United States
Environmental Protection
Agency

ProUCL Version 4.0
Technical Guide

RESEARCH AND DEVELOPMENT

EPA/600/R-07/041
April 2007
www.epa.gov

ProUCL Version 4.0 Technical Guide

Prepared for

Brian Schumacher

U.S. Environmental Protection Agency
Office of Research and Development
National Exposure Research Laboratory
Environmental Sciences Division
Technology Support Center
Characterization and Monitoring Branch
944 E. Harmon Ave.
Las Vegas, NV 89119

Prepared by

Anita Singh, Ph.D.[1]
Ashok K. Singh, Ph.D.[2]

[1]Lockheed Martin Environmental Services
1050 E. Flamingo Road, Suite N240
Las Vegas, NV 89119

[2]Department of Hotel Management
University of Nevada, Las Vegas
Las Vegas, NV 89154

U.S. Environmental Protection Agency
Office of Research and Development
Washington, DC 20460

129cmb07

Notice

The United States Environmental Protection Agency (EPA) through its Office of Research and Development funded and managed the research described here. It has been peer reviewed by the EPA and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation by the EPA for use.

ProUCL software was developed by Lockheed Martin under a contract with the EPA and is made available through the EPA Technical Support Center in Las Vegas, Nevada.

Use of any portion of ProUCL that does not comply with the ProUCL Technical Guide is not recommended.

ProUCL contains embedded licensed software. Any modification of the ProUCL source code may violate the embedded licensed software agreements and is expressly forbidden.

ProUCL software provided by the EPA was scanned with McAfee VirusScan v4.5.1 SP1 and is certified free of viruses.

With respect to ProUCL distributed software and documentation, neither the EPA nor any of their employees, assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed. Furthermore, software and documentation are supplied “as-is” without guarantee or warranty, expressed or implied, including without limitation, any warranty of merchantability or fitness for a specific purpose.

Executive Summary

Statistical inference, including both estimation and hypotheses testing approaches, is routinely used to:

1. Estimate environmental parameters of interest, such as exposure point concentration (EPC) terms, not-to-exceed values, and background level threshold values (BTVs) for contaminants of potential concern (COPC),

2. Identify areas of concern (AOC) at a contaminated site,

3. Compare contaminant concentrations found at two or more AOCs of a contaminated site,

4. Compare contaminant concentrations found at an AOC with background or reference area contaminant concentrations, and

5. Compare site concentrations with a cleanup standard to verify the attainment of cleanup standards.

Several exposure and risk management and cleanup decisions in support of United States Environmental Protection Agency (EPA) projects are often made based upon the mean concentrations of the COPCs. A 95% upper confidence limit (UCL95) of the unknown population (e.g., an AOC) arithmetic mean (AM), μ_1, can be used to:

- Estimate the EPC term of the AOC under investigation,
- Determine the attainment of cleanup standards,
- Compare site mean concentrations with reference area mean concentrations, and
- Estimate background level mean contaminant concentrations. The background mean contaminant concentration level may be used to compare the mean of an area of concern. It should be noted that it is not appropriate to compare individual point-by-point site observations with the background mean concentration level.

It is important to compute a reliable and stable *UCL95* of the population mean using the available data. The *UCL95* should approximately provide the 95% coverage for the unknown population mean, μ_1. Based upon the available background data, it is equally important to compute reliable and stable upper percentiles, upper prediction limits (*UPLs*), or upper tolerance limits (*UTLs*). These upper limits based upon background (or reference) data are used as estimates of BTVs, compliance limits (*CL*), or not-to-exceed values. These upper limits are often used in site (point-by-point) versus background comparison evaluations.

Environmental scientists often encounter trace level concentrations of COPCs when evaluating sample analytical results. Those low level analytical results cannot be measured accurately and, therefore, are typically reported as less than one or more detection limit (*DL*) values (also called nondetects). However, practitioners need to obtain reliable estimates of the population mean, μ_1, and the population standard deviation, σ_1, and upper limits including the *UCL* of the population mass or mean, the *UPL*, and the *UTL* based upon data sets with nondetect (*ND*) observations. Additionally, they may have to use hypotheses testing approaches to verify the attainment of cleanup standards, and compare site and background concentrations of COPCs as mentioned above.

Background evaluation studies, BTVs, and not-to-exceed values should be estimated based upon defensible background data sets. The estimated BTVs or not-to-exceed values are then used to identify the COPCs, to identify the site AOCs or hot spots, and to compare the contaminant concentrations at a site with background concentrations. The use of appropriate statistical methods and limits for site versus background comparisons is based upon the following factors:

1. Objective of the study,
2. Environmental medium (e.g., soil, groundwater, sediment, air) of concern,
3. Quantity and quality of the available data,
4. Estimation of a not-to-exceed value or of a mean contaminant concentration,
5. Pre-established or unknown cleanup standards and BTVs, and
6. Sampling distributions (parametric or nonparametric) of the concentration data sets collected from the site and background areas under investigation.

In background versus site comparison evaluations, the environmental population parameters of interest may include:

- Preliminary remediation goals (PRGs),
- Soil screening levels (SSLs),
- RBC standards,
- BTVs, not-to-exceed values, and
- Compliance limit, maximum concentration limit (MCL), or alternative concentration limit (ACL), frequently used in groundwater applications.

When the environmental parameters listed above are not known or pre-established, appropriate upper statistical limits are used to estimate those parameters. The UPL, UTL, and upper percentiles are used to estimate the BTVs and not-to-exceed values. Depending upon the site data availability, point-by-point site observations are compared with the estimated (or pre-established) BTVs and not-to-exceed values. If enough site and background data are available, two-sample hypotheses testing approaches are used to compare site concentrations with background concentrations levels. These statistical methods can also be used to compare contaminant concentrations of two site AOCs, surface and subsurface contaminant concentrations, or upgradient versus monitoring well contaminant concentrations.

ProUCL Version 4.0 (ProUCL 4.0) is an upgrade of ProUCL Version 3.0 (EPA, 2004). ProUCL 4.0 contains statistical methods to address various environmental issues for both full data sets without nondetects and for data sets with NDs (also known as left-censored data sets).

ProUCL 4.0 contains:

1. Rigorous parametric and nonparametric (including bootstrap methods) statistical methods (instead of simple ad hoc or substitution methods) that can be used on full data sets without nondetects and on data sets with below detection limit (BDL) or ND observations.

2. State-of-the-art parametric and nonparametric UCL, UPL, and UTL computation methods. These methods can be used on full-uncensored data sets without nondetects and also on data sets with BDL observations. Some of the methods (e.g., Kaplan-Meier method, ROS methods) are applicable on left-censored data sets having multiple detection limits. The UCL and other upper limit computation methods cover a wide range of skewed data sets with and without the BDLs.

3. Single sample (e.g., Student's t-test, sign test, proportion test, Wilcoxon Singed Rank test) and two-sample (Student's t-test, Wilcoxon-Mann-Whitney test, Gehan test, quantile test) parametric and nonparametric hypotheses testing approaches for data sets with and without ND observations. These hypothesis testing approaches can be used to: verify the attainment of cleanup standards, perform site versus background comparisons, and compare two or more AOCs, monitoring wells (MWs).

4. The single sample hypotheses testing approaches are used to compare site mean, site median, site proportion, or a site percentile (e.g., 95^{th}) to a compliance limit (action level, regularity limit). The hypotheses testing approaches can handle both full-uncensored data sets without nondetects, and left-censored data sets with nondetects. Simple two-sample hypotheses testing methods to compare two populations are available in ProUCL 4.0, such as two-sample t-tests, Wilcoxon-Mann-Whitney (WMW) Rank Sum test, quantile test, Gehan's test, and dispersion test. Variations of hypothesis testing methods (e.g., Levene's method to compare dispersions, generalized WRS test) are easily available in most commercial and freely available software packages (e.g., MINITAB, R).

5. ProUCL 4.0 also includes graphical methods (e.g., box plots, multiple Q-Q plots, histogram) to compare two or more populations. ProUCL 4.0 can also be used to display a box plot of one population (e.g., site data) with compliance limits or upper limits (e.g., UPL) of other population (background area) superimposed on the same graph. This kind of graph provides a useful visual comparison of site data with a compliance limit or BTVs. Graphical displays of a data set (e.g., Q-Q plot) should be used to gain insight knowledge contained in a data set that may not otherwise be clear by looking at simple test statistics such as t-test, Dixon test statistic, or Shapiro-Wilk (S-W) test statistic.

6. ProUCL 4.0 can process multiple contaminants (variables) simultaneously and has the capability of processing data by groups. A valid group column should be included in the data file.

7. ProUCL 4.0 provides GOF test for data sets with nondetects. The user can create additional columns to store extrapolated (estimated) values for nondetects based upon normal ROS, gamma ROS, and lognormal ROS (robust ROS) methods.

ProUCL 4.0 retains all of the capabilities of ProUCL 3.0, including goodness-of-fit (GOF) tests for a normal, lognormal, and a gamma distribution and computation of UCLs based upon full data sets without nondetects. Graphical displays and GOF tests for data sets with BDL observations have also been included in ProUCL 4.0. It is re-emphasized that the computation of appropriate UCLs, UPLs, and other limits is based upon the assumption that the data set under study represents a single a single population. This means that the data set used to compute the limits should represent a single statistical population. For example, a background data set should represent a defensible background data set free of outlying observations. ProUCL 4.0 includes simple and commonly used classical outlier identification procedures,

such as the Dixon test and the Rosner test. These procedures are included as an aid to identify outliers. These simple classical outlier tests often suffer from masking effects in the presence of multiple outliers. Description and use of robust and resistant outlier procedures is beyond the scope of ProUCL 4.0.

It is suggested that the classical outlier procedures should always be accompanied by graphical displays including box plots and Q-Q plots. The use of a Q-Q plot is useful to identify multiple or mixture samples that might be present in a data set. However, the decision regarding the proper disposition of outliers (e.g., to include or not to include outliers in statistical analyses; or to collect additional verification samples) should be made by members of the project team and experts familiar with site and background conditions. Guidance on the disposition of outliers and their accommodation in a data set by using a transformation (e.g., lognormal distribution) is discussed in Chapter 1 of this Technical Guide.

ProUCL 4.0 has improved graphical methods, which may be used to compare the concentrations of two or more populations such as:

1. Site versus background populations,
2. Surface versus subsurface concentrations,
3. Concentrations of two or more AOCs, and
4. Identification of mixture samples and/or potential outliers

These graphical methods include multiple quantile-quantile (Q-Q) plots, side-by-side box plots, and histograms. Whenever possible, it is desirable to supplement statistical results with useful visual displays of data sets. There is no substitute for graphical displays of a data set. For example, in addition to providing information about the data distribution, a normal Q-Q plot can also help identify outliers and multiple populations that may be present in a data set. On a Q-Q plot, observations well separated from the majority of the data may represent potential outliers, and jumps and breaks of significant magnitude may suggest the presence of observations from multiple populations in the data set. It is suggested that analytical outlier tests (e.g., Rosner test) and goodness-of-fit (G.O.F.) tests (e.g., SW test) should always be supplemented with the graphical displays such as Q-Q plot and box plot.

ProUCL 4.0 serves as a companion software package for the *Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites* (EPA, 2002a) and the *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b). ProUCL 4.0 is also useful to verify the attainment of cleanup standards (EPA, 1989). ProUCL 4.0 can also be used to perform two-sample hypotheses tests and to compute various upper limits often needed in groundwater monitoring applications (EPA, 1992 and EPA, 2004).

Acronyms and Abbreviations

ACL	alternative concentration limit
A-D, AD	Anderson-Darling test
AM	arithmetic mean
AOC	area(s) of concern
BC	Box-Cox transformation
BCA	bias-corrected accelerated bootstrap method
BD	binomial distribution
BDL	below detection limit
BTV	background threshold value
CC	confidence coefficient
CDF, cdf	cumulative distribution function
CERCLA	Comprehensive Environmental Recovery, Compensation, and Liability Act
CL	compliance limit
CLT	central limit theorem
CMLE	Cohen's maximum likelihood estimate
COPC	contaminant(s) of potential concern
C_S	cleanup standards
CV	coefficient of variation
DCGL	Design Concentration Guideline Level
df	degrees of freedom
DL, L	detection limit
DL/2 (t)	UCL based upon DL/2 method using Student's t-distribution cutoff value
DL/2 Estimates	estimates based upon data set with nondetects replaced by half of the respective detection limits
DOE	Department of Energy
DQO	data quality objectives

EA	exposure area
EDF	empirical distribution function
EM	expectation maximization
EU	exposure unit
EPA	Environmental Protection Agency
EPC	exposure point concentration
GOF, G.O.F.	goodness-of-fit
H-UCL	UCL based upon Land's H-statistic
IQR	interquartile range
KM (%)	UCL based upon Kaplan-Meier estimates using the percentile bootstrap method
KM (Chebyshev)	UCL based upon Kaplan-Meier estimates using the Chebyshev inequality
KM (t)	UCL based upon Kaplan-Meier estimates using the Student's t-distribution cutoff value
KM (z)	UCL based upon Kaplan-Meier estimates using standard normal distribution cutoff value
K-M, KM	Kaplan-Meier
K-S, KS	Kolmogorov-Smirnov
LBGR	Lower Bound of the Gray Region
LN	lognormal distribution
MAD	median absolute deviation
MCL	maximum concentration limit, maximum compliance limit
MDD	minimum detectable difference
ML	maximum likelihood
MLE	maximum likelihood estimate
MLE (t)	UCL based upon maximum likelihood estimates using Student's t-distribution cutoff value
MLE (Tiku)	UCL based upon maximum likelihood estimates using the Tiku's method
Multi Q-Q	multiple quantile-quantile plot

MV	minimum variance
MVUE	minimum variance unbiased estimate
MW	monitoring well
ND	nondetect
NRC	Nuclear Regulatory Commission
OLS	ordinary least squares
ORD	Office of Research and Development
PDF, pdf	probability density function
PLE	product limit estimate
PRG	preliminary remediation goals
Q-Q	quantile-quantile
RBC	risk-based cleanup
RCRA	Resource Conservation and Recovery Act
RL	reporting limit
RMLE	restricted maximum likelihood estimate
ROS	regression on order statistics
RSD	relative standard deviation
RU	remediation unit
RV	random variable
S	substantial difference
SD, *Sd, sd*	standard deviation
SE	Standard error
SND	standard normal distribution
SSL	soil screening levels
S-W, SW	Shapiro-Wilk

U.S. EPA, USEPA	United States Environmental Protection Agency
UCL	upper confidence limit
UCL95	a 95% upper confidence limit
UMLE	unbiased maximum likelihood estimate method
UPL	upper prediction limit
UTL	upper tolerance limit
WMW	Wilcoxon-Mann-Whitney
WRS	Wilcoxon Rank Sum
WSR	Wilcoxon Signed Rank

Glossary

Anderson-Darling (AD) test: The Anderson-Darling test assesses whether known data come from a specified distribution.

Background Measurements: The measurements that are not related to the site. Background sources can be naturally occurring or anthropogenic (man-made).

Bias: The systematic or persistent distortion of a measured value from its true value (this can occur during sampling design, the sampling process, or laboratory analysis).

Bootstrap Method: The bootstrap method is a computer-based method for assigning measures of accuracy to sample estimates. This technique allows estimation of the sample distribution of almost any statistic using only very simple methods. Bootstrap methods are generally superior to ANOVA for small data sets or where sample distributions are non-normal.

Central Limit Theorem (CLT): The central limit theorem states that given a distribution with a mean μ and variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean (μ) and a variance σ^2/N as N, the sample size, increases.

Coefficient of Variation (CV): A dimensionless quantity used to measure the spread of data relative to the size of the numbers. For a normal distribution, the coefficient of variation is given by s/xBar. Also known as the relative standard deviation (RSD).

Confidence Coefficient: The confidence coefficient (a number in the closed interval [0, 1]) associated with a confidence interval for a population parameter is the probability that the random interval constructed from a random sample (data set) contains the true value of the parameter. The confidence coefficient is related to the significance level of an associated hypothesis test by the equality: level of significance = 1 – confidence coefficient.

Confidence Interval: Based upon the sampled data set, a confidence interval for a parameter is a random interval within which the unknown population parameter, such as the mean, or a future observation, x0, falls.

Confidence Limit: The lower or an upper boundary of a confidence interval. For example, the 95% upper confidence limit (UCL) is given by the upper bound of the associated confidence interval.

Coverage, Coverage Probability: The coverage probability (e.g., = 0.95) of an upper confidence limit (UCL) of the population mean represents the confidence coefficient associated with the UCL.

Data Quality Objectives (DQOs): Qualitative and quantitative statements derived from the DQO process that clarify study technical and quality objectives, define the appropriate type of data, and specify tolerable levels of potential decision errors that will be used as the basis for establishing the quality and quantity of data needed to support decisions.

Detection Limit: A measure of the capability of an analytical method to distinguish samples that do not contain a specific analyte from samples that contain low concentrations of the analyte. The lowest concentration or amount of the target analyte that can be determined to be different from zero by a single

measurement at a stated level of probability. Detection limits are analyte- and matrix-specific and may be laboratory-dependent.

Empirical Distribution Function (EDF): In statistics, an empirical distribution function is a cumulative probability distribution function that concentrates probability 1/*n* at each of the *n* numbers in a sample.

Estimate: A numerical value computed using a random data set (sample), and is used to guess (estimate) the population parameter of interest (e.g., mean). For example, a sample mean represents an estimate of the unknown population mean.

Expectation Maximization (EM): The EM algorithm is used to approximate a probability function (p.f. or p.d.f.). EM is typically used to compute maximum likelihood estimates given incomplete samples.

Exposure Point Concentration (EPC): The contaminant concentration within an exposure unit to which the receptors are exposed. Estimates of the EPC represent the concentration term used in exposure assessment.

Extreme Values: The minimum and the maximum values.

Goodness-of-Fit (GOF): In general, the level of agreement between an observed set of values and a set wholly or partly derived from a model of the data.

Gray Region: A range of values of the population parameter of interest (such as mean contaminant concentration) within which the consequences of making a decision error are relatively minor. The gray region is bounded on one side by the action level. The width of the gray region is denoted by the Greek letter delta in this guidance.

H-Statistic: The unique symmetric unbiased estimator of the central moment of a distribution.

H-UCL: UCL based on Land's H-Statistic.

Hypothesis: Hypothesis is a statement about the population parameter(s) that may be supported or rejected by examining the data set collected for this purpose. There are two hypotheses: a null hypothesis, (H_0), representing a testable presumption (often set up to be rejected based upon the sampled data), and an alternative hypothesis (H_A), representing the logical opposite of the null hypothesis.

Jackknife Method: A statistical procedure in which, in its simplest form, estimates are formed of a parameter based on a set of N observations by deleting each observation in turn to obtain, in addition to the usual estimate base d on N observations, N estimates each based on N-1 observations.

Kolmogorov-Smirnov (KS) test: The Kolmogorov-Smirnov test is used to decide if a sample comes from a population with a specific distribution. The Kolmogorov-Smirnov test is based on the empirical distribution function (EDF).

Level of Significance: The error probability (also known as false positive error rate) tolerated of falsely rejecting the null hypothesis and accepting the alternative hypothesis.

Lilliefors test: A test of normality for large data sets when the mean and variance are unknown.

Maximum Likelihood Estimates (MLE): Maximum likelihood estimation (MLE) is a popular statistical method used to make inferences about parameters of the underlying probability distribution of a given data set.

Mean: The sum of all the values of a set of measurements divided by the number of values in the set; a measure of central tendency.

Median: The middle value for an ordered set of n values. Represented by the central value when n is odd or by the average of the two most central values when n is even. The median is the 50th percentile.

Minimum Detectable Difference (MDD): The minimum detectable difference (MDD) is the smallest difference in means that the statistical test can resolve. The MDD depends on sample-to-sample variability, the number of samples, and the power of the statistical test.

Minimum Variance Unbiased Estimates (MVUE): A minimum variance unbiased estimator (MVUE or MVU estimator) is an unbiased estimator of parameters, whose variance is minimized for all values of the parameters. If an estimator is unbiased, then its mean squared error is equal to its variance.

Nondetect (ND): Censored data values.

Nonparametric: A term describing statistical methods that do not assume a particular population probability distribution, and are therefore valid for data from any population with any probability distribution, which can remain unknown.

Optimum: An interval is optimum if it possesses optimal properties as defined in the statistical literature. This may mean that it is the shortest interval providing the specified coverage (e.g., 0.95) to the population mean. For example, for normally distributed data sets, the UCL of the population mean based upon Student's t distribution is optimum.

Outlier: Measurements (usually larger or smaller than the majority of the data values in a sample) that are not representative of the population from which they were drawn. The presence of outliers distorts most statistics if used in any calculations.

p-value: In statistical hypothesis testing, the p-value of an observed value $t_{observed}$ of some random variable T used as a test statistic is the probability that, given that the null hypothesis is true, T will assume a value as or more unfavorable to the null hypothesis as the observed value $t_{observed}$.

Parameter: A parameter is an unknown constant associated with a population.

Parametric: A term describing statistical methods that assume a normal distribution.

Population: The total collection of N objects, media, or people to be studied and from which a sample is to be drawn. The totality of items or units under consideration.

Prediction Interval: The interval (based upon historical data, or a background well) within which a newly and independently obtained (often labeled as a future observation) site observation (from a compliance well) of the predicted variable (lead) falls with a given probability (or confidence coefficient).

Probability of Type 2 Error (=β): The probability, referred to as β (beta), that the null hypothesis will not be rejected when in fact it is false (false negative).

Probability of Type I Error = Level of Significance (= α): The probability, referred to as α (alpha), that the null hypothesis will be rejected when in fact it is true (false positive).

p^{th} Percentile: The specific value, X_p of a distribution that partitions a data set of measurements in such a way that the p percent (a number between 0 and 100) of the measurements fall at or below this value, and (100-p) percent of the measurements exceed this value, X_p.)

p^{th} Quantile: The specific value of a distribution that divides the set of measurements in such a way that the proportion, p, of the measurements falls below (or are equal to) this value, and the proportion (1-p) of the measurements exceed this value.

Quality Assurance: An integrated system of management activities involving planning, implementation, assessment, reporting, and quality improvement to ensure that a process, item, or service is of the type and quality needed and expected by the client.

Quality Assurance Project Plan: A formal document describing, in comprehensive detail, the necessary QA, QC, and other technical activities that must be implemented to ensure that the results of the work performed will satisfy the stated performance criteria.

Quantile Plot: A graph that displays the entire distribution of a data set, ranging from the lowest to the highest value. The vertical axis represents the measured concentrations, and the horizontal axis is used to plot the percentiles of the distribution.

Range: The numerical difference between the minimum and maximum of a set of values.

Regression on Order Statistics (ROS): A regression line is fit to the normal scores of the order statistics for the uncensored observations and then to fill in values extrapolated from the straight line for the observations below the detection limit.

Resampling: The repeated process of obtaining representative samples and/or measurements of a population of interest.

Reliable UCL: This is similar to a stable UCL.

Robustness: Robustness is used to compare statistical tests. A robust test is the one with good performance (that is not unduly affected by outliers) for a wide variety of data distributions.

Sample: A sample here represents a random sample (data set) obtained from the population of interest (e.g., a site area, a reference area, or a monitoring well). The sample is supposed to be a representative sample of the population under study. The sample is used to draw inferences about the population parameter(s).

Shapiro-Wilk (SW) test: In statistics, the Shapiro-Wilk test tests the null hypothesis that a sample $x_1, ..., x_n$ came from a normally distributed population.

Skewness: A measure of asymmetry of the distribution of the characteristic under study (e.g., lead concentrations). It can also be measured in terms of the standard deviation of log-transformed data. The higher is the standard deviation, the higher is the skewness.

Stable UCL: The UCL of a population mean is a stable UCL if it represents a number of practical merits, which also has some physical meaning. That is, a stable UCL represents a realistic number (e.g., contaminant concentration) that can occur in practice. Also, a stable UCL provides the specified (at least approximately, as much as possible, as close as possible to the specified value) coverage (e.g., ~0.95) to the population mean.

Standard Deviation (sd): A measure of variation (or spread) from an average value of the sample data values.

Standard Error (SE): A measure of an estimate's variability (or precision). The greater the standard error in relation to the size of the estimate, the less reliable the estimate. Standard errors are needed to construct confidence intervals for the parameters of interests such as the population mean and population percentiles.

Tolerance Limit: A confidence limit on a percentile of the population rather than a confidence limit on the mean. For example, a 95 percent one-sided TL for 95 percent coverage represents the value below which 95 percent of the population values are expected to fall with 95 percent confidence. In other words, a 95% UTL with coverage coefficient 95% represents a 95% upper confidence limit for the 95^{th} percentile.

Unreliable UCL, Unstable UCL, Unrealistic UCL: The UCL of a population mean is unstable, unrealistic, or unreliable if it is orders of magnitude higher than the other UCLs of population mean. It represents an impractically large value that cannot be achieved in practice. For example, the use of Land's H statistic often results in impractically large inflated UCL value. Some other UCLs, such as the bootstrap t UCL and Hall's UCL, can be inflated by outliers resulting in an impractically large and unstable value. All such impractically large UCL values are called unstable, unrealistic, unreliable, or inflated UCLs.

Upper Confidence Limit (UCL): The upper boundary (or limit) of a confidence interval of a parameter of interest such as the population mean.

Upper Prediction Limit (UPL): The upper boundary of a prediction interval for an independently obtained observation (or an independent future observation).

Upper Tolerance Limit (UTL): The upper boundary of a tolerance interval.

Winsorization method: The Winsorization method is a procedure that replaces the n extreme values with the preset cut-off value. This method is sensitive to the number of outliers, but not to their actual values.

Table of Contents

Introduction

The Need for ProUCL Software

Statistical inferences about the sampled populations and their parameters are made based upon defensible and representative data sets of appropriate sizes collected from the populations under investigation. Statistical inference, including both estimation and hypotheses testing approaches, is routinely used to:

1. Estimate environmental parameters of interest such as exposure point concentration (EPC) terms, not-to-exceed values, and background level threshold values (BTVs) for contaminants of potential concern (COPC),

2. Identify areas of concern (AOC) at a contaminated site,

3. Compare contaminant concentrations found at two or more AOCs of a contaminated site,

4. Compare contaminant concentrations found at an AOC with background or reference area contaminant concentrations,

5. Compare site concentrations with a cleanup standard to verify the attainment of cleanup standards.

Statistical inference about the sampled populations and their parameters are made based upon defensible and representative data sets of appropriate sizes collected from the populations under investigation. Environmental data sets originated from the Superfund and RCRA sites often consist of observations below one or more detection limits (DLs). In order to address the statistical issues arising in exposure and risk assessment applications; background versus site comparison and evaluation studies; and various other environmental applications, several graphical, parametric, and nonparametric statistical methods for data sets with nondetects and without nondetects have been incorporated into ProUCL Version 4.0 (ProUCL 4.0).

Exposure and risk management and cleanup decisions in support of United States Environmental Protection Agency (EPA) projects are often made based upon the mean concentrations of the COPCs. A 95% upper confidence limit (UCL95) of the unknown population (e.g., an AOC) arithmetic mean (AM), μ_1, can be used to:

- Estimate the EPC term of the AOC under investigation,
- Determine the attainment of cleanup standards,
- Compare site mean concentrations with reference area mean concentrations, and
- Estimate background level mean contaminant concentrations. The background mean contaminant concentration level may be used to compare the mean of an AOC. It should be noted that it is not appropriate to compare individual point-by-point site observations with the background mean concentration level.

It is important to compute a reliable and stable *UCL95* of the population mean using the available data. The *UCL95* should approximately provide the 95% coverage for the unknown population mean, μ_1. Based upon the available background data, it is equally important to compute reliable and stable upper percentiles, upper prediction limits (*UPLs*), or upper tolerance limits (*UTLs*). These upper limits based upon background (or reference) data are used as estimates of BTVs, compliance limits (*CL*), or not-to-

exceed values. These upper limits are often used in site (point-by-point) versus background comparison evaluations.

Environmental scientists often encounter trace level concentrations of COPCs when evaluating sample analytical results. Those low level analytical results cannot be measured accurately, and therefore are typically reported as less than one or more detection limit (*DL*) values (also called nondetects). However, practitioners often need to obtain reliable estimates of the population mean, μ_1, the population standard deviation, σ_1, and upper limits, including the upper confidence limit (*UCL)* of the population mass or mean, the *UPL*, and the *UTL* based upon data sets with nondetect (*ND*) observations. Hypotheses testing approaches are often used to verify the attainment of cleanup standards, and compare site and background concentrations of COPCs.

Background evaluation studies, BTVs, and not-to-exceed values should be estimated based upon defensible background data sets. The estimated BTVs or not-to-exceed values are then used to identify the COPCs, to identify the site AOCs or hot spots, and to compare the contaminant concentrations at a site with background concentrations. The use of appropriate statistical methods and limits for site versus background comparisons is based upon the following factors:

1. Objective of the study,
2. Environmental medium (e.g., soil, groundwater, sediment, air) of concern,
3. Quantity and quality of the available data,
4. Estimation of a not-to-exceed value or of a mean contaminant concentration,
5. Pre-established or unknown cleanup standards and BTVs, and
6. Sampling distributions (parametric or nonparametric) of the concentration data sets collected from the site and background areas under investigation.

In background versus site comparison evaluations, the environmental population parameters of interest may include:

- Preliminary remediation goals (PRGs),
- Soil screening levels (SSLs),
- Risk-based cleanup (RBC) standards,
- BTVs, not-to-exceed values, and
- Compliance limit, maximum concentration limit (MCL), or alternative concentration limit (ACL), frequently used in groundwater applications.

When the environmental parameters listed above are not known or have not been pre-established, appropriate upper statistical limits are used to estimate the parameters. The UPL, UTL, and upper percentiles are used to estimate the BTVs and not-to-exceed values. Depending upon the site data availability, point-by-point site observations are compared with the estimated (or pre-established) BTVs and not-to-exceed values. If enough site and background data are available, two-sample hypotheses testing approaches are used to compare site concentrations with background concentrations levels. These statistical methods can also be used to compare contaminant concentrations of two site AOCs, surface and subsurface contaminant concentrations, or upgradient versus monitoring well contaminant concentrations.

ProUCL 4.0 Capabilities

ProUCL 4.0 is an upgrade of ProUCL Version 3.0 (EPA, 2004). ProUCL 4.0 contains statistical methods to address various environmental issues for both full data sets without nondetects and for data sets with NDs (also known as left-censored data sets).

ProUCL 4.0 contains:

1. Rigorous parametric and nonparametric (including bootstrap methods) statistical methods (instead of simple ad hoc or substitution methods) that can be used on full data sets without nondetects and on data sets with below detection limit (BDL) or nondetect (ND) observations.

2. State-of-the-art parametric and nonparametric UCL, UPL, and UTL computation methods. These methods can be used on full-uncensored data sets without nondetects and also on data sets with BDL observations. Some of the methods (e.g., Kaplan-Meier method, ROS methods) are applicable on left-censored data sets having multiple detection limits. The UCL and other upper limit computation methods cover a wide range of skewed data sets with and without the BDLs.

3. Single sample (e.g., Student's t-test, sign test, proportion test, Wilcoxon Singed Rank test) and two-sample (Student's t-test, Wilcoxon-Mann-Whitney test, Gehan test, quantile test) parametric and nonparametric hypotheses testing approaches for data sets with and without ND observations. These hypothesis testing approaches can be used to: verify the attainment of cleanup standards, perform site versus background comparisons, and compare two or more AOCs, monitoring wells (MWs).

