpha}$ depends on the confidence level α, the standard deviation, and the sample size.

This method of calculating the 95% UCL is implemented in the SAS macro named **logconf**. It applies a table of H values that was derived for every sample size from 3 to 1001, and for sample standard deviations ranging from 0.1 to 10.0 (after the log transformation) at intervals of 0.1. The H-values were calculated using a computer program (Lyon and Land 1999) that is an implementation of methods described by Land (1971, 1972).

When an exact H value was not located in the table, the H value to apply was derived using linear interpolation from values present in the table. If the sample standard deviation was less than 0.1, the H value for a sample standard deviation equal to 0.1 was used. Likewise if the sample standard deviation was greater than 10.0, the H value calculated for a sample standard deviation equal to 10.0 was used.

When the sample size was greater than 1001, the macro **logconf** applies "Cox's Method" (Land 1971) to calculate the UCL as shown below

$$95\%\ UCL = e^{\hat{\mu}+\frac{\hat{\sigma}^2}{2}+1.645\sqrt{\frac{\hat{\sigma}^2}{n}+\frac{\hat{\sigma}^4}{2(n+1)}}}$$

in which 1.645 is the 95 percentile of the standard normal cumulative distribution function.

This equation can be used as long as $\hat{\mu}+\frac{\hat{\sigma}^2}{2}$ is approximately normally distributed. For sample sizes of 1000 or more, this holds by the Central Limit Theorem. Unless standard deviations are unusual, Land's method and Cox's method generally yield similar 95% UCLs for a sample size of 1000. As noted, the macro

logconf automatically applies the method appropriate for the data. The **logconf** macro simplifies the calculation of summary statistics of lognormal data that are not censored by eliminating the present need of looking up tabulated values and of interpolating between them. If there are extensive data, and if they contain no more than a few nondetects, it would still be of interest to apply the macro **logconf** to see how its results compare to the results from the applications (described in detail below) that are preferred when data are left censored.

4. CALCULATION OF THE MEAN, CONFIDENCE LIMITS, AND TOLERANCE BOUNDS THAT APPLY WHEN DATA ARE EITHER UNCENSORED OR LEFT CENSORED

4.1 LOGNORMAL AND NORMAL MODEL ESTIMATES

This section deals specifically with left-censored data that fit either a normal or a lognormal distribution. If examination of the data shows them to be consistent with the normal distribution, most of the derivations shown below apply without transforming the data logarithmically. On the other hand, if the data appear to be lognormal, then the derivations are exactly as shown below. Estimation under the normal model is simpler and will not be discussed here. For a discussion of estimation under both models, see Lawless (1982).

Regardless of whether the normal or lognormal approach is used, it seems advisable to apply the product limit estimate methods (using SAS macro **ple** described in the next subsection) for comparison with other values when there is an appreciable number of nondetects. The difficulty of determining the underlying distribution of the data becomes more extreme when the data contain many nondetects. Application of the PLE avoids the problem of tying the statistical estimates to a particular distribution when the true distribution may be unknown. The lognormal distribution is the most commonly used distribution for modeling environmental contaminant data (EPA 1995b). As will be shown below, in some situations the lognormal estimates may appear absurdly large, and in other instances they may not seem conservative enough (i.e., the PLE 95% UCL is much higher).

The following discussion points out the statistical foundation for the method applied in the macro named **lnor**. While this method has some similarity to what Land (1971) called "Cox's Method", it goes beyond that method by providing a mechanism to account for nondetects. While the macro is suitable when the data are left censored, it can also be applied when there are not any nondetects. However, for uncensored data, known to be lognormal, for which the sample sizes are smaller than 1001, the macro **logconf** should be applied.

It should be noted that the macro **lnor** calculates both the lognormal and normal model estimates. As will be described in detail later, the actual concentration values (not their logs) are input to the macro. To derive the lognormal estimates, the macro, of course, begins by calculating the logarithms. For the normal model estimates, it does the calculation without a transformation to logarithms. While comparison of the results of the two methods is useful, knowledge about the underlying distribution, when available, indicates which result is applicable.

Again let μ denote the log-scale mean, and let σ^2 denote the log-scale variance. Because the lognormal mean is $e^{\mu + \frac{\sigma^2}{2}}$, if (L, U) is a $1-\alpha$ confidence interval for $\mu+\sigma^2/2$, $e^{(L, U)}$ is a $1-\alpha$ confidence interval for

the lognormal mean. This argument provides the basis, in both the left-censored and all-detects cases, for the lognormal confidence limits and tolerance bounds discussed here. For the case of all detects, the approach is called Cox's direct method (Land 1972). (This approach was developed in the previous section for application in the macro **logconf** when the sample size exceeds 1001.)

First consider the case of all detects. Let $\hat{\mu}$ and $\hat{\sigma}^2$ denote the sample mean and variance of the data after logarithms have been taken. Then $\hat{\mu}$ and $\hat{\sigma}^2$ are known to be optimal in a sense: they are unbiased complete sufficient statistics for μ and σ^2 (see, for example, Wilks 1962), and $\hat{\mu}+\hat{\sigma}^2/2$ is the minimum variance unbiased estimate (MVUE) of $\mu+\sigma^2/2$.

The variance of $\hat{\mu} + \hat{\sigma}^2/2$ is $Var(\hat{\mu})+Var(\hat{\sigma}^2/2)$ because $\hat{\mu}$ and $\hat{\sigma}^2$ are (statistically) independent. It is known that $(n-1)\hat{\sigma}^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom. From this it can be shown that $Var(\hat{\sigma}^2) = 2\sigma^4/(n-1)$, and thus

$$Var\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) = \frac{\sigma^2}{n} + \frac{\sigma^4}{2(n-1)}.$$

In the above equation, $\frac{\sigma^2}{n} = Var(\hat{\mu})$ and $\frac{\sigma^4}{2(n-1)} = Var(\frac{\hat{\sigma}^2}{2})$. Because as logarithms the data are normally distributed, the sample mean and variance are independent and there is no need to deal with an estimate of the covariance.

Because $E[\hat{\sigma}^4] = (n+1)\sigma^4/(n-1)$, the value of $(n-1)\hat{\sigma}^4/(n+1)$ is unbiased for σ^4, and

$$\frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^4}{2(n+1)}$$

becomes the MVUE of $Var(\hat{\mu}+\hat{\sigma}^2/2)$. In Cox's direct method, this variance estimate is then used with $\hat{\mu}+\hat{\sigma}^2/2$ (i.e., the point estimate) to compute confidence limits for, or tests about, $\mu+\sigma^2/2$. By the Central Limit Theorem, the distribution of $\hat{\mu}+\hat{\sigma}^2/2$ is approximately normal. The confidence limits (when expressed as logarithms) are symmetrical about the mean just as they are for the non-logarithmic methods. The confidence limits reported in the output are asymmetrical because they are found by taking the antilogarithms.

With left-censored data, an analogous approach is used. Let x_i ,..., x_n denote the n observations with χ_i denoting the detection limit for the i^{th} observation if it is a nondetect. The maximum likelihood estimates (MLEs) of μ and σ are computed by maximizing the likelihood (L) of the following equation

$$L = \prod_{x \text{ detect}} \frac{1}{\sigma x}\phi\left(\frac{\log(x)-\mu}{\sigma}\right) \prod_{x \text{ detection limit}} \Phi\left(\frac{\log(x)-\mu}{\sigma}\right),$$

where ϕ and Φ are the standard normal density and distribution function (see Lawless). The values of μ and σ that result in the maximum value of L are called the maximum likelihood estimates (MLEs), and they are called $\tilde{\mu}$ and $\tilde{\sigma}$, respectively. The SAS Lifereg procedure (SAS/STAT PROC LIFEREG) computes these MLE's.

When there is censoring, $\tilde{\mu}$ and $\tilde{\sigma}$ are not independent, and thus the covariance of $\tilde{\mu}$ and $\tilde{\sigma}$ is nonzero. The variances and covariance of $\tilde{\mu}$ and $\tilde{\sigma}$ can be estimated by inverting the *information matrix*.

$$-\begin{pmatrix} \dfrac{\partial^2 \log(L)}{\partial \mu^2} & \dfrac{\partial^2 \log(L)}{\partial \mu \partial \sigma} \\ \dfrac{\partial^2 \log(L)}{\partial \mu \partial \sigma} & \dfrac{\partial^2 \log(L)}{\partial \sigma^2} \end{pmatrix}_{\tilde{\mu},\, \tilde{\sigma}}.$$

The inverse of the above information matrix provides an estimate of the covariance matrix of the MLE's (see Wilks 1962). Denoting the parts of the above information matrix with letters as follows

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix},$$

it can be shown by direct matrix multiplication that its inverse is

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix} = \frac{1}{ac - b^2} \begin{pmatrix} c & -b \\ -b & a \end{pmatrix}.$$

The SAS Lifereg procedure is used to compute this inverse, which provides the estimate of the covariance matrix of $\tilde{\mu}$ and $\tilde{\sigma}$. With this information on the variability of the estimates, it is possible to calculate confidence limits.

The goal is to estimate $\mu + \sigma^2/2$. The variance of $\tilde{\sigma}$ can be estimated as above (using variable C). However, the variance of $\tilde{\sigma}^2$ is needed in order to compute confidence limits for $\mu + \sigma^2/2$. To compute the variance of $\tilde{\sigma}^2$, it is necessary to move beyond the regular SAS routines and to parameterize the likelihood, alternatively, in terms of μ and τ instead of μ and σ, where $\tau = \sigma^2$. The MLE $\tilde{\tau}$ is just $\tilde{\sigma}^2$, and the covariance estimates for the MLE's $\tilde{\mu}$ and $\tilde{\tau}$ can be inferred from the μ and σ parameterization results using the chain rule of partial differentiation:

$$\frac{\partial^2 \log(L)}{\partial \mu \partial \tau} = \frac{\partial^2 \log(L)}{\partial \mu \partial \sigma} \frac{\partial \sigma}{\partial \tau} = \frac{1}{2} \frac{\partial^2 \log(L)}{\partial \mu \partial \sigma} \tau^{-1/2},$$

and

$$\frac{\partial^2 \log(L)}{\partial \tau^2} = \frac{\partial^2 \log(L)}{\partial \sigma^2}\frac{1}{4}\tau^{-1} - \frac{\partial \log(L)}{\partial \sigma}\frac{1}{4}\tau^{-3/2}.$$

(Notice that $\partial^2\log(L)/\partial\tau^2$ reduces to $(\partial^2\log(L)/\partial\sigma^2)\tau^{-1}/4$ at the MLE because $\frac{\partial \log(L)}{\partial \sigma}$ is zero.) In this way the following covariance matrix estimate

$$\begin{pmatrix} A & D \\ D & G \end{pmatrix}$$

is derived for the MLE's $\tilde{\mu}$ and $\tilde{\tau}$, where $D=2\tilde{\tau}^{-1/2}B$ and $G=4\tilde{\tau}^{1}C$.

It is straightforward to verify that in the uncensored case

$$\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \tilde{\tau} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2,$$

$A = \tilde{\tau}/n$, and $G = 2\tilde{\tau}^2/n$. (B and D are 0 in the uncensored case.)

However, because $E(\tilde{\tau}) = (n-1)\tau/n$, and $E(\tilde{\tau}^2) = (n^2-1)\tau^2/n^2$ (from similar results for $\hat{\sigma}^2$), the adjusted estimate of $\tau = \sigma^2$, $n\tilde{\tau}/(n-1)$, is used instead of $\tilde{\tau}$, the adjusted estimate $nA/(n-1)$ is used instead of A, and the adjusted estimate $[n/(n-1)][n^2/(n^2-1)]G$ is used instead of G. (Then the expectation of G becomes $2\tau^2/(n-1)$, which is the variance of $n\tilde{\tau}/(n-1)$.) In order that the censored case reduces continuously to Cox's direct method as the detection limits approach 0, these same adjustments are made in the censored case as well. Then, $\mu+\sigma^2/2$ is estimated by $\tilde{\mu}+\tilde{\tau}/2$, and its variance,

$$Var\left(\tilde{\mu} + \frac{\tilde{\tau}}{2}\right) = Var(\tilde{\mu}) + \frac{1}{4}Var(\tilde{\tau}) + Cov(\tilde{\mu}, \tilde{\tau}),$$

is estimated by the sum of the terms $A + G/4 + D$, which correspond to the three terms on the right side of the equation above.

Tolerance bounds, that is, confidence limits for lognormal quantiles, can be computed in much the same way. Let z_p denote the p^{th} quantile of the standard normal distribution. On the log scale, the p^{th} quantile is $\mu + z_p\sigma$, which can be estimated by plugging in estimates of μ and σ. Tolerance bounds for the estimate $\tilde{\mu} + z_p\tilde{\sigma}$ can be computed using an estimate of

$$Var(\tilde{\mu} + z_p\tilde{\sigma}) = Var(\tilde{\mu}) + z_p^2 Var(\tilde{\sigma}) + 2\,z_p Cov(\tilde{\mu}, \tilde{\sigma}),$$

namely $A + z_p^2 C + 2\,z_p B$, and by treating $\tilde{\mu} + z_p\tilde{\sigma}$ as (approximately) normal. Tolerance bounds for the same quantiles on the original scale can then be obtained by exponentiating the log-scale confidence bounds as follows:

$$e^{\tilde{\mu} + z_\alpha\tilde{\sigma} + t_{(1-\alpha)}\sqrt{A + z_p^2 B + 2z_p C}},$$

where $1-\alpha$ is the confidence level, and $t_{(1-\alpha)}$ is the $1-\alpha$ quantile of the *t*-distribution (with $n-1$ degrees of freedom). For $p = .50$, $z_p = 0$, and the quantiles (on either scale) are medians. When there are only detects, the median estimate on the original scale is the geometric mean (on the log scale).

To illustrate how quantile estimates and tolerance bounds can be used, confidence limits will be computed for the underlying lognormal distribution itself. Suppose that X_p is the p^{th} quantile (on the original scale), and that $U(p)$ is a $1-\alpha$ upper confidence limit for X_p. (Note that the 90^{th} percentile is the same as the 0.9^{th} quantile.) Then

$$P(X_p \le U(p)) = 1-\alpha,$$

where, P is the probability. From this, a lower confidence limit for $F(x)$, the cumulative distribution function at x, can be derived as follows.

$$P[\, x \le U(F(x))\,] = 1-\alpha,$$

hence

$$P[\, U^{-1}(x) \le F(x)\,] = 1-\alpha,$$

where U^{-1} denotes the (functional) inverse of U. That is, $U^{-1}(x)$ is a $1-\alpha$ lower tolerance bound for $F(x)$. Likewise, upper tolerance bounds for $F(x)$ can be derived from the lower tolerance bounds for X_p.