4. The single sample hypotheses testing approaches are used to compare site mean, site median, site proportion, or a site percentile (e.g., 95^{th}) to a compliance limit (action level, regularity limit). The hypotheses testing approaches can handle both full-uncensored data sets without nondetects, and left-censored data sets with nondetects. Simple two-sample hypotheses testing methods to compare two populations are available in ProUCL 4.0, such as two-sample t-tests, Wilcoxon-Mann-Whitney (WMW) Rank Sum test, quantile test, Gehan's test, and dispersion test. Variations of hypothesis testing methods (e.g., Levene's method to compare dispersions, generalized WRS test) are easily available in most commercial and freely available software packages (e.g., MINITAB, R).

5. ProUCL 4.0 includes graphical methods (e.g., box plots, multiple Q-Q plots, histogram) to compare two or more populations. Additionally, ProUCL 4.0 can also be used to display a box plot of one population (e.g., site data) with compliance limits or upper limits (e.g., UPL) of other population (background area) superimposed on the same graph. This kind of graph provides a useful visual comparison of site data with a compliance limit or BTVs. Graphical displays of a data set (e.g., Q-Q plot) should be used to gain insight knowledge contained in a data set that may not otherwise be clear by looking at simple test statistics such as t-test, Dixon test statistic, or Shapiro-Wilk (S-W) test statistic.

6. ProUCL 4.0 can process multiple contaminants (variables) simultaneously and has the capability of processing data by groups. A valid group column should be included in the data file.

7. ProUCL 4.0 provides a GOF test for data sets with nondetects. The user can create additional columns to store extrapolated (estimated) values for nondetects based upon normal ROS, gamma ROS, and lognormal ROS (robust ROS) methods.

ProUCL Applications

The methods incorporated in ProUCL 4.0 can be used on data sets with and without BDL and ND observations. Methods and recommendations as incorporated in ProUCL 4.0 are based upon the results and findings of the extensive simulation studies as summarized in Singh and Singh (2003), and Singh, Maichle, and Lee (EPA, 2006). It is anticipated that ProUCL 4.0 will serve as a companion software package for the following EPA documents:

- *Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites* (EPA, 2002a), and
- *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b).

Methods included in ProUCL 4.0 can be used in various other environmental applications including the verification of cleanup standards (EPA, 1989), and computation of upper limits needed in groundwater monitoring applications (EPA, 1992 and EPA, 2004).

In 2002, EPA issued guidance for calculating the UCLs of the unknown population means for contaminant concentrations at hazardous waste sites. The ProUCL 3.0 software package (EPA, 2004) has served as a companion software package for the EPA (2002a) guidance document for calculating UCLs of mean contaminant concentrations at hazardous waste sites. ProUCL 3.0 has several parametric and nonparametric statistical methods that can be used to compute appropriate UCLs based upon full-uncensored data sets without any ND observations. ProUCL 4.0 retains the capabilities of ProUCL 3.0, including goodness-of-fit (GOF) and the UCL computation methods for data sets without any BDL observations. However, ProUCL 4.0 has the additional capability to perform GOF tests and computing UCLs and other upper limits based upon data sets with BDL observations.

ProUCL 4.0 defines log-transform *(log)* as the natural logarithm *(ln)* to the base e. ProUCL 4.0 also computes the maximum likelihood estimates (MLEs) and the minimum variance unbiased estimates (MVUEs) of unknown population parameters of normal, lognormal, and gamma distributions. This, of course, depends upon the underlying data distribution. ProUCL 4.0 computes the $(1 - \alpha)100\%$ UCLs of the unknown population mean, μ_1, using 5 parametric and 10 nonparametric methods. It should be pointed out that ProUCL 4.0 computes the ***simple*** summary statistics for detected raw and log-transformed data for full data sets without NDs, as well as for data sets with BDL observations. It is noted that estimates of mean and *sd* for data sets with NDs based upon rigorous statistical methods (e.g., MLE, ROS, K-M methods) are note provided in the summary statistics. Those estimates and the associated upper limits for data sets with NDs are provided under the menu options: Background and UCL.

It is emphasized that throughout this Technical Guide, and in the ProUCL 4.0 software, it is assumed that one is dealing with a single population. If multiple populations (e.g., background and site data mixed together) are present, it is recommended to first separate them out (e.g., using appropriate statistical

population partitioning techniques), and then compute appropriate respective 95% UCLs separately for each of the identified populations. Outliers, if any, should be identified and thoroughly investigated. ProUCL 4.0 provides two commonly used simple classical outlier identification procedures: 1) Dixon test, and 2) Rosner test. Outliers distort most parametric statistics (e.g., mean, UCLs, upper prediction limits (UPLs), test statistics) of interest. Moreover, it should be noted that even though outliers might have minimal influence on hypotheses testing statistics based upon ranks (e.g., WMW test), outliers do distort those nonparametric statistics (including bootstrap methods), which are based upon higher order statistics such as UPLs and UTLs. Decisions about the disposition (exclusion or inclusion) of outliers in a data set used to estimate the EPC terms or BTVs should be made by all parties involved (e.g., project team, EPA, local agency, potentially responsible party, etc.) in the decision making process.

The presence of outlying observations also distorts statistics based upon bootstrap re-samples. The use of higher order values (quantiles) of the distorted statistics for the computation of the UCLs or UPLs based upon bootstrap t and Hall's bootstrap methods may yield unstable and erratic UCL values. This is especially true for the upper limits providing higher confidence coefficients such as 95%, 97.5%, or 99%. Similar behavior of the bootstrap t UCL is observed for data sets having BDL observations. Therefore, the bootstrap t and Hall's bootstrap methods should be used with caution. It is suggested that the user should examine various other UCL results and determine if the UCLs based upon the bootstrap t and Hall's bootstrap methods represent reasonable and reliable UCL values of practical merit. If the results based upon these two bootstrap methods are much higher than the rest of methods, then this could be an indication of erratic behavior of those bootstrap UCL values, perhaps distorted by outlying observations. In case these two bootstrap methods yield erratic and inflated UCLs, the UCL of the mean should be computed using the adjusted or the approximate gamma UCL computation method for highly skewed gamma distributed data sets of small sizes. Alternatively, one may use a 97.5% or 99% Chebyshev UCL to estimate the mean of a highly skewed population. It should be noted that typically, a Chebyshev UCL may yield conservative and higher values of the UCLs than other methods available in ProUCL 4.0 This is especially true when data are moderately skewed and sample size is large. In such cases, when the sample size is large, one may want to use a 95% Chebyshev UCL or a Chebyshev UCL with lower confidence coefficient such as 92.5% or 90% as estimate of the population mean.

ProUCL Methods

ProUCL 4.0 provides 15 UCL computation methods for full data sets without any BDL observations; 5 are parametric and 10 are nonparametric methods. The nonparametric methods do not depend upon any assumptions about the data distributions. The five parametric UCL computation methods are:

1. Student's t-UCL,
2. Approximate gamma UCL using chi-square approximation,
3. Adjusted gamma UCL (adjusted for level significance),
4. Land's H-UCL, and
5. Chebyshev inequality-based UCL (using MVUEs of parameters of a lognormal distribution).

The 10 nonparametric methods are:

1. The central limit theorem (CLT)-based UCL,
2. Modified t-statistic (adjusted for skewness)-based UCL,
3. Adjusted-CLT (adjusted for Skewness)-based UCL,
4. Chebyshev inequality based-UCL (using sample mean and sample standard deviation),
5. Jackknife method-based UCL,
6. UCL based upon standard bootstrap,
7. UCL based upon percentile bootstrap,
8. UCL based upon bias-corrected accelerated (BCA) bootstrap,
9. UCL based upon bootstrap t, and
10. UCL based upon Hall's bootstrap.

Environmental scientists often encounter trace level concentrations of COPCs when evaluating sample analytical results. Those low level analytical results cannot be measured accurately, and therefore are typically reported as less than one or more DL values. However, the practitioners need to obtain reliable estimates of the population mean, μ_1, and the population standard deviation, σ_1, and upper limits including the UCL of the population mass (measure of central tendency) or mean, UPL, and UTL. Several methods are available and cited in the environmental literature (Helsel (2005), Singh and Nocerino (2002), Millard and Neerchal (2001)) that can be used to estimate the population mean and variance. However, till to date, no specific recommendations are available for the use of appropriate methods that can be used to compute upper limits (e.g., UCLs, UPLs) based upon data sets with BDL observations. Singh, Maichle, and Lee (EPA 2006) extensively studied the performance of several parametric and nonparametric UCL computation methods for data sets with BDL observations. Based upon their results and findings, several methods to compute upper limits (UCLs, UPLs, and UTLs) needed to estimate the EPC terms and BTVs have been incorporated in ProUCL 4.0.

In 2002, EPA issued another *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b). This EPA (2002b) background guidance document is currently being revised to include statistical methods that can be used to estimate the BTVs and not-to-exceed values based upon data sets with and without the BDL observations. In background evaluation studies, BTVs, compliance limits, or not-to-exceed values often need to be estimated based upon defensible background data sets. The estimated BTVs or not-to-exceed values are then used for screening the COPCs, to identify the site AOCs or hot spots, and also to determine if the site concentrations (perhaps after a remediation activity) are comparable to background concentrations, or are approaching the background level concentrations. Individual point-by-point site observations (composite samples preferred) are sometimes compared with those not-to-exceed values or BTVs. It should be pointed out that in practice, it is preferred to use hypotheses testing approaches to compare site versus background concentrations provided enough (e.g., at least 8-10 detected observations from each of the two populations) site and background data are available. Chapter 1 provides practical guidance on the minimum sample size

requirements to estimate and use the BTVs, single and two-sample hypotheses testing approaches to perform background evaluations and background versus site comparisons. Chapter 1 also briefly discusses the differences in the definitions and uses of the various upper limits as incorporated in ProUCL 4.0. Detailed discussion of the various methods to estimate the BTVs and other not-to-exceed values for full-uncensored data sets (Chapter 5) without any nondetect values and for left-censored data sets (Chapter 6) with nondetect values are given in the revised background guidance document.

ProUCL 4.0 includes statistical methods to compute UCLs of the mean, upper limits to estimate the BTVs, other not-to-exceed values, and compliance limits based upon data sets with one or more detection limits. The use of appropriate statistical methods and limits for exposure and risk assessment, and site versus background comparisons, is based upon several factors:

1. Objective of the study;

2. Environmental medium (e.g., soil, groundwater, sediment, air) of concern;

3. Quantity and quality of the available data;

4. Estimation of a not-to-exceed value or of a mean contaminant concentration;

5. Pre-established or unknown cleanup standards and BTVs; and

6. Sampling distributions (parametric or nonparametric) of the concentration data sets collected from the site and background areas under investigation.

In background versus site comparison studies, the population parameters of interest are typically represented by *upper threshold limits* (e.g., upper percentiles, upper confidence limits of an upper percentile, upper prediction limit) of the background data distribution. It should be noted that the upper threshold values are estimated and represented by upper percentiles and other values from the upper tail of the background data distribution. These background upper threshold values do not represent measures of central tendency such as the mean, the median, or their upper confidence limits. These environmental parameters may include:

- Preliminary remediation goals (PRGs), compliance limits,
- Soil screening levels (SSLs),
- Risk-based cleanup (RBC) standards,
- BTVs, compliance limits, or not-to-exceed values, and
- Maximum concentration limit (MCL) or alternative concentration limit (ACL) used in Groundwater applications.

When the environmental parameters listed above are not known or pre-established, appropriate upper statistical limits are used to estimate those parameters. The UPL, UTL, and upper percentiles are typically used to estimate the BTVs, not-to-exceed values, and other parameters listed above. Depending upon the availability of site data, point-by-point site observations are compared with the estimated (or pre-established) BTVs and not-to-exceed values. If enough site and background data are available, two-sample hypotheses testing approaches (preferred method to compare two populations) are used to compare site concentrations with background concentrations levels. The hypotheses testing methods can also be used to compare contaminant concentrations of two site AOCs, surface and subsurface contaminant concentrations, or upgradient versus monitoring well contaminant concentrations.

Background versus Site Comparison Evaluations

The following statistical limits have been incorporated in ProUCL 4.0 to assist in background versus site comparison evaluations:

Parametric Limits for Full-Uncensored Data Sets without Nondetect Observations

- UPL for a single observation (Normal, Lognormal) not belonging to the original data set
- UPL for next k (k is user specified) or k future observations (Normal, Lognormal)
- UTL, an upper confidence limit of a percentile (Normal, Lognormal)
- Upper percentiles (Normal, Lognormal, and Gamma)

Nonparametric Limits for Full-Uncensored Data Sets without Nondetect Observations

Nonparametric limits are typically based upon order statistics of a data set such as a background or a reference data set. Depending upon the size of the data set, higher order statistics (maximum, second largest, third largest, and so on) are used as these upper limits (e.g., UPLs, UTLs). The details of these methods with sample size requirements can be found in Chapter 5 of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b). It should be, noted that the following statistics might get distorted by the presence of outliers (if any) in the data set under study.

- UPL for a single observation not belonging to the original data set
- UTL, an upper confidence limit of a percentile
- Upper percentiles
- Upper limit based upon interquartile range (IQR)
- Upper limits based upon bootstrap methods

For data sets with BDL observations, the following parametric and nonparametric methods to compute the upper limits were studied and evaluated by Singh, Maichle, and Lee (EPA, 2006) via Monte Carlo Simulation Experiments. Depending upon the performances of those methods, only some of the methods have been incorporated in ProUCL 4.0. Methods (e.g., Delta method, DL method, uniform (0, DL) generation method) not included in ProUCL 4.0 do not perform well in comparison with other methods.

Note: *When the percentage of NDs in a data set is high (e.g., > 40%-50%), especially when multiple detection limits might be present, it is hard to reliably perform GOF tests (to determine data distribution) on those data sets with many NDs. The uncertainty associated with those GOF tests will be high, especially when the data sets are of small sizes (< 10-20). It should also be noted that the parametric MLE methods (e.g., for normal and lognormal distributions) often yield unstable estimates of mean and sd. This is especially true when the number of nondetects exceeds 40%-50%. In such situations, it is preferable to use nonparametric (e.g., KM method) methods to compute statistics of interest such as UCLs, UPLs, and UTLs. Nonparametric methods do not require any distributional assumptions about the data sets under investigation. Singh, Maichle, and Lee (EPA, 2006) also concluded that the performance of the KM estimation method is better (in terms of coverage probabilities) than various other parametric estimation (e.g., MLE, EM, ROS) methods.*

Parametric Methods to Compute Upper Limits for Data Sets with Nondetect Observations

- Simple substitution (proxy) methods (0, DL/2, DL)
- MLE method, often known as Cohen's MLE method – single detection limit
- Restricted MLE method – single detection limit – not in ProUCL 4.0
- Expectation Maximization (EM) method – single detection limit – not in ProUCL 4.0
- EPA Delta log method – single detection limit – not in ProUCL 4.0
- Regression method on detected data and using slope and intercept of the OLS regression line as estimates of standard deviation, *sd,* and mean (not a recommended method)
- Robust ROS (regression on order statistics) on log-transformed data – nondetects extrapolated (estimated) using robust ROS; mean, *sd, UCLs,* and other statistics computed using the detected and extrapolated data in original scale – multiple detection limits
- Normal ROS – nondetects extrapolated (estimated) using normal distribution, mean, *sd, UCLs,* and other statistics computed using the detected and extrapolated data – multiple detection limits.
- It is noted that the estimated NDs often become negative and even larger than the detection limits (not a recommended method)
- Gamma ROS – nondetects extrapolated (estimated) using gamma distribution, mean, *sd, UCLs,* and other statistics computed using the detected and extrapolated data – multiple detection limits

Nonparametric Methods to Compute Upper Limits for Data Sets with Nondetect Observations

- Bootstrap Methods
  - Percentile Bootstrap on robust ROS
  - Percentile Bootstrap
  - BCA Bootstrap
  - Bootstrap t
- Jackknife Method
  - Jackknife on robust ROS
- Kaplan-Meier (KM) Method
  - Bootstrap (percentile, BCA) using KM estimates
  - Jackknife using KM estimates
  - Chebyshev Method using KM estimates
- Winsorization Method

For uncensored full data sets without any NDs, the performance (in terms of coverage for the mean) of the various UCL computation methods was evaluated by Singh and Singh (2003). The performance of the parametric and nonparametric UCL methods based upon data sets with nondetect observations was studied by Singh, Maichle, and Lee (EPA 2006). Several of the methods listed above have been incorporated in ProUCL 4.0 to compute the estimates of EPC terms (95% UCL), and of BTVs (UPLs, UTLs, upper percentiles). Methods that did not perform well (e.g., poor coverage or unrealistically large values, infeasible and biased estimates) are not included in ProUCL 4.0. Methods not incorporated in

ProUCL 4.0 are: EPA Delta Log method, Restricted MLE method, and EM method, substitution method (0, and DL), and Regression method.

Note: *It should be noted that for data sets with NDs, the DL/2 substitution method has been incorporated in ProUCL 4.0 only for historical reasons and also for its current default use. It is well known that the DL/2 method (with NDs replaced by DL/2) does not perform well (e.g., Singh, Maichle, and Lee (EPA 2006)) even when the percentage of NDs is only 5%-10%. It is strongly suggested to avoid the use of DL/2 method for estimation and hypothesis testing approaches used in various environmental applications. Also, when the % of NDs becomes high (e.g., > 40%-50%), it is suggested to avoid the use of parametric MLE methods. For data sets with high percentage of NDs (e.g., > 40%), the distributional assumptions needed to use parametric methods are hard to verify; and those parametric MLE methods may yield unstable results.*

It should also be noted that even though the lognormal distribution and some statistics based upon lognormal assumption (e.g., Robust ROS, DL/2 method) are available in ProUCL 4.0, ProUCL 4.0 does not compute MLEs of mean and sd based upon a lognormal distribution. The main reason is that the estimates need to be computed in the original scale via back-transformation (Shaarawi, 1989, Singh, Maichle, and Lee (EPA 2006)). Those back-transformed estimates often suffer from an unknown amount of significant bias. Hence, it is also suggested to avoid the use of a lognormal distribution to compute MLEs of mean and sd, and associated upper limits, especially UCLs based upon those MLEs obtained using a lognormal distribution.

ProUCL 4.0 recommends the use of an appropriate UCL to estimate the EPC terms. It is desirable that the user consults with the project team and experts familiar with the site before using those recommendations. Furthermore, there does not seem to be a general agreement about the use of an upper limit (e.g., UPL, percentile, or UTL) to estimate not-to-exceed values or BTVs to be used for screening of the COPCs and in site versus background comparison studies. ProUCL 4.0 can compute both parametric and nonparametric upper percentiles, UPLs, and UTLs for uncensored and censored data sets. However, no specific recommendations have been made regarding the use of UPLs, UTLs, or upper percentiles to estimate the BTVs, compliance limits, and other related background or reference parameters. However, the developers of ProUCL 4.0 prefer the use of UPLs or upper percentiles to estimate the background population parameters (e.g., BTVs, not-to-exceed values) that may be needed to perform point-by-point site versus background comparisons.

The standard bootstrap and the percentile bootstrap UCL computation methods do not perform well (do not provide adequate coverage to population mean) for skewed data sets. For skewed distributions, the bootstrap t and Hall's bootstrap (meant to adjust for skewness) methods do perform better (in terms of coverage for the population mean) than the other bootstrap methods. However, it has been noted (e.g., Efron and Tibshirani (1993) and Singh, Singh, and Iaci (2002b)) that these two bootstrap methods sometimes yield erratic and inflated *UCL* values (orders of magnitude higher than the other *UCLs*). This may occur when outliers are present in a data set. Similar behavior of the bootstrap t UCL is observed based upon data sets with NDs. Therefore, whenever applicable, ProUCL 4.0 provides cautionary statements regarding the use of bootstrap methods.

ProUCL 4.0 provides several state-of-the-art parametric and nonparametric UCL, UPL, and UTL computation methods that can be used on uncensored data sets (full data sets) and on data sets with BDL observations. Some of the methods (e.g., Kaplan-Meier method, ROS methods) incorporated in ProUCL 4.0 are applicable on left-censored data sets having multiple detection limits. The UCLs and other upper

limits computation methods in ProUCL 4.0 cover a wide range of skewed data distributions with and without the BDLs arising from the environmental applications.

ProUCL 4.0 also has parametric and nonparametric single and two-sample hypotheses testing approaches required to: compare site location (e.g., mean, median) to a specified cleanup standard; perform site versus background comparisons; or compare of two or more AOCs. These hypotheses testing methods can handle both full (uncensored data sets without NDs) and left-censored (with nondetects) data sets. Specifically, two-sample tests such as t-test, Wilcoxon-Mann-Whitney (WMW) Rank Sum test, quantile test, and Gehan's test are available in ProUCL 4.0 to compare concentrations of two populations.

Single sample parametric (Student's t-test) and nonparametric (sign test, Wilcoxon Signed Rank (WSR) test, tests for proportions and percentiles) hypotheses testing approaches are also available in ProUCL 4.0. The single sample hypotheses tests are useful when the environmental parameters such as the clean standard, action level, or compliance limits are known, and the objective is to compare site concentrations with those known threshold values. Specifically, a t-test (or a sign test) may be used to verify the attainment of cleanup levels at an AOC after a remediation activity; and a test for proportion may be used to verify if the proportion of exceedances of an action level (or a compliance limit) by sample concentrations collected from the AOC (or a MW) exceeds a certain specified proportion (e.g., 1%, 5%, 10%). As mentioned before, ProUCL 4.0 can perform these hypotheses on data sets with and without nondetect observations.

Note: *It should be noted that as cited in the literature, some of the hypotheses testing approaches (e.g., nonparametric two-sample WMW) deal with the single detection limit scenario. If multiple detection limits are present, all NDs below the largest detection limit need to be considered as NDs (Gilbert, 1987, Helsel, 2005). This in turn may reduce the power and increase uncertainty associated with test. As mentioned before, it is always desirable to supplement the test statistics and test conclusions with graphical displays such as the multiple Q-Q plots and side-by-side box plots. ProUCL 4.0 can graph box plots and Q-Q plots for data sets with nondetect observations. Gehan test as available in ProUCL 4.0 should be used in case multiple detection limits are present. ProUCL 4.0 can draw Q-Q plots and box plots for data sets with and without nondetect observations.*

It should be pointed out that when using two-sample hypotheses approaches (WMW test, Gehan test, and quantile test) on data sets with NDs, both samples and variables (e.g., site-As, Back-As) should be specified as having nondetects. This means, a ND column (0= ND, and 1 = detect) should be provided for each variable (here D_site-As, and D_Back-As) to be used in this comparison. If a variable (e.g., site-As) does not have any nondetects, still a column with label D_site-As should be included in the data set with all entries = 1 (detected values).

Moreover, in single sample hypotheses tests (e.g., sign test, proportion test) used to compare site mean/median concentration level with a cleanup standard, $C_{s,}$ or compliance limit (e.g., proportion test), all NDs (if any) should lie below the cleanup standard, $C_{s.}$

The differences between these tests should be noted and understood. Specifically, a t-test or a Wilcoxon Signed Rank (WSR) test are used to compare the measures of location and central tendencies (e.g., mean, median) of a site area (e.g., AOC) to a cleanup standard, C_s or action level also representing a measure of central tendency (e.g., mean, median); whereas, a proportion test compares if the proportion of site observations from an AOC exceeding a compliance limit (*CL*) exceeds a specified proportion, P_0 (e.g., 5%, 10%). The percentile test compares a specified percentile (e.g., 95^{th}) of the site data to a pre-specified upper threshold (e.g., reporting limit, action level). All of these tests have been incorporated in ProUCL

4.0. Most of the single sample and two-sample hypotheses tests also report associated p-values. For some of the hypotheses tests (e.g., WMW test, WSR test, proportion test), large sample approximate p-values are computed using continuity correction factors.

Graphical Capabilities

ProUCL 4.0 has useful exploratory graphical methods that may be used to visually compare the concentrations of:

1. A site area of concern (AOC) with an action level. This can be done using a box plot of site data with action level superimposed on that graph,

2. Two or more populations, including site versus background populations, surface versus subsurface concentrations, and

3. Two or more AOCs.

The graphical methods include double and multiple quantile-quantile (Q-Q) plots, side-by-side box plots, and histograms. Whenever possible, it is desirable to supplement statistical test results and statistics with visual graphical displays of data sets. There is no substitute for graphical displays of a data set as the visual displays often provide useful information about a data set, which cannot be revealed by simple test statistics such as t-test, SW test, Rosner test, WMW test. For example, in addition to providing information about the data distribution, a normal Q-Q plot can also help identify outliers and multiple populations that might be present in a data set. This kind of information cannot be revealed by simple test statistics such as a Shapiro-Wilk (SW) test or Rosner's outlier test statistic. Specifically, the SW test may lead to the conclusion that a mixture data set (representing two or more populations) can be modeled by a normal (or lognormal) distribution, whereas the occurrence of obvious breaks and jumps in the associated Q-Q plot may suggest the presence of multiple populations in the mixture data set. It is suggested that the user should use exploratory tools to gain necessary insight into a data set and the underlying assumptions (e.g., distributional, single population) that may not be revealed by simple test statistics.

Note: *On a Q-Q plot, observations well separated from the majority of the data may represent potential outliers, and obvious jumps and breaks of significant magnitude may suggest the presence of observations from multiple populations in the data set.*

The analyses of data categorized by a group ID variable such as: 1) Surface vs. Subsurface; 2) AOC1 vs. AOC2; 3) Site vs. Background; and 4) Upgradient vs. Downgradient monitoring wells are quite common in many environmental applications. ProUCL 4.0 offers this option for data sets with and without nondetects. The Group Option provides a powerful tool to perform various statistical tests and methods (including graphical displays) separately for each of the group (samples from different populations) that may be present in a data set. For an example, the same data set may consist of samples from the various groups or populations representing site, background, two or more AOCs, surface, subsurface, monitoring wells. The graphical displays (e.g., box plots, Q-Q plots) and statistics (computations of background statistics, UCLs, hypotheses testing approaches) of interest can be computed separately for each group by using this option.

Chapter 1

Guidance on the Use of Statistical Methods as Incorporated in ProUCL 4.0 & Associated Minimum Sample Size Requirements

This chapter describes the differences between the various statistical limits (e.g., UCLs, UPLs, UTLs) often used to estimate the environmental parameters of interest including exposure point concentration (EPC) terms and background threshold values (BTVs). Some suggestions about the minimum sample size requirements needed to use statistical inferential methods to estimate the environmental parameters: EPC terms, BTVs, and not-to-exceed values, and to compare site data with background data or with some pre-established reference limits (e.g., preliminary remediation goals (PRGs), action levels, compliance limits) have also been provided. It is noted that, several EPA guidance documents (e.g., EPA, 1997, 2002a, 2006) discuss in details about data quality objectives (DQOs) and sample size determinations based upon those DQOs needed for the various statistical methods used in environmental applications. Also, appropriate sample collection methods (e.g., instruments, sample weights, discrete or composite, analytical methods) depend upon the medium (e.g., soil, sediment, water) under consideration. For an example, Gerlach and Nocerino (EPA, 2003) describe optimal soil sample (based upon Gy theory) collection methods. Therefore, the topics of sample size determination based upon DQOs, data validation, and appropriate sample collection methods for the various environmental media are not considered in ProUCL 4.0 and the associated technical and technical guides. It is assumed that data sets to be used in ProUCL are of good quality, and whenever possible have been obtained using the guidance provided in various EPA (2003, 2006) documents. It is users' responsibility to assure that adequate amount of good quality data have been collected.

Note: *Here, emphasis is given on the practical applicability and appropriate use of statistical methods needed to address statistical issues arising in risk management, background versus site evaluation studies, and various other environmental applications. Specifically, guidance on minimum sample size requirements as provided in this chapter is useful when data have already been collected, or it is not possible (e.g., due to resource limitations) to collect the number of samples obtained using DQO processes as described in EPA (2006).*

Decisions based upon statistics obtained using data sets of small sizes (e.g., 4 to 6 detected observations) cannot be considered reliable enough to make a remediation decision that affects human health and the environment. For an example, a background data set of size 4 to 6 is not large enough to characterize background population, to compute BTV values, or to perform background versus site comparisons. In order to perform reliable and meaningful statistical inference (estimation and hypothesis testing), one should determine the sample sizes that need to be collected from the populations under investigation using appropriate DQO processes and decision error rates (EPA, 2006). However, in some cases, it may not be possible (e.g., resource constraints) to collect the same number of samples recommended by the DQO process. In order to address such cases, some minimum sample size requirements for background and site data sets are described in this chapter.

The use of an appropriate statistical method depends upon the environmental parameter(s) being estimated or compared. The measures of central tendency (e.g., means, medians, or their upper confidence limits (UCLs)) are often used to compare site mean concentrations (e.g., after remediation activity) with a cleanup standard, C_s, representing some central tendency measure of a reference area or some other known threshold representing a measure of central tendency. The upper threshold values, such as the compliance limits (e.g., alternative concentration limit (ACL), maximum concentration limit (MCL)), or

not-to-exceed values, are used when individual point-by-point observations are compared with those not-to-exceed values or some other compliance limit. It should be noted that depending upon whether the environmental parameters (e.g., BTVs, not-to-exceed value, EPC term, or cleanup standards) are known or unknown, different statistical methods with different data requirements are needed to compare site concentrations with pre-established (known) or estimated (unknown) standards and BTVs.

ProUCL 4.0 has been developed to address issues arising in exposure assessment, risk- assessment, and background versus site comparison applications. Several upper limits, and single- and two-sample hypotheses testing approaches, for both full-uncensored and left-censored data sets, are available in ProUCL 4.0. The details of the statistical and graphical methods included in ProUCL 4.0 can be found in the ProUCL Technical guidance. In order to make sure that the methods in ProUCL 4.0 are properly used, this chapter provides guidance on:

1. analysis of site and background areas and data sets,
2. collection of discrete or composite samples,
3. appropriate use of the various upper limits,
4. guidance regarding minimum sample sizes,
5. point-by-point comparison of site observations with BTVs,
6. use of hypotheses testing approaches,
7. using small data sample sets,
8. use of maximum detected value, and
9. discussion of ProUCL usage for special cases.

1.1 Background Data Sets

The project team familiar with the site should identify and chose a background area. Depending upon the site activities and the pollutants, the background area can be site-specific or a general reference area. An appropriate random sample of independent observations should be collected from the background area. A defensible background data set should represent a “single” background population (e.g., representing pristine site conditions before any of the industrial site activities) free of contaminating observations such as outliers. In a background data set, outliers may represent potentially contaminated observations from impacted site areas under study or possibly from other polluted site(s). This scenario is common when background samples are obtained from the various onsite areas (e.g., large federal facilities). Outlying observations should not be included in the estimation (or hypotheses testing procedures) of the BTVs. The presence of outliers in the background data set will yield distorted estimates of the BTVs and hypothesis testing statistics. The proper disposition of outliers to include or not include them in the data set should be decided by the project team.