On the log scale, tolerance bounds for X_p are of the form

$$TB(p) = \tilde{\mu} + z_p\tilde{\sigma} \pm t_{1-\alpha}\sqrt{A + z_p^2 B + 2z_p C}.$$

where TB(p) stands for the tolerance bound for X_p. This reduces algebraically to

$$[\tilde{\sigma}^2 - Ct_{1-\alpha}^2]z_p^2 - 2[(TB(p) - \tilde{\mu})\tilde{\sigma} + t_{1-\alpha}^2 B]z_p + (TB(p) - \tilde{\mu})^2 - t_{1-\alpha}^2 A = 0,$$

which is a quadratic equation in z_p. The two solutions are

$$\frac{-g - \sqrt{g^2 - 4dh}}{2d} \text{ and } \frac{-g + \sqrt{g^2 - 4dh}}{2d}$$

where $d = \tilde{\sigma}^2 - Ct_{1-\alpha}^2$, $g = -2[(TB(p) - \tilde{\mu})\tilde{\sigma} + t_{1-\alpha}^2 B]$, and $h = (TB(p) - \tilde{\mu})^2 - t_{1-\alpha}^2 A$. The following shows how to invert TB, and thus how to compute tolerance bounds for $F(x)$: substitute $\log(x)$ for $TB(p)$ in g and h, and compute

$$\Phi\left[\frac{-g - \sqrt{g^2 - 4dh}}{2d}\right] \text{ and } \Phi\left[\frac{-g + \sqrt{g^2 - 4dh}}{2d}\right],$$

which are the lower and upper $1-\alpha$ confidence bounds for $F(x)$, respectively. These tolerance bounds are illustrated in Figure 1 for lead in groundwater in the Upper East Fork Poplar Creek watershed. These data are discussed more in section 6.

4.2 PRODUCT LIMIT ESTIMATES

The product limit estimate (PLE) is a statistical distribution function estimate, like the sample distribution function, except that the PLE adjusts for censoring. Like the sample distribution function, the PLE is not premised on any underlying distribution model, and the PLE reduces to the sample distribution function as detection limits (for nondetects) approach zero. As discussed below, the PLE can be used to compute mean estimates as well as standard errors for those estimates. Besides being a good mean estimate in its own right, the PLE-based mean estimate and its standard error provide a good reference for parametric (e.g., lognormal) estimates. It is often impossible to know what underlying distribution is appropriate for a particular data analysis. Goodness-of-fit tests are often used to determine whether data fit a particular distribution. When samples are small, however, as is often the case for environmental data, it is impossible to test adequately whether the data fit a particular distribution. Another well-known problem with such tests is that they have a tendency to reject even adequate models when the sample size is large. For these reasons, it is a good idea to compare any parametric estimate with the comparable PLE-based estimate.

The PLE is calculated using the macro **ple** from a set of observations, which are measurements in the case of detects and detection limits in the case of nondetects. Even if detection limits are known for detected values, they are not used in calculations. It is important to keep in mind that the macro PLE is adapted for left-censored data and non-censored data, but not for right-censored data.

The method of the PLE and confidence limits will be presented in two ways, first mathematically and then with a simple example that is explained geometrically. This approach is being taken because the underlying concept is straightforward, and those features can be overlooked if there is only a mathematical presentation, which of necessity requires a large number of variables to explain. The mathematical presentation is important for completeness and because, for some, it will be easy to understand.

4.2.1 Mathematical Explanation of the PLE and its Application

The PLE, which is called $\tilde{F}$ below, is defined as follows. For n observations, $x_1, \ldots, x_n$ (detection limits or actual measurements), let $x_1' < \ldots < x_{n'}'$ denote the (say) n' distinct values at which detects are observed. For $j = 1, \ldots, n'$, let m_j denote the number of detects at x_j', and let n_j denote the number of $x_i \le x_j'$. Also let $x_{(1)}$ denote the smallest x_i. Then for $x \ge x_{n'}'$, $\tilde{F}(x) = 1.$; for $x_1' \le x < x_{n'}'$,

$$\tilde{F}(x) = \prod_{j \text{ such that } x_j' > x} \frac{n_j - m_j}{n_j};$$

for $x_{(1)} \le x < x_1'$, $\tilde{F}(x) = \tilde{F}(x_1')$; and for $0 \le x < x_{(1)}$, $\tilde{F}(x)$ is either 0 or undefined, the latter if there is a nondetect at $x_{(1)}$.

To understand the PLE, it is necessary to understand the concept of conditional probability. Let $P(A|B)$—read "probability of A given B"—denote the conditional probability of event A given event B. Introductory probability texts explain that

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

where $P(A \cap B)$ is the probability that the events *A and B* both occur.

As its name suggests, the PLE is a limit of a product of probabilities. Consider k arbitrary points $0 \le y_1 \le \ldots \le y_k$. Then for $i = 1, \ldots, k$, $F(y_i) =$

$$P(X \le y_i) = P(X \le y_i | X \le y_{i+1}) \times P(X \le y_{i+1}) = P(X \le y_i | X \le y_{i+1}) \times P(X \le y_{i+1} | X \le y_{i+2}) \times P(X \le y_{i+2})$$
$$= \ldots = P(X \le y_i | X \le y_{i+1}) \times P(X \le y_{i+1} | X \le y_{i+2}) \times \ldots \times P(X \le y_{k-1} | X \le y_k) \times P(X \le y_k).$$

Consider estimating $F(y_i)$. When there is no censoring, the proportion of observations less than or equal to y_i is used. When there is censoring, however, that proportion may be indeterminate because it is ambiguous whether any nondetects with detection limits greater than y_i actually exceed y_i. To account for censoring, the individual factors in the above product are estimated.

For each j, let c_j denote the number of $x_i \le y_j$ unambiguously. Only nondetects whose detection limits are less than or equal to y_j and detects whose values are less than or equal to y_j are counted in c_j. Let d_j be the number of detects between y_{j-1} (exclusive) and y_j (inclusive). To estimate $P(X \le y_j | X \le y_{j+1})$, c_j / c_{j+1} might be used. That would be a good estimate unless there are nondetects with detection limits between y_j (inclusive) and y_{j+1} (exclusive). Then $c_{j+1} > c_j$ even if the actual (but censored) values corresponding to the detection limits are all less than y_j (as well as y_{j+1}) and d_{j+1} is 0.

Instead, to estimate $P(X \le y_j | X \le y_{j+1})$, it is better to use $(c_{j+1} - d_{j+1})/c_{j+1}$. That is, those nondetects are ignored whose detection limit is between y_j and y_{j+1}, because their actual values relative to y_j are ambiguous. Note that $c_{j+1} - d_{j+1}$ is c_j, except possibly for any nondetects whose detection limits are between y_j and y_{j+1}. To minimize the effect of nondetects whose detection limits are between y_j and y_{j+1}, the y_j partition is made as fine as possible. That is, y_k is taken to be the largest detect value, y_0 is set to 0, k is taken to be large (i.e., $k \to \infty$), and the y_i's are taken so that

$$\max_{j=1,\ldots,k} |y_j - y_{j-1}| \to 0.$$

It can be shown (see Kaplan and Meier 1958) that in the limit as $k \to \infty$, the estimate obtained by substituting these estimates for the individual factors in the product expression for $F(y_i)$, is the PLE.

As with the sample distribution function, a mean estimate can be computed from the PLE:

$$\hat{\mu} = \sum_{i=1}^{n'} x_i' [\tilde{F}(x_i') - \tilde{F}(x_{i-1}')],$$

where $x_0 = 0$. An estimate $\hat{V}$ of the variance of $\hat{\mu}$ can be determined similarly as follows

$$\hat{V} = \frac{D}{D-1} \sum_{i=1}^{n'-1} a_i^2 \frac{m_{i+1}}{n_{i+1}(n_{i+1} - m_{i+1})},$$

where D is the total number of detects, and

$$a_i = \sum_{j=1}^{i} (x_{j+1}' - x_j')\tilde{F}(x_j'),$$

for $i = 1, \ldots, n'-1$. For details see Kaplan and Meier (1958, Sections 2.3 and 6.2). A geometric interpretation of $\hat{V}$ is given in the next section.

Once the variance of the mean has been determined, the calculation of the confidence limits, as shown below, resembles that of the ordinary mean, except that there is no need to divide the square root of the variance by the square root of the sample size to obtain the standard error of the mean. This is because the method of calculation of $\hat{V}$already accounts for the number of samples. Thus, in the output of the macros, the "PLE mean standard error" is simply $\sqrt{\hat{V}}$.

$$95\%\ UCL = \hat{\mu} + t_{n-1}(0.95)\sqrt{\hat{V}}$$

$$95\%\ LCL = \hat{\mu} - t_{n-1}(0.95)\sqrt{\hat{V}}$$

The values multiplied by the standard error of the mean are taken from the *t*-distribution with n-1 degrees of freedom, just as they were when calculating confidence limits for the ordinary mean.

Thus far attention has been on the PLE for mean estimation; however, the PLE itself is a distribution function estimate. Confidence limits for the PLE as a distribution function estimate are discussed in Kaplan and Meier (1958, Sections 2.2 and 6.1) and Lawless (1982, Section 2.3.2). The confidence limits are derived from an estimate of the variance of the PLE at each x. The variance estimate (in the left-censored case) is

$$\hat{Var}(\tilde{F}(x)) = \tilde{F}(x)^2 \sum_{j \text{ such that } x_j' > x} \frac{m_j}{n_j(n_j - m_j)}.$$

Confidence limits for $F(x)$, based on this variance estimate, are computed in the usual way (i.e., by assuming $\tilde{F}(x)$ to be approximately normal). $\tilde{F}$ and 95% tolerance bounds for F are illustrated in Figure 4 for the groundwater lead data. (The same data are considered in Figure 3 based on macro **lnor**.)

Figures 3 and 4 contrast the lognormal and PLE cumulative distribution function estimates. The lognormal (parametric) analysis yields a smooth curve, which contrasts sharply with the nonparametric PLE analysis which yields a granular step-function. The granularity of the PLE is especially striking when the sample size (or the number of distinct values at which detects occur) is small. The granularity should not be considered a drawback unless the smooth alternative is also better, which would be true only if the underlying distribution is approximately lognormal.

Recall that for the lognormal model tolerance bounds were derived from which it was possible to compute confidence limits for the estimates for the probability distribution function F. For the PLE, steps are carried out in reverse order. That is, both quantile estimates and tolerance bounds are derived from the distribution function estimate (i.e., the PLE) and confidence bounds just discussed. For the p^{th} quantile X_p, and for $U(x)$, the $1-\alpha$ upper confidence bound for $F(x)$

$$P[\, F(X_p) \le U(X_p)\,] = P[\, p \le U(X_p)\,] = P[U^{-1}(p) \le X_p\,] = 1-\alpha.$$

That is, $U^{-1}(p)$ is a $1-\alpha$ lower tolerance bound for X_p. Figure 2 shows clearly that the PLE and its confidence limits are step functions—that is, they have flat spots. For this reason their inverses, which are needed to calculate tolerance bounds, are not uniquely defined. To invert a function with flat spots (i.e., to find the value on the x-axis that corresponds to the value p on the y-axis), it is necessary to choose from multiple values (i.e., from the inverse image). The following rule was applied to deal with this problem in the calculation of quantile estimates and confidence tolerance bounds in the **ple** macro. When the choice is to result in a lower tolerance bound (for X_p), the smallest value is chosen; when the choice is to result in an upper tolerance bound, the largest value is chosen. When the choice is to result in a point estimate, the average of the largest and smallest values is taken. This approach is conservative in that it leads to the widest tolerance bounds.

4.2.2 Simple Example of Calculation of the PLE

In this section the steps for computing the PLE mean, its standard error, and its 95% confidence limits are demonstrated using an example. The following randomly sampled measurements of an analyte were reported (concentrations in μgram of analyte/gram of soil): 0.10U, .20, 1.30, 0.70, 0.40, 0.70, 0.10U, 0.26, 0.31U, 0.80, and 1.10. The U following a number indicates that it is a nondetect. To compute the PLE, first arrange the data in decreasing order, as shown below.

1.30
1.10
.80
.70
.70
.40
.31U
.26

.20
.10U
.10U

These data are shown in Table 1, with one row for each of the measurements for which a detected measurement was found. The lowest concentration was included even though it was a nondetect.

Table 1. The steps in the calculation of the product limit estimate (PLE)

Concentration of sample, in μgram/gram of soil	(B) No. of detects or nondetects ≤ this concentration	(C) No. of detects at this concentration	B - C / B	Terms multiplied together to to calculate the PLE	Value of PLE just to the left of the concentration
1.30	11	1	10/11 = 0.909	1 x 0.909	0.909
1.10	10	1	9/10 = 0.9	0.909 x 0.9	0.818
0.80	9	1	8/9 = 0.889	0.818 x 0.889	0.727
0.70	8	2	6/8 = 0.75	0.727 x 0.75	0.545
0.40	6	1	5/6 = 0.833	0.545 x 0.833	0.454
0.26	4	1	3/4 = 0.75	0.454 x 0.75	0.341
0.20	3	1	2/3 = 0.667	0.341 x 0.667	0.227
0.10U	2	0	2/2 = 1	0.227 x 1	0.*

*When the smallest concentration is a detect, this calculated value is 0. Otherwise it must be set equal to zero, as it has been here.

Values from Table 1 should then be graphed, as shown in Fig. 1. The values on the X-axis come from the first column in Table 1, and the values on the Y-axis from the last column.

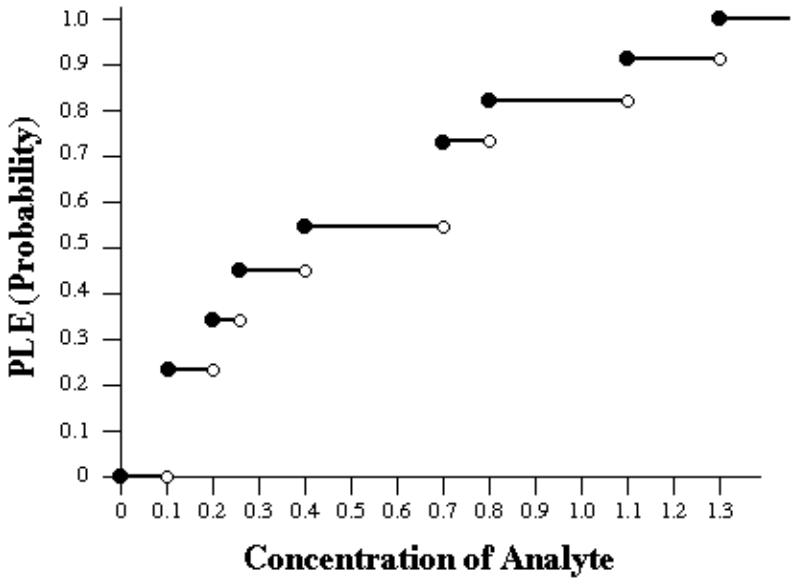

Fig. 1. The step function which is the cumulative distribution function for the sample used in the example.

The value of $\hat{\mu}$ is calculated by finding the area between the step function and a line drawn horizontally above the step function at the probability value of 1.0, as shown below. The value of both the SE and the 95% confidence limits of $\hat{\mu}$ are calculated using the steps which are explicitly explained below. Begin by drawing the rectangles shown in Fig. 2 using the same figure constructed above.

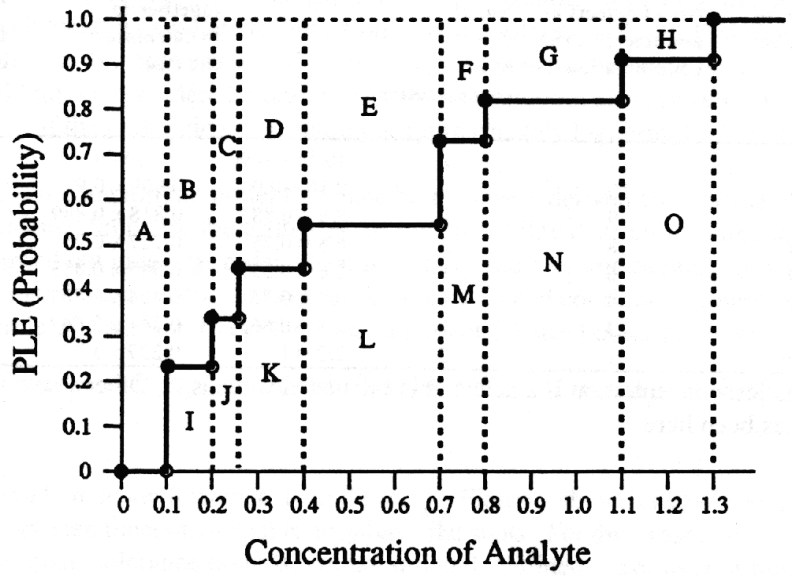

Fig. 2. The step function shown in Fig. 1 with the addition of labeled rectangles.

The steps to be carried out are as follows.

1. Calculate the areas of all rectangles.
2. The $\hat{\mu}$ of the PLE is equal to the sum of all rectangles A-H, that is, of all rectangles above the step function.

3a. If the smallest concentration is a **nondetect**, as in the example above, find the areas of the following rectangles or groups of rectangles (all of those below the step function except for the left-most one, which in this figure is I): J, J+K, J+K+L, J+K+L+M, J+K+L+M+N, and lastly for this particular step function, J+K+L+M+N+O.

or

3b. If the smallest concentration is a **detect**, find the areas of the following rectangles or groups of rectangles (all of those below the step function): I, I+J, I+J+K, I+J+K+L, I+J+K+L+M, I+J+K+L+M+N, and lastly for this particular step function, I+J+K+L+M+N+O.

4. Square the grouped areas.
5. When the smallest concentration is a **nondetect**, as in the example, start with rectangle J (i.e, the next to the left-most rectangle under the step function), divide each squared area calculated in step 3 by the product of B x (B-C) found in Table 1 for the row of the PLE (shown in the last column) that matches the top of the tallest rectangle included in the area. For example, the top of rectangle J is at the PLE of 0.341. Since B is 4 and C is 1 in the row for the PLE of 0.341, divide the squared area of rectangle J by 12, which is the product of 4 times 3, which is 4 x (4 - 1). Working toward the top of the Table 1, the top of the combined rectangles (J+K)

is at the PLE of 0.454. Following the same rule, divide the squared area of combined rectangles (J+K) by 30. Continue this procedure until reaching the top of Table 1. If the smallest concentration is a detect, the procedure is identical except that you begin with the left-most rectangle under the step function, or, in a figure like this one, with rectangle I.

6. Multiply each of the resulting quantities by the value of C from the same row of the same PLE in Table 1. In this example, C is always 1 except in the row of PLE 0.545, for which C is 2. This means that the quantity calculated in step 5 for combined rectangles J+K+L must be multiplied by 2.
7. The sum of all terms found in step 6 is the variance uncorrected for the degrees of freedom.
8. The degrees of freedom correction involves multiplying the uncorrected variance by the number of detects and dividing by (the number of detects - 1).
9. The resulting term is $\hat{V}$.
10. Take the square root of $\hat{V}$ to obtain the SE of $\hat{\mu}$.
11. Use the *t*-table to find the correct number to multiply times the SE to calculate the 95% LCL and UCL. One of many places where this table can be found is on page A-11 of EPA (1995b). Select the value for column .95 on the row for the proper number of degrees of freedom. To be consistent with the macro **ple**, the number of degrees of freedom for this step should equal the total sample size minus 1. That is, this step is not restricted to the number of detects.

Tables 2 and 3 show the calculations for the above steps for the example.

Table 2. Calculations of areas of rectangles in Figure 2

Rectangle	Height (units)	Width (units)	Area (units-squared)
A	1	.1	.1
B	.773	.1	.0772
C	.659	.06	.0395
D	.545	.14	.0764
E	.455	.3	.1364
F	.273	.1	.0273
G	.182	.3	.0545
H	.091	.2	.0182
I	.227	.1	.0227
J	.341	.06	.0205
K	.454	.14	.0640
L	.545	.3	.1636
M	.727	.1	.0727
N	.818	.3	.2455
O	.909	.2	.1818

The sum of the areas of all rectangles above the step function, which in this example includes rectangles A-H, is 0.530. This sum is the $\hat{\mu}$ of the PLE.

Table 3. Remaining steps involved in the calculation of the SE of the mean of the PLE

Rectangles included in area	A = area (units squared)	A^2	(D) This is product of detects from Table 1	(C) No. of numbers shown	$A^2 \times C$
J	.0205	.00042	4 x 3	1	.000035
J+K	.0841	.00071	6 x 5	1	.000236
J+K+L	.2477	.06137	8 x 6	2	.002557
J+K+L+M	.3205	.10269	9 x 8	1	.001426
J+K+L+M+N	.5659	.32025	10 x 9	1	.003558
J+K+L+M+N+O	.7477	.55910	11 x 10	1	.005083

Sum of the values in the right column of Table 3 = uncorrected $\hat{V}$ = .012895
Degrees of freedom correction = 8 ÷ (8-1) = 1.1429

$\hat{V}$ = 1.1429 x .012895 = 0.01474

SE = $\sqrt{\hat{V}}$ = $\sqrt{0.01474}$ = 0.1214

The sample size was 11. There are thus 10 degrees of freedom. To calculate the 95% confidence limits, the appropriate term from the *t*-distribution is 1.812.

Accordingly,

$$95\%\ UCL = 0.530 + (1.812\ x\ 0.1214) = 0.530 + 0.220 = 0.750$$

$$95\%\ LCL = 0.530 - (1.812\ x\ 0.1214) = 0.530 - 0.220 = 0.310$$

It is interesting to note that when there are no nondetects present in a sample, the ordinary mean and its confidence limits are identical to those computed for the PLE.

5. USING THE SAS MACROS

The objective here is to discuss how to use the three SAS macros described in the report. Examples are presented in Section 6. The macros themselves are in Appendices A, B, and C.. It is assumed that the user has a basic familiarity with SAS, particularly in regard to what SAS data sets are like. For an introduction to SAS see SAS (1990).

To use the SAS macros, a SAS data set containing the input variables must be created. The variables are (1) a *result* variable, which is the analysis result, or in the case of a nondetect, the detection limit; (2) a

qualifier variable, "U" or blank, which indicates whether the observation is a nondetect ("U") or detect (blank) ; (3) *group* variables (if any), which define the groups for which statistics are to be computed; (4) a *parameter* variable, which names the chemical analyte to which the data apply (e.g., Aluminum); and (5) any *ID* variables that the user may wish to carry along with the statistics computed for each group and analyte.

For the **lnor** macro only, an additional variable is needed. It is termed *lower*, and it provides a lower bound for each observation. Usually the value of *lower* is 0 for a non-detect and identical to the *result* variable for a detect. In certain cases, however, the value for *lower* might reasonably be taken as neither 0 nor the *result*. This occurs, for example when several duplicates at a site are to be combined into a single observation for that site, with the recorded observation being the (mathematical) composite observation. To illustrate, suppose there are two observations at a site, one a detect, say D, and the other, say L, the detection limit for a nondetect. The detect is a single value, but the nondetect, if not for the censoring, could have been any value between 0 and L. Therefore, if not for the censoring, the average of the two observations could have been anywhere between $D/2$ and $(D+L)/2$. The average (or composite) is between $D/2$ and $(D+L)/2$, and is said to be interval censored. In the **lnor** macro, the values of $(D+L)/2$ and $D/2$ are assigned to the variables *result* and *lower*, respectively. When the macro **lnor** is applied to interval censored data, it applies these two variables when it uses the SAS Lifereg procedure.

The SAS macros must be accessible from the SAS program that calls them. They can either be included as part of the program code, or they can be called using the SAS autocall facility. To use the autocall facility, include

sasautos = 'directory'

in a SAS options statement. The macros should be in files in the directory named, with file name **lnor**.sas (for the **lnor** macro), **ple**.sas (for the **ple** macro), or **logconf**.sas (for the **logconf** macro).

To use the macros in a SAS program, include the statements

%lnor(input, output, group, result, lower, qual, parm, id, confid, toler),
or
%ple(input, output, group, result, qual, parm, id, confid, toler),
or
%logconf(input, output, group, result, parm, id).

Here "input" is the input data set, which can have any valid data set name, not just "input"; "output" refers to the name given to the data set of summary statistics computed with the macro (again, any name, not just "output"), and so on. The term "confid" refers to the desired confidence level for upper and lower confidence bounds, for example, .95. For "toler" it is necessary to substitute a space-delimited list of values for which quantile estimates and tolerance bounds are to be computed (e.g., .75 .90 .99).

The calling of the macros is illustrated in the next section. At present macro **logconf** computes confidence limits of means and tolerance bounds of percentiles at only the 95% level. In order to run the macro **logconf,** the user must also have a copy of, or access to, the **hfun** SAS data set. A copy of that data set can be found at the same website as this report.

6. AN EXAMPLE

Here the macros **logconf**, **lnor**, and **ple** are applied to several samples of lead concentrations (mg/L) from groundwater in wells at the Y-12 fuel station in the Upper East Fork Poplar Creek watershed at Oak Ridge, Tennessee. The data, which are in Table 4 (located at end of text), consist of time series, one for each station (i.e., well). It is assumed that there is no trend over time and that it is reasonable to treat the data as a random sample. The data illustrate the structure appropriate for input to the three SAS macros. The processing groups are the different stations. The U's indicate observations that are nondetects. Figures 3 and 4 (located at end of text) were computed from this data set for station ST-006. The lead concentration data used for illustration are taken from an early listing of the data, which was before they were subjected to careful quality assurance examination.

Table 5 (located at end of text) contains a description of the contents (SAS proc contents output) of the data set produced by calling **lnor**.sas for the data in Table 4. Table 6 (located at end of text) lists the output (i.e., the parameters and summary statistics) produced by the **lnor** macro for the data in Table 4. Table 7 (located at end of text) contains a description of the contents of the data set produced by calling **ple**.sas for the data in Table 4. Table 8 (at end of text) lists the output produced by the **ple** macro for the data in Table 4. The ordinary mean statistics found in tables containing summary statistics are computed by straightforward substitution of detection limits for nondetects and by proceeding with mean-and-standard-error calculations that are usual for the all-detects case.

Tables 9 (at end of text) contains additional data from groundwater in wells at the same source. These data, which lack nondetects, are in the format required by the **logconf** macro. Table 10 (at end of text) contains a description of the contents (SAS proc contents output) of the data set produced by calling **logconf**.sas for the Y-12 fuel station data. Table 11 (at end of text) lists the output produced by the **logconf** macro for the data in Table 9. The SAS program used to print the data in Tables 4 and 9, to call the SAS macros, and to produce Tables 5-8 and 10-11 is listed in Appendix D.

7. DISCUSSION

The three SAS macros reported here provide a variety of methods for computing confidence limits for environmental data. They also provide several approaches for dealing with left-censored data, which are common when making environmental measurements.

A recent publication (Schmoyer et al., 1996) has considerable relevance to the application of these programs. It addresses the uncertainty regarding whether the lognormal distribution is the best model for mean estimation of concentrations of analytes in the environment. The authors simulated data sets for lognormal, truncated normal, and gamma data for a range of sample sizes and coefficients of variation. They found that with the small sample sizes typical of environmental data, when using the Shapiro-Wilk test it was difficult to detect departures from lognormality that were important in the sense of appreciably degrading the performance of lognormal-based estimates and tests. They concluded that "(i) the lognormal distribution may be too heavy tailed to be a reasonable statistical model, and (ii) alternatives may be better than the lognormal-based methods." They pointed out that there is "usually no physical basis for lognormality, normality, or any other distribution." They concluded from their simulations that "lognormal-based statistics might not be as good as the ordinary sample mean and *t*-test, if there are complete (i.e. uncensored) data, or as good as means or tests

computed from the product limit estimate (PLE; Kaplan and Meier, 1958), when there is random left censoring."

Because the lognormal approach is standard in the analysis of concentrations of analytes in the environment, it would be remiss not to provide statistical methods for the lognormal distribution. The macros **lnor** and **logconf** serve this purpose. The implementation in the **lnor** macro of modifications, both for the lognormal and normal models, that address left-censoring of data make application of the lognormal and normal model approaches much more reliable for environmental data. These methods are more solidly grounded in theory than such methods as substituting 0 for nondetects, substituting the detection level divided by 2 for nondetects, or procedures which involve graphing the data and replacing the nondetects with values that fit assumed underlying distributions. The **logconf** macro simplifies the calculation of summary statistics for uncensored lognormal data by eliminating the need of looking up tabulated values and of interpolating between them.

A feature of the PLE that is of practical importance when dealing with left-censored data is that it works well even when there are nondetects at different detection levels that overlap detected concentrations. In some other methods, which are based on percentiles or trimming, measurements and detection limits cannot be arbitrarily interleaved.

As noted earlier, all of the macros except for **logconf** provide both upper and lower confidence limits and tolerance bounds. It is perhaps more obvious why the upper limits are important, and especially the 95% UCL, because of the EPA guidance that in risk analysis either the 95% UCL or the highest concentration reported should be used, depending on which is lower. It is important to realize that the LCLs also have important applications. For example, if the 95% LCL is less than zero, there is reason to question whether the contaminant is even present. In such a case, and if the 95% UCL were high enough to cause concern, the great width of the confidence limits would strongly suggest the need for additional sampling before recommending that action be take. A second example of the importance of having LCLs occurs in situations in which the 95% UCL is high enough to exceed an action level for a contaminant at a site. When that happens, if the 95% LCL also exceeds the action level, it is clear that action should be considered.