Decisions based upon distorted statistics can be incorrect, misleading, and expensive. It should be noted that the objective is to compute background statistics based upon the majority of the data set representing the dominant background population, and not to accommodate a few low probability outliers that may

also be present in the background data set. A couple of simple classical outlier tests (Dixon and Rosner tests) are available in ProUCL 4.0. Since these classical tests suffer from masking effects (e.g., some extreme outliers may mask the occurrence of other intermediate outliers), it is suggested that these classical outlier tests should always be supplemented with graphical displays such as a box plot or a Q-Q plot. The use of robust and resistant outlier identification procedures (Singh and Nocerino, 1995, Rousseeuw and Leroy, 1987) is recommended when multiple outliers may be present in a data set. Those methods are beyond the scope of ProUCL 4.0. However, several robust outlier identification and estimation are available in Scout (EPA, 2000), which is currently under revision and upgrade.

An appropriate background data set of a reasonable size (preferably computed using DQO processes) is needed to characterize a background area including computation of upper limits (e.g., estimates of BTVs, not-to-exceed values) based upon background data sets and also to compare site and background data sets using hypotheses testing approaches. As mentioned before, a small background data set of size 4 to 6 is not large enough to compute BTVs or to perform background versus site comparisons. At the minimum, a background sample should have at least 8 to 10 (more observations are preferable) detected observations to estimate BTVs or to use hypotheses testing approaches.

1.2 Site Data Sets

A defensible data set from a site population (e.g., AOC, EA, RU, group of monitoring wells) should be representative of the site area under investigation. Depending upon the site areas under investigation, different soil depths and soil types may be considered as representing different statistical populations. In such cases, background versus site comparisons may have to be conducted separately for each of those site sub-populations (e.g., surface and sub-surface layers of an AOC, clay and sandy site areas). These issues, such as comparing depths and soil types, should also be considered in a planning and sampling design before starting to collect samples from the various site areas under investigation. Specifically, the availability of an adequate amount of representative site data is required from each of those site sub-populations defined by sample depths, soil types, and the various other characteristics. For detailed guidance on soil sample collections, the reader is referred to Gerlach and Nocerino (EPA (2003)).

The site data collection requirements depend upon the objective of the study. Specifically, in background versus site comparisons, site data are needed to perform:

- Individual point-by-point site observation comparisons with pre-established or estimated BTVs, PRGs, cleanup standards, and not-to-exceed-values. Typically, this approach is used when only a small number (e.g., < 4 to 6) of detected site observations (preferably based upon composite samples) are available which need to be compared with BTVs and not-to-exceed values. Some applications of the point-by-point site observation comparisons are described later in this chapter.

- Single sample hypotheses tests to compare site data with pre-established cleanup standards, C_s (e.g., representing a measure of central tendency); or with BTVs and not-to-exceed values (used for tests for proportions and percentiles). The hypotheses testing approaches are used when enough site data are available. Specifically, when at least 8 to 10 detected (more are desirable) site observations are available, it is preferable to use hypotheses testing approaches to compare site observations with specified threshold values. The use of hypotheses testing approaches can control the two types (Type 1 and Type 2) of error rates more efficiently than the point-by-point individual observation comparisons. This is especially true as the number of point-by-point comparisons

increases. This issue is illustrated by the following table summarizing the probabilities of exceedances (false positive error rate) of the background threshold value (e.g., 95th percentile) by site observations, even when the site and background populations have comparable distributions. The probabilities of these chance exceedances increase as the sample size increases.

Sample Size	Probability of Exceedance
1	0.05
2	0.10
5	0.23
8	0.34
10	0.40
12	0.46
64	0.96

- Two-sample hypotheses testing to compare site data distribution with background data distribution to determine if the site concentrations are comparable to background concentrations. Adequate amount of data need to be made available from the site as well as the background populations. It is preferable to collect at least 8 to 10 detected observations from each of the population under comparison.

1.3 Discrete Samples or Composite Samples?

In a data set (background or site), collected samples should be either all discrete or all composite. In general, both discrete and composite site samples may be used for individual point-by-point site comparisons with a threshold value, and for single and two-sample hypotheses testing applications.

- If possible, the use of composite site samples is preferred when comparing individual point-by-point site observations from an area (e.g., area of concern (AOC), remediation unit (RU), exposure area (EA)) with some pre-established or estimated BTV, compliance limit (CL), or some other not-to-exceed value. This comparison approach is useful when few (< 4 to 6) detected site observations are compared with a pre-established or estimated BTV or some other not-to-exceed threshold.

- When using a single sample hypothesis testing approach, site data can be obtained by collecting all discrete or all composite samples. The hypothesis testing approach is used when many (e.g., exceeding 8 to 10) detected site observations are available. Details of the single sample hypothesis approaches are widely available in EPA documents (1989, 1997, and 2006). Some of those single sample hypotheses testing procedures are also available in ProUCL 4.0.

- If a two-sample hypotheses testing approach is used to perform site versus background comparisons, then samples from both of the populations should be either all discrete samples, or all composite samples. The two-sample hypothesis testing approach is used when many (e.g., exceeding 8 to 10) site, as well as background, observations are available. For better and more accurate results with higher statistical power, the availability of more observations (e.g., exceeding 10-15) from each of the two populations is desirable, perhaps based upon an appropriate DQO process, as described in an EPA guidance document (2006).

1.4 Upper Limits and Their Use

The appropriate computation and use of statistical limits depend upon their applications and the parameters (e.g., EPC term, not-to-exceed value) they are supposed to be estimating. Depending upon the objective of the study, a pre-specified cleanup standard, $C_{s,}$ or a risk-based cleanup (RBC) can be viewed as to represent: 1) as average contaminant concentration; or 2) a not-to-exceed upper threshold value. These two threshold values, an average value, μ_0, and a not-to-exceed value, A_0, represent two significantly different parameters, and different statistical methods and limits are used to compare the site data with these two different parameters or threshold values. Statistical limits, such as an upper confidence limit (UCL) of the population mean, an upper prediction limit (UPL) for an independently obtained "single" observation, or independently obtained k observations (also called future k observations, next k observations, or k different observations), upper percentiles, and upper tolerance limits (UTLs), are often used to estimate the environmental parameters, including the EPC terms, compliance limits (e.g., ACL, MCL), BTVs, and other not-to-exceed values. Here, UTL95%-95% represents a 95% confidence limit of the 95th percentile of the distribution of the contaminant under study.

It is important to understand and note the differences between the uses and numerical values of these statistical limits so that they can be properly used. Specifically, the differences between UCLs and UPLs (or upper percentiles), and UCLs and UTLs should be clearly understood and acknowledged. A UCL with a 95% confidence limit (UCL95) of the mean represents an estimate of the population mean (measure of the central tendency of a data distribution), whereas a UPL95, a UTL95%-95%, and an upper 95th percentile represent estimates of a threshold value in the upper tail of the data distribution. Therefore, a UCL95 should represent a smaller number than an upper percentile or an upper prediction limit. Also, since a UTL 95%-95% represents a 95% UCL of the upper 95th percentile, a UTL should be ≥ the corresponding UPL95 and the 95th upper percentile. Typically, it is expected that the numerical values of these limits should follow the order given as follows:

Sample Mean ≤ UCL95 of Mean ≤ Upper 95th Percentile ≤ UPL95 of a Single Observation ≤ UTL95%-95%

It should also be pointed out that as the sample size increases, a UCL95 of the mean approaches (converges to) the population mean, and a UPL95 approaches the 95th percentile. The differences among the various upper limits are further illustrated in Example 1-1 below. It should be noted that, in some cases, these limits might not follow the natural order described above. This is especially true when the upper limits are computed based upon a lognormal distribution (Singh, Singh, and Engelhardt, 1997). It is well known that a lognormal distribution based H-UCL95 (Land's UCL95) often yields unstable and impractically large UCL values. An H-UCL95 often becomes larger than UPL95 and even larger than a UTL 95%-95%. This is especially true when dealing with skewed data sets of smaller sizes. Moreover, it should also be noted that in some cases, a H-UCL95 becomes smaller than the sample mean, especially when the data are mildly skewed to moderately skewed, and the sample size is large (e.g., > 50, 100). Some of these issues, related to a lognormal distribution and H-UCL95 based upon Land's (1975) statistic are discussed in Chapter 3 of the revised background document for CERCLA sites.

1.4.1 Example 1-1

Consider a simple site-specific background data set associated with a Superfund site. The data set (given in Appendix 5 of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b)) has several inorganic contaminants of potential concern, including aluminum, arsenic, chromium, iron, and lead. It is noted that iron concentrations follow a normal

distribution. Some upper limits for the iron data set are summarized as follows. It is noted that the various upper limits do follow the order as described above.

Table 1-1. Computation of Upper Limits for Iron (Normally Distributed)

Mean	Median	Min	Max	UCL95	UPL95 for a Single Observation	UPL95 for 4 Observations	UTL95/95	95% Upper Percentile
9618	9615	3060	18700	11478	18145	21618	21149	17534

A 95% UCL (UCL95) of the mean is the most commonly used limit in environmental applications. For an example, a 95% UCL of mean is used as an estimate of the EPC. A UCL95 should not be used to estimate a background threshold value (a value in the upper tail of the background data distribution) to be compared with individual site observations. There are many instances in background evaluations and background versus site comparison studies, when it is not appropriate to use a 95% UCL. Specifically, when point-by-point site observations are to be compared with a BTV, then that BTV should be estimated (or represented) by a limit from the upper tail of the reference set (background) data distribution.

A brief discussion about the differences between the applications and uses of the various statistical limits is provided below. This will assist a typical user in determining which upper limit (e.g., UCL95 or UPL95) to use to estimate the parameter of interest (e.g., EPC or BTV).

- A UCL represents an average value that should be compared with a threshold value also representing an average value (pre-established or estimated), such as a mean cleanup standard, $C_{s.}$ For an example, a site 95% UCL exceeding a cleanup value, $C_{s,}$ may lead to the conclusion that the cleanup level, C_s, has not been attained by the site area. It should be noted that UCLs of means are typically computed based upon the site data set.

- When site averages (and not individual site observations) are compared with a threshold value (pre-determined or estimated), such as a PRG or a RBC, or with some other cleanup standard, C_s, then that threshold should represent an average value, and not a not-to-exceed threshold value for individual observation comparisons.

- A UCL represents a “collective” measure of central tendency, and it is not appropriate to compare individual site observations with a UCL. Depending upon data availability, single or two-sample hypotheses testing approaches are used to compare site averages: with a specified or pre-established cleanup standard (single sample hypothesis), or with the background population averages (two-sample hypothesis).

- A UPL, an upper percentile, or an UTL represents an upper limit to be used for point-by-point individual site observation comparisons. UPLs and UTLs are computed based upon background data sets, and individual site observations are compared with those limits. A site observation for a contaminant exceeding a background UTL or UPL may lead to the conclusion that the contaminant is a contaminant of potential concern (COPC) to be included in further risk evaluation and risk management studies.

- When individual point-by-point site observations are compared with a threshold value (pre-determined or estimated) of a background population or some other threshold and compliance limit value, such as a PRG, MCL, or ACL, then that threshold value should represent a not-to-exceed value. Such BTVs or not-to-exceed values are often estimated

by a 95% UPL, UTL95%-95%, or by an upper percentile. ProUCL 4.0 can be used to compute any of these upper limits based upon uncensored data sets as well as data sets with nondetect values.

- As the sample size increases, a UCL approaches the sample mean, and a UPL95 approaches the corresponding 95th upper percentile.

- It is pointed out that the developers of ProUCL 4.0 prefer the use of a 95% UPL (UPL95) as an estimate of BTV or a not-to-exceed value. As mentioned before, the option of comparing individual site observations with a BTV (specified or estimated) should be used when few (< 4 to 6) detected site observations (preferably composite values) are to be compared with a BTV.

- When enough (e.g., > 8 to 10) detected site observations are available, it is preferred to use hypotheses testing approaches. Specifically, single sample hypotheses testing (comparing site to a specified threshold) approaches should be used to perform site versus a known threshold comparison; and two-sample hypotheses testing (provided enough background data are also available) approaches should be used to perform site versus background comparison. Several parametric and nonparametric single and two-sample hypotheses testing approaches are available in ProUCL 4.0.

It is re-emphasized that only averages should be compared with the averages or UCLs, and individual site observations should be compared with UPLs, upper percentiles, or UTLs. For an example, the comparison of a 95% UCL of one population (e.g., site) with a 90% or 95% upper percentile of another population (e.g., background) cannot be considered fair and reasonable as these limits (e.g., UCL and UPL) estimate and represent different parameters. It is hard to justify comparing a UCL of one population with a UPL of the other population. Conclusions (e.g., site dirty or site clean) derived by comparing UCLs and UPLs, or UCLs and upper percentiles as suggested in Wyoming DEQ, Fact Sheet #24 (2005), cannot be considered fair and reliable. Specifically, the decision error rates associated with such comparisons can be significantly different from the specified (e.g., Type I error = 0.1, Type II error = 0.1) decision errors.

1.5 Point-by-Point Comparison of Site Observations with BTVs, Compliance Limits, and Other Threshold Values

Point-by-point observation comparison method is used when a small number (e.g., 4 to 6 locations) of detected site observations are compared with pre-established or estimated BTVs, screening levels, or preliminary remediation goals (PRGs). In this case, individual point-by-point site observations (preferably based upon composite samples from various site locations) are compared with estimated or pre-established background (e.g., USGS values) values, PRGs, or some other not-to-exceed value. Typically, a single exceedance of the BTV, PRG, or a not-to-exceed value by a site (or from a monitoring well) observation may be considered as an indication of contamination at the site area under investigation. The conclusion of an exceedance by a site value is Sometimes confirmed by re-sampling (taking a few more collocated samples) that site location (or a monitoring well) exhibiting contaminant concentration in excess of the BTV or PRG. If all collocated (or collected during the same time period) sample observations collected from the same site location (or well) exceed the PRG (or MCL) or a not-to-exceed value, then it may be concluded that the location (well) requires further investigation (e.g., continuing treatment and monitoring) and cleanup.

When BTV contaminant concentrations are not known or pre-established, one has to collect, obtain, or extract a data set of an appropriate size that can be considered as representative of the site related background. Statistical upper limits are computed using the data set thus obtained, which are used as estimates of BTVs and not-to-exceed values. It should be noted that in order to compute reasonably reliable and accurate estimates of BTVs and not-to-exceed values based upon a background (or reference) data set, enough background observations (minimum of 8 to 10) should be collected, perhaps using an appropriate DQO process as described in EPA (2006). Typically, background samples are collected from a comparable general reference area or site-specific areas that are known to be free of contamination due to any of the site related activities. Several statistical limits can be used to estimate the BTVs based upon a defensible data set of an adequate size. A detailed description of the computation and estimation of BTVs is given in Chapter 5 (for uncensored data sets) and in Chapter 6 for data sets with nondetects of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b). Once again, the use of this point-by-point comparison method is recommended when not many (e.g., < 4 to 6) site observations are to be compared with estimated BTVs or PRGs. An exceedance of the estimated BTV by a site value may be considered as an indication of the existing or continuing contamination at the site.

Note: *When BTVs are not known, it is suggested that at least 8 to 10 (more are preferable) detected representative background observations be made available to compute reasonably reliable estimates of BTVs and other not-to-exceed values.*

The point-by-point comparison method is also useful when quick turnaround comparisons are required. Specifically, when the decisions have to be made in real time by a sampling or screening crew, or when few detected site samples are available, then individual point-by-point site concentrations are compared either with pre-established PRGs, cleanup goals and standards, or with estimated BTVs and not-to-exceed values. The crew can use these comparisons to make the following informative decisions:

1. screen and identify the COPCs,
2. identify the polluted site AOCs,
3. continue or stop remediation or excavation at a site AOC or a RU, or
4. move the cleanup apparatus and crew to the next AOC or RU.

During the screening phase, an exceedance of a compliance limit, action level, a BTV, or a PRG by site values for a contaminant may declare that contaminant as a COPC. Those COPCs are then included in future site remediation and risk management studies. During the remediation phase, an exceedance of the threshold value such as a compliance limit (CL) or a BTV by sample values collected from a site area (or a monitoring well (MW)) may declare that site area as a polluted AOC, or a hot spot requiring further sampling and cleanup. This comparison method can also be used to verify if the site concentrations (e.g., from the base or side walls of an excavated site area) are approaching or meeting PRG, BTV, or a cleanup standard after some excavation has been conducted at that site area.

If a larger number of detected samples (e.g., greater than 8 to10) are available from the site locations representing the site area under investigation (e.g., RU, AOC, EA), then the use of hypotheses testing approaches (both single sample and a two-sample) is preferred. The use of a hypothesis testing approach will control the error rates more tightly and efficiently than the individual point-by-point site observations

versus BTV comparisons, especially when many site observations are compared with a BTV or a not-to-exceed value.

Note: *In background versus site comparison evaluations, scientists usually prefer the use of hypotheses testing approaches over point-by-point site observation comparisons with BTVs or not-to-exceed values. Hypotheses testing approaches require the availability of larger data sets from the populations under investigation. Both single sample (used when BTVs, not-to-exceed values, compliance limits, or cleanup standards are known and pre-established) and two-sample (used when BTVs and compliance limits are unknown) hypotheses testing approaches are available in ProUCL 4.0.*

1.6 Hypothesis Testing Approaches and Their Use

Both single sample and two-sample hypotheses testing approaches are used to make cleanup decisions at polluted sites, and also to compare contaminant concentrations of two (e.g., site versus background) or more (several monitoring wells (MWs)) populations. The uses of hypotheses testing approaches in those environmental applications are described as follows.

1.6.1 Single Sample Hypotheses – BTVs and Not-to-Exceed Values are Known (Pre-established)

When pre-established BTVs and not-to-exceed values are used, such as the USGS background values (Shacklette and Boerngen (1984)), thresholds obtained from similar sites, pre-established threshold and not-to-exceed values, PRGs, or RBCs, there is no need to extract, establish, or collect a background or reference data set. When the BTVs and cleanup standards are known, one-sample hypotheses are used to compare site data (provided enough site data are available) with known and pre-established threshold values. It is suggested that the project team determine (e.g., using DQO) or decide (depending upon resources) about the number of site observations that should be collected and compared with the "pre-established" standards before coming to a conclusion about the status (clean or polluted) of the site area (e.g., RU, AOC) under investigation. When the number of available detected site samples is less than 4 to 6, one might perform point-by-point site observation comparisons with a BTV; and when enough detected site observations (> 8 to 10, more are preferable) are available, it is desirable to use single sample hypothesis testing approaches.

Depending upon the parameter (e.g., the average value, μ_0, or a not-to-exceed value, A_0), represented by the known threshold value, one can use single sample hypothesis tests for population mean (t-test, sign test) or single sample tests for proportions and percentiles. The details of the single sample hypotheses testing approaches can be found in EPA (2006) and the Technical Guide for ProUCL 4.0. Several single sample tests listed as follows are available in ProUCL 4.0.

One-Sample t-Test: This test is used to compare the site mean, μ, with some specified cleanup standard, $C_{s,}$ where the cleanup standard, $C_{s,}$ represents an average threshold value, μ_0. The Student's t-test (or a UCL of mean) is often used (assuming normality of site data or when site sample size is large such as larger than 30, 50) to determine the attainment of cleanup levels at a polluted site after some remediation activities.

One-Sample Sign Test or Wilcoxon Signed Rank (WSR) Test: These tests are nonparametric tests and can also handle nondetect observations provided all nondetects (e.g., associated detection limits) fall below the specified threshold value, $C_{s.}$ These tests are used to compare the site location (e.g., median, mean) with some specified cleanup standard, $C_{s,}$ representing a similar location measure.

One-Sample Proportion Test or Percentile Test: When a specified cleanup standard, $A_{0,}$ such as a PRG or a BTV represents an upper threshold value of a contaminant concentration distribution (e.g., not-to-exceed value, compliance limit) rather than the mean threshold value, μ_0, of the contaminant concentration distribution, then a test for proportion or a test for percentile (or equivalently a UTL 95%-95%, UTL 95%-90%) may be used to compare site proportion or site percentile with the specified threshold or action level, $A_{0.}$ This test can also handle ND observations provided all NDs are below the compliance limit.

In order to obtain reasonably reliable estimates and test statistics, an adequate amount of representative site data (8 to 10 detected observations) is needed to perform the hypotheses tests. As mentioned before, in case only a few (e.g., < 4 to 6) detected site observations are available, then point-by-point site concentrations may be compared with the specified action level, A_0.

1.6.2 Two-Sample Hypotheses – When BTVs and Not-to-Exceed Values are Unknown

When BTVs, not-to-exceed values, and other cleanup standards are not available, then site data are compared directly with the background data. In such cases, a two-sample hypothesis testing approach can be used to perform site versus background comparisons. Note that this approach can be used to compare concentrations of any two populations including two different site areas or two different monitoring wells (MWs). In order to use and perform a two-sample hypothesis testing approach, enough data should be available (collected) from each of the two populations under investigation. Site and background data requirements (e.g., based upon DQOs) to perform two-sample hypothesis test approaches are described in EPA (1989b, 2006), Breckenridge and Crockett (1995), and the VSP (2005) software package. While collecting site and background data, for better representation of populations under investigation, one may also want to account for the size of the background area (and site area for site samples) into sample size determination. That is, a larger number (>10 to 15) of representative background (or site) samples should be collected from larger background (or site) areas. As mentioned before, every effort should be made to collect as many samples as determined using DQO processes as described in EPA documents (2006).

The two-sample (or more) hypotheses approaches are used when the site parameters (e.g., mean, shape, distribution) are being compared with the background parameters (e.g., mean, shape, distribution). The two-sample hypotheses testing approach is also used when the cleanup standards or screening levels are not known *a priori*, and they need to be estimated based upon a data set from a background or reference population. Specifically, two-sample hypotheses testing approaches are used to compare: 1) the average contaminant concentrations of two or more populations such as the background population and the potentially contaminated site areas, or 2) the proportions of site and background observations exceeding a pre-established compliance limit, $A_{0.}$ In order to derive reliable conclusions with higher statistical power based upon hypothesis testing approaches, enough data (e.g., minimum of 8 to 10 samples) should be available from all of the populations under investigation. It is also desirable to supplement statistical methods with graphical displays, such as the double Q-Q plots, or side-by-side multiple box plots, as available in ProUCL 4.0. Two-sample hypotheses testing approaches as incorporated in ProUCL 4.0 are listed as follows:

1. Student t-test - with equal dispersions and unequal variances – Parametric test
2. Wilcoxon-Mann-Whitney (WMW) test – Nonparametric test
3. Gehan test – Nonparametric test

Some details of these approaches are described later in this Technical Guide. It should be noted that the WMW, Gehan, and quantile tests are also available for data sets with NDs. Gehan's test is specifically meant to be used on data sets with multiple detection limits. It is also suggested that for best and reliable conclusions, both the WMW and quantile tests should be used on the same data set. The details of these two tests with examples are given in EPA (1994, 2006).

The samples collected from the two (or more) populations should all be of the same type obtained using similar analytical methods and apparatus. In other words, the collected site and background samples should be all discrete or all composite (obtained using the same design and pattern), and be collected from the same medium (soil) at similar depths (e.g., all surface samples or all subsurface samples) and time (e.g., during the same quarter in groundwater applications) using comparable (preferably same) analytical methods. Good sample collection methods and sampling strategies are given in EPA (1996, 2003) guidance documents.

1.7 Minimum Sample Size Requirements

Due to resource limitations, it may not be possible (nor needed) to sample the entire population (e.g., background area, site area, areas of concern, exposure areas) under study. Statistics is used to draw inference(s) about the populations (clean, dirty) and their known or unknown parameters (e.g., comparability of population means, not-to-exceed values, upper percentiles, and spreads) based upon much smaller data sets (samples) collected from those populations under study. In order to determine and establish BTVs, not-to-exceed values, or site-specific screening levels, defensible data set(s) of appropriate size(s) needs to be collected from background areas (e.g., site-specific, general reference or pristine area, or historical data). The project team and site experts should decide what represents a site population and what represents a background population. The project team should determine the population size and boundaries based upon all current and future objectives for the data collection. The size and area of the population (e.g., a remediation unit, area of concern, or an exposure unit) may be determined based upon the potential land use, and other exposure and risk management objectives and decisions. Moreover, appropriate effort should be made to properly collect soil samples (e.g., methods based upon Gy sampling theory), as described in Gerlach and Nocerino (2003).

Using the collected site and background data sets, statistical methods supplemented with graphical displays are used to perform site versus background comparisons. The test results and statistics obtained by performing such site versus background comparisons are used to determine if the site and background level contaminant concentration are comparable; or if the site concentrations exceed the background threshold concentration level; or if an adequate amount of cleanup and remediation approaching the BTV or some cleanup level have been performed at polluted areas (e.g., AOC, RU) of the site under study.

In order to perform statistical inference (estimation and hypothesis testing), one needs to determine the sample sizes that need to be collected from the populations (e.g., site and background) under investigation using appropriate DQO processes (EPA 2006). However, in some cases, it may not be possible to collect the same number of samples as determined by using a DQO process. For example, the data might have already been collected (often is the case in practice) without using a DQO process, or due to resource constraints, it may not be possible to collect as many samples as determined by using a DQO based sample size formula. It is observed that, in practice, the project team and the decision makers may not collect enough background samples, perhaps due to various resource constraints. However, every effort should be made to collect at least 8 to 10 (more are desirable) background observations before using methods as incorporated in ProUCL 4.0. The minimum sample size recommendations as described here are useful when resources are limited (as often is the case), and it may not be possible to collect as many

background and site (e.g., AOC, EU) samples as computed using DQOs and the sample size determination formulae given in the EPA (2006). Some minimum sample size requirements are also given in Fact Sheet #24, prepared by Wyoming Department of Environmental Quality (June 2005).

As mentioned before, the topics of DQO processes and the sample size determination are described in detail in the EPA (2006) guidance document. Therefore, the sample size determination formulae based upon DQO processes are not included in ProUCL 4.0 and its Technical Guide. However, some guidance and suggestions on the minimum number of background and site samples needed to be able to use statistical methods for the computation of upper limits, and to perform single sample tests, two-sample tests such as t-test and Wilcoxon- Mann-Whitney (WMW) test, and various other tests are provided here. The minimum sample size recommendations (requirements) as described here are made so that reasonably reliable estimates of EPC terms and BTVs, and defensible values of test statistics for single or two-sample hypotheses tests (e.g., t-test, WMW test), can be computed.

1.7.1 Minimum Sample Size for Estimation and Point-by-Point Site Observation Comparisons

- Point-by-point observation comparison method is used when a small number (e.g., 4 to 6 locations) of detected site observations are compared with pre-established or estimated BTVs, screening levels, or PRGs. In this case, individual point-by-point site observations (preferably based upon composite samples from various site locations) are compared with estimated or pre-established background (e.g., USGS values) values, PRGs, or some other not-to-exceed value.

- When BTV contaminant concentrations are not known or pre-established, one has to collect, obtain, or extract a data set of an appropriate size that can be considered as representative of the site related background. Statistical upper limits are computed using the data set thus obtained; which are used as estimates of BTVs and not-to-exceed values. It should be noted that in order to compute reasonably reliable and accurate estimates of BTVs and not-to-exceed values based upon a background (or reference) data set, enough background observations (minimum of 8 to 10) should be collected perhaps using an appropriate DQO process as described in EPA (2006). Typically, background samples are collected from a comparable general reference area or a site-specific area.

- When enough (e.g., > 8 to 10) detected site observations are available, it is preferred to use hypotheses testing approaches. Specifically, single sample hypotheses testing (comparing site to a specified threshold) approaches should be used to perform site versus a known threshold comparison and two-sample hypotheses testing (provided enough background data are also available) approaches should be used to perform site versus background comparison.

1.7.2 Minimum Sample Sizes for Hypothesis Testing

Statistical methods (as in ProUCL 4.0) used to estimate EPC terms, BTVs, PRGs, or to compare the site contaminant concentration data distribution with the background data distribution can be computed based upon small site and background data sets (e.g., of sizes 3, 4, 5, or 6). However, those statistics cannot be considered representative and reliable enough to make important cleanup and remediation decisions. It is recommended not to use those statistics to draw cleanup and remediation decisions potentially impacting the human health and the environment. It is suggested that the estimation and hypothesis testing methods as incorporated in ProUCL 4.0 may not be used on background data sets with fewer than 8 to 10 detected

observations. Also, when using hypotheses testing approaches, it is suggested that the site and background data be obtained using an appropriate DQO process as described in EPA (2006). In case that is not possible, it is suggested that the project team at least collect 8 to 10 observations from each of the populations (e.g., site area, MWs, background area) under investigation.

Site versus background comparisons and computation of the BTVs depend upon many factors, some of which cannot be controlled. These factors include the site conditions, lack of historical information, site medium, lack of adequate resources, measurement and analytical errors, and accessibility of the site areas. Therefore, whenever possible, it is desirable to use more than one statistical method to perform site versus background comparison. The use of statistical methods should always be supplemented with appropriate graphical displays.

1.7.3 Sample Sizes for Bootstrap Methods

Several parametric and nonparametric (including bootstrap methods) UCL, UPL, and other limits computation methods for both full-uncensored data sets (without nondetects) and left-censored data sets with nondetects are available in ProUCL 4.0. It should be noted that bootstrap resampling methods are useful when not too few (e.g., < 10-15) and not too many (e.g., > 500-1000) detected observations are available. For bootstrap methods (e.g., percentile method, BCA bootstrap method, bootstrap t method), a large number (e.g., 1000, 2000) of bootstrap resamples (with replacement) are drawn from the same data set. Therefore, in order to obtain bootstrap resamples with at least some distinct values (so that statistics can be computed from each resample), it is suggested that a bootstrap method should not be used when dealing with small data sets of sizes less than 10-15. Also, it is not required to bootstrap a large data set of size greater than 500 or 1000; that is when a data set of a large size (e.g., > 1000) is available, there is no need to obtain bootstrap resamples to compute statistics of interest (e.g., UCLs). One can simply use a statistical method on the original large data set. Moreover, bootstrapping a large data set of size greater than 500 or 1000 will be time consuming.

1.8 Statistical Analyses by a Group ID

The analyses of data categorized by a group ID variable such as: 1) Surface vs. Subsurface; 2) AOC1 vs. AOC2; 3) Site vs. Background; and 4) Upgradient vs. Downgradient monitoring wells are quite common in many environmental applications. ProUCL 4.0 offers this option for data sets with and without nondetects. The Group Option provides a powerful tool to perform various statistical tests and methods (including graphical displays) separately for each of the group (samples from different populations) that may be present in a data set. For an example, the same data set may consist of samples from the various groups or populations representing site, background, two or more AOCs, surface, subsurface, monitoring wells. The graphical displays (e.g., box plots, Q-Q plots) and statistics (background statistics, UCLs, hypotheses testing approaches) of interest can be computed separately for each group by using this option.

It should be pointed out that it is the users' responsibility to provide adequate amount of detected data to perform the group operations. For an example, if the user desires to produce a graphical Q-Q plot (using only detected data) with regression lines displayed, then there should be at least two detected points (to compute slope, intercept, *sd*) in the data set. Similarly if the graphs are desired for each of the group specified by the group ID variable, there should be at least two detected observations in each group specified by the group variable. ProUCL 4.0 generates a warning message (in orange color) in the lower panel of the ProUCL 4.0 screen. Specifically, the user should make sure that a variable with nondetects

and categorized by a group variable should have enough detected data in each group to perform the various methods (e.g., GOF tests, Q-Q plots with regression lines) as incorporated in ProUCL 4.0.

1.9 Use of Maximum Detected Value as Estimates of Upper Limits

Some practitioners tend to use the maximum detected value as an estimate of the EPC term. This is especially true when the sample size is small such as ≤ 5, or when a UCL95 exceeds the maximum detected values (EPA, 1992b). Also, many times in practice, the BTVs and not-to-exceed values are estimated by the maximum detected value. This section discusses the appropriateness of using the maximum detected value as estimates of the EPC term, BTVs, or other nor-to-exceed values.

1.9.1 Use of Maximum Detected Value to Estimate BTVs and Not-to-Exceed Values

It is noted that BTVs and not-to-exceed values represent upper threshold values in the upper tail of a data distribution; therefore, depending upon the data distribution and sample size, the BTVs and other not-to-exceed values may be estimated by the maximum detected value. As described earlier, upper limits, such as UPLs, UTLs, and upper percentiles, are used to estimate the BTVs and not-to-exceed values. It is noted that a nonparametric UPL or UTL is often estimated by higher order statistics such as the maximum value or the second largest value (EPA 1992a, RCRA Guidance Addendum). The use of higher order statistics to estimate the UTLs depends upon the sample size. For an example: 1) 59 to 92 samples, a nonparametric UTL95%-95 is given by the maximum detected value; 2) 93 to 123 samples, a nonparametric UTL95%-95 is given by the second largest maximum detected value; and 3) 124 to 152 samples, a UTL95%-95 is given by the third largest detected value in the sample.

Note: *Therefore, when a data set does not follow a discernable distribution, the maximum observed value (or other high order statistics such as the second largest, third largest) may be used as an estimate of BTV or a not-to-exceed value, provided the maximum value does not represent an outlier or a contaminating observation perhaps representing a hot location. The selection of a higher order statistic (e.g., largest, second largest, third largest) to estimate BTV depends upon the sample size and confidence coefficient.*

1.9.2 Use of Maximum Detected Value to Estimate EPC Terms

This issue was also discussed in the ProUCL 3.0 Technical Guide (EPA, 2004). Some practitioners tend to use the maximum detected value as an estimate of the EPC term. This is especially true when the sample size is small such as ≤ 5, or when a UCL95 exceeds the maximum detected values (EPA, 1992b). Specifically, a RAGS document (EPA, 1992) suggests the use of the maximum detected value as a default value to estimate the EPC term when a 95% *UCL* (e.g., the *H-UCL*) exceeded the maximum value. ProUCL 3.0 and ProUCL 4.0 can compute a 95% *UCL* of mean using several methods based upon normal, Gamma, lognormal, and non-discernable distributions. In past (e.g., EPA, 1992b), only two methods were used to estimate the EPC term based upon: 1) Student's t-statistic and a normal distribution, and 2) Land's H-statistic (1975) and a lognormal model. The use of H-statistic often yields unstable and impractically large UCL95 of the mean (Singh, Singh, and Iaci, 2002). For skewed data sets of smaller sizes (e.g., < 30, < 50), H-UCL often exceeds the maximum detected value. This is especially true when some extreme high outliers may be present in the data set. Since the use of a lognormal distribution has been quite common (e.g., suggested as a default model in a RAGS document (EPA, 1992)), the exceedance of the maximum detected value by H-UCL95 is frequent for many skewed data sets of smaller sizes (e.g., < 30, < 50). It is also be noted that for highly skewed data sets, the sample

mean indeed can even exceed the upper 90%, 95%, etc., percentiles, and consequently, a 95% *UCL* of mean can exceed the maximum observed value of a data set.

All of these occurrences result in the possibility of using the maximum detected value as an estimate of the EPC term. It should be pointed out that in some cases, the maximum observed value actually might represent a highly polluted outlying observation. Obviously, it is not desirable to use a highly polluted value as an estimate of average exposure (EPC term) for an exposure area. This is especially true when one is dealing with lognormally distributed data sets of small sizes. As mentioned before, for such highly skewed data sets that cannot be modeled by a gamma distribution, a 95% *UCL* of the mean should be computed using an appropriate distribution-free nonparametric method.

It should be pointed out that the EPC term represents the average exposure contracted by an individual over an exposure area (EA) during a long period of time; therefore, the EPC term should be estimated by using an average value (such as an appropriate 95% *UCL* of the mean) and not by the maximum observed concentration. One needs to compute an average exposure and not the maximum exposure. It is unlikely that an individual will visit the location (e.g., in an EA) of the maximum detected value all of the time. One can argue that the use of this practice results in a conservative (higher) estimate of the EPC term. The objective is to compute an accurate estimate of the EPC term. Today, several other methods (instead of H-UCL) as described in EPA (2002), and included in ProUCL 3.0 (EPA 2004) and ProUCL 4.0 (EPA 2007), are available which can be used to estimate the EPC terms. It is unlikely (but possible with outliers) that the UCLs based upon those methods will exceed the maximum detected value, unless some outliers are present in the data set. ProUCL 4.0 displays a warning message when the recommended 95% *UCL* (e.g., Hall's or bootstrap t *UCL* with outliers) of the mean exceeds the observed maximum concentration. When a 95% *UCL* does exceed the maximum observed value, ProUCL4.0 recommends the use of an alternative *UCL* computation method based upon the Chebyshev inequality. One may use a 97.5% or 99% Chebyshev UCL to estimate the mean of a highly skewed population. It should be noted that typically, a Chebyshev UCL yield conservative (but stable) and higher values of the UCLs than other methods available in ProUCL 4.0. This is especially true when data are moderately skewed and sample size is large. In such cases, when the sample size is large (and other UCL methods such as bootstrap t method yield unrealistically high values), one may want to use a 95% Chebyshev UCL or a Chebyshev UCL with lower confidence coefficient such as 92.5% or 90% as estimate of the population mean, especially when the sample size is large (e.g., >100, 150). The detailed recommendations (as functions of sample size and skewness) for the use of those UCLs are summarized in ProUCL 3.0 Technical Guide (EPA, 2004).

Singh and Singh (2003) studied the performance of the max test (using the maximum observed value as an estimate of the EPC term) via Monte Carlo simulation experiments. They noted that for skewed data sets of small sizes (e.g., < 10-20), the max test does not provide the specified 95% coverage to the population mean, and for larger data sets, it overestimates the EPC term, which may require unnecessary further remediation. This can also be viewed in the graphs presented in ProUCL 3.0 Technical Guide. The use of the maximum value as an estimate of the EPC term also ignores most (except for maximum value) of the information contained in the data set.

With the availability of so many *UCL* computation methods (15 of them), the developers of ProUCL 4.0 do not feel any need to use the maximum observed value as an estimate of the EPC term representing an average exposure by an individual over an EA. Also, for the distributions considered, the maximum value is not a sufficient statistic for the unknown population mean.

Note: *It is recommended that the maximum observed value NOT be used as an estimate of the EPC term representing average exposure contracted by an individual over an EA. For the sake of interested users, ProUCL displays a warning message when the recommended 95% UCL (e.g., Hall's bootstrap UCL, etc.) of the mean exceeds the observed maximum concentration. For such scenarios (when a 95% UCL does exceed the maximum observed value), an alternative 95% UCL computation method is recommended by ProUCL 4.0.*

1.10 Samples with Nondetect Observations

Nondetect observations (or less than obvious values) are inevitable in most environmental data sets. Singh, Maichle, and Lee (EPA, 2006) studied the performances (in terms of coverages) of the various UCL95 computation methods including the simple substitution methods (such as the DL/2 and DL methods) for data sets with nondetect observations. They concluded that the UCLs obtained using the substitution methods, including the replacement of nondetects by respective DL/2, do not perform well even when the percentage of nondetect observations is low, such as 5%-10%. They recommended avoiding the use of substitution methods to compute UCL95 based upon data sets with nondetect observations.

1.10.1 Avoid the Use of DL/2 Method to Compute UCL95

Based upon the results of the report by Singh, Maichle, and Lee (EPA, 2006), it is strongly recommended to avoid the use of the DL/2 method to perform GOF test, and to compute the summary statistics and various other limits (e.g., UCL, UPL) often used to estimate the EPC terms and BTVs. Until recently, the DL/2 method has been the most commonly used method to compute the various statistics of interest for data sets with BDL observations. The main reason of its common use has been the lack of the availability of other defensible methods and associated programs that can be used to estimate the various environmental parameters of interest. Today, several other methods (e.g., KM method, bootstrap methods) with better performances are available that can be used to compute the various upper limits of interest. Some of those parametric and nonparametric methods are now available in ProUCL 4.0. *Even though the DL/2 method (to compute UCLs, UPLs, and for goodness-of-fit test) has also been incorporated in ProUCL 4.0, its use is not recommended due to its poor performance.* The DL/2 method is included in ProUCL 4.0 only for historical reasons as it had been the most commonly used and recommended method until recently (EPA, 2006). Some of the reviewers of ProUCL 4.0 suggested and requested the inclusion of DL/2 method in ProUCL for comparison purposes.

Note: *The DL/2 method has been incorporated in ProUCL 4.0 for historical reasons only. NERL-EPA, Las Vegas strongly recommends avoiding the use of DL/2 method even when the percentage (%) of NDs is as low as 5%-10%. There are other methods available in ProUCL 4.0 that should be used to compute the various summary statistics and upper limits based upon data sets with single and multiple detection limits.*

1.11 Samples with Low Frequency of Detection

When all of the sampled data values are reported as nondetects, the EPC term should also be reported as a nondetect value, perhaps by the maximum reporting limit (RL) or maximum RL/2. Statistics (e.g., UCL95) computed based upon only a few detected values (e.g., < 4 to 6) cannot be considered reliable enough to estimate the EPC terms having potential impact on the human heath and the environment. When the number of detected data is small, it is preferable to use simple ad hoc methods rather than using statistical methods to compute the EPC terms and other upper limits. Specifically, it is suggested that in

cases when the detection frequency is low (e.g., < 4%-5%) and the number of detected observations is low, the project team and the decision makers together should make a decision on site-specific basis on how to estimate the average exposure (EPC term) for the contaminant and area under consideration. For such data sets with low detection frequencies, other measures such as the median or mode represent better estimates (with lesser uncertainty) of the population measure of central tendency.

Additionally, it is also suggested that when most (e.g., > %95) of the observations for a contaminant lie below the detection limit(s) or reporting limits (RLs), the sample median or the sample mode (rather than the sample average which cannot be computed accurately) may be used as an estimate the EPC term. Note that when the majority of the data are nondetects, the median and the mode will also be a nondetect. The uncertainty associated with such estimates will be high. It is noted that the statistical properties, such as the bias, accuracy, and precision of such estimates, would remain unknown. In order to be able to compute defensible estimates, it is always desirable to collect more samples.

Note: *In case the number of available detected samples is small (< 5), it is suggested that the project team decide about the estimation of the EPC term on site-specific basis. For such small data sets with very few detected values (< 5), the final decision ("policy decision") on how to estimate the EPC term should be determined by the project team and decision makers.*

1.12 Some Other Applications of Methods in ProUCL 4.0

In addition to performing background versus site comparisons for CERCLA and RCRA sites, and estimating the EPC terms in exposure and risk evaluation studies, the statistical methods as incorporated in ProUCL 4.0 can be used to address other issues dealing with environmental investigations that are conducted at Superfund or RCRA sites.

1.12.1 Identification of COPCs

Risk assessors and RPMs often use screening levels or BTVs to identify the COPCs during the screening phase of a cleanup project to be conducted at a contaminated site. The screening for the COPCs is performed prior to any characterization and remediation activities that may have to be conducted at the site under investigation. This comparison is performed to screen out those contaminants that may be present in the site medium of interest at low levels (e.g., at or below the background levels or some pre-established screening levels) and may not pose any threat and concern to human health and the environment. Those contaminants may be eliminated from all future site investigations, and risk assessment and risk management studies.

In order to identify the COPCs, point-by-point site observations (preferably composite samples) are compared with some pre-established screening levels, SSL, or estimated BTVs. This is especially true when the comparisons of site concentrations with screening levels or BTVs are conducted in real time by the sampling or cleanup crew right there in the site field. The project team should decide about the type of site samples (discrete or composite) and the number of detected site observations (not more than 4 to 6) that should be collected and compared with the screening levels or the BTVs. In case BTVs, screening levels, or not-to-exceed values are not known, the availability of a defensible background or reference data set of reasonable size (e.g., > 8 to 10, more are preferable) is required to obtain reliable estimates of BTVs and screening levels. When a reasonable number of detected site observations are available, it is preferable to use hypotheses testing approaches. The contaminants with concentrations exceeding the respective screening values or BTVs may be considered as COPCs, whereas contaminants with

concentrations (in all collected samples) lower than the screening value, PRG, or an estimated BTV may be omitted from all future evaluations including the risk assessment and risk management investigations.

1.12.2 Identification of Non-Compliance Monitoring Wells

In monitoring well (MW) compliance assessment applications, individual (often discrete) contaminant concentrations from a MW are compared with some pre-established ACL, MCL, or an estimated compliance limit (CL) based upon a group of upgradient wells representing the background population. An exceedance of the MCL or the BTV by a MW concentration may be considered as an indication of contamination in that MW. In such individual concentration comparisons, the presence of contamination (determined by an exceedance) may have to be confirmed by re-sampling from that MW. If concentrations of contaminants in both the original sample and the re-sample(s) exceed the MCL or BTV, then that MW may require closer scrutiny, perhaps triggering the remediation remedies as determined by the project team. If the concentration data from a MW for about 4 to 5 continuous quarters (or some other designated time period determined by the project team) are below the MCL or BTV level, then that MW may be considered as complying with (achieving) the pre-established or estimated standards. Statistical methods as described in Chapters 5 and 6 of the revised Guidance for *Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b) can be used to estimate the not-to-exceed values or BTVs based upon background or upgradient wells in case the ACLs or MCLs are not pre-determined.

1.12.3 Verification of the Attainment of Cleanup Standards, C_s

Hypothesis testing approaches may be used to verify the attainment of the cleanup standard, C_s, at polluted site areas of concern after conducting remediation and cleanup at the site AOC (EPA, 2006). In order to properly address this scenario, a site data set of adequate size (minimum of 8 to 10 detected site observations) needs to be made available from the remediated or excavated areas of the site under investigation. The sample size should also account for the size of the remediated site area; meaning that larger site areas should be sampled more (with more observations) to obtain a representative sample of the site under investigation.

Typically, the null hypothesis of interest is H_0: Site Mean, $\mu_s >= C_s$ versus the alternative hypothesis, H_1: Site Mean, $\mu_s < C_s$, where the cleanup standard, C_s, is known *a priori*. The sample size needed to perform such single sample hypotheses tests can be obtained using the DQO process-based sample size formula as given in the EPA (2006) documents. In any case, in order to use this test, a minimum of 8 to 10 detected site samples should be collected. The details of the statistical methods used to perform single sample hypothesis as described above can be found in EPA (2006).

1.12.4 Using BTVs (Upper Limits) to Identify Hot Spots

The use of upper limits (e.g., UTLs) to identify hot spot(s) has also been mentioned in the *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b). Point-by-point site observations (preferably using composite samples representing a site location) are compared with a pre-established or estimated BTV. Exceedances of the BTV by site observations may be considered as representing locations with elevated concentrations (hot spots). Chapters 5 and 6 of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b) describe several methods to estimate the BTVs based upon data sets without nondetects (NDs) and left-censored data sets with NDs.

The rest of the chapters of this Technical Guide briefly describe the various statistical methods as incorporated in ProUCL 4.0. Those methods are useful to analyze environmental data sets with and without the nondetect observations. It should be noted that ProUCL 4.0 is the first software package equipped with single sample and two-sample hypotheses testing approaches that can be used on data sets with nondetect observations.

Note: *It should be pointed out that while developing ProUCL 4.0, emphasis is given to the practical applicability of the estimation and hypotheses testing methods as incorporated in ProUCL 4.0. Also, it should be noted that ProUCL 4.0 does provide many graphical and statistical methods often used in the various statistical applications. ProUCL 4.0 does not provide statistical methods that may be used to compute sample sizes based upon DQO processes (EPA, 2006). Those sample size determination methods are available in other freeware packages such as VSP (2005) and DataQUEST (EPA, 1997). However, as mentioned before, some practical guidance on the minimum sample size requirements to be able to use methods as available in ProUCL 4.0 has been provided in Chapter 1. Similar statements and suggestions have been made throughout this Technical Guide.*

Chapter 2

Methods for Computing (1 – α)100% UCL of Mean for Data Sets without Nondetect Observations as Incorporated in ProUCL 4.0 Software

2.1 Introduction

Exposure assessment and cleanup decisions in support of U.S. EPA projects are often made based upon the mean concentrations of the contaminants of potential concern. A 95% upper confidence limit (*UCL*) of the unknown population arithmetic mean (*AM*), μ_1, is often used to: estimate the exposure point concentration (EPC) term (EPA, 1992; EPA, 2002), determine the attainment of cleanup standards (EPA, 1989, and EPA, 1991), estimate background level contaminant concentrations, or compare the soil concentrations with site-specific soil screening levels (EPA, 1996). It is, therefore, important to compute a reliable, conservative, and stable 95% *UCL* of the population mean using the available data. The 95% UCL should approximately provide the 95% coverage for the unknown population mean, μ_1. EPA (2002a) has developed a guidance document for calculating upper confidence limits based upon full data sets without nondetect observations. Most of those *UCL* computation methods as described in the EPA (2002a) guidance document are available in ProUCL 3.0. ProUCL 3.0 can also compute 95% UCLs of the mean based upon the gamma distribution which is better suited to model positively skewed environmental data sets. ProUCL 4.0 represents an upgrade of ProUCL 3.0. Specifically, ProUCL 4.0 provides several parametric and nonparametric UCL computation methods for data sets with nondetect (ND) observations. Therefore, this Technical Guide is an upgrade of the technical guide associated with ProUCL 3.0. The capabilities and methods as incorporated in ProUCL 3.0 are also available in ProUCL 4.0. Parametric and nonparametric UCL computation methods as incorporated in ProUCL 4.0 for data sets with nondetect observations are described in Chapter 4 of this Technical Guide. The details of those UCL computation methods can be found in Singh, Maichle, and Lee (EPA, 2006).

Chapter 2 describes the UCL methods for full data sets without ND observations as incorporated in ProUCL 3.0 Technical Guide. Computation of a *(1 – α)100% UCL* of the population mean depends upon the data distribution. Typically, environmental data are positively skewed, and a default lognormal distribution (EPA, 1992) is often used to model such data distributions. The H-statistic-based Land's (Land, 1971, 1975) *H-UCL* of the mean is used in these applications. Hardin and Gilbert (1993), Singh, Singh, and Engelhardt (1997, 1999), Schultz and Griffin (1999), and Singh, Singh, and Iaci (2002b) pointed out several problems associated with the use of the lognormal distribution and the *H-UCL* of the population *AM*. In practice, for lognormal data sets with high standard deviation (*Sd*), σ, of the natural log-transformed data (e.g., σ exceeding 2.0), the *H-UCL* can become unacceptably large, exceeding the 95% and 99% data quantiles, and even the maximum observed concentration, by orders of magnitude (Singh, Singh, and Engelhardt, 1997). This is especially true for skewed data sets of sizes smaller than $n <$ 50-70.

The *H-UCL* is also very sensitive to a few low or high values. For example, the addition of a sample with below detection limit measurement can cause the *H-UCL* to increase by a large amount (Singh, Singh, and Iaci, 2002b). Realizing that the use of H-statistic can result in unreasonably large *UCL*, it has been recommended (EPA, 1992) to use the maximum observed value as an estimate of the *UCL* (EPC term) in cases where the *H-UCL* exceeds the maximum observed value. The issue of the use of the maximum detected value as an estimate of the EPC term has been discussed earlier in Chapter 1. There are 15 *UCL* computation methods available in ProUCL; 5 are parametric and 10 are nonparametric. The

nonparametric methods do not depend upon any of the data distributions. For full data sets without any nondetect observations, ProUCL 4.0 (and also ProUCL 3.0) makes recommendations on how to compute an appropriate UCL95. Those recommendations are made based upon the findings of extensive simulation study conducted by Singh and Singh (2003).

It is noted that both lognormal and gamma distributions can be used to model positively skewed data sets. It is also noted that it is hard to distinguish between a lognormal and a gamma distribution, especially when the sample size is small, such as $n < 50$ - 70. In practice, many skewed data sets follow a lognormal as well as a gamma distribution. Singh, Singh, and Iaci (2002b) observed that the *UCL* based upon a gamma distribution results in reliable and stable values of practical merit. It is, therefore, desirable to test if an environmental data set follows a gamma distribution. For data sets (of all sizes) which follow a gamma distribution, the EPC should be computed using an adjusted gamma *UCL* (when $0.1 \leq k < 0.5$) of the mean or an approximate gamma *UCL* (when $k \geq 0.5$) of the mean, as these *UCLs* approximately provide the specified 95% coverage to the population mean, $\mu_1 = k\theta$, gamma distribution. For values of $k < 0.1$, a 95% *UCL* may be obtained using bootstrap t-method or Hall's bootstrap method when the sample size, n is less than 15, and for larger samples, a *UCL* of the mean may be computed using the adjusted or approximate gamma *UCL*. Here, k is the shape parameter of a gamma distribution as described in later in this chapter.

It should be pointed out that both bootstrap t and Hall's bootstrap methods sometimes result in erratic, inflated, and unstable *UCL* values, especially in the presence of outliers (Efron and Tibshirani, 1993). Therefore, these two methods should be used with caution. The user should examine the various *UCL* results and determine if the *UCLs* based upon the bootstrap t and Hall's bootstrap methods represent reasonable and reliable *UCL* values of practical merit. If the results based upon these two methods are much higher than the rest of methods (except for the *UCLs* based upon lognormal distribution), then this could be an indication of erratic *UCL* values. ProUCL prints out a warning message whenever the use of these two bootstrap methods is recommended. In case these two bootstrap methods yield erratic, unstable, and inflated *UCLs*, the *UCL* of the mean may be computed using the adjusted or the approximate gamma *UCL* computation method, or based upon the Chebyshev inequality.

ProUCL 4.0 has goodness-of-fit (GOF) methods to test for normality, lognormality, and a gamma distribution of a data set with and without nondetect observations. Depending upon the data distribution, ProUCL 4.0 can be used to compute a conservative and stable *95% UCL* of the population mean, μ_1, and various other upper limits (e.g., UPLs, UTLs) for data sets with and without the nondetect observations. The critical values of the Anderson-Darling test statistic and the Kolmogorov-Smirnov test statistic to test for gamma distribution were generated using Monte Carlo simulation experiments. Those critical values are tabulated in Appendix A for various levels of significance. Singh, Singh, and Engelhardt (1997, 1999); Singh, Singh, and Iaci (2002b); and Singh and Singh (2003) studied several parametric and nonparametric *UCL* computation methods that have been included in ProUCL 4.0. Most of the mathematical algorithms and formulae used in ProUCL to compute the various statistics are summarized in this chapter. ProUCL computes the various summary statistics for raw, as well as log-transformed data sets with and without nondetect observations. In this Technical Guide and in ProUCL, log-transform (*log*) stands for the natural logarithm (*ln*) to the base e. ProUCL also computes the maximum likelihood estimates (*MLEs)* and the minimum variance unbiased estimates (*MVUEs*) of various unknown population parameters of normal, lognormal, and gamma distributions. For full data sets without nondetect observations, ProUCL 4.0 (and also ProUCL 3.0) computes the *$(1 - \alpha)$100% UCLs* of the unknown population mean, μ_1, using five (5) parametric and ten (10) nonparametric methods, which are described in section 2.4 of this chapter.

For data sets without NDs, comparisons of the performances of the UCL computation methods (in terms of coverage probabilities) were performed by Singh and Singh (2003) and Singh *et al.* (2006). It is also well known that the Jackknife method (with sample mean as an estimator) and Student's t-method yield identical *UCL* values. Moreover, it is noted that the standard bootstrap method and the percentile bootstrap method do not perform well (do not provide adequate coverage) for skewed data sets. However, for the sake of completeness, all of the parametric as well as nonparametric methods have been included in ProUCL 4.0. Also, it was noted that the omission of a method such as the Jackknife method or the bias-corrected accelerated (BCA) bootstrap method triggers the curiosity of some of the users as they may think that the omitted method might perform better than the various other methods already incorporated in ProUCL. In order to satisfy all users, ProUCL 4.0 provides most of the bootstrap *UCL* computation methods.

2.2 Goodness-of-Fit (GOF) Test Procedures to Test for a Data Distribution

Let $x_1, x_2, \ldots, x_n$ be a random sample (e.g., representing lead concentrations) from the underlying population (e.g., remediated part of a site) with unknown mean, μ_1, and variance, σ_1^2. Let μ and σ represent the population mean and the population standard deviation (*Sd*) of the log-transformed (natural log to the base e) data. Let $\bar{y}$ and s_y ($= \hat{\sigma}$) be the sample mean and sample *Sd*, respectively, of the log-transformed data, $y_i = \log(x_i)$; $i = 1, 2, \ldots, n$. Specifically, let

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (2\text{-}1)$$

$$\hat{\sigma}^2 = s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (2\text{-}2)$$

Similarly, let $\bar{x}$ and s_x be the sample mean and *Sd* of the raw data, $x_1, x_2, \ldots, x_n$, obtained by replacing y by x in equations (2-1) and (2-2), respectively. In this Technical Guide, irrespective of the underlying distribution, μ_1, and σ_1^2 represent the mean and variance of the random variable X (in original units), whereas μ and σ^2 represent the mean and variance of its logarithm, given by $Y = \log_e(X)$ = natural logarithm.

Three data distributions have been considered in ProUCL 4.0. These include the normal, lognormal, and the gamma distributions. Shapiro-Wilk ($n \leq 50$) and Lilliefors ($n > 50$) test statistics are used to test for normality or lognormality of a data set. It should be noted that even though Shapiro-Wilk (S-W) test has been extended up to samples of size 2000 (Royston, 1982), ProUCL 4.0 provides S-W test only for samples of sizes up to 50. Lilliefors test (along with graphical Q-Q plot) seems to perform fairly well for samples of size 50 and higher. The empirical distribution function (EDF)-based methods: the Kolmogorov-Smirnov (K-S) test and the Anderson-Darling (A-D) test are used to test for a gamma distribution. Extensive critical values for these two test statistics have been obtained via Monte Carlo simulation experiments. For interested users, these critical values are given in the Appendix for various levels of significance. In addition to these formal tests, the informal histogram and quantile-quantile (Q-Q) plot are also available to visually test data distributions. Q-Q plots also provide useful information about the presence of potential outliers and multiple populations. A brief description of these GOF tests follows.

2.2.1 Test Normality and Lognormality of a Data Set

ProUCL tests for normality or lognormality of a data set using three different methods described below. The program tests normality or lognormality at three different levels of significance, namely, 0.01, 0.05, and 0.1. The details of those methods can be found in the cited references below.

2.2.1.1 Normal Quantile-Quantile (Q-Q) Plot

This is a simple informal graphical method to test for an approximate normality or lognormality of a data distribution (Hoaglin, Mosteller, and Tukey 1983; Singh 1993). A linear pattern displayed by the bulk of the data suggests approximate normality or lognormality (performed on log-transformed data) of the data distribution. For example, a high value (e.g., 0.95 or greater) of the correlation coefficient of the linear pattern may suggest approximate normality (or lognormality) of the data set under study. However, it should be noted that on this graphical display, observations well separated (sticking out) from the linear pattern displayed by the bulk data represent the outlying observations. Also, apparent jumps and breaks in the Q-Q plot suggest the presence of multiple populations. The correlation coefficient of such a Q-Q plot can still be high, which does not necessarily imply that the data follow a normal (or lognormal) distribution. Therefore, the informal graphical Q-Q plot test should always be accompanied by other more powerful tests, such as the Shapiro-Wilk test or the Lilliefors test. The goodness-of-fit (GOF) test of a data set should always be judged based upon the formal as well as informal graphical displays. The normal Q-Q plot may be used as an aid to identify outliers or to identify multiple populations. ProUCL performs the graphical Q-Q plot test on raw data as well as on standardized data. All relevant statistics, such as the correlation coefficient, are also displayed on the Q-Q plot.

2.2.1.2 Shapiro-Wilk W Test

This is a powerful test and is often used to test the normality or lognormality of the data set under study (Gilbert 1987). ProUCL performs this test for samples of size 50 or smaller. The Scout (EPA 1999) software package (currently under revision and upgrade) will extend the S-W test for samples of size up to 2000. Based upon the selected level of significance and the computed test statistic, ProUCL also informs the user if the data are normally (or lognormally) distributed. This information should be used to obtain an appropriate *UCL* of the mean. The program prints the relevant statistics (such as the S-W test statistic, slope, and correlation) on the Q-Q plot of the data. For convenience, normality, lognormality, or gamma distribution test results for 0.05 level of significance are also displayed on the Excel-type output summary sheets.

2.2.1.3 Lilliefors Test

This test is useful for data sets of larger size (Dudewicz and Misra, 1988, Conover, 1999). ProUCL performs this test for samples of sizes up to 1000. Based upon the selected level of significance and the computed test statistic, ProUCL informs the user if the data are normally (or lognormally) distributed. The user should use this information to obtain an appropriate *UCL* of the mean. The program prints the relevant statistics on the Q-Q plot of data. For convenience, normality, lognormality, or gamma distribution test results for 0.05 level of significance are also displayed on the *UCL* output summary sheets. It should be pointed out that sometimes, in practice, these two goodness-of-fit tests could lead to different conclusions.

2.2.2 Gamma Distribution

Singh, Singh, and Iaci (2002b) studied gamma distributions to model positively skewed environmental data sets and to compute a *UCL* of the mean based upon a gamma distribution. They studied several *UCL* computation methods using Monte Carlo simulation experiments. A continuous random variable, X (e.g., concentration of a contaminant), is said to follow a gamma distribution, G (k, θ) with parameters $k > 0$ (shape parameter) and $\theta > 0$ (scale parameter), if its probability density function is given by the following equation:

$$f(x;k,\theta) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-x/\theta}; \qquad x > 0$$
$$= 0; \qquad \textit{otherwise} \tag{2-3}$$

The parameter, k, is the shape parameter, and θ is the scale parameter. Many positively skewed data sets follow a lognormal as well as a gamma distribution. Gamma distributions can be used to model positively skewed environmental data sets. It is observed that the use of a gamma distribution results in reliable and stable 95% *UCL* values. It is therefore, desirable to test if an environmental data set follows a gamma distribution. If a skewed data set does follow a gamma model, then a 95% *UCL* of the population mean should be computed using a gamma distribution. For details of the two gamma goodness-of-fit tests, maximum likelihood estimation of gamma parameters, and the computation of a 95% *UCL* of the mean based upon a gamma distribution, refer to D'Agostino and Stephens (1986), and Singh, Singh, and Iaci (2002). These methods are briefly described as follows.