Some additional comparisons between the methods applied in the macros are worth considering when deciding which of the 95% UCLs of means are reasonable candidates as input concentrations in risk analysis. Ordinary means are conservative (upwardly biased) concentration estimates, and the 95% UCLs of the ordinary means are likewise conservative. They are upwardly biased because the nondetects are treated as actual concentrations, which means that some very small values are likely to be treated as substantially larger ones in the analysis. In the eight examples using actual environmental data, which were compared in this report for both the **lnor** and **ple** macros, this upward bias could be substantial because usually a high proportion of the samples were nondetects.

In the eight examples, the 95% UCL of the normal model mean calculated by the **lnor** macro, when compared to that of the ordinary mean (Table 6), was always smaller with a mean ratio (i.e., normal ÷ ordinary) between them being 0.679 with a minimum of 0.341 and a maximum of 0.923. In our examples, a lot of the normal-model estimates are negative, suggesting that the normal-model is not appropriate in this situation. (Indeed, 5 of the 8 estimates of the mean were negative.)

In the eight examples, the 95% UCLs of the PLE, which corrects for nondetection, were always smaller than those of the ordinary mean (Table 8) . The mean ratio of the 95% UCLs between them (i.e., ordinary ÷

PLE) was 1.38 with a minimum of 1.01 and a maximum of 2.80. Since the ordinary UCL is upwardly biased, ratios no larger than this provide little reason for concern that use of the 95% UCL of the PLE would lead to any important underestimation of risk.

Much larger differences were found between the 95% UCLs of the ordinary mean and the lognormal mean, as calculated using the **lnor** macro (Table 6). In the examples, the **median** ratio of the 95% UCLs between them (i.e., lognormal ÷ ordinary) was 3.86 with a minimum of 1.15 and a maximum of 6146. Three of the 8 examples had a ratio of 12.8 or higher. The lognormal 95% UCL at station ST-007, which yielded the ratio of 6146, is enormous compared to the ordinary UCL. It is important to realize that that estimate is not a mistake. Instead it illustrates a fundamental problem and practical difficulty of the lognormal model. The problem is that the right tail of the lognormal distribution is extremely heavy. For mean estimates and confidence limits, this leads to occasional anomalous results such as this. This difficulty is discussed in detail by Schmoyer et al. (1996), where—as noted earlier—advantages of the PLE over the lognormal approach are demonstrated in computer simulations. For the ST-007 UCL the problem can be seen (though not so glaringly) in the log-scale standard deviation (_LS_STD), which is much bigger in this case than for the other stations. Even the two ratios between 12 and 24 represent probable overestimates that could seriously impact risk estimates. This is because there is usually a linear relationship between exposure concentrations and risk estimates, and thus a decision to apply uncritically the lognormal estimate of the 95% UCL could result in risk estimates over an order of magnitude too high. There is always the possibility, however remote, in such a situation that the underlying distribution is lognormal, in which case the large estimate might be valid. For this reason, unless the lognormal distribution can be discounted on statistical grounds, the large estimate should be reported. However, there would seem to be good reason in such a case to apply instead in risk estimation the 95% UCL of the PLE mean, which, as shown above, is likely to have a magnitude much more similar to that of the 95% UCL of the ordinary mean.

In our example, for both the **ple** and **lnor** macros, 50, 75, and 90th percentiles were estimated. For the **lnor** macro, both lognormal and normal-based percentiles are calculated. These estimates and confidence (tolerance) bounds may be compared. The PLE estimates and tolerance bounds generally assume data values, and are thus coarser. In many cases, the upper tolerance bounds for the 90th percentiles cannot be computed at all. That happens when the lower confidence bound for PLE distribution estimate is below .90. It could happen for other quantiles as well.

In view of the many uncertainties in the calculation of means, confidence limits, and tolerance bounds for environmental data, it is advisable to compare the output for the three macros. When the data are left censored to more than a trivial extent, the **logconf** macro is inappropriate. Unless the data to be used in a risk assessment are clearly not lognormal, the results of the statistics computed for the lognormal distribution should be reported. However, it is strongly recommended that those confidence limits and tolerance bounds be compared with those found using the **ple** macro. In many cases the upper 95% UCL of the PLE mean would be the most appropriate concentration to apply in risk analysis.

The SAS macros described in this report should provide the basis for development of exposure concentrations for all ORO risk assessments in which the sample selection procedure emulates simple random sampling. See report BJC/OR-271 (Bechtel Jacobs 1999) for an overview of the data evaluation process.

8. REFERENCES

Bechtel Jacobs (Bechtel Jacobs Company LLC) 1999. *Guidance for Conducting Risk Assessments and Related Risk Activities for the DOE-ORO Environmental Management Program*, BJC/OR-271, Oak Ridge, TN.

Casella, G. and Berger R.L. 1990. *Statistical Inference*. Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.

EPA (U. S. Environmental Protection Agency) 1989. *Risk Assessment Guidance for Superfund, Volume I, Human Health Evaluation Manual (Part A)*, December 1989.

EPA (U. S. Environmental Protection Agency) 1992. *Supplemental Guidance to RAGS: Calculating the Concentration Term*, May 1992.

EPA (U. S. Environmental Protection Agency) 1995a. *Data Collection and Evaluation, Human Health Risk Assessment*, in *Supplemental Guidance to RAGS*: Region 4 Bulletins, Bulletin No. 1, November 1995.

EPA (U. S. Environmental Protection Agency) 1995b. *Guidance for Data Quality Assessment, EPA QA/G-9, External Working Draft*, March 27, 1995.

Kaplan, E.L. and Meier, P. 1958. "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, **53**, 457-481.

Land, C.E. 1971. "Confidence intervals for linear functions of the normal mean and variance," *Annals of Mathematical Statistics*, **42**, 1187-1205.

Land, C.E. 1972. "An evaluation of approximate confidence interval estimation methods for lognormal means," *Technometrics*, **14**, 145-158.

Lawless, J.F. 1982. *Statistical Models and Methods for Lifetime Data*. Wiley, New York.

Lyon, B. F. and Land C. E. 1999. *Computation of Confidence Limits for Linear Functions of the Normal Mean and Variance*, ORNL/TM–1999/245, September 1999, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

SAS 1990. SAS Institute Inc., *SAS® Language: Reference, Version 6, First Edition* (and other volumes), Cary, NC: SAS Institute Inc.

Schmoyer, R.L. et al. 1996. "Difficulties with the lognormal model in mean estimation and testing," *Environmental and Ecological Statistics*, **3**, pp. 81-97.

Wilks, S.S. 1962. *Mathematical Statistics*, Wiley, New York.

Table 4. Groundwater lead concentrations (mg/L) from Y-12 Fuel Station[1]

Monitoring station	Date collected	Analytical result/detection limit	lower (for lnor.sas)	Qualifier
ST-001	19JUN90	.0073	.0073	
	26SEP90	.0054	.0054	
	06DEC90	.0020	.	U
	08MAR91	.0020	.	U
	18JUN91	.0020	.	U
	25SEP91	.0070	.0070	
	14DEC91	.0020	.	U
	08MAR92	.0020	.	U
	07MAY92	.0020	.	U
	19AUG92	.0020	.	U
	09NOV92	.0020	.	U
	10MAR93	.0020	.	U
	21JUN93	.0020	.	U
	22SEP93	.0250	.	U
	16NOV93	.0250	.	U
ST-002	04MAY89	.0120	.0120	
	24AUG89	.0020	.	U
	03NOV89	.0054	.0054	
	26FEB90	.0020	.	U
	17MAY90	.0020	.	U
	04AUG90	.0020	.	U
	23OCT90	.0020	.	U
	24JAN91	.0020	.	U
	19APR91	.0078	.0078	
	29JUL91	.0200	.0200	
	09OCT91	.0020	.	U
	11JAN92	.0160	.0160	
	14APR92	.0020	.	U
	27JUL92	.0020	.	U
	20OCT92	.0020	.	U
	02FEB93	.0020	.	U
	16APR93	.0056	.0056	
	04AUG93	.0250	.	U
	14OCT93	.0250	.	U
ST-003	08MAY89	.0350	.0350	
	25AUG89	.0091	.0091	
	04NOV89	.0110	.0110	
	27FEB90	.0057	.0057	
	17MAY90	.0210	.0210	
	04AUG90	.0020	.	U
	23OCT90	.0042	.0042	
	25JAN91	.0020	.	U
	19APR91	.0020	.	U
	30JUL91	.0020	.	U
	09OCT91	.0067	.0067	
	11JAN92	.0020	.	U
	14APR92	.0020	.	U
	29JUL92	.0020	.	U

Table 4 (continued)

Monitoring station	Date collected	Analytical result/detection limit	lower (for lnor.sas)	Qualifier
ST-004	04MAY89	.0071	.0071	
	24AUG89	.0020	.	U
	04NOV89	.0020	.	U
	27FEB90	.0042	.0042	
	17MAY90	.0360	.0360	
	04AUG90	.0020	.	U
	23OCT90	.0020	.	U
	24JAN91	.0120	.0120	
	19APR91	.0020	.	U
	29JUL91	.0020	.	U
	09OCT91	.0046	.0046	
	11JAN92	.0020	.	U
	14APR92	.0020	.	U
	27JUL92	.0020	.	U
	22OCT92	.0071	.0071	
	03FEB93	.0020	.	U
	19APR93	.0020	.	U
	06AUG93	.0250	.	U
	14OCT93	.0250	.	U
ST-005	23OCT90	.0820	.0820	
	08MAY89	.0110	.0110	
	24AUG89	.0020	.	U
	04NOV89	.0020	.	U
	26FEB90	.0055	.0055	
	18MAY90	.0310	.0310	
	06AUG90	.0020	.	U
	24JAN91	.0020	.	U
	19APR91	.0020	.	U
	29JUL91	.0150	.0150	
	09OCT91	.0280	.0280	
	11JAN92	.0020	.	U
	14APR92	.0020	.	U
	27JUL92	.0100	.0100	
	22OCT92	.0020	.	U
	02FEB93	.0020	.	U
	16APR93	.0020	.	U
	05AUG93	.0250	.	U
	14OCT93	.0250	.	U
ST-006	04MAY89	.0640	.0640	
	25JAN91	.0610	.0610	
	09OCT91	.0700	.0700	
	14APR92	.0620	.0620	
	24AUG89	.0049	.0049	
	07NOV89	.0020	.	U
	26FEB90	.0020	.	U
	18MAY90	.0085	.0085	
	06AUG90	.0060	.0060	
	23OCT90	.0020	.	U
	19APR91	.0020	.	U

Table 4 (continued)

Monitoring station	Date collected	Analytical result/detection limit	lower (for lnor.sas)	Qualifier
ST-006	29JUL91	.0020	.	U
continued	11JAN92	.0020	.	U
	27JUL92	.0042	.0042	
	20OCT92	.0020	.	U
	02FEB93	.0020	.	U
	19APR93	.0110	.0110	
	05AUG93	.0250	.	U
	14OCT93	.0250	.	U
ST-007	08MAY92	.1400	.1400	
	20AUG92	.0800	.0800	
	09MAR91	.0020	.	U
	18JUN91	.0020	.	U
	26SEP91	.0190	.0190	
	14DEC91	.0020	.	U
	12MAR92	.0020	.	U
	10NOV92	.0068	.0068	
	11MAR93	.0020	.	U
	23JUN93	.0020	.	U
	27SEP93	.0250	.	U
	19NOV93	.0250	.	U
ST-008	08MAR91	.0064	.0064	
	20JUN91	.0054	.0054	
	25SEP91	.0020	.	U
	14DEC91	.0170	.0170	
	08MAR92	.0140	.0140	
	06MAY92	.0062	.0062	
	19AUG92	.0220	.0220	
	09NOV92	.0210	.0210	
	10MAR93	.0056	.0056	
	18JUN93	.0170	.0170	
	22SEP93	.0250	.	U
	15NOV93	.0250	.	U

[1]The data in this table are from an early listing of the data, which was before they had been subjected to careful quality assurance examination.

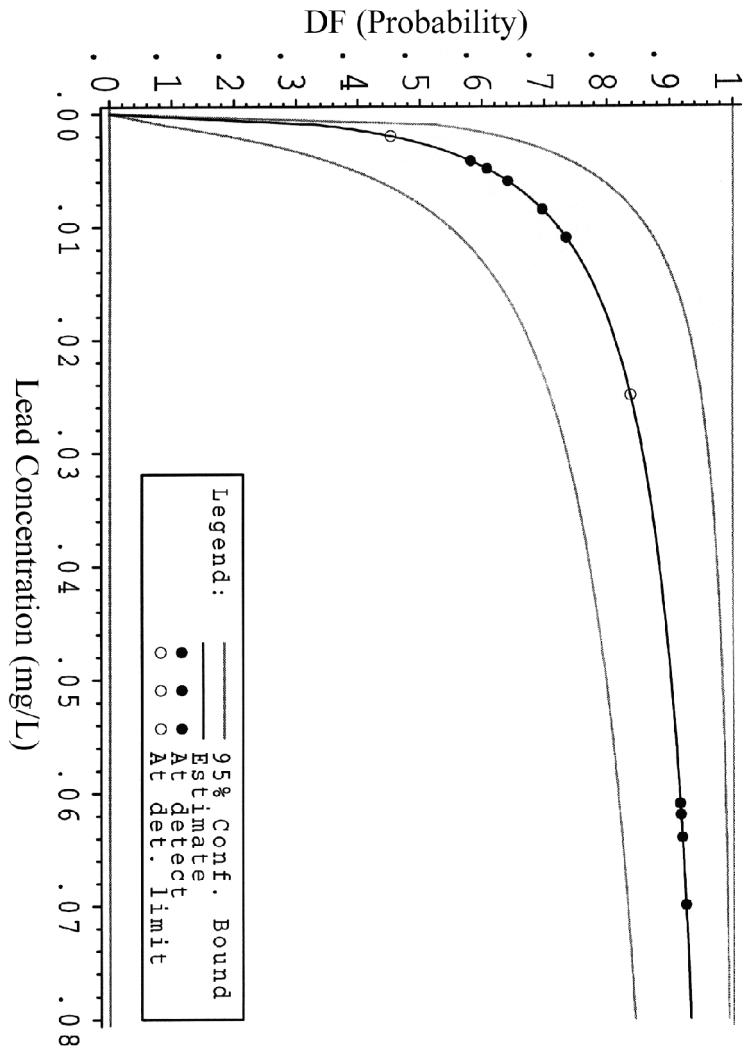

Figure 3. Lognormal distribution function (DF) estimate and 95% confidence bounds for groundwater lead at Station ST-006.