For data sets which follow a gamma distribution, the adjusted 95% *UCL* of the mean based upon a gamma distribution is optimal and approximately provides the specified 95% coverage to population mean, $\mu_1 = k\theta$ (Singh, Singh, and Iaci (2002)). Moreover, this adjusted gamma *UCL* yields reasonable numbers of practical merit. The two test statistics used for testing for a gamma distribution are based upon the empirical distribution function (EDF). The two EDF tests included in ProUCL are the Kolmogorov-Smirnov (K-S) test and Anderson-Darling (A-D) test, which are described in D'Agostino and Stephens (1986) and Stephens (1970). The graphical Q-Q plot for gamma distribution has also been included in ProUCL. The critical values for the two EDF tests are not easily available, especially when the shape parameter, k, is small ($k < 1$). Therefore, the associated critical values have been obtained via extensive Monte Carlo simulation experiments. These critical values for the two test statistics are given in Appendix A. The 1%, 5%, and 10% critical values of these two test statistics have been incorporated in ProUCL 4.0. It should be noted that the goodness-of-fit tests for gamma distribution depend upon the *MLEs* of gamma parameters, k and θ, which should be computed first before performing the goodness-of-fit tests. It is noted that the information about estimation of gamma parameters, gamma GOF tests, and construction of gamma Q-Q plots is not easily available in statistical textbooks. Therefore, the detailed description of these methods for gamma distribution is provided as follows.

2.2.2.1 Quantile-Quantile (Q-Q) Plot for a Gamma Distribution

Let $x_1, x_2, \ldots, x_n$ be a random sample from the gamma distribution, G(k,θ). Let $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ represent the ordered sample. Let $\hat{k}$ and $\hat{\theta}$ represent the maximum likelihood estimates (*MLEs*) of k and θ, respectively. For details of the computation of the *MLEs* of k and θ, refer to Singh, Singh, and Iaci (2002). Estimations of the gamma parameters are also briefly described later in this Technical Guide. The Q-Q plot for gamma distribution is obtained by plotting the scatter plot of pairs $(x_{0i}, x_{(i)})$ $i := 1, 2, \ldots, n$.

Here the quantiles, x_{0i}, are given by the equation, $x_{0i} = z_{0i}\hat{\theta}/2;\ i := 1, 2, \ldots, n$, where the quantiles Z_{0i} (already ordered) are obtained by using the inverse chi-square distribution and are given as follows.

$$\int_0^{z_{0i}} f(\chi^2_{2\hat{k}})d\chi^2_{2\hat{k}} = (i-1/2)/n; \quad i := 1, 2, \ldots, n \tag{2-4}$$

In (2-4), $\chi^2_{2\hat{k}}$ represents a chi-square random variable with $2\hat{k}$ degrees of freedom (*df*). The program, PPCHI2 (Algorithm AS91) as given in Best and Roberts (1975), Applied Statistics (1975, Vol. 24, No. 3) has been used to compute the inverse chi-square percentage points, as given by the above equation given by (2-4). This represents an informal graphical method to test for a gamma distribution. All relevant statistics including the *MLE* of k are also displayed on the gamma Q-Q plot.

This informal test should always be accompanied by the formal Anderson-Darling (A-D) test or Kolmogorov-Smirnov (K-S) test and vice versa. A linear pattern displayed by the scatter plot of bulk of the data may suggest an approximate gamma distribution. For example, a high value (e.g., 0.95 or greater) of the correlation coefficient of the linear pattern may suggest approximate gamma distribution of the data set under study. However, on this Q-Q plot points well separated from the bulk of data may represent outliers. Apparent breaks and jumps in the gamma Q-Q plot suggest the presence of multiple populations. Thus, Q-Q plots are also useful to identify outliers or the presence of multiple populations. The correlation coefficient of such a Q-Q plot (e.g., with outliers) can still be high which does not necessarily imply that the data follow a gamma distribution. Therefore, graphical Q-Q plot and other formal EDF tests, such as the Anderson-Darling (A-D) test or the Kolmogorov-Smirnov (K-S) test should be used on the same data set. A formal statistical test such as a K-S test or A-D test may lead to conclusion of a gamma distribution even for a data set with potential outliers and multiple populations. The final conclusion about the data distribution should be based upon the formal goodness-of-fit tests. This statement is true for all GOF tests (e.g., normal, lognormal, and gamma distributions) as incorporated in ProUCL 4.0.

2.2.2.2 Empirical Distribution Function (EDF)-Based Goodness-of-Fit Tests

Next, the two formal empirical distribution function (EDF)-based test statistics to test for a gamma distribution are briefly described here. Let $F(x)$ be the cumulative distribution function (CDF) of the gamma random variable X. Let $Z = F(X)$, then Z represents a uniform U(0,1) random variable. For each x_i, compute z_i by using the incomplete gamma function given by the equation $z_i = F(x_i);\ i := 1, 2, \ldots, n$. The algorithm as given in Numerical Recipes book (Press *et al.*, 1990) has been used to compute the incomplete gamma function. Arrange the resulting z_i in ascending order as $z_{(1)} \le z_{(2)} \le \ldots \le z_{(n)}$. Let $\bar{z} = \left(\sum_{i=1}^{n} z_i\right)/n$ be the mean of the z_i; $i := 1, 2, \ldots, n$.

Compute the following two test statistics.

$$D^+ = \max_i\{1/n - z_{(i)}\}, \text{ and } D^- = \max_i\{z_{(i)} - (i-1)/n\} \tag{2-5}$$

The Kolmogorov-Smirnov test statistic is given by $D = \max(D^+, D^-)$.

The Anderson-Darling test statistic is given by the following equation.

$$A^2 = -n - (1/n)\sum_{i=1}^{n}\{(2i-1)[\log z_{(i)} + \log(1 - z_{(n+1-i)})]\} \qquad (2\text{-}6)$$

The critical values for these two statistics, D and A^2, are not readily available. For the Anderson-Darling test, only the asymptotic critical values are available in the statistical literature (D'Agostino and Stephens (1986)). Some raw critical values for K-S test are given in Schneider (1978), and Schneider and Clickner (1976). For these two tests, ExpertFit (2001) software and Law and Kelton (2000) use generic critical values for all completely specified distributions as given in D'Agostino and Stephens (1986). It is observed that the conclusions derived using these generic critical values for completely specified distributions and the simulated critical values for the gamma distribution with unknown parameters can be different. Therefore, to test for a gamma distribution, it is preferred and advised to use the critical values of these test statistics specifically obtained for gamma distributions with unknown parameters.

In practice, the distributions are not completely specified and exact critical values for these two test statistics are needed. It should be noted that the distributions of the K-S test statistic, D, and the A-D test statistic, A^2, do not depend upon the scale parameter, θ; therefore, the scale parameter, θ, has been set equal to 1 in all of the simulation experiments. The critical values for these two statistics have been obtained via extensive Monte Carlo simulation experiments for several small and large values of the shape parameter, k, and with $\theta = 1$. These critical are included in Appendix A. In order to generate the critical values, random samples from gamma distributions were generated using the algorithm as given in Whittaker (1974). It is observed that the critical values thus obtained are in close agreement with all available published critical values. The generated critical values for the two test statistics have been incorporated in ProUCL for three levels of significance, 0.1, 0.05, and 0.01. For each of the two tests, if the test statistic exceeds the corresponding critical value, then the hypothesis that the data follow a gamma distribution is rejected. ProUCL computes these test statistics and prints them on the gamma Q-Q plot and also on the *UCL* summary output sheets generated by ProUCL.

2.3 Estimation of Parameters of the Three Distributions as Incorporated in ProUCL

Throughout this Technical Guide, μ_1 and σ_1^2 are the mean and variance of the random variable, X, and μ and σ^2 are the mean and variance of the random variable, $Y = \log(X)$. Also, $\hat{\sigma}$ represents the standard deviation of the log-transformed data. It should be noted that for both lognormal and gamma distributions, the associated random variable can take only positive values. This is typical of environmental data sets to consist of only positive values.

2.3.1 Normal Distribution

Let X be a continuous random variable (e.g., concentration of COPC), which follows a normal distribution, N (μ_1, σ_1^2) with mean, μ_1, and variance, σ_1^2. The probability density function of a normal distribution is given by the following equation:

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-(x-\mu)^2/2\sigma^2]; \qquad -\infty < x < \infty \qquad (2\text{-}7)$$

For normally distributed data sets, it is well known (Hogg and Craig, 1978) that the minimum variance unbiased estimates (*MVUEs*) of the mean, μ_1, and the variance, σ_1^2, are respectively given by the sample mean, $\bar{x}$, and sample variance, s_x^2. It is also well known that for normally distributed data sets, a *UCL* of the unknown mean, μ_1, base upon Student's t-distribution is optimal. It is observed via Monte Carlo simulation experiments (Singh and Singh (2003) Draft EPA Report) that for normally distributed data sets, the modified t-*UCL* and *UCL* based upon bootstrap t method also provide the exact 95% coverage to the population mean. For normally distributed data sets, the *UCLs* based upon these three methods are very similar.

Lognormal Distribution

If $Y = \log(X)$ is normally distributed with the mean, μ, and variance, σ^2, then X is said to be lognormally distributed with parameters μ and σ^2 and is denoted by LN(μ, σ^2). It should be noted that μ and σ^2 are not the mean and variance of the lognormal random variable, X, but they are the mean and variance of the log-transformed random variable, Y, whereas μ_1, and σ_1^2 represent the mean and variance of X. Some parameters of interest of a two-parameter lognormal distribution, LN(μ, σ^2), are given as follows:

$$\text{Mean} = \mu_1 = \exp(\mu + 0.5\sigma^2) \tag{2-8}$$

$$\text{Median} = M = \exp(\mu) \tag{2-9}$$

$$\text{Variance} = \sigma_1^2 = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1] \tag{2-10}$$

$$\text{Coefficient of Variation} = CV = \sigma_1 / \mu_1 = \sqrt{\exp(\sigma^2) - 1} \tag{2-11}$$

$$\text{Coefficient of Skewness} = CV^3 + 3CV \tag{2-12}$$

2.3.2.1 MLEs of the Parameters of a Lognormal Distribution

For lognormal distributions, note that $\bar{y}$ and s_y ($= \hat{\sigma}$) are the maximum likelihood estimators (*MLEs*) of μ and σ, respectively. The *MLE* of any function of the parameters μ and σ^2 is obtained by simply substituting these *MLEs* in place of the parameters (Hogg and Craig 1978). Therefore, replacing μ and σ by their *MLEs* in equations (2-8) through (2-12) will result in the *MLEs* (but biased) of the respective parameters of the lognormal distribution. The program ProUCL computes all of these *MLEs* for lognormally distributed data sets. These *MLEs* are printed on the Excel-type output spreadsheet generated by ProUCL.

2.3.2.2 Relationship between Skewness and Standard Deviation, σ

Note that for a lognormal distribution, the *CV* (given by equation (2-11) above) and the coefficient of skewness (given by equation (2-12)) depend only on σ. Therefore, in this Technical Guide and also in ProUCL, the standard deviation, σ (*Sd* of log-transformed variable, *Y*), or its *MLE*, s_y ($= \hat{\sigma}$), has been used as a measure of the skewness of lognormal and also of other skewed data sets with positive values. The larger is the *Sd*, the larger are the *CV* and the skewness. For example, for a lognormal distribution: with σ = *0.5*, the skewness = 1.75; with σ =1.0, the skewness = 6.185; with σ =*1.5*, the skewness = 33.468; and with σ = 2.0, the skewness = 414.36. Thus, the skewness of a lognormal distribution becomes

unreasonably large as σ starts approaching and exceeding 2.0. Note that for a gamma distribution, the skewness is a function of the shape parameter, *k.* As *k* decreases, the skewness increases.

It is observed (Singh, Singh, Engelhardt (1997) and Singh, Singh, and Iaci (2002b)) that for smaller sample sizes (such as smaller than 50), and for values of σ approaching 2.0 (and skewness approaching 414), the use of the H-statistic-based *UCL* results in impractical and unacceptably large values. For simplicity, the various levels of skewness of a positive data set as used in ProUCL and in this Technical Guide are summarized as follows:

Table 2-1. Skewness as a Function of σ (or its *MLE*, $s_y = \hat{\sigma}$), *sd* of log(*X*)

Standard Deviation	Skewness
$\sigma < 0.5$	Symmetric to mild skewness
$0.5 \le \sigma < 1.0$	Mild skewness to moderate skewness
$1.0 \le \sigma < 1.5$	Moderate skewness to high skewness
$1.5 \le \sigma < 2.0$	High skewness
$2.0 \le \sigma < 3.0$	Extremely high skewness
$\sigma \le 3.0$	Provides poor coverage

These values of σ (or its estimate, *Sd* of log-transformed data) are used to define the skewness levels of lognormal and skewed non-discernable data distributions, as used in Tables 2-2 and 2-3.

2.3.2.3 MLEs of the Quantiles of a Lognormal Distribution

For highly skewed (e.g., σ exceeding 1.5), lognormally distributed populations, the population mean, μ_1, often exceeds the higher quantiles (e.g., 80%, 90%, 95%) of the distribution. Therefore, the computation of these quantiles is also of interest. This is especially true when one may want to use the *MLEs* of the higher order quantiles (e.g., 95%, 97.5%, etc.) as an estimate of the EPC term. The formulae to compute these quantiles are briefly described here.

The p^{th} quantile (or 100 p^{th} percentile), x_p, of the distribution of a random variable, *X*, is defined by the probability statement, $P(X \le x_p) = p$. If z_p is the p^{th} quantile of the standard normal random variable, *Z*, with $P(Z \le z_p) = p$, then the p^{th} quantile of a lognormal distribution is given by $x_p = \exp(\mu + z_p\sigma)$. Thus the *MLE* of the p^{th} quantile is given by

$$\hat{x}_p = \exp(\hat{\mu} + z_p\hat{\sigma}) \qquad (2\text{-}13)$$

For example, on the average, 95% of the observations from a lognormal LN(μ, σ^2) distribution would lie below $\exp(\mu + 1.65\sigma)$. The 0.5^{th} quantile of the standard normal distribution is $z_{0.5} = 0$, and the 0.5^{th} quantile (or median) of a lognormal distribution is $M = \exp(\mu)$, which is obviously smaller than the mean, μ_1, as given by equation (1-8). Also, note that the mean, μ_1, is greater than x_p if and only if $\sigma > 2z_p$. For example, when $p = 0.80$, $z_p = 0.845$, μ_1 exceeds $x_{0.80}$, the 80^{th} percentile if and only if $\sigma > 1.69$, and, similarly, the mean, μ_1, will exceed the 95^{th} percentile if and only if $\sigma > 3.29$. ProUCL computes the *MLEs* of the 50% (median), 90%, 95%, and 99% percentiles of lognormally distributed data sets. For lognormally distributed background data sets, a 95% or 99% percentile may be used as an estimate of the background threshold value; that is the background level contaminant concentration.

2.3.2.4 MVUEs of Parameters of a Lognormal Distribution

Even though the sample *AM,* $\bar{x}$, is an unbiased estimator of the population *AM*, μ_1, it does not have the minimum variance (*MV*). The *MV unbiased estimates (MVUEs)* of μ_1 and ${\sigma_1}^2$ of a lognormal distribution are given as follows:

$$\hat{\mu}_1 = \exp(\bar{y}) g_n(s_y^2/2) \tag{2-14}$$

$$\hat{\sigma}_1^2 = \exp(2\bar{y})[g_n(2s_y^2) - g_n((n-2)s_y^2/(n-1))] \tag{2-15}$$

The series expansion of the function $g_n(\mu)$ is given in Bradu and Mundlak (1970), and Aitchison and Brown (1976). Tabulations of this function are also provided by Gilbert (1987). Bradu and Mundlak (1970) give the *MVUE* of the variance of the estimate, $\hat{\mu}_1$,

$$\hat{\sigma}^2(\hat{\mu}_1) = \exp(2\bar{y})[(g_n(2s_y^2))^2 - g_n((n-2)s_y^2/(n-1))] \tag{2-16}$$

The square root of the variance given by equation (1-16) is called the standard error (*SE*) of the estimate, $\hat{\mu}_1$, given by equation (2-14). Similarly, a *MVUE* of the median of a lognormal distribution is given by

$$\hat{M} = \exp(\bar{y}) g_n[-s_y^2/(2(n-1))] \tag{2-17}$$

For a lognormally distributed data set, ProUCL also computes these *MVUEs* given by equations (2-14) through (2-17).

2.3.2 Estimation of the Parameters of a Gamma Distribution

Next, we consider the estimation of parameters of a gamma distribution. Since the estimation of gamma parameters is typically not included in standard statistical textbooks, this has been described in some detail in this Technical Guide. The population mean and variance of a two-parameter gamma distribution, G(k, θ), are functions of both parameters, k and θ. In order to estimate the mean, one has to obtain estimates of k and θ. The computation of the maximum likelihood estimate (*MLE*) of k is quite complex and requires the computation of Digamma and Trigamma functions. Several authors (Choi and Wette, 1969, Bowman and Shenton 1988, Johnson, Kotz, and Balakrishnan, 1994) have studied the estimation of the shape and scale parameters of a gamma distribution. The maximum likelihood estimation method to estimate the shape and scale parameters of a gamma distribution is described below.

Let $x_1, x_2, \ldots, x_n$ be a random sample (e.g., representing contaminant concentrations) of size n from a gamma distribution, G(k, θ), with unknown shape and scale parameters, k and θ, respectively. The log likelihood function (obtained using equation (2-3)) is given as follows:

$$LogL(x_1, x_2, \ldots, x_n; k, \theta) = -nk\log(\theta) - n\log\Gamma(k) + (k-1)\sum\log(x_i) - \sum x_i \big/ \theta \tag{2-18}$$

To find the MLEs of k and θ, we differentiate the log likelihood function as given in (1-18) with respect to k and θ, and set the derivatives to zero. This results in the following two equations:

$$Log(\hat{\theta}) + \frac{\Gamma'(\hat{k})}{\Gamma(\hat{k})} = \frac{1}{n}\sum \log(x_i) \text{ , and} \tag{2-19}$$

$$\hat{k}\hat{\theta} = \frac{1}{n}\sum x_i = \bar{x} \tag{2-20}$$

Solving equation (2-20) for $\hat{\theta}$, and substituting the result in (2-19), we get following equation:

$$\frac{\Gamma'(\hat{k})}{\Gamma(\hat{k})} - \log(\hat{k}) = \frac{1}{n}\sum \log(x_i) - \log\left(\frac{1}{n}\sum x_i\right) \tag{2-21}$$

There does not exist a closed form solution of equation (2-21). This equation needs to be solved numerically for $\hat{k}$, which requires the use of Digamma and Trigamma functions. This is quite easy to do using a personal computer. An estimate of *k* can be computed iteratively by using the Newton-Raphson (Faires and Burden 1993) method, leading to the following iterative equation:

$$\hat{k}_l = \hat{k}_{l-1} - \frac{\log(\hat{k}_{l-1}) - \Psi(\hat{k}_{l-1}) - M}{1/\hat{k}_{l-1} - \Psi'(\hat{k}_{l-1})} \tag{2-22}$$

The iterative process stops when $\hat{k}$ starts to converge. In practice, convergence is typically achieved in fewer than 10 iterations. In equation (2-22),

$$M = \log(\bar{x}) - \sum \log(x_i)/n,\ \Psi(k) = \frac{d}{dk}\left(\log\Gamma(k)\right), \text{ and } \Psi'(k) = \frac{d}{dk}\left(\Psi(k)\right)$$

Here $\Psi(k)$ is the Digamma function and $\Psi'(k)$ is the Trigamma function. In order to obtain the *MLEs* of *k* and θ, one needs to compute the Digamma and Trigamma functions. Good approximate values for these two functions (Choi and Wette 1969) can be obtained using the following approximations. For $k \geq 8$, these functions are approximated by

$$\Psi(k) \approx \log(k) - \left\{1 + \left[1 - (1/10 - 1/(21k^2))/k^2\right]/(6k)\right\}/(2k), \text{ and} \tag{2-23}$$

$$\Psi'(k) \approx \left\{1 + \left\{1 + \left[1 - (1/5 - 1/(7k^2))/k^2\right]/(3k)\right\}/(2k)\right\}/k \tag{2-24}$$

For $k < 8$, one can use the following recurrence relation to compute these functions:

$$\Psi(k) = \Psi(k+1) - 1/k \text{ , and} \tag{2-25}$$

$$\Psi'(k) = \Psi'(k+1) + 1/k^2 \tag{2-26}$$

In ProUCL, equations (2-23) - (2-26) have been used to estimate *k*. The iterative process requires an initial estimate of *k*. A good starting value for *k* in this iterative process is given by $k_0 = 1 / (2M)$. Thom (1968) suggested the following approximation as an estimate of *k*:

$$\hat{k} \approx \frac{1}{4M}\left(1+\sqrt{1+\frac{4}{3}M}\right) \tag{2-27}$$

Bowman and Shenton (1988) suggested using $\hat{k}$, as given by (2-27), to be a starting value of *k* for an iterative procedure, calculating $\hat{k}_l$ at the l^{th} iteration from the following formula:

$$\hat{k}_l = \frac{\hat{k}_{l-1}\{\log(\hat{k}_{l-1}) - \Psi(\hat{k}_{l-1})}{M} \tag{2-28}$$

Both equations (2-22) and (2-28) have been used to compute the *MLE* of *k*. It is observed that the estimate, $\hat{k}$, based upon Newton-Raphson method, as given by equation (2-22), is in close agreement with that obtained using equation (2-28) with Thom's approximation as an initial estimate. Choi and Wette (1969) further concluded that the *MLE* of *k*, $\hat{k}$, is biased high. A bias-corrected (Johnson, Kotz, and Balakrishnan 1994) estimate of *k* is given by:

$$\hat{k}^* = (n-3)\hat{k}/n + 2/(3n) \tag{2-29}$$

In (2-29), $\hat{k}$ is the *MLE* of *k* obtained using either (2-22) or (2-28). Substitution of equation (2-29) in equation (2-20) yields an estimate of the scale parameter, θ, given as follows:

$$\hat{\theta}^* = \bar{x}/\hat{k}^* \tag{2-30}$$

ProUCL computes simple *MLE*s of *k* and θ, and also bias-corrected estimates of *k* and θ. The bias-corrected estimate of k as given by (2-29) has been used in the computation of the *UCLs* (as given by equations (2-34) and (2-35)) of the mean of a gamma distribution.

2.4 Methods for Computing a UCL of the Unknown Population Mean

ProUCL computes a *(1 – α)100% UCL* of the population mean, μ_1, using the following 5 parametric and 10 nonparametric methods. Five of the 10 nonparametric methods are based upon the bootstrap method. Modified t and adjusted central limit theorem adjust for skewness for skewed data sets. However, it is noted that (Singh, Singh, and Iaci (2002b) and Singh and Singh (2003)) this adjustment is not adequate enough for moderately skewed to highly skewed data sets. For details, interested users are referred to graphical displays of coverage probability comparisons for normal, gamma, and lognormal distributions given in Singh and Singh (2003).

Parametric Methods

1. Student's t-statistic – assumes normality or approximate normality
2. Approximate gamma *UCL* – assumes gamma distribution of the data set
3. Adjusted gamma *UCL* – assumes gamma distribution of the data set

4. Land's H-Statistic – assumes lognormality

5. Chebyshev Theorem using the *MVUE* of the parameters of a lognormal distribution (denoted by Chebyshev (*MVUE*)) – assumes lognormality

Nonparametric Methods

1. Modified t-statistic – modified for skewed distributions

2. Central limit theorem (*CLT*) – to be used for large samples

3. Adjusted central limit theorem (adjusted-*CLT*) – adjusted for skewed distributions and to be used for large samples

4. Chebyshev Theorem using the sample arithmetic mean and *Sd* (denoted by Chebyshev (Mean, *Sd*))

5. Jackknife method – yields the same result as Student's t-statistic for the *UCL* of the population mean

6. Standard bootstrap

7. Percentile bootstrap

8. Bias-corrected accelerated (BCA) bootstrap

9. Bootstrap t

10. Hall's bootstrap

Even though it is well known that some of the methods (e.g., *CLT*, *UCL* based upon Jackknife method (same as Student's t-*UCL*), standard bootstrap and percentile bootstrap methods) do not perform well enough to provide the adequate coverage to the population mean of skewed distributions, these methods have been included in ProUCL to satisfy the curiosity of all users.

ProUCL can compute a *(1 – α)100% UCL* (except for the *H-UCL and adjusted gamma UCL*) of the mean for any confidence coefficient *(1 – α)* value lying in the interval [0.5, 1.0). For the computation of the *H-UCL*, only two confidence levels, namely, 0.90 and 0.95 are supported by ProUCL. For adjusted gamma *UCL*, three confidence levels, namely: 0.90, 0.95, and 0.99 are supported by ProUCL 4.0. An approximate gamma *UCL* can be computed for any level of significance in the interval [0.5,1). Whenever, ProUCL 4.0 cannot compute a UCL for a specified confidence coefficient (e.g., 0.99 for H-UCL), ProUCL 4.0 prints out "N/A." Based upon sample size, *n*, skewness, and data distribution, ProUCL 4.0 also makes recommendations on how to obtain an appropriate *95% UCL* of the unknown population mean, μ_1.

2.4.1 (1 – α)100% UCL of the Mean Based Upon Student's t-Statistic

The widely used well-known Student's t-statistic is given by,

$$t = \frac{\bar{x} - \mu_1}{s_x / \sqrt{n}} \tag{2-31}$$

where $\bar{x}$ and s_x are, respectively, the sample mean and sample standard deviation obtained using the raw data. If the data are a random sample from a normal population with mean, μ_1, and standard deviation, σ_1, then the distribution of this statistic is the familiar Student's t-distribution with (n -1) degrees of freedom (*df*). Let $t_{\alpha,n-1}$ be the upper α^{th} quantile of the Student's t-distribution with (n -1) *df*.

A *(1 – α)100% UCL* of the population mean, μ_1, is given by,

$$UCL = \bar{x} + t_{\alpha,n-1} s_x / \sqrt{n} \tag{2-32}$$

For a normally (when the skewness is about ~0) distributed population, equation (2-32) provides the best (optimal) way of computing a *UCL* of the mean. Equation (2-32) may also be used to compute a *UCL* of the mean based upon very mildly skewed (e.g., |skewness|<0.5) data sets, where the skewness is given by equation (2-43). It should be pointed out that even for mildly to moderately skewed data sets (e.g., when *σ, the Sd of log-transformed data,* starts approaching and exceeding 0.5), the *UCL* given by (2-32) might not provide the desired coverage (e.g., = 0.95) to the population mean. This is especially true when the sample size is smaller than 20-25 (Singh and Singh (2003)). The situation gets worse (coverage much smaller than 0.95) for higher values of the *Sd*, *σ*, or its *MLE*, s_y.

2.4.2 Computation of the UCL of the Mean of a Gamma, G (k, θ), Distribution

In the statistical literature, even though methods exist to compute a *UCL* of the mean of a gamma distribution (Grice and Bain 1980, Wong 1993), those methods have not become popular due to their computational complexity. Those approximate and adjusted methods depend upon the chi-square distribution and an estimate of the shape parameter, *k*. As seen above, computation of an *MLE* of *k* is quite involved, and this works as a deterrent to the use of a gamma distribution-based *UCL* of the mean. However, the computation of a gamma *UCL* currently should not be a problem due to easy availability of personal computers.

Given a random sample, $x_1, x_2, \ldots, x_n$, of size *n* from a gamma, G(k, θ), distribution, it can be shown that $2n\bar{x}/\theta$ follows a chi-square distribution, χ^2_{2nk}, with 2*nk* degrees of freedom (*df*). When the shape parameter, *k*, is known, a uniformly most powerful test of size of the null hypothesis, H_0: $\mu_1 \geq C_s$, against the alternative hypothesis, H_1: $\mu_1 < C_s$, is to reject H_0 if $\bar{x}/C_s < \chi^2_{2nk}(\alpha)/2nk$. The corresponding (1 – α) 100% uniformly most accurate *UCL* for the mean, μ_1, is then given by the probability statement.

$$P(2nk\bar{x}/\chi^2_{2nk}(\alpha) \geq \mu_1) = 1 - \alpha \tag{2-33}$$

where χ^2_υ denotes the cumulative percentage point of the chi-square distribution (e.g., α is the area in the left tail). That is, if *Y* follows χ^2_υ, then $P(Y \leq \chi^2_\upsilon(\alpha)) = \alpha$. In practice, *k* is not known and needs to be estimated from data. A reasonable method is to replace *k* by its bias-corrected estimate, $\hat{k}^*$, as given by equation (2-29). This yields the following approximate *(1 – α)100% UCL* of the mean, μ_1.

$$Approximate - UCL = 2n\hat{k}^{*}\bar{x}\Big/\chi^{2}_{2n\hat{k}^{*}}(\alpha) \quad (2\text{-}34)$$

It should be pointed out that the *UCL* given by equation (2-34) is an approximate *UCL* and there is no guarantee that the confidence level of (1 – α) will be achieved by this *UCL*. However, it does provide a way of computing a *UCL* of the mean of a gamma distribution. Simulation studies conducted in Singh, Singh, and Iaci (2002b) and in Singh and Singh (2003) suggest that an approximate gamma *UCL* thus obtained provides the specified coverage (95%) as the shape parameter, *k*, approaches 0.5. Therefore, when $k \geq 0.5$, one can always use the approximate *UCL* given by equation (2-34) to estimate the EPC term. This approximation is good even for smaller (e.g., $n = 5$) sample sizes as shown in Singh, Singh, Iaci (2002b), and in Singh and Singh (2003).

Grice and Bain (1980) computed an adjusted probability level, β (adjusted level of significance), which can be used in (2-34) to achieve the specified confidence level of (1 – α). For α = 0.05 (confidence coefficient of 0.95), α = 0.1, and α = 0.01, these probability levels are given below in Table 2-2 for some values of the sample size *n*. One can use interpolation to obtain an adjusted β for values of *n* not covered in the table. The adjusted *(1 – α)100% UCL* of the gamma mean, $\mu_1 = k\theta$, is given by the following equation.

$$Adjusted - UCL = 2n\hat{k}^{*}\bar{x}\Big/\chi^{2}_{2n\hat{k}^{*}}(\beta) \quad (2\text{-}35)$$

where β is given in Table 2-1 for α = 0.05, 0.1, and 0.01. Note that as the sample size, *n*, becomes large, the adjusted probability level, β, approaches the specified level of significance, α. Except for the computation of the *MLE* of *k*, equations (2-34) and (2-35) provide simple chi-square-distribution-based *UCLs* of the mean of a gamma distribution. It should also be noted that the *UCLs* as given by (2-34) and (2-35) only depend upon the estimate of the shape parameter, *k*, and are independent of the scale parameter, θ, and its ML estimate. Consequently, as expected, it is observed that coverage probabilities for the mean associated with these *UCLs* do not depend upon the values of the scale parameter, θ. It should also be noted that gamma *UCLs* do not depend upon the standard deviation of data which gets distorted by the presence of outliers. Thus, outliers will have reduced influence on the computation of the gamma distribution based upon *UCLs* of the mean, μ_1.