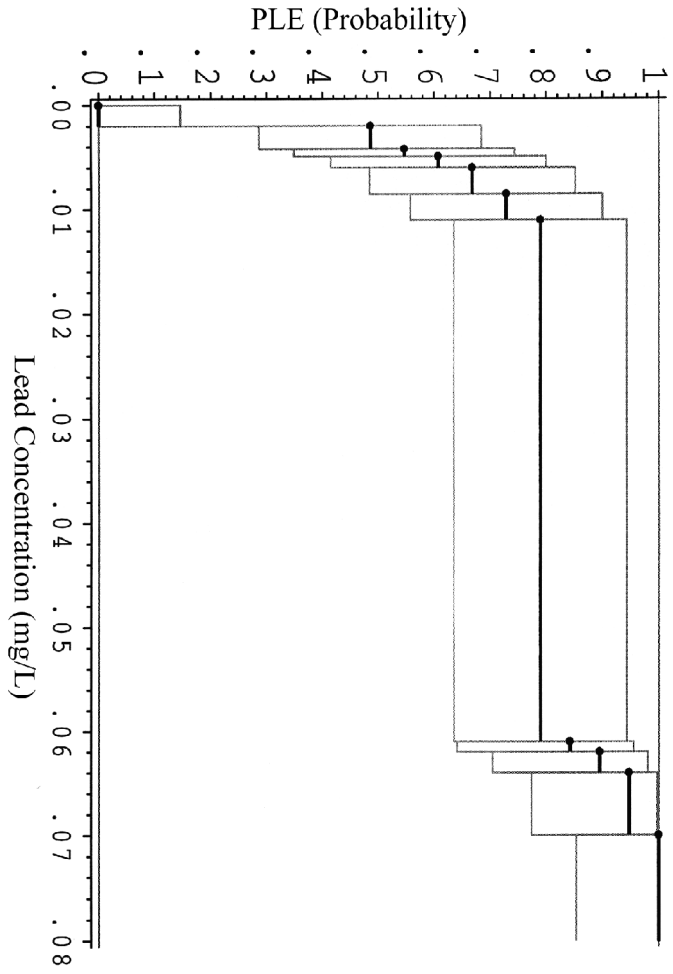

Figure 4. PLE (black lines) and 95% confidence bounds (gray lines) for groundwater lead at Station-006.

Table 5. SAS macro lnor.sas output data set contents for groundwater lead (Y-12 Fuel Station)

CONTENTS PROCEDURE

Data Set Name:	WORK.OUTPUT	Observations:	8
Member Type:	DATA	Variables:	40
Engine:	V611	Indexes:	0
Created:	7:19 Monday, Sep 23, 1996	Observation Length:	349
Last Modified:	7:19 Monday, Sep 23, 1996	Deleted Observations:	0
Protection:		Compressed:	NO
Data Set Type:		Sorted:	YES
Label:			

-----Engine/Host Dependent Information-----

Data Set Page Size:	32768
Number of Data Set Pages:	1
File Format:	607
First Data Page:	1
Max Obs per Page:	93
Obs in First Data Page:	8
File Name:	/usr/tmp/SAS_worka00000548/outputl.ssd01
Inode Number:	5282
Access Permission:	rw-r--r--
Owner Name:	schmoyer
File Size (bytes):	40960

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Format	Label
2	ANALYSIS	Char	35	10	$35.	
11	LOWER	Num	8	109		
1	STATION	Char	10	0	$10.	
9	_AR_LCL	Num	8	93		Ordinary mean, LCL (p=.95)
4	_AR_MN	Num	8	53		Ordinary mean
7	_AR_SEM	Num	8	77		Ordinary mean, std. err.
10	_AR_UCL	Num	8	101		Ordinary mean, UCL (p=.95)
8	_DET	Num	8	85		Detects
25	_LN_LCL	Num	8	221		Lognormal mean, LCL (p=.95)
24	_LN_MN	Num	8	213		Lognormal mean
26	_LN_UCL	Num	8	229		Lognormal mean, UCL (p=.95)
21	_LS_MN	Num	8	189		Ln-scale mean
23	_LS_SEM	Num	8	205		Ln-scale mean, std. err.
22	_LS_STD	Num	8	197		Ln-scale std. dev.
12	_LT500	Num	8	117		
13	_LT750	Num	8	125		
14	_LT900	Num	8	133		
6	_MAX	Num	8	69		Maximum
5	_MIN	Num	8	61		Minimum
27	_NL500	Num	8	237		
28	_NL750	Num	8	245		
29	_NL900	Num	8	253		
30	_NQ500	Num	8	261		
31	_NQ750	Num	8	269		
32	_NQ900	Num	8	277		

Table 5 (continued)

#	Variable	Type	Len	Pos	Format	Label
33	_NU500	Num	8	285		
34	_NU750	Num	8	293		
35	_NU900	Num	8	301		
39	_N_LCL	Num	8	333		Normal mean, LCL (p=.95)
36	_N_MN	Num	8	309		Normal mean
38	_N_SEM	Num	8	325		Normal mean, std. err.
37	_N_STD	Num	8	317		Normal scale std. dev.
40	_N_UCL	Num	8	341		Normal mean, UCL (p=.95)
3	_OBS	Num	8	45		Observed
15	_QU500	Num	8	141		
16	_QU750	Num	8	149		
17	_QU900	Num	8	157		
18	_UT500	Num	8	165		
19	_UT750	Num	8	173		
20	_UT900	Num	8	181		

-----Sort Information-----

Sortedby:	STATION ANALYSIS
Validated:	YES
Character Set:	ASCII

Table 6. Output of sas macro lnor.sas for groundwater lead (Y-12 Fuel Station)[1]

Station	Observed	Detects	Minimum	Maximum	Ordinary mean LCL (p=.95)	Ordinary mean
ST-001	15	3	.002	0.025	.0023642	0.005980
ST-002	19	6	.002	0.025	.0040552	0.007305
ST-003	14	7	.002	0.035	.0031129	0.007621
ST-004	19	6	.002	0.036	.0035510	0.007526
ST-005	19	7	.002	0.082	.0055121	0.013289
ST-006	19	9	.002	0.070	.0088167	0.018821
ST-007	12	4	.002	0.140	.0036627	0.025650
ST-008	12	9	.002	0.025	.0095179	0.013883

Station	Ordinary mean UCL (p=.95)	Ordinary mean std. err.	Lognormal mean LCL (p=.95)	Lognormal mean
ST-001	0.009596	0.002053	.0005677	0.00269
ST-002	0.010555	0.001874	.0013973	0.00726
ST-003	0.012130	0.002546	.0024582	0.01112
ST-004	0.011502	0.002292	.0013533	0.00656
ST-005	0.021067	0.004485	.0020880	0.03244
ST-006	0.028825	0.005769	.0045854	0.04117
ST-007	0.047637	0.012243	.0003124	0.30242
ST-008	0.018249	0.002431	.0073280	0.01228

Station	Lognormal mean UCL (p=.95)	95% LTB 50%-ile	Estimate 50%-ile	95% UTB 50%-ile	95% LTB 75%-ile
ST-001	0.013	.0001093	.0006545	0.003919	.0006132
ST-002	0.038	.0003122	.0011131	0.003969	.0015330
ST-003	0.050	.0008549	.0023732	0.006588	.0029288
ST-004	0.032	.0003091	.0010737	0.003730	.0014653
ST-005	0.504	.0003024	.0013421	0.005957	.0021501
ST-006	0.370	.0008040	.0026223	0.008552	.0042257
ST-007	292.770	.0000542	.0009075	0.015196	.0009633
ST-008	0.021	.0054525	.0086904	0.013851	.0090535

Station	Estimate 75%-ile	95% UTB 75%-ile	95% LTB 90%-ile	Estimate 90%-ile	95% UTB 90%-ile	Ln-scale mean
ST-001	0.001958	0.006252	0.001681	0.005250	0.01640	-7.33166
ST-002	0.003969	0.010273	0.004077	0.012460	0.03808	-6.80057
ST-003	0.007437	0.018886	0.006335	0.020792	0.06824	-6.04352
ST-004	0.003743	0.009561	0.003848	0.011517	0.03447	-6.83666
ST-005	0.007037	0.023031	0.007388	0.031265	0.13230	-6.61352
ST-006	0.012240	0.035454	0.012934	0.048978	0.18546	-5.94371
ST-007	0.008199	0.069790	0.004728	0.059449	0.74758	-7.00481
ST-008	0.014869	0.024420	0.013140	0.024110	0.04424	-4.74553

Table 6 (continued)

Station	Ln-scale mean std. err.	Ln-scale std. dev.	Normal mean LCL (p=.95)	Normal mean	Normal mean UCL (p=.95)
ST-001	1.01618	1.68173	-0.00485	-0.000791	0.003269
ST-002	0.73317	1.93638	-0.01034	-0.002213	0.005916
ST-003	0.57657	1.75745	-0.00805	0.001326	0.010699
ST-004	0.71807	1.90218	-0.01650	-0.005310	0.005877
ST-005	0.85940	2.52391	-0.03219	-0.009634	0.012920
ST-006	0.68174	2.34679	-0.02148	0.000005	0.021494
ST-007	1.56921	3.40848	-0.11127	-0.033658	0.043953
ST-008	0.25957	0.83164	0.00675	0.011187	0.015621

Station	Normal mean std. err.	Normal scale std. dev.	Normal Model 95% LTB 50%-ile	Normal Model Estimate 50%-ile	Normal Model 95% UTB 50%-ile
ST-001	0.002305	0.004408	-0.00485	-0.000791	0.003269
ST-002	0.004688	0.012515	-0.01034	-0.002213	0.005916
ST-003	0.005293	0.016053	-0.00805	0.001326	0.010699
ST-004	0.006452	0.017235	-0.01650	-0.005310	0.005877
ST-005	0.013006	0.038182	-0.03219	-0.009634	0.012920
ST-006	0.012392	0.042469	-0.02148	0.000005	0.021494
ST-007	0.043216	0.093615	-0.11127	-0.033658	0.043953
ST-008	0.002469	0.007811	0.00675	0.011187	0.015621

Station	Normal Model 95% LTB 75%-ile	Normal Model Estimate 75%-ile	Normal Model 95% UTB 75%-ile	Normal Model 95% LTB 90%-ile
ST-001	-0.000989	0.002081	0.005152	0.001859
ST-002	-0.000213	0.006004	0.012220	0.006077
ST-003	0.003196	0.011760	0.020324	0.010429
ST-004	-0.002609	0.006005	0.014618	0.006295
ST-005	-0.002651	0.015432	0.033514	0.016478
ST-006	0.008650	0.027886	0.047122	0.029159
ST-007	-0.032186	0.026796	0.085778	0.012416
ST-008	0.011479	0.016231	0.020983	0.014972

Station	Normal Model Estimate 90%-ile	Normal Model 95% UTB 90%-ile
ST-001	0.004667	0.00747
ST-002	0.013399	0.02072
ST-003	0.021151	0.03187
ST-004	0.016189	0.02608
ST-005	0.037992	0.05951
ST-006	0.052980	0.07680
ST-007	0.081207	0.15000
ST-008	0.020771	0.02657

[1]A few of the columns of lesser important information have been omitted from this table to save space.

Table 7 . SAS macro ple.sas output data set contents for groundwater lead (Y-12 Fuel Station)

CONTENTS PROCEDURE

Data Set Name:	WORK.OUTPUT	Observations:	8
Member Type:	DATA	Variables:	23
Engine:	V611	Indexes:	0
Created:	7:20 Monday, Sep 23, 1996	Observation Length:	213
Last Modified:	7:20 Monday, Sep 23, 1996	Deleted Observations:	0
Protection:		Compressed:	NO
Data Set Type:		Sorted:	YES
Label:			

-----Engine/Host Dependent Information-----

Data Set Page Size:	24576
Number of Data Set Pages:	1
File Format:	607
First Data Page:	1
Max Obs per Page:	115
Obs in First Data Page:	8
File Name:	/usr/tmp/SAS_worka00000548/outputp.ssd01
Inode Number:	5290
Access Permission:	rw-r--r--
Owner Name:	schmoyer
File Size (bytes):	32768

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Format	Label
2	ANALYSIS	Char	35	10	$35.	
1	STATION	Char	10	0	$10.	
9	_AR_LCL	Num	8	93		Ordinary mean, LCL (p=.95)
4	_AR_MN	Num	8	53		Ordinary mean
7	_AR_SEM	Num	8	77		Ordinary mean, std. err.
10	_AR_UCL	Num	8	101		Ordinary mean, UCL (p=.95)
8	_DET	Num	8	85		Detects
16	_LT50	Num	8	149		
17	_LT75	Num	8	157		
18	_LT90	Num	8	165		
6	_MAX	Num	8	69		Maximum
5	_MIN	Num	8	61		Minimum
3	_OBS	Num	8	45		Observed
22	_PL_LCL	Num	8	197		PLE mean, LCL (p=.95)
11	_PL_MN	Num	8	109		PLE mean
12	_PL_SEM	Num	8	117		PLE mean, std. err.
23	_PL_UCL	Num	8	205		PLE mean, UCL (p=.95)
13	_QU50	Num	8	125		
14	_QU75	Num	8	133		
15	_QU90	Num	8	141		
19	_UT50	Num	8	173		
20	_UT75	Num	8	181		
21	_UT90	Num	8	189		

Table 7 (continued)

-----Sort Information-----

Sortedby:	STATION ANALYSIS
Validated:	YES
Character Set:	ASCII

Table 8. Output of SAS macro ple.sas for groundwater lead (Y-12 Fuel Station)[1]

Station	Observed	Detects	Minimum	Maximum	Ordinary mean LCL (p=.95)	Ordinary mean
ST-001	15	3	.002	0.025	.0023642	0.005980
ST-002	19	6	.002	0.025	.0040552	0.007305
ST-003	14	7	.002	0.035	.0031129	0.007621
ST-004	19	6	.002	0.036	.0035510	0.007526
ST-005	19	7	.002	0.082	.0055121	0.013289
ST-006	19	9	.002	0.070	.0088167	0.018821
ST-007	12	4	.002	0.140	.0036627	0.025650
ST-008	12	9	.002	0.025	.0095179	0.013883

Station	Ordinary mean UCL (p=.95)	Ordinary mean std. err.	PLE mean LCL (p=.95)	PLE mean
ST-001	0.009596	0.002053	.0026744	0.003054
ST-002	0.010555	0.001874	.0032712	0.005224
ST-003	0.012130	0.002546	.0032505	0.007621
ST-004	0.011502	0.002292	.0021256	0.005270
ST-005	0.021067	0.004485	.0034500	0.011120
ST-006	0.028825	0.005769	.0063935	0.016599
ST-007	0.047637	0.012243	.0000000	0.022271
ST-008	0.018249	0.002431	.0077189	0.011660

Station	PLE mean UCL (p=.95)	PLE mean std. err.	95%LTB 50%-ile	Estimate 50%-ile	95% UTB 50%-ile	95% LTB 75%-ile
ST-001	0.003433	0.000215	.0020	0.0020	0.0020	.0020
ST-002	0.007176	0.001126	.0020	0.0020	0.0054	.0020
ST-003	0.011992	0.002468	.0020	0.0031	0.0091	.0042
ST-004	0.008414	0.001813	.0020	0.0020	0.0042	.0020
ST-005	0.018791	0.004423	.0020	0.0020	0.0100	.0020
ST-006	0.026805	0.005885	.0020	0.0042	0.0085	.0049
ST-007	0.046237	0.013345	.0020	0.0020	0.0068	.0020
ST-008	0.015601	0.002195	.0056	0.0102	0.0170	.0064

Station	Estimate 75%-ile	95% UTB 75%-ile	95% LTB 90%-ile	Estimate 90%-ile	95% UTB 90%-ile
ST-001	0.0020	0.0070	0.0020	0.0070	0.0073
ST-002	0.0056	0.0120	0.0056	0.0160	0.0200
ST-003	0.0091	0.0350	0.0091	0.0210	.
ST-004	0.0046	0.0071	0.0046	0.0120	.
ST-005	0.0110	0.0310	0.0110	0.0310	.
ST-006	0.0110	0.0640	0.0085	0.0640	.
ST-007	0.0190	0.1400	0.0068	0.0800	.
ST-008	0.0170	0.0220	0.0170	0.0215	0.0220

[1]A few of the columns of lesser important information have been omitted from this table to save space.