Table 2-2. Adjusted Level of Significance, β

n	α = 0.05 probability level, β	α = 0.1 probability level, β	α = 0.01 probability level, β
5	0.0086	0.0432	0.0000
10	0.0267	0.0724	0.0015
20	0.0380	0.0866	0.0046
40	0.0440	0.0934	0.0070
--	0.0500	0.1000	0.0100

2.4.3 (1 – α)100% UCL of the Mean Based Upon H-Statistic (H-UCL)

The one-sided *(1 – α)100% UCL* for the mean, μ_1, of a lognormal distribution as derived by Land (1971, 1975) is given as follows:

$$UCL = \exp\left(\bar{y} + 0.5s_y^2 + s_y H_{1-\alpha}\Big/\sqrt{n-1}\right) \quad (2\text{-}36)$$

Tables of H-statistic critical values can be found in Land (1975) and also in Gilbert (1987). Theoretically, when the population is lognormal, Land (1971) showed that the *UCL* given by equation (2-36) possesses optimal properties and is the uniformly most accurate unbiased confidence limit. However, it is noticed that, in practice, the H-statistic-based results can be quite disappointing and misleading, especially when the data set consists of outliers, or is a mixture from two or more distributions (Singh, Singh, and Engelhardt, 1997, 1999 and Singh, Singh, and Iaci, 2002b). Even a minor increase in the *Sd*, s_y, drastically inflates the *MVUE* of μ_1 and the associated *H-UCL*. The presence of low as well as high data values increases the *Sd*, s_y, which in turn inflates the *H-UCL*. Furthermore, it is observed (Singh, Singh, Engelhardt, and Nocerino 2002a) that for samples of sizes smaller than 15-25, and for values of σ approaching 1.0 and higher (for moderately skewed to highly skewed data sets), the use of H-statistic-based *UCL* results in impractical and unacceptably large *UCL* values.

In practice, many data sets follow a lognormal as well as gamma model. However, the population mean based upon a lognormal model can be significantly greater (often unrealistically large) than the population mean based upon a gamma model. In order to provide the specified 95% coverage for an inflated mean based upon a lognormal model, the resulting *UCL* based upon H-statistic also yield impractical *UCL* values. The use of a gamma model results in practical estimates (e.g., *UCL*) of the population mean. Therefore, for positively skewed data sets, it is recommended to test for a gamma model first. If data follow a gamma distribution, then the *UCL* of the mean should be computed using a gamma distribution. The gamma distribution is better suited to model positively skewed environmental data sets.

2.4.4 (1 – α)100% UCL of the Mean Based Upon Modified t-Statistic for Asymmetrical Populations

Chen (1995), Johnson (1978), Kleijnen, Kloppenburg, and Meeuwsen (1986), and Sutton (1993) suggested the use of the modified t-statistic for testing the mean of a positively skewed distribution (including the lognormal distribution). The *(1 – α)100% UCL* of the mean thus obtained is given by

$$UCL = \bar{x} + \hat{\mu}_3 \big/ (6 s_x^2 n) + t_{\alpha,n-1} s_x \big/ \sqrt{n} \qquad (2\text{-}37)$$

Where $\hat{\mu}_3$, an unbiased moment estimate (Kleijnen, Kloppenburg, and Meeuwsen 1986) of the third central moment, is given as follows,

$$\hat{\mu}_3 = n \sum_{i=1}^{n} (x_i - \bar{x})^3 \Big/ (n-1)(n-2) \qquad (2\text{-}38)$$

It should be pointed out that this modification for a skewed distribution does not perform well even for mildly to moderately skewed data sets (e.g., when σ starts approaching and exceeding 0.75). Specifically, it is observed that the *UCL* given by equation (2-37) may not provide the desired coverage of the population mean, μ_1, when σ starts approaching and exceeding 0.75 (Singh, Singh, and Iaci, 2002b). This is especially true when the sample size is smaller than 20-25. This small sample size requirement increases as σ increases. For example, when σ starts approaching and exceeding 1.5, the *UCL* given by equation (2-37) does not provide the specified coverage (e.g., 95%), even for samples as large as 100. Since this method does not require any distributional assumptions, it is a nonparametric method.

2.4.5 (1 – α)100% UCL of the Mean Based Upon the Central Limit Theorem

The central limit theorem *(CLT)* states that the asymptotic distribution, as *n* approaches infinity, of the sample mean, $\overline{x}_n$, is normally distributed with mean, μ_1, and variance, σ_1^2/n. More precisely, the sequence of random variables given by

$$z_n = \frac{\overline{x}_n - \mu_1}{\sigma / \sqrt{n}} \tag{2-39}$$

has a standard normal limiting distribution. In practice, for large sample sizes, *n*, the sample mean, $\overline{x}$, has an approximate normal distribution irrespective of the underlying distribution function. Since the *CLT* method requires no distributional assumptions, this is a nonparametric method.

As noted by Hogg and Craig (1978), if σ_1 is replaced by the sample standard deviation, s_x, the normal approximation for large *n* is still valid. This leads to the following approximate large sample nonparametric *(1 – α)100% UCL* of the mean,

$$UCL = \overline{x} + z_\alpha s_x / \sqrt{n} \tag{2-40}$$

An often cited and used rule of thumb for a sample size associated with the CLT method is $n \geq 30$. However, this may not be adequate enough if the population is skewed, specifically when σ (*Sd* of log-transformed variable) starts exceeding 0.5 (Singh, Singh, Iaci, 2002b). In practice, for skewed data sets, even a sample as large as 100 is not large enough to provide adequate coverage to the mean of skewed populations (even for mildly skewed populations). A refinement of the CLT approach, which makes an adjustment for skewness by Chen (1995), is given as follows.

2.4.6 (1 – α)100% UCL of the Mean Based Upon the Adjusted Central Limit Theorem (Adjusted-CLT)

The "*adjusted-CLT" UCL* is obtained if the standard normal quantile, z_α, in the upper limit of equation (2-40) is replaced by (Chen, 1995)

$$z_{\alpha,adj} = z_\alpha + \frac{\hat{k}_3}{6\sqrt{n}}(1 + 2z_\alpha^2) \tag{2-41}$$

Thus, the adjusted (1 – α)100% *UCL* for the mean, μ_1, is given by

$$UCL = \overline{x} + \left[z_\alpha + \hat{k}_3(1 + 2z_\alpha^2)/(6\sqrt{n}) \right] s_x / \sqrt{n} \tag{2-42}$$

Here $\hat{k}_3$, the coefficient of skewness (raw data), is given by

$$\text{Skewness (raw data) } \hat{k}_3 = \hat{\mu}_3 / s_x^3 \tag{2-43}$$

where $\hat{\mu}_3$, an unbiased estimate of the third moment, is given by equation (2-38). This is another large sample approximation for the *UCL* of the mean of skewed distributions. This is a nonparametric method, as it does not depend upon any of the distributional assumptions.

As with the modified t-*UCL*, it is observed that this adjusted-*CLT UCL* does not provide adequate coverage to the population mean when the population is skewed, specifically when σ starts approaching and exceeding 0.75 (Singh, Singh, and Iaci, 2002b, and Singh and Singh, 2003). This is especially true when the sample size is smaller than 20-25. This small sample size requirement increases as σ increases. For example, when σ starts approaching and exceeding 1.5, the *UCL* given by equation (2-42) does not provide the specified coverage (e.g., 95%), even for samples as large as 100. It is noted that *UCL* as given by (2-42) does not provide adequate coverage to the mean of a gamma distribution, especially when $k \leq$ 1.0 and the sample size is small.

Thus, the *UCLs* based upon these skewness adjusted methods, such as the Johnson's modified t and Chen's adjusted-*CLT*, do not provide the specified coverage to the population mean for mildly to moderately skewed (e.g., σ in (0.5, 1.0)) data sets, even for samples as large as 100 (Singh, Singh, and Iaci, 2002b). The coverage of the population mean provided by these *UCLs* becomes worse (much smaller than the specified coverage) for highly skewed data sets.

2.4.7 Chebyshev (1 – α)100% UCL of the Mean Using Sample Mean and Sample sd

The Chebyshev inequality can be used to obtain a reasonably conservative but stable estimate of the *UCL* of the mean, μ_1. The two-sided Chebyshev theorem (Hogg and Craig, 1978) states that given a random variable, *X*, with finite mean and standard deviation, μ_1 and σ_1, we have

$$P(-k\sigma_1 \leq x - \mu_1 \leq k\sigma_1) \geq 1 - 1/k^2 \qquad (2\text{-}44)$$

This result can be applied on the sample mean, $\bar{x}$ (with mean, μ_1 and variance, σ_1^2/n), to obtain a conservative *UCL* for the population mean, μ_1. For example, if the right side of equation (2-44) is equated to 0.95, then $k = 4.47$, and $UCL = \bar{x} + 4.47\sigma_1/\sqrt{n}$ is a conservative 95% upper confidence limit for the population mean, μ_1. Of course, this would require the user to know the value of σ_1. The obvious modification would be to replace σ_1 with the sample standard deviation, s_x, but since this is estimated from data, the result is no longer guaranteed to be conservative. In general, the following equation can be used to obtain a *(1 – α)100% UCL* of the population mean, μ_1:

$$UCL = \bar{x} + \sqrt{(1/\alpha)}s_x/\sqrt{n} \qquad (2\text{-}45)$$

A slight refinement of equation (2-45) is given (suggested by S. Ferson) as follows,

$$UCL = \bar{x} + \sqrt{((1/\alpha) - 1)}s_x/\sqrt{n} \qquad (2\text{-}46)$$

ProUCL computes the Chebyshev *(1 – α)100% UCL* of the population mean using equation (2-46). This *UCL* is labeled as *Chebyshev (Mean, Sd)* on the output sheets generated by ProUCL. Since this Chebyshev method requires no distributional assumptions about the data set under study, this is a nonparametric method. This *UCL* may be used as an estimate of the upper confidence limit of the

population mean, μ_1, when the data are not normal, lognormal, or gamma distributed, especially when *Sd*, σ (or its estimate, s_y), starts approaching and exceeding 1.5.

2.4.8 Chebyshev (1 – α)100% UCL of the Mean of a Lognormal Population Using the MVUE of the Mean and its Standard Error

ProUCL uses equation (2-44) on the *MVUEs* of the lognormal mean and *Sd* to compute a *UCL* (denoted by *(1 – α)100% Chebyshev* (*MVUE)*) of the population mean of a lognormal population. In general, if μ_1 is an unknown mean, $\hat{\mu}_1$ is an estimate, and $\hat{\sigma}_1(\hat{\mu}_1)$ is an estimate of the standard error of $\hat{\mu}_1$, then the following equation,

$$UCL = \hat{\mu}_1 + \sqrt{((1/\alpha)-1)}\hat{\sigma}_1(\hat{\mu}_1) \quad (2\text{-}47)$$

yields an approximate *(1 – α)100% UCL* for μ_1, which should tend to be conservative, but this is not assured. For example, for a lognormally distributed data set, a 95% (with α = 0.05) Chebyshev *(MVUE) UCL* of the mean can be obtained using the following equation,

$$UCL = \hat{\mu}_1 + 4.359\hat{\sigma}_1(\hat{\mu}_1) \quad (2\text{-}48)$$

Here $\hat{\mu}_1$ and $\hat{\sigma}_1(\hat{\mu}_1)$ are given by equations (2-14) and (2-16), respectively. Thus, for lognormally distributed data sets, ProUCL also uses equation (2-48) to compute a *(1 – α)100% Chebyshev (MVUE) UCL* of the mean. It should be noted that for lognormally distributed data sets, some recommendations to compute a 95% *UCL* of the population mean are summarized later in this chapter. It is recommended that for skewed data sets, one should always perform gamma goodness-of-fit (GOF) test. Many times, a skewed data set can be modeled both by a lognormal distribution as well as a gamma distribution. However, since, the use of a lognormal distribution often yields inflated and unstable upper limits including *UCLs* (Singh, Singh, and Engelhardt, 1997) and *UPLs* (Gibbons, 1994), it is suggested that if a data set follows a gamma distribution (even when data may also be lognormally distributed), then the *UCL* of mean, μ_1 (and other upper limits) should be computed using a gamma distribution. This is especially true when the data are highly skewed with *sd* of log-transformed data exceeding 1.5, 2.0, and the sample size is small such as < 50, < 70, < 100.

On the other hand, it is also noticed that the use of a lognormal distribution based H-UCL (based upon Land's H-statistic) often yields a UCL that is lower than the sample mean. This is especially true for mildly skewed to moderately skewed data sets of larger sizes (e.g., >50, 100). Some examples illustrating this issue are given in Chapter 3 of the revised background document for CERCLA sites (EPA, 2002). Therefore, it is suggested to avoid the use of a lognormal distribution to model environmental data sets.

From the Monte-Carlo results discussed in Singh, Singh, and Iaci (2002b), and in Singh and Singh (2003), it is observed that for highly skewed gamma distributed data sets (with $k < 0.5$), the coverage provided by the Chebyshev 95% *UCL* (given by (2-46)) is smaller than the specified coverage of 0.95. This is especially true when the sample size is smaller than 10-20. As expected, for larger samples sizes, the coverage provided by the 95% Chebyshev *UCL* is at least 95%. For larger samples, the Chebyshev 95% *UCL* will result in a higher (but stable) *UCL* of the mean of positively skewed gamma distributions.

It is observed (Singh and Singh 2003) that for moderately skewed to highly skewed lognormally distributed data sets (e.g., with σ exceeding 1), 95% Chebyshev *MVUE UCL* does not provide the specified coverage to the population mean. This is true when the sample size is less than 10-50. The

details and graphical displays can be found in Singh and Singh (2003). For highly skewed (e.g., $\sigma > 2$), lognormal data sets of sizes, *n* less than 50-70, the *H-UCL* results in unstable (impractical values which are orders of magnitude higher than other *UCLs*) unjustifiably large *UCL* values (Singh *et al.* 2002a). For such highly skewed lognormally distributed data sets of sizes less than 50-70, one may want to use 97.5% or 99% Chebyshev *MVUE UCL* of the mean as an estimate of the EPC term (Singh and Singh 2003). It should also be noted that for skewed data sets, the coverage provided by a 95% *UCL* based upon Chebyshev inequality is higher than those based upon the percentile bootstrap method or the BCA bootstrap method. Thus for skewed data sets, the Chebyshev inequality-based 95% *UCL* of the mean (samples of all sizes from both lognormal and gamma distributions) performs better than the 95% *UCL* based upon the BCA bootstrap method. Also, when data are lognormally distributed, the coverage provided by Chebyshev *MVUE UCL* (Singh and Singh 2003) is better than the one based upon Hall's bootstrap or bootstrap t method. This is especially true when the sample size starts exceeding 10-15. However, for highly skewed data sets of sizes less than 10-15, it is noted that Hall's bootstrap method provides slightly better coverage than the Chebyshev *MVUE UCL* method. Just as for the gamma distribution, it is observed that for lognormally distributed data sets, the coverage provided by Hall's and bootstrap t methods do not increase much with the sample size.

2.4.9 (1 – α)100% UCL of the Mean Using the Jackknife and Bootstrap Methods

Bootstrap and jackknife methods as discussed by Efron (1982) are nonparametric statistical resampling techniques which can be used to reduce the bias of point estimates and construct approximate confidence intervals for parameters, such as the population mean. These two methods require no assumptions regarding the statistical distribution (e.g., normal, lognormal, or gamma) of the underlying population, and can be applied to a variety of situations no matter how complicated. There exists in the literature of statistics an extensive array of different bootstrap methods for constructing confidence intervals for the population mean, μ_1. In the ProUCL 4.0 software package, five bootstrap methods have been incorporated:

1. The standard bootstrap method,
2. Bootstrap t method (Efron, 1982 and Hall, 1988),
3. Hall's bootstrap method (Hall, 1992 and Manly, 1997),
4. Simple bootstrap percentile method (Manly, 1997), and
5. Bias-corrected accelerated (BCA) percentile bootstrap method (Efron and Tibshirani, 1993 and Many, 1997).

Let $x_1, x_2, \ldots, x_n$ be a random sample of size *n* from a population with an unknown parameter, θ (*e.g.*, $\theta = \mu_1$), and let $\hat{\theta}$ be an estimate of θ, which is a function of all *n* observations. For example, the parameter, θ, could be the population mean and a reasonable choice for the estimate, $\hat{\theta}$, might be the sample mean, $\bar{x}$. Another choice for $\hat{\theta}$ is the *MVUE* of the mean of a lognormal population, especially when dealing with lognormal data sets.

2.4.9.1 (1 – α)100% UCL of the Mean Based Upon the Jackknife Method

In the jackknife approach, *n* estimates of θ are computed by deleting one observation at a time (Dudewicz and Misra 1988). Specifically, for each index, *i*, denote by $\hat{\theta}_{(i)}$, the estimate of θ (computed similarly as $\hat{\theta}$) when the i^{th} observation is omitted from the original sample of size *n*, and let the arithmetic mean of these estimates be given by:

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \qquad (2\text{-}49)$$

A quantity known as the i^{th} "pseudo-value" is defined by

$$J_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)} \qquad (2\text{-}50)$$

The jackknife estimator of θ is given by the following equation.

$$J(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} J_i = n\hat{\theta} - (n-1)\tilde{\theta} \qquad (2\text{-}51)$$

If the original estimate $\hat{\theta}$ is biased, then under certain conditions, part of the bias is removed by the jackknife method, and an estimate of the *SE* of the jackknife estimate, $J(\hat{\theta})$, is given by

$$\hat{\sigma}_{J(\hat{\theta})} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(J_i - J\left(\hat{\theta}\right)\right)^2} \qquad (2\text{-}52)$$

Next, consider the t-type statistic given by

$$t = \frac{J(\hat{\theta}) - \theta}{\hat{\sigma}_{J(\hat{\theta})}} \qquad (2\text{-}53)$$

The t-type statistic given above has an approximate Student's t- distribution with $n - 1$ degrees of freedom, which can be used to derive the following approximate (1–α) 100% UCL for θ,

$$UCL = J(\hat{\theta}) + t_{\alpha,n-1}\hat{\sigma}_{J(\hat{\theta})} \qquad (2\text{-}54)$$

If the sample size, *n*, is large, then the upper α^{th} *t*-quantile in the above equation can be replaced with the corresponding upper α^{th} standard normal quantile, z_α. Observe, also, that when $\hat{\theta}$ is the sample mean, $\bar{x}$, then the jackknife estimate is also the sample mean, $J(\bar{x}) = \bar{x}$, and the estimate of the standard error given by equation (2-52) simplifies to $s_x/n^{1/2}$, and the *UCL* in equation (2-54) reduces to the familiar *t*-statistic based *UCL* given by equation (2-32). ProUCL uses the jackknife estimate as the sample mean, that yields $J(\bar{x}) = \bar{x}$, which in turn translates equation (2-54) to Student's t- *UCL* given by equation (2-

32). This method has been included in ProUCL to satisfy the curiosity of those users who do not recognize that this jackknife method (with sample mean as the estimator) yields a *UCL* of the population mean identical to the *UCL* based upon the Student's t- statistic as given by equation (2-32).

Note: *It is well known that the Jackknife method (with sample mean as an estimator) and Student's t-method yield identical UCL values. However, a typical user may be unaware of this fact, and some researchers may want to see these issues described and discussed at one place. It is also noted that it has been suggested that a 95% UCL based upon the Jackknife method on the full data set obtained using robust ROS may provide adequate coverage (e.g., Shumway, Kayhanian, and Azari (2002)) to the population mean of skewed distributions, which of course is not true. It is well known (Singh, Singh, and Nocerino, 2003) that Student's t-UCL (and therefore, Jackknife UCL) fails to provide adequate coverage to the population mean of moderate to highly skewed distributions.*

2.4.9.2 (1 – α)100% UCL of the Mean Based Upon the Standard Bootstrap Method

In bootstrap resampling methods, repeated samples of size *n* are drawn with replacement from a given set of observations. The process is repeated a large number of times (e.g., 2000 times), and each time an estimate, $\hat{\theta}_i$, of θ is computed. The estimates thus obtained are used to compute an estimate of the *SE* of $\hat{\theta}$. A description of the bootstrap method, illustrated by application to the population mean, μ_1, and the sample mean, $\bar{x}$, is given as follows.

Step 1. Let (x_{i1}, x_{i2}, ... , x_{in}) represent the i^{th} sample of size *n* with replacement from the original data set, (x_1, x_2, ..., x_n). Then compute the sample mean and denote it by $\bar{x}_i$.

Step 2. Repeat Step 1 independently *N* times (e.g., 1000-2000), each time calculating a new estimate. Denote these estimates (KM means, RMLE means) by $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N$. The bootstrap estimate of the population mean is the arithmetic mean, $\bar{x}_B$, of the *N* estimates $\bar{x}_i : i := 1, 2, \ldots, N$. The bootstrap estimate of the standard error of the estimate, $\bar{x}$, is given by:

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{x}_i - \bar{x}_B)^2} \qquad (2\text{-}55)$$

If some parameter, θ (say, the population median), other than the mean is of concern with an associated estimate (e.g., the sample median), then the same steps described above could be applied with the parameter and its estimates used in place of μ_1 and $\bar{x}$. Specifically, the estimate, $\hat{\theta}_i$, would be computed, instead of $\bar{x}_i$, for each of the *N* bootstrap samples. The general bootstrap estimate, denoted by $\bar{\theta}_B$, is the arithmetic mean of the *N* estimates. The difference, $\bar{\theta}_B - \hat{\theta}$, provides an estimate of the bias of the estimate, $\hat{\theta}$, and an estimate of the *SE* of $\hat{\theta}$ is given by

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\hat{\theta}_i - \bar{\theta}_B)^2} \qquad (2\text{-}56)$$

A (1–α)100% standard bootstrap UCL for θ is given by

$$UCL = \hat{\theta} + z_{\alpha}\hat{\sigma}_{B} \quad (2\text{-}57)$$

ProUCL computes the standard bootstrap *UCL* by using the population *AM* and sample *AM*, respectively given by μ_1 and $\bar{x}$. It is observed that the *UCL* obtained using the standard bootstrap method is quite similar to the *UCL* obtained using the Student's t-statistic as given by equation (2-32), and, as such, does not adequately adjust for skewness. For skewed data sets, the coverage provided by standard bootstrap *UCL* is much lower than the specified coverage.

2.4.9.3 (1 – α)100% UCL of the Mean Based Upon the Simple Percentile Bootstrap Method

Bootstrap resampling of the original data set is used to generate the bootstrap distribution of the unknown population mean (Manly 1997). In this method, $\bar{x}_i$, the sample mean, is computed from the i^{th} resampling (i =1,2,..., N) of the original data. These $\bar{x}_i$;, i:=1,2,...,N are arranged in ascending order as $\bar{x}_{(1)} \le \bar{x}_{(2)} \le \cdots \le \bar{x}_{(N)}$. The *(1 – α)100% UCL* of the population mean, μ_1, is given by the value that exceeds the (1 – α)*100% of the generated mean values. The 95% *UCL* of the mean is the 95th percentile of the generated means and is given by:

$$95\%\ Percentile - UCL = 95^{th}\%\ \bar{x}_i;\ i{:}\ =1, 2, ..., N \quad (2\text{-}58)$$

For example, when N = 1000, a simple bootstrap 95% percentile-*UCL* is given by the 950th ordered mean value given by $\bar{x}_{(950)}$.

Singh and Singh (2003) observed that for skewed data sets, the coverage provided by this simple percentile bootstrap method is much lower than the coverage provided by the bootstrap t and Hall's bootstrap methods. It is observed that for skewed (lognormal and gamma) data sets, the BCA bootstrap method performs slightly better (in terms of coverage probability) than the simple percentile method.

2.4.9.4 (1 – α)100% UCL of the Mean Based Upon the Bias-Corrected Accelerated (BCA) Percentile Bootstrap Method

The BCA bootstrap method is also a percentile bootstrap method adjusts for bias in the estimate (Efron and Tibshirani 1993 and Manly 1997). The performance of this method for skewed distributions (e.g., lognormal and gamma) is not well studied. It was conjectured that the BCA method would perform better than the various other methods. Singh and Singh (2003) investigated and compared its performance (in terms of coverage probabilities) with parametric methods and other bootstrap methods. For skewed data sets, this method does represent a slight improvement (in terms of coverage probability) over the simple percentile method. However, this improvement is not adequate enough and yields *UCLs* with a coverage probability much lower than the specified coverage of 0.95. The BCA upper confidence limit of intended (1 – α)100% coverage is given by the following equation:

$$BCA - UCL = \bar{x}^{(\alpha_2)} \quad (2\text{-}59)$$

Here $\bar{x}^{(\alpha_2)}$ is the $\alpha_2 100^{th}$ percentile of the distribution of the $\bar{x}_i$; i: = *1, 2, …, N*. For example, when N = 2000, $\bar{x}^{(\alpha_2)}$ = $(\alpha_2 N)^{th}$ ordered statistic of $\bar{x}_i$; i: = *1, 2, …, N* given by $\bar{x}_{(\alpha_2 N)}$.

Here α_2 is given by the following probability statement:

$$\alpha_2 = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{\alpha}(\hat{z}_0 + z^{(1-\alpha)})}\right] \tag{2-60}$$

Here $\Phi(.)$ is the standard normal cumulative distribution function and $z^{(1-\alpha)}$ is the $100(1-\alpha)^{th}$ percentile of a standard normal distribution. For example, $z^{(0.95)} = 1.645$, and $\Phi(1.645) = 0.95$. Also in the equation (2-60), $\hat{z}_0$ (bias correction) and $\hat{\alpha}$ (acceleration factor) are given by

$$\hat{z}_0 = \Phi^{-1}\left[\frac{\#(\bar{x}_i < \bar{x})}{N}\right] \tag{2-61}$$

Here $\Phi^{-1}(.)$ is the inverse function of a standard normal cumulative distribution function, e.g., $\Phi^{-1}(0.95) = 1.645$. $\hat{\alpha}$ is the acceleration factor and is given by the following equation.

$$\hat{\alpha} = \frac{\sum(\bar{x} - \bar{x}_{-i})^3}{6[\sum(\bar{x} - \bar{x}_{-i})^2]^{1.5}} \tag{2-62}$$

Here the summation in (2-62) is being carried from $i = 1$ to $i = n$, the sample size. $\bar{x}$ is the sample mean based upon all n observation and $\bar{x}_{-i}$ is the mean of (n-1) observations without the i^{th} observation, i: = 1, 2, …, n.

Singh and Singh (2003) observed that for skewed data sets (e.g., gamma and lognormal), the coverage provided by this BCA percentile method is much lower than the coverage provided by the bootstrap t and Hall's bootstrap methods. This is especially true when the sample size is small. The BCA method does provide an improvement over the simple percentile method and the standard bootstrap method. However, bootstrap t and Hall's bootstrap methods perform better (in terms of coverage probabilities) than the BCA method. For skewed data sets, the BCA method also performs better than the modified t-*UCL*. For gamma distributions, the coverage provided by BCA 95% *UCL* approaches 0.95 as the sample size increases. For lognormal distributions, the coverage provided by the BCA 95% *UCL* is much lower than the specified coverage of 0.95.

2.4.9.5 (1 – α)100% UCL of the Mean Based Upon the Bootstrap t Method

Another variation of the bootstrap method, called the "bootstrap t" by Efron (1982), is a nonparametric method that uses the bootstrap methodology to estimate quantiles of the pivotal quantity, t statistic, given by equation (2-31). Rather than using the quantiles of the familiar Student's t-statistic, Hall (1988) proposed to compute estimates of the quantiles of the statistic given by equation (2-31) directly from the data.

Specifically, in Steps 1 and 2 described above in Section 2.4.9.2, if $\bar{x}$ is the sample mean computed from the original data, and $\bar{x}_i$ and $s_{x,I}$ are the sample mean and sample standard deviation computed from the i^{th} resampling of the original data, the N quantities $t_i = \sqrt{n}[(\bar{x}_i - \bar{x})/s_{x,i}]$ are computed and sorted, yielding ordered quantities, $t(1) \le t(2) \le \ldots \le t(N)$. The estimate of the lower α^{th} quantile of the pivotal

quantity in equation (2-31) is $t\alpha_{,B} = t(\alpha_N)$. For example, if N = 1000 bootstrap samples are generated, then the 50th ordered value, $t_{(50)}$, would be the bootstrap estimate of the lower 0.05^{th} quantile of the pivotal quantity in equation (2-31). Then a (1–α)100% *UCL* of the mean based upon the bootstrap t-method is given as follows.

$$\text{UCL} = \bar{x} - t_{(\alpha N)} \frac{s_x}{\sqrt{n}} \qquad (2\text{-}63)$$

Note the " – " sign in equation (2-63). ProUCL computes the bootstrap t *UCL* based upon the quantiles obtained using the sample mean, $\bar{x}$. It is observed that the *UCL* based upon the bootstrap t method is more conservative than the other *UCLs* obtained using the Student's t, modified-t, adjusted-*CLT*, and the standard bootstrap methods. This is especially true for skewed data sets. This method seems to adjust for skewness to some extent.

It is observed that for skewed data sets (e.g., gamma, lognormal), the 95% UCL based upon the bootstrap t method performs better than the 95% *UCLs* based upon the simple percentile and the BCA percentile methods (Singh and Singh (2003)). For highly skewed (k < 0.1 or σ > 2.5-3.0) data sets of small sizes (e.g., n < 10), the bootstrap t method performs better than other (adjusted gamma *UCL*, or Chebyshev inequality *UCL*) *UCL* computation methods. It is noted that for the gamma distribution, the performances (coverages provided by the respective *UCLs*) of the bootstrap t and Hall's bootstrap methods are very similar. It is also noted that for larger samples, these two methods (bootstrap t and Hall's bootstrap) approximately provide the specified 95% coverage to the mean, $k\theta$, of the gamma distribution. For gamma distributed data sets, the coverage provided by a bootstrap t (and Hall's bootstrap *UCL*) 95% *UCL* approaches 95% as sample size increases for all values of k considered (k = 0.05-5.0) in Singh and Singh (2003). However, it is noted that the coverage provided by these two bootstrap methods is slightly lower than 0.95 for samples of smaller sizes.

For lognormally distributed data sets, the coverage provided by the bootstrap t 95% *UCL* is a little bit lower than the coverage provided by the 95% *UCL* based upon Hall's bootstrap method. However, it should be noted that for lognormally distributed data sets, for samples of all sizes, the coverage provided by these two methods (bootstrap t and Hall's bootstrap) is significantly lower than the specified 0.95 coverage. This is especially true for moderately skewed to highly skewed (e.g., σ >1.0) lognormally distributed data sets. It should be pointed out that the bootstrap t and Hall's bootstrap methods sometimes result in unstable, erratic, and unreasonably inflated *UCL* values especially in the presence of outliers (Efron and Tibshirani, 1993). Therefore, these two methods should be used with caution. If this is the case, and these two methods result in erratic and inflated *UCL* values, then an appropriate Chebyshev inequality based *UCL* may be used to estimate the EPC term for nonparametric skewed data sets.