Table 9. Groundwater barium concentrations (mg/L) from Y-12 Fuel Station[1]

Monitoring station	Date collected	Analytical result
STB-01	19JUN90	0.32
	26SEP90	0.26
	06DEC90	0.24
	08MAR91	0.22
	18JUN91	0.22
	25SEP91	0.27
	14DEC91	0.22
	08MAR92	0.23
	07MAY92	0.21
	19AUG92	0.24
	09NOV92	0.20
	10MAR93	0.23
	21JUN93	0.22
	22SEP93	0.24
	16NOV93	0.21
STB-02	08MAY89	0.09
	25AUG89	0.08
	04NOV89	0.08
	27FEB90	0.06
	17MAY90	0.12
	04AUG90	0.06
	23OCT90	0.10
	25JAN91	0.09
	19APR91	0.04
	30JUL91	0.06
	09OCT91	0.11
	11JAN92	0.04
	14APR92	0.05
	29JUL92	0.07
STB-03	04MAY89	0.59
	24AUG89	0.58
	04NOV89	0.58
	27FEB90	0.58
	17MAY90	1.30
	04AUG90	0.50
	23OCT90	0.52
	24JAN91	0.97
	19APR91	0.50
	29JUL91	0.51
	09OCT91	0.53
	11JAN92	0.47
	14APR92	0.47
	27JUL92	0.45
	22OCT92	0.57
	03FEB93	0.54
	19APR93	0.48
	06AUG93	0.57
	14OCT93	0.57

[1]This early listing of the barium data is used for demonstration purposes only, recognizing that barium is not considered to be a site-related contaminant.

Table 10. SAS macro logconf.sas output data set contents for groundwater barium (Y-12 Fuel Station)

CONTENTS PROCEDURE

Data Set Name:	WORK.OUTPUTL	Observations:	3
Member Type:	DATA	Variables:	6
Engine:	V611	Indexes:	0
Created:	7:30 Wednesday, Sep 25, 1996	Observation Length:	77
Last Modified:	7:30 Wednesday, Sep 25, 1996	Deleted Observations:	0
Protection:		Compressed: NO	
Data Set Type:		Sorted: NO	
Label:			

-----Engine/Host Dependent Information-----

Data Set Page Size:	8192
Number of Data Set Pages:	1
File Format:	607
First Data Page:	1
Max Obs per Page:	106
Obs in First Data Page:	3
File Name:	/tmp/SAS_worka00002577/outputl.ssd01
Inode Number:	7681
Access Permission:	rw-r--r--
Owner Name:	schmoyer
File Size (bytes):	16384

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos	Format	Label
2	ANALYSIS	Char	35	10	$35.	
4	MEAN	Num	8	53		Log-scale mean
5	STAND	Num	8	61		Log-scale standard deviation
1	STATION	Char	10	0	$10.	
6	UCL_95	Num	8	69		Land's lognormal 95% UCL
3	_OBS	Num	8	45		Observed

Table 11. Output of SAS macro logconf.sas for groundwater barium (Y-12 Fuel Station)

Station	Observed	Log-scale mean	Log-scale standard deviation	Land's lognormal 95% UCL
STB-02	14	-2.64814	0.35203	0.09089
STB-01	15	-1.45359	0.11812	0.24876
STB-03	19	-0.55913	0.25686	0.65937

APPENDIX A

THE LNOR MACRO

The lnor macro is listed below.

```
/*************************************************************************\
*     SAS macro lnor: Calculate lognormal and normal-based statistics for analytes
*      that have one or more detects.
*
*    © 1999
*     Lockheed Martin Energy Research Corporation
*     All rights reserved
*
*    Neither the Government nor LMER makes any warrantee, express or implied, or
*    assumes any liability or responsibility for the use of this software.                *

\*************************************************************************/

*NOTE:  Nonpositive or missing &RESULT values, validation rejects, and any
other data that is not wanted should not be entered into these calculations.
Validation rejects will be treated as detects.  Proc lifereg will drop
nonpositive values from the lognormal model analyses, but proc means will
include them when it calculates the number of observations (_obs), which
will then be used incorrectly to adjust the lifereg estimates.;

%macro
LNOR (input, output, group, result, lower, qual, parm, id, confid, toler);

*MACRO ARGUMENTS:
input--name of input data set.
output--name of output data set.
group--list of variables (delimited by spaces) that defines groups (e.g.,
  sites) over which statistics are to be computed.
result--variable that gives analysis result or, in the case of a nondetect,
  the detection limit.
lower--same as result for detects.  For simple nondetects, missing (.).  Can
  also be a numeric value less than result, namely, for interval censoring.
  For example, if an observation represents an average for two duplicates,
  one of which is a detect, at say X, the other a nondetect at say L, then
  the average is between (X+0)/2 and (X+L)/2.  Then take lower=X/2,  and
  result=(X+L)/2.
qual--qualifier variable, "U" for nondetect, "I" for interval-censored.
  Anything else is treated as a detect.
parm--variable that names the analytes (e.g., Aluminium, Arsensic).
id--list of ID variables to be carried along.
confid--confidence level, between 0 and 1 (e.g., .95), for confidence limits.
toler--space delimited list of values for which tolerance bounds should be
  computed.
;
```

```
data _null_;
length toll $ 100;
if index(' '||'&TOLER','.') > 0 then do;
toll='q='||trim(left('&TOLER'));
call symput('TOLER',trim(left(toll)));
end;
run;


 proc sort  data = &INPUT  out = detects;
  where &RESULT ^= .;
  by &GROUP &PARM &QUAL;
 run;


 data detects
     nondet;

  set detects;
  by &GROUP &PARM;

  if &QUAL in ('U','I') then cen=1;
  else cen=0;

  retain _keep_ 0;

  if first.&PARM then do;

    if &QUAL = ' ' then _keep_ = 1;  /* At least one detect */
    else _keep_ = 0;                 /* All non-detects */

  end;

  if _keep_ then output detects;
  else output nondet;
  drop _keep_;

 run;


 /* Calculate mean, minimum, maximum, number of observations for detects */

 proc means  data = detects  noprint;
  by &GROUP &PARM;
  var cen &RESULT;
  id &ID;
  output out = stat  (drop = _freq_ _type_)
        sum    (cen) = _ndt
```

```
        n      (&RESULT) = _obs
        mean   (&RESULT) = _ar_mn
        min    (&RESULT) = _min
        max    (&RESULT) = _max
        stderr (&RESULT) = _ar_sem;
run;


data stat;

  set stat;

  label _det    = "Detects"
        _ar_lcl = "Ordinary mean, LCL (p=&CONFID)"
        _ar_ucl = "Ordinary mean, UCL (p=&CONFID)"
  ;

  _det = _obs - _ndt;  /* Number of detects */

  if _obs > 1 then do;

    _ar_ucl = _ar_mn + _ar_sem * tinv (&CONFID, _obs - 1);
    _ar_lcl = max (0, _ar_mn - _ar_sem * tinv (&CONFID, _obs - 1));

  end;

run;


/* Calculate mean, minimum, maximum, number of observations for
nondetects */

proc means  data = nondet  noprint;
  by &GROUP &PARM;
  var &RESULT;
  id &ID;
  output out = nonstat (drop = _freq_ _type_)
        n    (&RESULT) = _obs
        mean (&RESULT) = _ar_mn
        min  (&RESULT) = _min
        max  (&RESULT) = _max;
run;



/* Proc lifereg is used to compute lognormal-based maximum likelihood
   estimates.  The information required to bias-adjust the estimate
   (subsequent data step) is output.  Bias is due to inequality Ef(X)
```

```
    ne f(EX), for nonlinear f and nondegenerate X.*/

  proc lifereg  data = detects  covout  outest = estm  noprint;
    by &GROUP &PARM;
    model (&LOWER, &RESULT) = / dist = lnormal  covb;
*censoring: 0=uncensored, 1=right, 2=left, 3=interval;
output out=toler
&TOLER
predicted=pred
std_err=se_pred;
  run;

data toler;
merge toler stat (keep = &GROUP &PARM _obs);
by &GROUP &PARM;

proc sort data=toler;
by &GROUP &PARM _prob_;

data toler; set toler;
by &GROUP &PARM _prob_;
if first._prob_;

     if _obs > 1 then df_adj = sqrt( _obs / (_obs - 1));
     else df_adj = .;
*degrees-of-freedom adjustment same as below for all stats;

_ut=pred*exp(tinv(&CONFID,_obs-1)*se_pred*df_adj/pred);
_lt=pred*exp(tinv(1- &CONFID,_obs-1)*se_pred*df_adj/pred);
rename pred=_qu;
prob_lab=_prob_;
_prob_=round(1000*_prob_,1);
keep &PARM &GROUP pred _ut _lt _prob_ prob_lab;

data _lt;
set toler;
length label $40;
label='Lognormal '||trim(left(prob_lab))||' LTB'||'' (p=&CONFID)";
rename label=prob_lab;
drop prob_lab;

proc transpose data=_lt prefix=_lt out=_lt;
by &GROUP &PARM;
var _lt;
id _prob_;
idlabel prob_lab;

data _qu;
```

```
set toler;
length label $40;
label='Lognormal '||trim(left(prob_lab))||' Quantile Estimate';
rename label=prob_lab;
drop prob_lab;

proc transpose data=_qu prefix=_qu out=_qu;
by &GROUP &PARM;
var _qu;
id _prob_;
idlabel prob_lab;

data _ut;
set toler;
length label $40;
label='Lognormal '||trim(left(prob_lab))||' UTB'||'' (p=&CONFID)';
rename label=prob_lab;
drop prob_lab;

proc transpose data=_ut prefix=_ut out=_ut;
by &GROUP &PARM;
var _ut;
id _prob_;
idlabel prob_lab;

data tol;
merge _lt _qu _ut;
by &GROUP &PARM;

proc datasets nolist;
delete _lt _qu _ut;

 /* Merge LIFEREG output data set with stat. Extract parameter
    estimates from estm and adjust the estimates if the data
    are uncensored. Basically, the adjustments involve removing
    the bias inherent in the MLE of the variance in the case of
    uncensored data. Calculate arithmetic mean and confidence
    limits and output estimates to the SAS data set &OUTPUT. */

 data output;

   merge stat estm tol;
   by &GROUP &PARM;

   drop _name_  _type_  intercep _scale_ _shape1_ _ndt
        _ls_var _ls_cov _ls_sev  _cin_   _dist_   _lnlike_ _model_ _label_;
   retain _ls_mn _ls_var _ls_std _ls_sem _ls_sev _ls_cov;
```

```
label _ln_mn  = "Lognormal mean"
      _ln_lcl = "Lognormal mean, LCL (p=&CONFID)"
      _ln_ucl = "Lognormal mean, UCL (p=&CONFID)"
      _ls_mn  = "Ln-scale mean"
      _ls_std = "Ln-scale std. dev."
      _ls_sem = "Ln-scale mean, std. err."
      _obs    = "Observed"
      _ar_mn  = "Ordinary mean"
      _min    = "Minimum"
      _max    = "Maximum"
      _ar_sem = "Ordinary mean, std. err."
;

if _type_ = "PARMS" then do;
  _ls_mn  = intercep;
  * Log-scale Mean--mu tilde in report;
  _ls_var = _scale_ ** 2;
  * Log-scale sample variance--tau tilde in report;
end;

else if _type_ = "COV" and _name_ = "INTERCPT" then do;
  _ls_sem = intercep;
  * Variance estimate for mean--A in report;

  _ls_cov = 2*sqrt (_ls_var)*_scale_;
  * Covariance estimate for mean and variance--D in report;
end;

else if _type_ = "COV" and _name_ = "SCALE" then
  _ls_sev = 4 * _ls_var * _scale_;
  * Variance estimate for variance--G in report;

if last.&PARM then do;

  if _obs > 1 then do;
  _ls_var = _ls_var * _obs / (_obs - 1);
  _ls_sem = _ls_sem * _obs / (_obs - 1);
  _ls_sev = _ls_sev*(_obs*_obs*_obs / ((_obs-1)*(_obs-1)*(_obs+1)));

  _ln_mn = exp (_ls_mn + (_ls_var / 2));  /* lognormal mean estimate */
  _cin_=sqrt(_ls_sem+_ls_sev/4+_ls_cov) * tinv (&CONFID, (_obs - 1));
  _ln_lcl = _ln_mn * exp (-_cin_);  /* LCL */
  _ln_ucl = _ln_mn * exp ( _cin_);  /* UCL */

  if _min <= 0 then do;
    _ln_mn  = .;  _ln_lcl = .;  _ln_ucl = .;
                  _ls_mn  = .;  _ls_std = .;  _ls_sem = .;
  end;
```

```
      _ls_std = sqrt (_ls_var);
      _ls_sem=sqrt(_ls_sem);

      end;

      else do;
      _ls_var = .; _ls_sem = .; _ls_sev=.; _ls_std=.;
      end;

      output;

      _ls_mn  = .;  _ls_var = .;  _ls_sem = .;
      _ls_sev = .;  _ls_cov = .;

    end;

  run;

  /* Proc lifereg is also used to compute normal-based maximum likelihood
     estimates.*/

  proc lifereg  data = detects  covout  outest = estn  noprint;
    by &GROUP &PARM;
    model (&LOWER, &RESULT) = / dist = normal  covb;
*censoring: 0=uncensored, 1=right, 2=left, 3=interval;
output out=tolern
&TOLER
predicted=pred
std_err=se_pred;
  run;

data tolern;
merge tolern stat (keep = &GROUP &PARM _obs);
by &GROUP &PARM;

proc sort data=tolern;
by &GROUP &PARM _prob_;

data tolern; set tolern;
by &GROUP &PARM _prob_;
if first._prob_;

      if _obs > 1 then df_adj = sqrt( _obs / (_obs - 1));
      else df_adj = .;
*degrees-of-freedom adjustment same as below for all stats;

_ut=pred+tinv(&CONFID,_obs-1)*se_pred*df_adj;
_lt=pred+tinv(1- &CONFID,_obs-1)*se_pred*df_adj;
```

```
rename pred=_qu;
prob_lab=_prob_;
_prob_=round(1000*_prob_,1);
keep &PARM &GROUP pred _ut _lt _prob_ prob_lab;

data nltb;
set toler;
length label $40;
label='Normal '||trim(left(prob_lab))||' LTB'||'' (p=&CONFID)';
rename label=prob_lab;
drop prob_lab;

proc transpose data=nltb prefix=_nl out=nltb;
by &GROUP &PARM;
var _lt;
id _prob_;
idlabel prob_lab;

data nqua;
set toler;
length label $40;
label='Normal '||trim(left(prob_lab))||' Quantile Estimate';
rename label=prob_lab;
drop prob_lab;

proc transpose data=nqua prefix=_nq out=nqua;
by &GROUP &PARM;
var _qu;
id _prob_;
idlabel prob_lab;

data nutb;
set toler;
length label $40;
label='Normal '||trim(left(prob_lab))||' UTB'||'' (p=&CONFID)';
rename label=prob_lab;
drop prob_lab;

proc transpose data=nutb prefix=_nu out=nutb;
by &GROUP &PARM;
var _ut;
id _prob_;
idlabel prob_lab;

data ntol;
merge nltb nqua nutb;
by &GROUP &PARM;
```

```
proc datasets nolist;
delete nltb nqua nutb;


 /* Merge LIFEREG output data set with stat. Extract parameter
    estimates from estm and adjust the estimates if the data
    are uncensored. Basically, the adjustments involve removing
    the bias inherent in the MLE of the variance in the case of
    uncensored data. Calculate arithmetic mean and confidence
    limits and output estimates to the SAS data set &OUTPUT. */

 data outputn;

   merge stat estn ntol;
   by &GROUP &PARM;

   drop _name_  _type_  intercep _scale_ _shape1_ _ndt
        _n_var _dist_   _lnlike_ _model_ _label_
                        lower _ar_lcl _ar_mn _ar_sem _ar_ucl _det _max
                        _min _obs;

   retain _n_mn _n_var _n_std _n_sem;

   label _n_mn  = "Normal mean"
         _n_lcl = "Normal mean, LCL (p=&CONFID)"
         _n_ucl = "Normal mean, UCL (p=&CONFID)"
         _n_std = "Normal scale std. dev."
         _n_sem = "Normal mean, std. err."
   ;

   if _type_ = "PARMS" then do;

     _n_mn  = intercep;      /* Mean */
     _n_var = _scale_ ** 2;  /* Variance */

     if _obs > 1 then _n_var = _n_var * _obs / (_obs - 1);
     else _n_var = .;

     _n_std = sqrt (_n_var);

   end;

   else if _type_ = "COV" and _name_ = "INTERCPT" then do;

     _n_sem = sqrt (intercep);             /* Standard error of mean */

     if _obs > 1 then _n_sem = _n_sem * sqrt (_obs / (_obs - 1));
     else _n_sem = .;
```

```
    end;

    if last.&PARM then do;

      /* Compute normal lower and upper confidence limits */

        _n_ucl = _n_mn + tinv (&CONFID, (_obs - 1)) * _n_sem;
        _n_lcl = _n_mn - tinv (&CONFID, (_obs - 1)) * _n_sem;

      if _min <= 0 then do;
        _n_mn  = .;  _n_lcl = .;  _n_ucl = .;
        _n_std = .;  _n_sem = .;
      end;

      output;

      _n_mn  = .;  _n_var = .;  _n_sem = .;

    end;

  run;

data &OUTPUT;
merge output outputn;
by &GROUP &PARM;

  data &OUTPUT;

    set &OUTPUT (in = in1)  /* Summary statistics for detects */
        nonstat (in = in2); /* Summary statistics for nondetects */

  run;

  proc sort  data = &OUTPUT;
    by &GROUP &PARM;
  run;

%MEND lnor;
```