2.4.9.6 (1 – α)100% UCL of the Mean Based Upon Hall's Bootstrap Method

Hall (1992) proposed a bootstrap method that adjusts for bias as well as skewness. This method has been included in UCL guidance document (EPA 2002a). For highly skewed data sets (e.g., LN (5,4)), it performs slightly better (higher coverage) than the bootstrap t method. In this method, $\bar{x}_i$ and $s_{x,i}$, and $\hat{k}_{3i}$, the sample mean, the sample standard deviation, and the sample skewness, respectively, are computed from the i^{th} resampling (i = 1, 2,..., N) of the original data. Let $\bar{x}$ be the sample mean, s_x be the sample standard deviation, and $\hat{k}_3$ be the sample skewness (as given by equation (2-43)) computed from the original data. The quantities, W_i and Q_i, given as follows are computed for each of the N bootstrap samples:

$$W_i = (\bar{x}_i - \bar{x})/s_{x,i}\text{ , and } Q_i(W_i) = W_i + \hat{k}_{3i}W_i^2/3 + \hat{k}_{3i}^2 W_i^3/27 + \hat{k}_{3i}/(6n)$$

The quantities, $Q_i(W_i)$, given above are arranged in ascending order. For a specified (1 – α) confidence coefficient, compute the $(\alpha N)^{th}$ ordered value, q_α, of the quantities, $Q_i(W_i)$. Next, compute $W(q_\alpha)$ using the inverse function, which is given as follows:

$$W(q_\alpha) = 3\left(\left(1 + \hat{k}_3(q_\alpha - \hat{k}_3/(6n))\right)^{1/3} - 1\right)/\hat{k}_3 \qquad (2\text{-}64)$$

In equation (2-64), $\hat{k}_3$ is computed using equation (2-43). Finally, the *(1 – α)100% UCL* of the population mean based upon Hall's bootstrap method (Manly 1997) is given as follows:

$$UCL = \bar{x} - W(q_\alpha)s_x \qquad (2\text{-}65)$$

For gamma distributions, Singh and Singh (2003) observed that the coverage probabilities provided by the 95% *UCL*s based upon bootstrap t and Hall's bootstrap methods are in close agreement. For larger samples, these two methods approximately provide the specified 95% coverage to the population mean, $k\theta$, of a gamma distribution. For smaller sample sizes (from gamma distribution), the coverage provided by these two methods is slightly lower than the specified level of 0.95. For both lognormal and gamma distributions, these two methods (bootstrap t and Hall's bootstrap) perform better than the other bootstrap methods, namely, the standard bootstrap method, simple percentile, and bootstrap BCA percentile methods.

Just like the gamma distribution, for lognormally distributed data sets, it is noted that Hall's *UCL* and bootstrap t *UCL* provide similar coverages. However, for highly skewed lognormal data sets, the coverages based upon Hall's method and bootstrap t method are significantly lower than the specified coverage, 0.95 (Singh and Singh, 2003). This is true even in samples of larger sizes (e.g., $n = 100$). For lognormal data sets, the coverages provided by Hall's bootstrap and bootstrap t methods do not increase much with the sample size, *n*. For highly skewed (e.g.,> 2.0) data sets of small sizes (e.g., $n < 15$), Hall's bootstrap method (and also bootstrap t method) performs better than the Chebyshev *UCL*, and for larger samples, the Chebyshev *UCL* performs better than Hall's bootstrap method. Similar to the bootstrap t method, it should be noted that Hall's bootstrap method sometimes results in unstable, inflated, and erratic values, especially in the presence of outliers (Efron and Tibshirani, 1993). Therefore, these two methods should be used with caution. If outliers are present in a data set, then a 95% *UCL* of the mean should be computed using alternative *UCL* computation methods.

2.5 Recommendations and Summary

This section describes the recommendations and summary on the computation of a 95% *UCL* of the unknown population arithmetic mean, μ_1, of a contaminant data distribution without censoring. These recommendations are based upon the findings of Singh, Singh, and Engelhardt (1997, 1999); Singh *et al.* (2002a); Singh, Singh, and Iaci (2002b); Singh and Singh (2003); and Singh *et al.* (2006). Recommendations have been summarized for: 1) normally distributed data sets, 2) gamma distributed data sets, 3) lognormally distributed data sets, and 4) data sets which are nonparametric and do not follow any of the three distributions included in ProUCL.

For skewed parametric as well as nonparametric data sets, there is no simple solution to compute a *95% UCL* of the population mean, μ_1. Singh *et al*. (2002a), Singh, Singh, and Iaci (2002b), and Singh and Singh (2003) noted that the *UCLs* based upon the skewness adjusted methods, such as the Johnson's modified t and Chen's adjusted-*CLT*, do not provide the specified coverage (e.g., 95%) to the population mean even for mildly to moderately skewed (e.g., $\hat{\sigma}$ in the interval [0.5, 1.0)) data sets for samples of sizes as large as 100. The coverage of the population mean by these skewness-adjusted *UCL* gets poorer (much smaller than the specified coverage of 0.95) for highly skewed data sets, where the skewness levels have been defined earlier as a function of σ or $\hat{\sigma}$ (standard deviation of log-transformed data).

2.5.1 Recommendations to Compute a 95% UCL of the Unknown Population Mean, μ_1, Using Symmetric and Positively Skewed Data Sets

Interested users may want to consult graphs as given in Singh and Singh (2003) for a better understanding of the summary and recommendations made in this section.

2.5.1.1 Normally or Approximately Normally Distributed Data Sets

As expected, for a normal distribution, N (μ_1, σ_1^2), the Student's t-statistic, modified t-statistic, and bootstrap t 95% *UCL* computation methods result in *UCLs* which provide coverage probabilities close to the nominal level, 0.95. Contrary to the general conjecture, the bootstrap BCA method does not perform better than the other bootstrap methods (e.g., bootstrap t). Actually, for normally distributed data sets, the coverages for the population mean, μ_1, provided by the *UCL*s based upon the BCA method and Hall's bootstrap method are lower than the specified 95% coverage. This is especially true when the sample size, *n* is less than 30. For details refer to Singh and Singh (2003).

- For normally distributed data sets, a *UCL* based upon the Student's t-statistic, as given by equation (2-32), provides the optimal *UCL* of the population mean. Therefore, for normally distributed data sets, one should always use a 95% *UCL* based upon the Student's t-statistic.

- The *95% UCL* of the mean given by equation (2-32) based upon Student's t-statistic may also be used when the *Sd*, s_y of the log-transformed data is less than 0.5, or when the data set approximately follows a normal distribution. A data set is approximately normal when the normal Q-Q plot displays a linear pattern (without outliers and jumps) and the resulting correlation coefficient is high (e.g., 0.95 or higher).

- Student's t-*UCL* may also be used when the data set is symmetric (but possibly not normally distributed). A measure of symmetry (or skewness) is $\hat{k}_3$, which is given by equation (2-43). A value of $\hat{k}_3$ close to zero (e.g., if absolute value of skewness is roughly less than 0.2 or 0.3) suggests approximate symmetry. The approximate symmetry of a data distribution can also be judged by looking at histogram of data sets.

2.5.1.2 Gamma Distributed Skewed Data Sets

In practice, many skewed data sets can be modeled both by a lognormal distribution and a gamma distribution, especially when the sample size is smaller than 70-100. As is well known, the 95% *H-UCL* of the mean based upon a lognormal model often results in unjustifiably large and impractical 95% *UCL*

values. In such cases, a gamma model, G (k, θ), may be used to compute a reliable 95% *UCL* of the unknown population mean, μ_1.

- Many skewed data sets follow a lognormal as well as a gamma distribution. It should be noted that the population means based upon the two models could differ significantly. Lognormal model based upon a highly skewed (e.g., $\hat{\sigma} \geq 2.5$) data set will have an unjustifiably large and impractical population mean, μ_1, and its associated *UCL*. The gamma distribution is better suited to model positively skewed environmental data sets.

  One should always first check if a given skewed data set follows a gamma distribution. If a data set does follow a gamma distribution or an approximate gamma distribution, one should compute a 95% UCL based upon a gamma distribution. Use of highly skewed (e.g., $\hat{\sigma} \geq 2.5$-3.0) lognormal distributions should be avoided. For such highly skewed lognormally distributed data sets that cannot be modeled by a gamma or an approximate gamma distribution, nonparametric *UCL* computation methods based upon the Chebyshev inequality may be used.

- The five bootstrap methods do not perform better than the two gamma *UCL* computation methods. It is noted that the performances (in terms of coverage probabilities) of bootstrap t and Hall's bootstrap methods are very similar. Out of the five bootstrap methods, bootstrap t and Hall's bootstrap methods perform the best (with coverage probabilities for population mean closer to the nominal level of 0. 95). This is especially true when the skewness is quite high (e.g., $\hat{k} < 0.1$) and the sample size is small (e.g., $n <$ 10-15). This can be seen from graphs presented in Appendix A of the Technical Guide for ProUCL 3.0 (EPA, 2004).

- The bootstrap BCA method does not perform better than the Hall's method or the bootstrap t method. The coverage for the population mean, μ_1, provided by the BCA method is much lower than the specified 95% coverage. This is especially true when the skewness is high (e.g., $\hat{k} < 1$) and sample size is small (Singh and Singh, 2003).

- From the results presented in Singh, Singh, and Iaci (2002b), and in Singh and Singh (2003), it is concluded that for data sets which follow a gamma distribution, a 95% *UCL* of the mean should be computed using the adjusted gamma *UCL* when the shape parameter, k, is: $0.1 \leq k < 0.5$, and for values of $k \geq 0.5$, a 95% *UCL* can be computed using an approximate gamma *UCL* of the mean, μ_1.

- For highly skewed gamma distributed data sets with $k < 0.1$, the bootstrap t *UCL* or Hall's bootstrap (Singh and Singh, 2003) may be used when the sample size is smaller than 15, and the adjusted gamma *UCL* should be used when sample size starts approaching and exceeding 15. The small sample size requirement increases as skewness increases (that is as k decreases, the required sample size, n, increases).

- The bootstrap t and Hall's bootstrap methods should be used with caution as sometimes these methods yield erratic, unreasonably inflated, and unstable *UCL* values, especially in the presence of outliers. In the case Hall's bootstrap and bootstrap t methods yield inflated and erratic *UCL* results, the 95% *UCL* of the mean should be computed based upon the adjusted gamma 95% *UCL*. ProUCL prints out a warning message associated

with the recommended use of the *UCLs* based upon the bootstrap t method or Hall's bootstrap method.

These recommendations for the use of gamma distribution are summarized in Table 2-3.

Table 2-3. Summary Table for the Computation of a 95% *UCL* of the Unknown Mean, μ_1, of a Gamma Distribution

$\hat{k}$	*Sample Size, n*	*Recommendation*
$\hat{k} \geq 0.5$	For all *n*	Approximate gamma 95% *UCL*
$0.1 \leq \hat{k} < 0.5$	For all *n*	Adjusted gamma 95% *UCL*
$\hat{k} < 0.1$	$n < 15$	95% *UCL* based upon bootstrap t or Hall's bootstrap method*
$\hat{k} < 0.1$	$n \geq 15$	Adjusted gamma 95% *UCL* if available, otherwise use approximate gamma 95% *UCL*

*In case bootstrap t or Hall's bootstrap methods yield erratic, inflated, and unstable *UCL* values, the *UCL* of the mean should be computed using adjusted gamma *UCL*.

2.5.1.3 Lognormally Distributed Skewed Data Sets

For lognormally, LN (μ, σ^2), distributed data sets, the H-statistic-based *UCL* does provide specified 0.95, coverage for the population mean for all values of σ. However, the H-statistic often results in unjustifiably large *UCL* values that do not occur in practice. This is especially true when skewness is high (e.g., $\sigma >$ 2.0). The use of a lognormal model unjustifiably accommodates large and impractical values of the mean concentration and its *UCLs*. The problem associated with the use of a lognormal distribution is that the population mean, μ_1, of a lognormal model becomes impractically large for larger values of σ, which in turn results in inflated *H-UCL* of the population mean, μ_1. Since the population mean of a lognormal model becomes too large, none of the other methods except for the *H-UCL* provides the specified 95% coverage for that inflated population mean, μ_1. This is especially true when the sample size is small and skewness is high. For extremely highly skewed data sets (with $\sigma >$ 2.5-3.0) of smaller sizes (e.g., < 70-100), the use of a lognormal distribution-based *H-UCL* should be avoided (e.g., see Singh *et al*., 2002a and Singh and Singh, 2003). Therefore, alternative *UCL* computation methods such as the use of a gamma distribution or use of a *UCL* based upon nonparametric bootstrap methods or Chebyshev inequality-based methods are desirable.

As expected for skewed (e.g., with σ (or $\hat{\sigma}$) ≥ 0.5) lognormally distributed data sets, the Student's t-*UCL*, modified t-*UCL*, adjusted-CLT *UCL*, the standard bootstrap methods all fail to provide the specified 0.95 coverage for the unknown population mean for samples of all sizes. Just like the gamma distribution, the performances (in terms of coverage probabilities) of bootstrap t and Hall's bootstrap methods are very similar (Singh and Singh, 2003). However, it is noted that the coverage provided by Hall's bootstrap (and also by bootstrap t) is much lower than the specified 95% coverage for the population mean, μ_1, for samples of all sizes of varying skewness. Moreover, the coverages provided by Hall's bootstrap or bootstrap t method do not increase much with the sample size.

Also the coverage provided by the BCA method is much lower than the coverage provided by Hall's method or the bootstrap t method. Thus, the BCA bootstrap method cannot be recommended to compute a 95% *UCL* of the mean of a lognormal population. For highly skewed data sets of small sizes (e.g., < 15) with σ exceeding 2.5-3.0, even the Chebyshev inequality-based *UCL*s fail to provide the specified 0.95 coverage for the population. However, as the sample size increases, the coverages provided by the Chebyshev inequality-based *UCL*s also increase. For such highly skewed data sets ($\hat{\sigma}$ > 2.5) of sizes less than 10-15, Hall's bootstrap or bootstrap t methods provide larger coverage than the coverage provided by the 99% Chebyshev *(MVUE) UCL*. Therefore, for highly skewed lognormally distributed data sets of small sizes, one may use Hall's method (bootstrap t UCL) to compute an estimate of the EPC term. The small sample size requirement increases with σ. This means that as skewness (σ) increases, the sample size, n needed to provide specified coverage (e.g., ~0.95) by Hall's bootstrap UCL also increases and becomes much larger than 20-30.

It should be noted that even a small increase in the *Sd*, σ, increases the skewness considerably. For example, for a lognormal distribution, when σ = 2.5, skewness ≈ 11825.1; and when σ = 3, skewness ≈ 729555. In practice, the occurrence of such highly skewed data sets (e.g., $\sigma \geq 3$) is not very common. Nevertheless, these highly skewed data sets can arise occasionally and, therefore, require separate attention. Singh *et al*. (2002a) observed that when the *Sd*, σ, starts approaching 2.5 (that is, for lognormal data, when *CV* > 22.74 and skewness > 11825.1), even a *99% Chebyshev (MVUE) UCL* fails to provide the desired 95% coverage for the population mean, μ_1. This is especially true when the sample size, *n*, is smaller than 30. For such extremely skewed data sets, the larger of the two *UCLs*: the *99% Chebyshev (MVUE) UCL* and the nonparametric *99% Chebyshev (Mean, Sd) UCL*, may be used as an estimate of the EPC.

It is also noted that, as the sample size increases, the *H-UCL* starts behaving in a stable manner. Therefore, depending upon the *Sd*, σ (*actually its MLE* $\hat{\sigma}$), for lognormally distributed data sets, one can use the *H-UCL* for samples of larger sizes such as greater than 70-100. This large sample size requirement increases as the *Sd,* $\hat{\sigma}$, increases, as can be seen in Table 2-4. ProUCL can compute an *H-UCL* for samples of sizes up to 1000. For lognormally distributed data sets of smaller sizes, some alternative methods to compute a 95% *UCL* of the population mean, μ_1, are summarized in Table 2-4.

Furthermore, it is noted that for moderately skewed (e.g., σ > 1, 1.25) data sets of larger sizes (e.g., *n* > 100-150), the *H-UCL* becomes even smaller than the sample mean or Student's t-*UCL* (and various other *UCLs*). It should be pointed out that the large sample behavior of *H-UCL* has not been investigated rigorously. For confirmation purposes (that is *H-UCL* does provide the 95% coverage for larger samples also), it is desirable to conduct such a study for samples of larger sizes.

Since skewness (as defined earlier) is a function of σ (or $\hat{\sigma}$), the recommendations for the computation of the *UCL* of the population mean are also summarized in Table 1-4 for various values of the *MLE* $\hat{\sigma}$ of σ and the sample size, *n*. Here $\hat{\sigma}$ is an *MLE* of σ, and is given by the *Sd* of log-transformed data given by equation (2-2). Note that Table 2-4 is applicable to the computation of a 95% *UCL* of the population mean based upon lognormally distributed data sets without nondetect observations. A procedure to compute a 95% *UCL* of the mean of a lognormal distribution is summarized in the following steps:

- Skewed data sets should be first tested for a gamma distribution. For lognormally distributed data sets (which can not be modeled by a gamma distribution), the method as summarized in Table 2-4 may be used to compute a 95% *UCL* of the mean.

Table 2-4. Summary Table for the Computation of a 95% *UCL* of the Unknown Mean, μ_1, of a Lognormal Population

$\hat{\sigma}$	***Sample Size, n***	***Recommendation***
$\hat{\sigma}$ < 0.5	For all *n*	Student's t, modified t, or *H-UCL*
0.5 ≤ $\hat{\sigma}$ < 1.0	For all *n*	*H-UCL*
1.0 ≤ $\hat{\sigma}$ < 1.5	*n* < 25	95% Chebyshev *(MVUE) UCL*
	n ≥ 25	*H-UCL*
1.5 ≤ $\hat{\sigma}$ < 2.0	*n* < 20	99% Chebyshev *(MVUE) UCL*
	20 ≤ *n* < 50	95% Chebyshev *(MVUE) UCL*
	n ≥ 50	*H-UCL*
1.5 ≤ $\hat{\sigma}$ < 2.0	*n* < 20	99% Chebyshev *(MVUE) UCL*
	20 ≤ *n* < 50	97.5% Chebyshev *(MVUE) UCL*
	50 ≤ *n* < 70	95% Chebyshev *(MVUE) UCL*
	n ≥ 70	*H-UCL*
2.5 ≤ $\hat{\sigma}$ < 3.0	*n* < 30	Larger of (99% Chebyshev *(MVUE) UCL* or 99% Chebyshev (Mean, Sd)
	30 ≤ *n* < 70	97.5% Chebyshev *(MVUE) UCL*
	70 ≤ *n* < 100	95% Chebyshev *(MVUE) UCL*
	n ≥ 100	*H-UCL*
3.0 ≤ $\hat{\sigma}$ ≤ 3.5	*n* < 15	Hall's bootstrap method*
	15 ≤ *n* < 50	Larger of (99% Chebyshev (*MVUE) UCL*, 99% Chebyshev(Mean, Sd)
	50 ≤ *n* < 100	97.5% Chebyshev *(MVUE) UCL*
	100 ≤ *n* < 150	95% Chebyshev *(MVUE) UCL*
	n ≥ 150	*H-UCL*
$\hat{\sigma}$ > 3.5	For all *n*	Use nonparametric methods*

*In the case that Hall's bootstrap method yields an erratic unrealistically large UCL value, then the UCL of the mean may be computed based upon the Chebyshev inequality.

- Specifically, for highly skewed (e.g., $1.5 < \sigma \leq 2.5$) data sets of small sizes (e.g., $n \leq 50$-70), the EPC term may be estimated by using a 97.5% or 99% *MVUE* Chebyshev *UCL* of the population mean (or mass). For larger samples (e.g., $n > 70$), the *H-UCL* may be used to estimate the EPC.

- For extremely highly skewed (e.g., $\sigma > 2.5$) lognormally distributed data sets, the population mean becomes unrealistically large. Therefore, the use of *H-UCL* should be avoided especially when the sample size is less than100. For such highly skewed data sets, Hall's bootstrap *UCL* may be used when the sample size is less than 10-15 (Singh and Singh 2003). The small sample size requirement increases with $\hat{\sigma}$. For example, n = 10 is considered small when $\hat{\sigma}$ = 3.0, and n = 15 is considered small when $\hat{\sigma}$ = 3.5.

- Hall's bootstrap and bootstrap t UCL methods should be used with caution as sometimes it yields erratic, inflated, and unstable *UCL* values, especially in the presence of outliers. For these highly skewed data sets of size, *n* (e.g., less than 10-15), in the case that Hall's

bootstrap method yields an erratic and inflated *UCL* value, the 99% Chebyshev *MVUE UCL* may be used to estimate the EPC term. ProUCL displays a warning message associated with the recommended use of Hall's bootstrap method.

2.5.1.4 Nonparametric Distribution-Free Skewed Data Sets without a Discernable Distribution

- The use of gamma and lognormal distributions as discussed here will cover a wide range of skewed data distributions. For skewed data sets which are neither gamma nor lognormal, one can use a nonparametric Chebyshev *UCL* or Hall's bootstrap *UCL* (for small samples) of the mean to estimate the EPC term.

- For skewed nonparametric data sets with negative and zero values, use a *95% Chebyshev (Mean, Sd) UCL* for the population mean, μ_1.

For all other nonparametric data sets with only positive values, the following procedure may be used to estimate the EPC term.

- For mildly skewed data sets with $\hat{\sigma} \leq 0.5$, one can use Student's t-statistic or modified t-statistic to compute a 95% *UCL* of mean, μ_1.

- For nonparametric moderately skewed data sets (e.g., σ or its estimate, $\hat{\sigma}$ in the interval (0.5, 1]), one may use a *95% Chebyshev (Mean, Sd) UCL* of the population mean, μ_1.

- For nonparametric moderately to highly skewed data sets (e.g., $\hat{\sigma}$ in the interval (1.0, 2.0]), one may use a *99% Chebyshev (Mean, Sd) UCL or 97.5% Chebyshev (Mean, Sd) UCL* of the population mean, μ_1, to obtain an estimate of the EPC term.

- For highly skewed to extremely highly skewed data sets with $\hat{\sigma}$ in the interval (2.0, 3.0], one may use Hall's *UCL* or the *99% Chebyshev (Mean, Sd) UCL* to compute the EPC term.

- Extremely skewed nonparametric data sets with σ exceeding 3.0 provide poor coverage. For such highly skewed data distributions, none of the methods considered provide the specified 95% coverage for the population mean, μ_1. The coverages provided by the various methods decrease as σ increases. For such data sets of sizes less than 30, a 95% *UCL* can be computed based upon Hall's bootstrap method or bootstrap t method. Hall's bootstrap method provides highest coverage (but less than 0.95) when the sample size is small. It is noted that the coverage for the population mean provided by Hall's method (and bootstrap t method) does not increase much as the sample size, *n*, increases. However, as the sample size increases, coverage provided by the *99% Chebyshev (Mean, Sd) UCL* method also increases. Therefore, for larger samples, a *UCL* should be computed based upon the *99% Chebyshev (Mean, Sd)* method. This large sample size requirement increases as $\hat{\sigma}$ increases. These recommendations are summarized in Table 2-5.

Table 2-5. Summary Table for the Computation of a 95% *UCL* of the Unknown Mean, μ_1, Based Upon a Skewed Data Set (with all Positive Values) without a Discernable Distribution, Where $\hat{\sigma}$ is the *sd* of Log-transformed Data

$\hat{\sigma}$	Sample Size, n	Recommendation
$\hat{\sigma} \leq 0.5$	For all *n*	95% *UCL* based on Student's t- or Modified-t statistic
$0.5 < \hat{\sigma} \leq 1.0$	For all *n*	95% Chebyshev (Mean, *Sd*) *UCL*
$1.0 < \hat{\sigma} \leq 2.0$	*n* < 50	99% Chebyshev (Mean, *Sd*) *UCL*
	n ≥ 50	97.5% Chebyshev (Mean, *Sd*) *UCL*
$2.0 < \hat{\sigma} \leq 3.0$	*n* < 10	Hall's Bootstrap *UCL**
	n ≥ 10	99% Chebyshev (Mean, *Sd*) *UCL*
$3.0 < \hat{\sigma} \leq 3.5$	*n* < 30	Hall's Bootstrap *UCL**
	n ≥ 30	99% Chebyshev (Mean, *Sd*) *UCL*
$\hat{\sigma} > 3.5$	*n* < 100	Hall's Bootstrap *UCL**
	n ≥ 100	99% Chebyshev (Mean, *Sd*) *UCL*

*If Hall's bootstrap method yields an erratic and unstable UCL value (e.g., happens when outliers are present), a *UCL* of the population mean may be computed based upon the *99% Chebyshev (Mean, Sd)* method.

2.5.2 Summary of the Procedure to Compute a 95% UCL of the Unknown Population Mean, μ_1, Based Upon Full Data Sets without Nondetect Observations

1. The first step in computing a *95% UCL* of a population arithmetic mean, μ_1, is to perform goodness-of-fit tests to test for normality, lognormality, or gamma distribution of the data set under study. ProUCL has three methods to test for normality or lognormality: the informal graphical test based upon a Q-Q plot, the Lilliefors test, and the Shapiro-Wilk W test. ProUCL also has three methods to test for a gamma distribution: the informal graphical Q-Q plot based upon gamma quantiles, the Kolmogorov-Smirnov (K-S) EDF test, and the Anderson-Darling (A-D) EDF test.

   ProUCL generates a quantile-quantile (Q-Q) plot to graphically test the normality, lognormality, or gamma distribution of the data. There is no substitute for graphical displays of a data set. On this graph, a linear pattern (e.g., with high correlation such as 0.95 or higher) displayed by bulk of data suggests approximate normality, lognormality, or gamma distribution. On this graph, points well separated from the majority of data may be potential outliers requiring special attention. Also, any visible jumps and breaks of significant magnitudes on a Q-Q plot suggest that more than one population may be present. In that case, each of the populations should be considered separately. That is, a separate EPC term should be computed for each of the populations.

2. It is, therefore, recommended to always use the graphical Q-Q plot as it provides useful information about the presence of multiple populations (e.g., site and background data mixed together) or outliers. Both graphical Q-Q plot and formal goodness-of-fit tests

should be used on the same data set before determining the distribution of the data set under investigation. A single test statistic such as the Shapiro-Wilk test (A-D test or some other GOF test) may lead to the incorrect conclusion that the data are normally (or gamma) distributed even when there are more than one population present. Only a graphical display, such as an appropriate Q-Q, can provide this kind of important information. Obviously, when multiple populations are present, those should be separated out and the EPC terms (the *UCLs*) or other estimates (e.g., BTVs) should be computed separately for each of those populations. Therefore, it is strongly recommended not to skip the GOF tests option in ProUCL 4.0. Since the computation of an appropriate *UCL* depends upon data distribution, it is advisable that the user should take his time (instead of blindly using a numerical value of a test statistic in an effort to automate the distribution selection process) to determine the data distribution. *Both graphical (e.g., Q-Q plots) and analytical procedures (Shapiro-Wilk test, K-S test) should be used on the same data set to determine the most appropriate distribution of the data set under study.*

3. After performing the goodness-of-fit test, ProUCL informs the user about the data distribution: normal, lognormal, gamma distribution, or a non-discernable distribution.

4. For a normally distributed (or approximately normally distributed) data set, the user is advised to use Student's t-distribution-based *UCL* of the mean. Student's t-distribution (or modified t-statistic) may also be used to compute the EPC term when the data set is symmetric (e.g., $\left|\hat{k}_3\right|$ is smaller than 0.2-0.3) or mildly skewed; that is, when σ or $\hat{\sigma}$ is less than 0.5.

5. For gamma distributed (or approximately gamma distributed) data sets, the user is advised to: use the approximate gamma *UCL* for $\hat{k} \geq 0.5$; use the adjusted gamma *UCL* for $0.1 \leq \hat{k} < 0.5$; use the bootstrap t method (or Hall's method) when $\hat{k} < 0.1$ and the sample size, $n < 15$; and use the adjusted gamma *UCL* (if available) for $\hat{k} < 0.1$ and sample size, $n \geq 15$. If the adjusted gamma *UCL* is not available, then use the approximate gamma *UCL* as an estimate of the EPC term. In the case that the bootstrap t method or Hall's bootstrap method yields an erratic inflated *UCL* (e.g., when outliers are present) result, the *UCL* should be computed using the adjusted gamma *UCL* (if available) or the approximate gamma *UCL*.

6. For lognormal data sets, ProUCL recommends (as summarized in Table 2-4) a method to estimate the EPC term based upon the sample size and standard deviation of the log-transformed data, $\hat{\sigma}$. ProUCL can compute an *H-UCL* of the mean for samples of sizes up to 1000. Nonparametric *UCL* computation methods such as the modified t, *CLT* method, adjusted-*CLT* method, bootstrap and jackknife methods are also included in ProUCL. However, it is noted that nonparametric *UCLs* based upon most of these methods do not provide adequate coverage to the population mean for moderately skewed to highly skewed data sets (e.g., Singh and Singh, 2003).

7. For data sets, which are not normally, lognormally, or gamma distributed, a nonparametric *UCL* of the mean based upon the Chebyshev inequality is preferred. The *Chebyshev (Mean, Sd) UCL* does not depend upon any distributional assumptions and can be used for moderately to highly skewed data sets which do not follow any of the three data distributions incorporated in ProUCL.

8. It should be noted that for extremely skewed data sets (e.g., with $\hat{\sigma}$ exceeding 3.0), even a Chebyshev inequality-based 99% *UCL* of the mean fails to provide the desired coverage (e.g., 0.95) of the population mean. A method to compute the EPC term for distribution-free data sets is summarized in Table 2-5. It should be pointed out that in the case that Hall's bootstrap method appears to yield erratic and inflated results (typically happens when outliers are present), the 99% Chebyshev *UCL* may be used as an estimate of the EPC term.

Chapter 3

Estimating Background Threshold Values or Establishing Site-Specific Background Concentrations Using Full Data Sets without Nondetect (ND) Observations

3.1 Introduction

Often in environmental applications, site-specific background level contaminant concentrations are needed to compare site concentrations (e.g., both before and after some remediation activities) with background level contaminant concentrations, also called as background statistics or background threshold values (BTVs). These BTVs are computed based upon the sampled data collected from the site-specific background as determined by all interested parties, including the potentially responsible parties, local, and federal government agencies. Many times, intermediate or future remediation decisions at a polluted site are made after performing such background versus site comparisons. A site observation exceeding a BTV can be viewed as coming from a contaminated area of the site under study. It is, therefore, important that these background statistics be computed using appropriate background data sets and defensible statistical methods. Some minimum sample size requirements (e.g., sample size >8-10) to estimate the BTVs based upon background data sets have been discussed in Chapter 1 of this guidance document. Chapter 1 also discusses situations when it may be appropriate to perform point-by-point site observations (preferably composite samples) comparisons with BTVs or with some pre-established threshold values. Specifically when not more than 4-6 site observations need to be compared individually with estimated or pre-established BTVs, one may compare point-by-point site observations with BTVs and other threshold values. If more than 8-10 (preferably more) site observations are available, then it is preferable to use single sample hypothesis (in case BTVs are pre-established) or two-sample hypothesis (in case BTVs need to be estimated using background data) testing approaches to perform site versus background comparisons. This chapter describes statistical limits that may be used to estimate the BTVs and other not-to-exceed values for full data sets without any nondetect (ND) observations. Statistical limits based upon data sets with nondetect observations are discussed in Chapter 5. Chapter 6 discusses the various single sample and two-sample hypotheses testing approaches for data sets with and without NDs as incorporated in ProUCL 4.0.

It should be pointed out that the availability of background statistics as discussed in this chapter is particularly useful when individual site observations from impacted areas of the site (perhaps after some remediation activities) are compared with some BTVs to determine if enough remediation (at the impacted areas of the site) has been performed yielding remediated site concentrations which are comparable to background level concentrations. This method of site versus background comparisons is also useful when not enough site data are available to perform two-sample comparisons such as the t-test or the nonparametric Wilcoxon Rank Sum (WRS) test. Moreover, in practice, during remediation activities, it is desirable to compare each individual site observation (collected during remediation phase) with some pre-determined or estimated background level threshold value(s). Sometimes pre-established screening levels are used as estimates of background threshold values. However, in practice, these BTVs need to be estimated based upon site-specific background (or reference) data sets collected using appropriate sampling methods and data quality objectives (DQOs). This chapter describes procedures, which can be used to compute relevant background statistics based upon an appropriate background data set without any nondetect observations. Methods to estimate the BTVs based upon data sets with NDs are described in Chapter 5 of this Technical manual.