APPENDIX B

THE PLE MACRO

The ple macro is listed below.

```
 /*************************************************************************\
 *
 *   ple.sas: Calculate statistics based on the product limit estimate
 *   (PLE), also known as Kaplan-Meier estimate, at least in the
 *   setting of right-censored (failure-time) data.
 *
 *    © 1999
 *     Lockheed Martin Energy Research Corporation
 *     All rights reserved
                                *
 *    Neither the Government nor LMER makes any warrantee, express or implied, or
 *    assumes any liability or responsibility for the use of this software.
 \*************************************************************************/


*NOTE:  Nonpositive or missing &RESULT values, validation rejects, and any
other data that is not wanted should not be entered into these calculations.
Validation rejects will be treated as detects.  The PLE calculations were
not designed to handle nonpositive &RESULT values, but proc means will
include them when it calculates the number of observations (_obs_), which
will then be used incorrectly to compute PLE-based estimates.;

*Labels for tolerance bounds could fail if there are too many, because
of length 200 limit for character variables in SAS.  (Limitation to be
removed in Version 7.);

%macro ple (input, output, group, result, qual, parm, id, confid, toler);

*MACRO ARGUMENTS:
input--name of input data set.
output--name of output data set.
group--list of variables (delimited by spaces) that defines groups (e.g.,
  sites) over which statistics are to be computed.
result--variable that gives analysis result or, in the case of a nondetect,
  the detection limit.
qual--"U" for nondetect, "I" for interval censored.  Otherwise, a detect.
  "I" treated as nondetect.
parm--variable that names the analytes (e.g., Aluminium, Arsensic).
id--list of ID variables to be carried along.
confid--condfidence level, between 0 and 1 (e.g., .95), for confidence limits.
toler--space delimited list of values for which tolerance bounds should be
  computed.
;

 %local P L U N TOL;
 %let TOL = 1.e-10;  /* Rounding tolerance */
```

```
  data _null_;
  length newt newltb newqua newutb $ 100 labell labelq labelu $200 g $ 8;
  newt=translate('.00001 '||"&TOLER",'.',' ');
  newltb='';
  newqua='';
  newutb='';
  labell='';
  labelq='';
  labelu='';
  n=0;
  g='0';
  do until (g=' ');
  number=put(g,8.);

  if number ne 0 then do;
    newltb=trim(left(newltb))||' _lt'||trim(left(g));
    newqua=trim(left(newqua))||' _qu'||trim(left(g));
    newutb=trim(left(newutb))||' _ut'||trim(left(g));
  end;
   if number gt 0 then do;
labell=
trim(left(labell))||' _lt'||trim(left(g))||"='"||trim(left(g))||' LTB '||"(p=&confid)"';

labelq=
trim(left(labelq))||' _qu'||trim(left(g))||"='"||trim(left(g))||" quantile estimate"';

labelu=
trim(left(labelu))||' _ut'||trim(left(g))||"='"||trim(left(g))||' UTB '||"(p=&confid)"';
  end;

  n=n+1;
  g=scan(newt,n,'.');
  end;
  call symput('L',trim(left(newltb)));
  call symput('P',trim(left(newqua)));
  call symput('U',trim(left(newutb)));
  call symput('TOLER','.00001 '||"&TOLER");
  call symput('labell',trim(left(labell)));
  call symput('labelq',trim(left(labelq)));
  call symput('labelu',trim(left(labelu)));
  run;

  data _null_;
  array qnt &P;
  call symput('N',trim(left(dim(qnt))));
  run;
```

```
proc sort  data = &INPUT  out = aggr;
  where &RESULT ^= .;
  by &GROUP &PARM &QUAL;
run;


data detects
     nondet;

  set aggr;
  by &GROUP &PARM;

  if &QUAL in ('U','I') then cen=1;
  else cen=0;

  retain _keep_ 0;

  if first.&PARM then do;

    if cen = 0 then _keep_ = 1;  /* At least one detect */
    else _keep_ = 0;                /* All non-detects */

  end;

  if _keep_ then output detects;
  else output nondet;

run;


/* Calculate mean, minimum, maximum, number of observations for detects */


proc means  data = detects  noprint;
  by &GROUP &PARM;
  var cen &RESULT;
  output out = stat  (drop = _freq_ _type_)
        sum    (CEN) = _ndt
        n      (&RESULT) = _obs
        mean   (&RESULT) = _ar_mn
        min    (&RESULT) = _min
        max    (&RESULT) = _max
        stderr (&RESULT) = _ar_sem;
run;


data stat;
```

```
  set stat;

  label _det    = "Detects"
        _ar_lcl = "Ordinary mean, LCL (p=&CONFID)"
        _ar_ucl = "Ordinary mean, UCL (p=&CONFID)"
  ;

  _det = _obs - _ndt;  /* Number of detects */

  if _obs > 1 then do;

    _ar_ucl = _ar_mn + _ar_sem * tinv (&CONFID, _obs - 1);
    _ar_lcl = max (0, _ar_mn - _ar_sem * tinv (&CONFID, _obs - 1));

  end;

run;


/* Calculate mean, minimum, maximum, number of observations for detects */

proc means  data = nondet  noprint;
  by &GROUP &PARM;
  var &RESULT;
  id &ID;
  output out = nonstat (drop = _freq_ _type_)
         n    (&RESULT) = _obs
         mean (&RESULT) = _ar_mn
         min  (&RESULT) = _min
         max  (&RESULT) = _max;
run;



/* Begin SAS code for product limit estimation.  Means and standard
   errors, upper confidence limits for means, and percentiles and their
   upper and lower confidence limits are computed, all based on PLEs.
   The percentiles are computed using the PLE with averaging (as in the
   SAS Proc Univariate default method - "empirical distribution function
   with averaging").  Because of discreteness, percentile estimates can
   be off when sample sizes are small.  (What is the fortieth percentile
   for a sample of size 1?)  Therefore percentile estimates should always
   be interpreted in light of sample sizes and confidence bounds. */


proc sort  data = detects;
  by &GROUP &PARM descending &RESULT &QUAL;
run;
```

```
    *PLE requires pass through data in order of descending &RESULT;
/* Pass through to compute PLE, PLE mean, and statistic "a" (defined on
   page 98 in STATISTICAL MODELS AND METHODS FOR LIFETIME DATA by J. F.
   Lawless (1982), John Wiley & Sons, New York) used to compute standard
   error of PLE mean (in subsequent pass through data) */

  data ple (keep = enter remain ple &RESULT last &QUAL &PARM &ID
                   &GROUP a se_ple lcb_ple ucb_ple)
       ple_mn (keep = &PARM &GROUP &QUAL _obs _ndt _det _pl_mn bias1);

   merge detects (in = in1)
         stat    (in = in2);
   by &GROUP &PARM;

   if in1 and in2;

    retain censored tot_cen enter remain ple _pl_mn se_ple last ss;

   label _pl_mn  = "PLE mean"
   ;

   &RESULT = max (&RESULT, 0);

   if first.&PARM then do;
     ss       = 0;
     censored = 0;
     tot_cen  = 0;
     enter    = _obs;
     remain   = _obs;
     ple      = 1;
     _pl_mn   = 0;
     last     = &RESULT;
   end;

   /* Note: if &QUAL = ' ', observation is treated as a detect */

   if &QUAL in ('U','I') then do;
     censored = censored + 1;  /* Sum number of nondetects */
     tot_cen  = tot_cen + 1;
   end;

   else do;

     ss     = ss + (enter - remain) / (enter * remain);
     _pl_mn = _pl_mn + (last - &RESULT) * (1 - ple);  /* ple mean */
     a      = (last - &RESULT) * ple;
     se_ple = ple * sqrt(ss);
```

```
  /* Calculate lower and upper confidence limits for ple */

  lcb_ple = max (0, ple - (se_ple * probit (&CONFID)));
  ucb_ple = min (1, ple + (se_ple * probit (&CONFID)));

  /* Compute exact confidence limits where censoring has no effect */

  if tot_cen = 0 then do;

    freq = round (_obs * ple, 1);

    if freq = _obs then ucb_ple = 1.0;
    else ucb_ple = betainv (&CONFID, freq + 1, _obs - freq);

    if freq = 0 then lcb_ple = 0.0;
    else lcb_ple = 1.0 - betainv (&CONFID, _obs - freq + 1, freq);

  end;

  /* Output standard error, lower and upper confidence limits
     for ple. */

  output ple;

  enter    = remain - censored;
  remain   = enter - 1;
  ple      = ple * remain / enter;
  censored = 0;
  last     = &RESULT;

end;

/* Calculate confidence limits at origin - needed for percentile
   estimates Also calculate mean */

if last.&PARM then do;

  a=0;
  bias1=0;

  if &QUAL in ('U','I') then do;

    ss     = ss + (enter - remain) / (enter * remain);
    se_ple = ple * sqrt(ss);
    /* Calculate lower and upper confidence limits for ple */

    lcb_ple = max (0, ple - (se_ple * probit (&CONFID)));
    ucb_ple = min (1, ple + (se_ple * probit (&CONFID)));
```

```
      /* Compute exact confidence limits where censoring has no effect */

      if tot_cen = 0 then do;

        freq = round (_obs * ple, 1);

        if freq = _obs then ucb_ple = 1.0;
        else ucb_ple = betainv (&CONFID, freq + 1, _obs - freq);

        if freq = 0 then lcb_ple = 0.0;
        else lcb_ple = 1.0 - betainv (&CONFID, _obs - freq + 1, freq);

      end;

    output ple;
    bias1=-(&RESULT-last)*ple;
    end;

    ple     = 0;
    &RESULT = 0;
    _pl_mn  = _pl_mn + (last - &RESULT) * (1 - ple);

    /* Output ple mean, ordinary mean, and std. error of ord. mean */

    output ple_mn;

    ucb_ple = 1 - (1 - &CONFID)**(1 / _obs);
    lcb_ple = 0;
    a       = 0;
    se_ple  = .;
    enter   = 0.1;
    remain  = 0.1;

    /* Output standard error, lower and upper confidence limits
       for ple */

    output ple;

  end;

run;


proc sort  data = ple;
  by &GROUP &PARM &RESULT ple;
run;

*print here for ple distribution function estimates;
```

```
*proc print data=ple;
*by &GROUP &PARM;
*id &GROUP &PARM;

 /* Compute standard errors of PLE means using statistic "a" computed
    above, as on page 98 in text by Lawless (see above)
    Also eliminate superfluous points to actual PLEs */

 data ple_se (keep = &PARM &GROUP &QUAL &ID _pl_sem)
      ple_cb (keep = &PARM &GROUP &QUAL &ID ple se_ple
                     &RESULT lcb_ple ucb_ple);

  set ple;
  by &GROUP &PARM &RESULT;

  retain sa _pl_sem 0;

  /* Output ple, and std. error, LCL, and UCL of ples */

  if last.&RESULT then output ple_cb;

  sa = sa + a;

  /* Compute standard error of ple mean */

  _pl_sem = _pl_sem + sa * sa * (enter - remain) / (enter * remain);

  /* Output standard error of ple mean */

  if last.&PARM then do;
    output ple_se;
    sa      = 0;
    _pl_sem = 0;
  end;

run;


/* Compute percentile estimates and their confidence bounds by
   inverting PLEs and their upper and lower confidence bounds.
   Where PLE is constant, use U-inverse (p) = inf{x|U(x) >= p}, and
   L-inverse (p) = sup{x|L(x) <= p}, and point estimate is average
   of upper and lower extremes. */

 data ple_cb;

  set ple_cb;
  by &GROUP &PARM;
```

```
array pct (i) &P;  /* Percentiles */
array lcb (i) &L;  /* LTB */
array ucb (i) &U; /* UTB */
array qnt (&N) _temporary_ (&TOLER);

/*Code variables indicate whether percentiles and corresponding
  lcb's and ucb's have been reached during passage through data.*/

array codex (i) codex1 - codex&N;
array codel (i) codel1 - codel&N;
array codeu (i) codeu1 - codeu&N;

keep &PARM &GROUP &ID
    &P &L &U;

retain lastx  codex1 - codex&N  codel1 - codel&N  codeu1 - codeu&N
           &P &L &U;

if first.&PARM then do;

  lastx = .;

  /* Set percentiles and confidence bounds to missing and zero
     to denote not defined */

  do over pct;
    pct   = .;
    ucb   = .;
    lcb   = .;
    codex = 0;
    codel = 0;
    codeu = 0;
  end;

end;

/* Rounding to ensure exact matches are not missed */

do over pct;

  /* code values = 0 denote not yet defined
     code values = 1 denote partially done
     code values = 2 denote done            */

  if codel = 0 and round (ucb_ple - qnt(i), &TOL) ge 0 then do;
    codel = 2;
    lcb   = &RESULT;  /* Set lower confidence bound */
  end;
```

```
    if round (lcb_ple - qnt(i), &TOL) > 0 then do;
      if codeu = 0 then do;
        codeu = 2;
        ucb   = &RESULT;  /* Set upper confidence bound */
      end;
    end;

    else do;
      ucb   = .;
      codeu = 0;
    end;

    if codex = 0 then do;

      if round (ple - qnt(i), &TOL) = 0 then codex = 1;

      else if round (ple - qnt(i), &TOL) > 0 then do;
        codex = 2;
        pct   = &RESULT;  /* Percentiles */
      end;

    end;

    else if codex = 1 and round (ple - qnt(i), &TOL) > 0 then do;
      codex = 2;
      pct   = (lastx + &RESULT) / 2;  /* Percentiles */
    end;

  end;

  lastx = &RESULT;

  if last.&PARM then output;

run;


data ple_mn;

  merge ple_mn
        ple_se
        ple_cb;
  by &GROUP &PARM;

  label _pl_sem = "PLE mean, std. err."
        _pl_lcl = "PLE mean, LCL (p=&CONFID)"
        _pl_ucl = "PLE mean, UCL (p=&CONFID)"
```

```
                        _det = 'Detects'
  ;

  /* Use t-distribution with n-1 df here. Alternatives might be
     normal distribution or t with (n-_ndt) df. */

  /* Standard error of ple mean */

  if _det > 1 then _pl_sem = sqrt (_pl_sem * _det / (_det - 1));
  else _pl_sem = .;

  /* Confidence limits */

  if _obs > 1 then do;
    _pl_ucl = _pl_mn + (_pl_sem * tinv (&CONFID, _obs - 1));
    _pl_lcl = _pl_mn - (_pl_sem * tinv (&CONFID, _obs - 1));
  end;

  drop _lt00001 _qu00001 _ut00001;

run;

/* End SAS code for PLE-based statistics. */



data detects;

  merge stat (in = in1)
        ple_mn  (in = in3);
  by &GROUP &PARM;

  drop _ndt _avg_ _var_ _cov_ _sem_ _sev_ bias1 &QUAL;
  retain _avg_ _var_ _sem_ _sev_ _cov_;

  label _obs    = "Observed"
        _ar_mn  = "Ordinary mean"
        _min    = "Minimum"
        _max    = "Maximum"
        _ar_sem = "Ordinary mean, std. err."
  ;

  if last.&PARM then do;

   /*  COMPUTE BIAS ADJUSTED PRODUCT LIMIT ESTIMATES & LIMITS   */

        _pl_mn = _pl_mn - bias1;
        _pl_lcl=max(0,_pl_lcl - bias1);
```

```
        _pl_ucl=_pl_ucl - bias1;

    ;

     output;

     _avg_ = .;  _var_ = .;  _sem_ = .;  _sev_ = .;  _cov_ = .;

   end;

 run;

 data &OUTPUT;

   set detects (in = in1)  /* Summary statistics for detects */
       nonstat (in = in2); /* Summary statistics for nondetects */
                           if in2 then _det=0;
       label &labell;
       label &labelq;
       label &labelu;

 run;

 proc sort  data = &OUTPUT;
   by &GROUP &PARM;
 run;

%MEND ple;
```

APPENDIX C

THE LOGCONF MACRO

The logconf macro is listed below.

```
  /*************************************************************************\
  *
  * logconf.sas: Calculate statistics based on the longnormal distribution when
  * there is no censoring.
                *
  *    © 1999
  *     Lockheed Martin Energy Research Corporation
  *     All rights reserved
  *
  *    Neither the Government nor LMER makes any warrantee, express or implied, or
  *    assumes any liability or responsibility for the use of this software.
  *
  \*************************************************************************/

%macro logconf(INPUT, OUTPUT, GROUP, RESULT, PARM, ID);

libname temp '/home/sun4/u5/schmoyer/lyon/project/macros';
*Put hfun.ssd01 SAS data set with Land H-function values
in same directory as macros;

data &output;
set &INPUT;
lresult=log(&RESULT);

proc sort data=&output;
by &GROUP &PARM;

proc means data=&output noprint;
var lresult;
by &GROUP &PARM;
output out=a (drop=_type_ _freq_)
n=n
mean=mean
std=stand;
id &ID;

proc sort;
by n;

data &output;
merge a (in=in1) temp.hfun;
array col {99} col1-col99;
by n;
if in1;

/* maximum number of samples.  Anything above
```

```
this number uses the theory that
(X-u)/(S/Sqrt(n)) follows N(0,1)*/
 _max_n_  = 1000;

temp=10*stand;

/* constants for the table of h values
used to calculate "Land's method" of UCL
for the 95% UCL.  This assumes constant
interval sized for the whole H table. */

 /* If the sigma is lower than what is found
    in the table, use the lowest interval */
 IF temp < 1 THEN temp = 1;

 /* If the sigma is higher than what is found
    in the table, use the highest interval */
 IF temp > 99 THEN temp = 99;

index=int(temp);

  /* Interpolate for new H value. */
h=col(index)*(index+1-temp)+col(index+1)*(temp-index);

IF N <= _max_n_ then
   UCL_95 = exp(mean + 0.5*stand*stand + stand * H/SQRT(N-1));
*Calculate Upper 95% Confidence Limit via "Land's Method".;

ELSE UCL_95 = exp(mean + stand*stand/2 + tinv(.95,n-1)*SQRT(stand*stand/N +
   stand*stand*stand*stand/(2*(N+1))));

   /* Calculate Upper 95% Confidence Limit via Cox's method
      exp((y+S^2/2) + Z.95*Beta)  Y = sample mean
      Beta = SQRT(S^2/(nu+1) + S^4/(2*(nu+2)))
      Z.95 = 95 percentile of N(0,1) = 1.645.
      t-95 substituted for z-95 to force exact agreement with
      lnor macro--RLS.*/

keep n &GROUP &PARM &ID UCL_95 mean stand;
label ucl_95="Land's lognormal 95% UCL"
n="Observed"
mean='Log-scale mean'
stand='Log-scale standard deviation';
rename n=_obs;

run;
%mend logconf;
```

APPENDIX D

THE SAS PROGRAM USED TO PRINT THE INPUT DATA, TO CALL THE SAS MACROS, AND TO PRODUCE THE OUTPUT FILES

The SAS program used to print the input data, to call the SAS macros, and to produce the output files is listed below.

```
  /*************************************************************************\
  *   SAS program that prints the data, calls the SAS macros, and prints the output
  *
  *   © 1999
  *    Lockheed Martin Energy Research Corporation
  *    All rights reserved
  *
  *    Neither the Government nor LMER makes any warrantee, express or implied, or
  *    assumes any liability or responsibility for the use of this software.
           *
  \*************************************************************************/
options ls=72 ps=60 sasautos='../macros' symbolgen mlogic mprint noovp nodate
nonumber;
libname there '/home/sun4/u5/schmoyer/lyon/project/transport.sam';

proc print data=there.subset label split='*';
title1 ' ';
title2 ' ';
title3
'Table 1. Groundwater Lead Concentrations (mg/L) from Y-12 Fuel Station';
var date_col result lower qual;
by station;
id station;
label station='Monitoring*Station'
result='Analytical*Result/Detection*Limit'
date_col='Date*Collected'
qual='Qualifier'
lower='Lower (for*lnor.sas)';
format date_col date7. lower result 5.4;

%lnor (there.subset, outputl, station, result, lower, qual, analysis, ,
.95,.50 .75 .90);
run;

options ls=72 ps=56;
proc contents  data = outputl;
title1 ' ';
title2 ' ';
title3 "Table 2. SAS Macro lnor.sas Output Data Set Contents";
title4 "for Groundwater Lead (Y-12 Fuel Station)";
run;

options ls=72 ps=63;
proc print data=outputl label split=',';
title1 ' ';
title2 ' ';
```

```
title3 'Table 3. Output of SAS Macro lnor.sas';
title4 'for Groundwater Lead (Y-12 Fuel Station)';
id station;
var _OBS _DET _MIN _MAX _AR_LCL _AR_MN _AR_UCL _AR_SEM
_LN_LCL _LN_MN _LN_UCL _LT500 _QU500 _UT500
_LT750 _QU750 _UT750 _LT900 _QU900 _UT900 _LS_MN _LS_SEM _LS_STD
_N_LCL _N_MN _N_UCL _N_SEM _N_STD
_NL500 _NQ500 _NU500 _NL750 _NQ750 _NU750 _NL900 _NQ900 _NU900;
label _LT500='95% LTB, 50%-ile'
_QU500='Estimate, 50%-ile'
_UT500='95% UTB, 50%-ile'
_LT750='95% LTB, 75%-ile'
_QU750='Estimate, 75%-ile'
_UT750='95% UTB, 75%-ile'
_LT900='95% LTB, 90%-ile'
_QU900='Estimate, 90%-ile'
_UT900='95% UTB, 90%-ile'
station='Station'
_ls_std = "Ln-scale,std. dev."
_NQ500='Normal Model, Estimate, 50%-ile'
_NQ750='Normal Model, Estimate, 75%-ile'
_NQ900='Normal Model, Estimate, 90%-ile'
_NL500='Normal Model, 95% LTB, 50%-ile'
_NL750='Normal Model, 95% LTB, 75%-ile'
_NL900='Normal Model, 95% LTB, 90%-ile'
_NU500='Normal Model, 95% UTB, 50%-ile'
_NU750='Normal Model, 95% UTB, 75%-ile'
_NU900='Normal Model, 95% UTB, 90%-ile';
run;

%ple (there.subset, outputp, station, result, qual, analysis, ,
.95,.50 .75 .90);

options ls=72 ps=56;
proc contents  data = outputp;
footnote;
title1 ' ';
title2 ' ';
title3 "Table 4. SAS Macro ple.sas Output Data Set Contents";
title4 "for Groundwater Lead (Y-12 Fuel Station)";
run;

options ls=72 ps=60;
proc print data=outputp label split=',';
title1 ' ';
title2 ' ';
title3 'Table 5. Output of SAS Macro ple.sas';
title4 'for Groundwater Lead (Y-12 Fuel Station)';
```

```
id station;
var _OBS _DET _MIN _MAX _AR_LCL _AR_MN _AR_UCL _AR_SEM _PL_LCL _PL_MN
_PL_UCL _PL_SEM _LT50 _QU50 _UT50 _LT75 _QU75 _UT75 _LT90 _QU90 _UT90;
label _LT50='95% LTB, 50%-ile'
_QU50='Estimate, 50%-ile'
_UT50='95% UTB, 50%-ile'
_LT75='95% LTB, 75%-ile'
_QU75='Estimate, 75%-ile'
_UT75='95% UTB, 75%-ile'
_LT90='95% LTB, 90%-ile'
_QU90='Estimate, 90%-ile'
_UT90='95% UTB, 90%-ile'
station='Station';
run;

proc print data=there.subsetd label split='*';
title1 ' ';
title2 ' ';
title3
'Table 6. Groundwater Barium Concentrations (mg/L) from Y-12 Fuel Station';
var date_col result;
by station;
id station;
label station='Monitoring*Station'
result='Analytical*Result'
date_col='Date*Collected';
format date_col date7. result 4.2;

%logconf(there.subsetd, outputl, station , RESULT, analysis, );

options ls=72 ps=56;
proc contents  data = outputl;
title1 ' ';
title2 ' ';
title3 "Table 7. SAS Macro logconf.sas Output Data Set Contents";
title4 "for Groundwater Barium (Y-12 Fuel Station)";
run;

options ls=72 ps=63;
proc print data=outputl label split=',';
title1 ' ';
title2 ' ';
title3 'Table 8. Output of SAS Macro logconf.sas';
title4 'for Groundwater Barium (Y-12 Fuel Station)';
var _obs mean stand ucl_95;
id station;
label station='Station';
run;
```