When enough site and background data are available, it is recommended to use two-sample tests (t-test, WRS test, etc.) to perform background versus site comparisons. Parametric and nonparametric procedures (hypotheses testing) can be used to compare the measures of central tendencies of the two populations (background versus site) when enough detected data are available from the two populations under consideration. Hypothesis testing approaches to perform site versus background comparisons are discussed in Chapter 6 of this Technical guidance document.

This chapter (and also Chapter 5) deals with the computation of background statistics (BTVs) when it is known/assumed that the underlying data set does represent a sample collected from some site-specific background area(s). That is, it is assumed a priori that all of the observations (at least most of them) come from a single background population. However, since outliers are inevitable in most environmental applications, some outliers may also be present in a background data set. These outlying observations need to be identified before computing the background statistics as outliers, when present, distort all of the statistics of interest (such as background statistics), which in turn may lead to incorrect remediation decisions for the site under study. The inclusion (or exclusion) of outliers in a background data set needs to be justified before performing other relevant statistical analyses. All interested parties should be involved in such decision making to determine the inclusion or exclusion of outliers in a background data set. The proper identification of multiple outliers is a complex issue and is beyond the scope of this document. A brief description of outlier identification is given in Section 1. A couple of outlier tests as incorporated in ProUCL 4.0 are given in Chapter 7 of this Technical document. Some discussions about the disposition of outliers are provided in Chapter 3 of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b).

A more complicated problem arises when the data set under study represents a mixture sample, which is inevitable in many environmental applications. In these cases, the data set under study may consist of samples from the background areas as well as from various areas of the site under study. In this situation, first one has to separate the background observations from other site related observations. After the background data set has been properly identified, one can proceed with the computation of background statistics as presented later in this chapter. However, separation of background data from a mixture sample is not an easy task. Using the population partitioning techniques, statisticians (e.g., see Singh, Singh, and Flatman, 1994) have developed some background separation methods from mixture samples. However, the topics of population partitioning and the identification of a valid background data set from the mixture sample are beyond the scope of ProUCL 4.0 and this guidance document. It requires developing a separate chapter, which will deal with the population partitioning methods including the identification of a valid background data set from a mixture sample. Throughout this chapter, it is assumed that one is dealing with a sample from a single population representing a valid site-related background data set.

The first step in establishing site-specific background level contaminant concentrations for site-related hazardous analytes is to perform background sampling to collect an appropriate number of samples from the designated site-specific background areas, perhaps using the input from experts and personnel familiar with the site operations and history. An appropriate DQO process should be followed to collect an adequate number of background samples. Once the adequate amount of data has been collected, the next step is to determine the data distribution. This is typically done using exploratory graphical tools as well as formal goodness-of-fit tests. These tests are described in several environmental documents (EPA 2006, ProUCL 2004, and Navy documents 1999, 2002a, 2002b). In addition to graphical displays, ProUCL 4.0 has goodness-of-fit (GOF) tests for normal, lognormal, and gamma distributions.

Once the data distribution of a background data set has been determined, one can use parametric or nonparametric statistical methods to compute background statistics. Several upper limits have been incorporated in ProUCL 4.0 that may be used as estimated of BTVs and other not-to-exceed values. A review of the environmental literature reveals that one or more of the following statistical limits are used to compute the background statistics; that is, to determine background level contaminant concentrations, BTVs. The BTVs are called upper background cutoff levels.

1. Upper percentiles
2. Upper prediction limits (UPLs)
3. Upper tolerance limits (UTLs)
4. IQR Upper Limit (upper end of the upper whisker in a box and whisker plot)

Depending upon the background data distribution, the background upper limits listed above can be computed by using parametric methods based upon probability distributions (e.g., normal, lognormal, or gamma) or by using nonparametric (distribution-free) methods. The background limits listed above are often used as background threshold values to compare individual site observations with background level contamination. Typically, a site observation (possibly based upon composite samples) in exceedance of a background threshold value can be considered as coming from a contaminated site area that may have been impacted by the site-related activities. In other words, such a site observation may be considered as exhibiting some evidence of contamination at the site due to site-related activities. In case of an exceedance of the BTV by a site location, some practitioners like to verify the possibility of contaminated site location by re-sampling that location, and comparing the sample value with the BTV.

The background threshold values are used when not enough site data (e.g., < 4-6 observations) are available to perform traditional two-sample comparisons (e.g., t-test, Wilcoxon Rank Sum test, Gehan's test, etc.) as described in Chapter 6 of this document. In the absence of adequate amount of site data, individual point-by-point site observations have to be compared with some BTVs to determine the presence or absence of contamination due to site related activities. This method of comparing site versus background level contamination is particularly helpful to use after some sort of remediation activities have taken place at the site; and the objective is to determine if the remediated site areas have been remediated enough to the background level contaminant concentrations. A brief discussion of identification and disposition of outliers is considered first.

3.2 Treatment of Outliers

While computing reliable background statistics, it is essential that one is dealing with a single population representing site background without potentially impacted observations (outliers). Therefore, a brief discussion on this topic is presented in this section. As well known, outliers, when present, typically represent observations from different populations(s), perhaps contaminated observations from the site under study. Outliers distort all of the statistics of interest, including the sample mean, the sample standard deviation, and, consequently, the parametric percentiles, and various upper limits such as UPLs, UTLs, and UCLs. It is noted that nonparametric upper percentiles are often represented by higher ordered statistics such as the largest value or the second largest value. In the case of extreme high observations, these higher order statistics may be outlying observations representing contaminated observations from the site (e.g., a large Federal Facility) under study. Decisions made based upon outliers or distorted

statistics can be incorrect and misleading. Therefore, special attention should be given to such outlying observations.

If justified, that is, if some outliers do represent observations from the contaminated areas of the site, then those observations should not be included in the computation of BTVs. This decision should be a team effort to determine whether or not an identified outlier does represent an observation from the contaminated part of the site. Such an outlying observation should not be part of the background data set. Specifically, such an observation should not be used in the computation of background statistics. All interested parties should be involved in making such decisions. Several classical (EPA, 2006) and robust (Singh and Nocerino, 1995) statistical procedures are available to identify multiple outliers. Robust and resistant outlier identification procedures are beyond the scope of ProUCL 4.0. In environmental applications (EPA, 2006 and Navy, 2002a, 2002b), classical procedures are used to identify outliers. A couple of those classical outlier tests are available in ProUCL 4.0. As mentioned before, classical outlier procedures suffer from masking effects as they get distorted by the same outlying observations that they are supposed to find! It is suggested to use robust and resistant statistical procedures to identify multiple outliers. Several robust outlier identification procedures are available in Scout (EPA, 1999) software package, which is currently under revision and upgrade. It is recommended to supplement the use of classical and robust procedures with graphical procedures such as box plots, quantile-quantile (Q-Q) plots.

Note: *It should be noted that the methods as incorporated in ProUCL 4.0 can be used on any data set (with or without nondetects) with or without the potential outliers. Specifically, it should not be misunderstood that ProUCL 4.0 is restricted to be used only on data sets without outliers. It is not a requirement to delete or omit the outliers before using estimation, UCL95, and various other limits computation methods (e.g., KM (BCA) UCL, MLE) as incorporated in ProUCL 4.0. The fact of the matter is that the user should be aware of the fact that the inclusion of a few outliers in the computations of these statistics may yield distorted estimates, UCL95, UPLs, UTLs, and various other statistics. Therefore, for more accurate and reliable statistics and results, the authors of this Technical Guide recommend that whenever justified, the low probability outlying observations (often coming from different population(s)) should not be included in the computation of the statistics used in the various decision making processes. The statistics (e.g., upper limits) of interest should be computed using the majority of the data set representing the dominant population (e.g., an AOC, a background area). The outlying observations should be separately investigated to determine the reasons for their occurrences (e.g., errors or contaminated locations). It is always a good practice to compute the statistics with and without the outliers, and compare the potential impact of outliers on the decision making processes.*

Throughout this chapter, $x_1, x_2, \ldots, x_n$ represent the background concentrations for a contaminant of potential concern (COPC) collected from some site-specific background or reference area. The objective is to estimate a BTV based upon this data set. The sample values are arranged in ascending order. The resulting ordered sample (called ordered statistics) is denoted by $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. The ordered statistics are often used as nonparametric estimates of upper percentiles, upper prediction limits (UPLs), and upper tolerance limits (UTLs). Also, let $y_i = \ln(x_i)$; $i = 1, 2, \ldots, n$, then $\bar{y}$ and s_y represent the mean and standard deviation (*sd*) of the log-transformed data. Some parametric and nonparametric upper limits often used to estimate BTVs are described throughout the rest of this chapter.

3.3 Upper p*100% Percentiles as Estimates of Background Threshold Values (BTVs)

Some Navy documents (1999, 2002a, 2002b) suggest the use of the 95th upper percentiles of the background distribution as estimates of the background level threshold values (e.g., pages 46, 137 Navy 2002b). However, explicit parametric formulae for the computation of the 95th percentiles are missing from the Navy (e.g., 2002a, 2002b) documents, making it difficult for a typical user to use the 95% percentiles as estimates of BTVs. Since these percentiles do represent a feasible method to compute background threshold values, (one of the objectives of the present document), for clarification, computation of both parametric as well as nonparametric percentiles are briefly described as follows.

In most statistical textbooks (e.g., Hogg and Craig, 1995), the pth (e.g., $p = 0.95$) sample percentile (of the measured sample values) is defined as that value, $\hat{x}_p$, such that p100% of the background data set lies at or below it. The carat sign over x_p, indicates that it represents a statistic (an estimate of the p^{th} population percentile) computed based upon the sampled data.

*3.3.1 Nonparametric p*100% Percentile*

It is quite simple to compute a nonparametric 95% percentile of a background data set. It should be pointed out that such nonparametric sample percentiles (for $0 < p < 1$) cannot exceed the maximum value in a background data set. These nonparametric 95% percentiles may be used when the background data (raw or transformed) do not follow a normal or a gamma distribution at some specified (e.g., d = 0.05, 0.1) level of significance.

It is noted that, the practitioners compute these nonparametric p^{th} percentiles (quantiles) in more than one way. Some users compute the p^{th} using the pn^{th} order statistic, which may be a whole number between 1 and n or a fraction lying between 1 and n. For example, if $n = 20$, and $p = 0.95$, then 20*0.95 = 19, thus the 19th ordered statistic represents the 95% percentile. If $n = 17$, and $p = 0.95$, then 17*0.95= 16.15, thus the 16.15th ordered value may be used as an estimate of the BTV. The 16.15th ordered value lies between the 16th and the 17th order statistics and can be computed by using simple linear interpolation given by:

$$x_{(16.15)} = x_{(16)} + 0.15\,(x_{(17)} - x_{(16)}). \qquad (3\text{-}1)$$

It is noted that some other users compute the p^{th} nonparametric percentile by the order statistic given by the $(pn+0.5)^{th}$ order statistic, while others compute the p^{th} nonparametric percentile by the order statistic given by the $(p*(n+1))^{th}$ order statistic. As mentioned above, if the number computed in not a whole number, then the percentile is computed using the linear interpolation illustrated above. In any case, if for a given value of p, the resulting number, $(p*(n+1))$ exceeds n, then that p^{th} percentile is estimated by the n^{th} order statistic, that is by the maximum value. In ProUCL 4.0, the p^{th} nonparametric percentile is estimated by the $(p*(n+1))^{th}$ order statistic. This formula is used on data sets with and without ND observations.

*3.3.2 Normal p*100% Percentile*

The computation of normal upper percentiles has been considered next. First, compute the sample mean, $\bar{x}$, and standard deviation (*sd*), s, using a defensible (e.g., outliers, multiple populations, mixture populations are not allowed) background data set without the outliers. For normally distributed data sets, the p*100th sample percentile is given by the following statement.

$$\hat{x}_p = \bar{x} + sz_p \tag{3-2}$$

Here z_p is the p*100th percentile of a standard normal, N (0,1), distribution, which means that the area (under the standard normal curve) to the left of z_p is p. If the distributions of the site data and the background data are comparable and similar (meaning no contamination due to the site related activities), then an observation coming from a population (e.g., site) similar (comparable) to that of the background population should lie at or below the p*100% upper percentile, $\hat{x}_p$, with probability p. Thus, the 95% percentile given by the above equation (for p = 0.95 or 0.99) may also be used as an estimate of the background threshold value when the background data are normally distributed.

*3.3.3 Lognormal p*100% Percentile*

To compute the p^{th} upper percentile, $\hat{x}_p$, of a lognormally distributed data set, the sample mean, $\bar{y}$, and standard deviation (*sd*), s_y, of log-transformed data are computed first using a defensible background data set without outliers. For lognormally distributed data sets, the p*100th percentile is given by the following statement,

$$\hat{x}_p = \exp(\bar{y} + s_y z_p), \tag{3-3}$$

where, as before, z_p is the upper p*100th percentile of a standard normal, N(0,1), distribution. A 95th percentile given by the above equation may be used as an estimate of the BTV for a COPC when the background data are lognormally distributed.

*3.3.4 Gamma p*100% Percentile*

Since the introduction of a gamma distribution, G (k, θ), is relatively new in environmental applications (e.g., Singh, Singh, and Iaci 2002), a brief description of the gamma distribution is given first. The equations giving the maximum likelihood estimates (MLEs) of the gamma parameters, k (= shape parameter) and θ (= scale parameter), can be found in Singh, Singh, and Iaci (2002) and also in the ProUCL 3.0 Technical Guide (EPA, 2004). A random variable (RV), X (e.g., Aroclor 1254 concentrations), follows a gamma distribution, G (k,θ), with parameters $k > 0$ and $\theta > 0$, if its probability density function is given by the following equation:

$$\begin{aligned} f(x;k,\theta) &= \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-x/\theta}; && x > 0 \\ &= 0; && \textit{otherwise} \end{aligned} \tag{3-4}$$

The mean, variance, and skewness of a gamma distribution are given by: mean = $\mu = k\theta$, variance = $\sigma^2 = k\theta^2$, and skewness = $2/\sqrt{k}$. Note that as k increases, the skewness decreases, and, consequently, a gamma distribution starts approaching a normal distribution for larger values of k (e.g., $k \geq 6$ - 8).

Let $\hat{k}$ and $\hat{\theta}$ represent the maximum likelihood estimates (MLEs) of k and θ respectively. Note the relationship between a chi-square and a gamma distribution. Specifically, the relationship between a gamma RV, X = G (k, θ), and a chi-square RV, Y, is given by X = Y * θ/2, where Y follows a chi-square distribution with 2k degrees of freedom. Thus, the percentiles of a chi-square distribution (as programmed

in ProUCL) can be used to determine the percentiles of a gamma distribution. In practice, *k* is replaced by its MLE. Thus, once an α*100% percentile, $y_{(\alpha)}$, of a chi-square distribution with 2*k* degrees of freedom is obtained, the α*100% percentile for a gamma distribution can be obtained by using the equation:

$$x_\alpha = y_\alpha * \theta / 2 \qquad (3\text{-}5)$$

3.3.5 Example 1

Consider a site-specific background data set associated with a Superfund site. The data set has several inorganic contaminants of potential concern, including aluminum, arsenic, chromium, and lead. The computation of background statistics obtained using ProUCL 4.0 are summarized in this example. The complete data set is given in Appendix 5 of the *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b).

3.3.5.1 Normal Percentile

Using the Shapiro-Wilk test as incorporated in ProUCL, it is determined that aluminum follows a normal distribution at 5% significance level. This can also be seen from the Q-Q plot as given in Figure 3-1. Since the data set follows a normal distribution, a normal 95% upper percentile may be used as an estimate of the BTV. The sample mean of aluminum data set is 7789.1667, the standard deviation, *s*, is 4263.8969, and $z_{0.95,}$ the upper 95% percentile of a standard normal distribution, is 1.6449. Thus normal 95% percentile for aluminum is:

$$\hat{x}_{0.95} = \bar{x} + sz_{0.95} = 7789.1667 + 4263.8969 * 1.6449 = 14802.8507$$

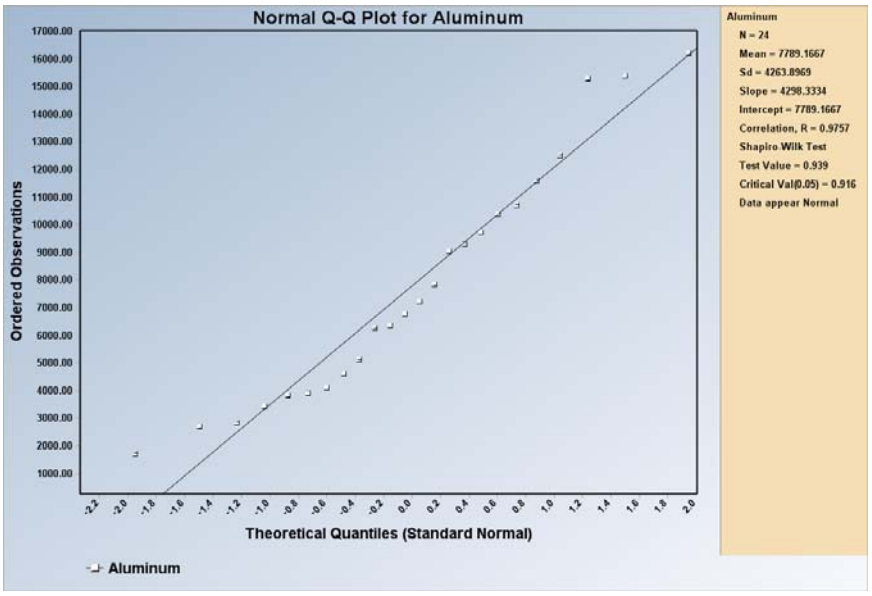

Figure 3-1. Shapiro-Wilk Normal Distribution Test for Aluminum

3.3.5.2 Lognormal Percentile

Using ProUCL 4.0, it is determined that chromium concentrations follow a lognormal distribution at 5% level of significance. This can also be seen from the chromium Q-Q as given in Figure 3-2. The sample mean and standard deviation of the log-transformed data are $\bar{y}$ = 2.3344, and s_y = 0.5678. Thus the 95% upper percentile for chromium is given by the following equation:

$$\hat{x}_{0.95} = \exp(\bar{y} + s_y z_{0.95}) = \exp(2.3344 + 0.5678 * 1.6449) = 26.26868$$

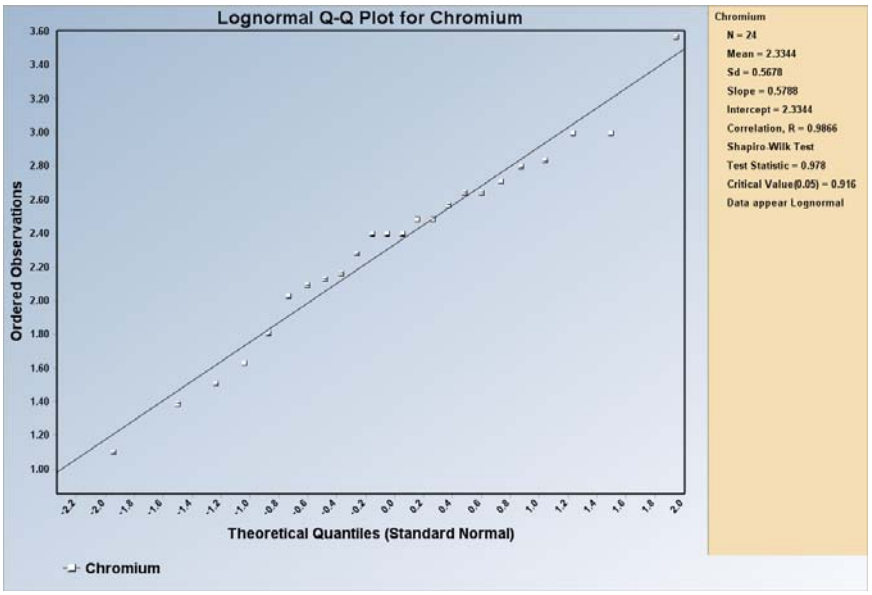

Figure 3-2. Shapiro-Wilk Lognormal Distribution Test for Chromium

3.3.5.3 Nonparametric Percentile

Using ProUCL, it is determined that Lead (Pb) concentrations do not follow any of the known distributions as incorporated in ProUCL 4.0. Therefore, an upper nonparametric 95th percentile may be used as an estimate of the BTV for lead concentrations. A 95% nonparametric upper percentile is given by the 0.95*n^{th} order statistic for n=24,

$$95\% \text{ Upper Percentile} = x_{(0.95n)} = x_{(22.8)} = x_{(22)} + 0.8(x_{(23)} - x_{(22)}).$$

The value for $x_{(22)}$ is 53.3 and for $x_{(23)}$ is 98.5. Thus, a 95% Upper Percentile = 53.3 + 0.8 * (98.5 – 53.3) = 89.46.

3.3.5.4 Gamma Percentile

Using ProUCL, it is determined that Arsenic (As) concentrations follow a gamma distribution. The gamma Q-Q plot displaying Anderson-Darling test statistic is given in Figure 3-3.

The bias-corrected MLE for k is 3.6616 and the bias-corrected MLE of θ is 0.5867. The 95% percentile for a chi-square distribution with degrees of freedom (df) $2\hat{k}$ is y_α = 14.5362. Using these values, one can derive the 95% gamma percentile as follows:

$$x_\alpha = y_\alpha = \theta / 2 = (14.5362 * 0.5867) / 2 = 4.2643.$$

It is noted that arsenic concentrations also follow a lognormal distribution. Therefore, for comparison, several upper percentiles are tabulated as follows in Table 3-1. This also includes the normal percentile even though arsenic does not follow a normal distribution.

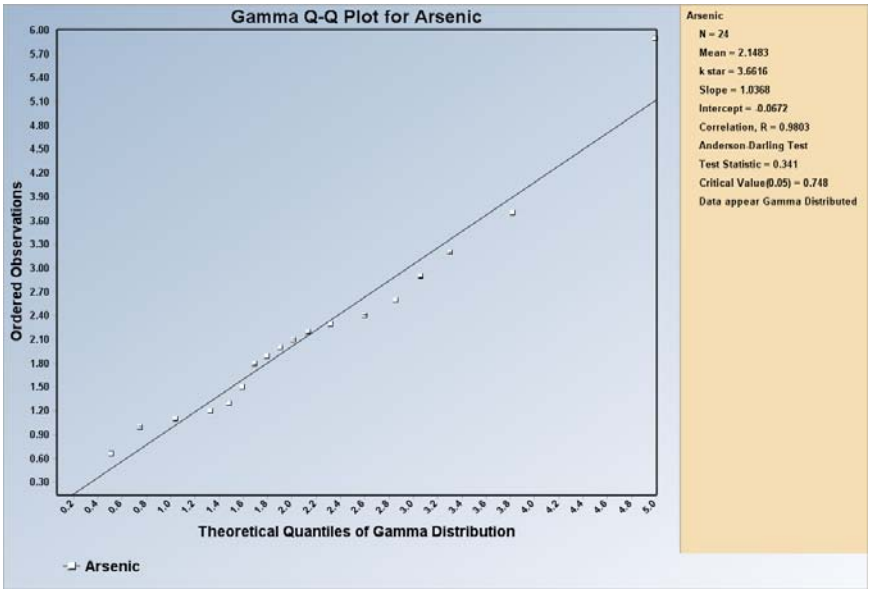

Figure 3-3. Anderson-Darling Gamma Distribution Test

Table 3-1. 95% Percentiles for Arsenic

Distribution	95% Percentile
Normal	4.0564
Gamma	4.2643
Lognormal	4.3877
Non-discernable	3.7000

The 1989 RCRA document is one of the early EPA guidance documents. The RCRA document talks about the use of both 95% UTLs and 95% UPLs to determine the presence or absence of contamination in compliance wells. The UPLs or UTLs are computed based upon the background data sets (e.g., upstream data, upgradient wells, site-specific background). Compliance well observations (or site data) are then compared with the background well (or background data) UTL or UPL. An exceedance in compliance well may suggest some evidence of contamination in that well. Similar arguments can be made when comparing concentrations of soil samples from an impacted site area with some site-specific background concentrations.

In 1992, EPA upgraded the 1989 RCRA document and came up with its addendum. This 1992 addendum also talks about the use of UTLs and UPLs as estimates of background level threshold values. The 1992 addendum modifies the formula for 95% UTL as given in the 1989 RCRA document (page 51, Chapter 4). The multipliers, *k* (used to compute a 95% UTL), as given in the 1989 document, are meant to provide at least 95% coverage. The 1992 addendum states that, in practice, the use of this factor, *k*, as given in Table 2-5 of the 1989 document, provides 98% (and not 95%) coverage. This is especially true when one is dealing with samples of small sizes. Thus, the use of factor *k* (to compute a UTL), as described above, may result in more false negatives (larger background statistics), which is not protective of the environment and human health. Therefore, a modified multiplier, which is the same as the prediction limit multiplier, has been suggested in the 1992 RCRA addendum. In the 1992 addendum, it is stated that this modified multiplier will on the average provide the specified coverage (= 95% here).

However, it desirable that these statements about the coverage's provided by the 95% UTLs and 95% UPLs be verified by Monte Carlo simulation experiments. As mentioned before, BTVs are often used to make remediation decisions at polluted sites. Therefore, these BTVs should be computed using defensible statistical procedures. This section describes the computation of UTLs as well as UPLs. Both parametric as well as nonparametric UTL and UPL computation procedures have been summarized in the following sections.

3.4 Upper Tolerance Limits

For many environmental applications, such as the groundwater monitoring applications, an upper tolerance limit (UTL) is often used to provide appropriate coverage for a proportion, *p*% (e.g., 80%, 90%, 95%, coverage, etc.) of future observations with confidence coefficient, (1 – α). It should be noted that an upper tolerance limit with confidence coefficient, (1 – α), covering a proportion of *p*% (*p* is also called the coverage coefficient), observations simply represent a (1 – α)100% upper confidence limit of the p^{th} percentile of the population under study (here the background population).

3.4.1 Normal Upper Tolerance Limits

First, compute the sample mean, $\bar{x}$, and standard deviation (*sd*), *s*, using a defensible (e.g., outliers, multiple populations, mixture populations not allowed) background data set without the outliers (e.g., see Example 2 below). For normally distributed data sets, an upper (1 – α)100% tolerance limit with tolerance or coverage coefficient = *p* (that is providing coverage to at least *p*100% proportion of observations) is given by the following statement.

$$UTL = \bar{x} + K * s \qquad (3\text{-}6)$$

Here, $K = K(n, \alpha, p)$ is the tolerance factor and depends upon the sample size, *n*, confidence coefficient = $(1 - \alpha)$, and the coverage proportion = *p*. The UTL given by the above equation represents a (1 – α)100% confidence interval for the p^{th} percentile of the underlying normal distributions. The values of the tolerance factor, *K*, have been tabulated extensively in the various statistical books (e.g., Hahn and Meeker 1991). Those *K* values are based upon non-central t-distributions. Also, some large sample approximations (e.g., Natrella, 1963) are available to compute the *K* values for one-sided tolerance intervals (same for both UTLs and lower tolerance limit). The approximate value of *K* is also a function of the sample size, *n*, coverage coefficient, *p*, and the confidence coefficient, $(1 - \alpha)$. In the ProUCL 4.0 software package, the values of *K* for samples of sizes ≤ 30, as given in Hahn and Meeker, have been directly programmed. For sample sizes larger than 30, the large sample approximations, as given in Natrella (1963), have been used to compute the *K* values. The Natrella's approximation seems to work

well for samples of sizes larger than 30 (Hahn and Meeker 1991). The program, ProUCL, seems to work well to compute these *K* values for sample of sizes as large as 5000 (and even larger).

3.4.2 Lognormal Upper Tolerance Limits

The procedure to compute UTLs for lognormally distributed data sets is similar to that for normally distributed data sets. First, the sample mean, $\bar{y}$, and standard deviation (*sd*), s_y, of the log-transformed data are computed using a defensible unique background data set without outliers. An upper (1 – α)100% tolerance limit with tolerance or coverage coefficient = *p* (that is, providing coverage to at least *p*100% proportion of observations) is given by the following statement.

$$\text{UTL} = \exp(\bar{y} + K * s_y) \tag{3-7}$$

Note that, just as for the normal distribution, the UTL given by the above equation represents a (1 – α)100% confidence interval for the p^{th} percentile of the lognormal distribution. The *K* factor used to compute the lognormal UTL is the same as the one used to compute the normal UTL. It should be noted that just as the upper confidence limits (UCLs) for the mean of lognormally distributed populations, the UTLs based upon lognormal distributions are typically higher (sometimes unrealistically higher as shown in the following example) than other parametric and nonparametric UTLs. The use of a lognormal UTL to estimate the BTV should be specifically avoided when skewness is high (e.g., *sd* of logged data > 1, 1.5) and sample size is small (e.g., < 30, 50).

3.4.3 Gamma Distribution Upper Tolerance Limits

Positively skewed environmental data sets can often be modeled by a gamma distribution. Also, ProUCL software has two goodness-of-fit tests (Anderson-Darling test and Kolmogorov-Smirnov test) to test for gamma distributions. Bootstrap methods: percentile (%) bootstrap and bias-corrected accelerated (BCA) bootstrap methods can be used to compute UTLs for gamma distributions. ProUCL 4.0, also have some bootstrap methods to compute UTLs.

3.4.4 Nonparametric Upper Tolerance Limits

The computation of nonparametric UTLs is somewhat messy as it is based upon binomial cumulative probabilities and order statistics. Just like parametric UTLs, a nonparametric UTL providing coverage to p100% observations with confidence coefficient (*CC)* (1 – α)100% represents an (1 – α)100% upper confidence limit for the p^{th} percentile of the population under study. Also, the nonparametric UTLs (given by order statistics) cannot exactly achieve the specified confidence coefficient, (1 – α). In most cases, only an approximate confidence coefficient can be achieved by nonparametric UTLs. One has to be satisfied with the achievable confidence coefficient, which is as close as possible to the specified CC of (1 – α). Thus, an appropriate UTL is chosen which provides coverage for the p^{th} percentile as close as possible to the specified confidence coefficient, (1 – α). Based upon binomial cumulative probabilities, the algorithms to compute nonparametric UTLs have been also incorporated in ProUCL 4.0 software.

It is noted that for simplicity and based upon professional judgment, the 1992 RCRA addendum recommends the use of either the largest value or the second largest value for both UPL and UTL. However, the use of the largest value as UPL (or UTL) may result in an overestimate, especially when there is a possibility of the presence of potential outliers in the background data set. Therefore, to be protective of the human health and the environment, it is preferable to use the second largest (or even a lower order statistic) data value as a nonparametric UTL or UPL. Whenever, these higher order statistics

represent potential outliers (e.g., well separated from the majority of the data on a normal Q-Q plot), one should avoid their use as estimates of the BTVs. The selection of these higher (e.g., largest, second largest) order statistics (as estimates of BTVs) also depends upon the sample size of the background data set. Specifically, for smaller data sets, it is likely that the higher order statistics (largest, or the second largest) can be chosen to estimate the BTV. As mentioned earlier, in practice, a few high observations (outliers) may represent contaminated observations (e.g., well separated from the majority of the data on a normal Q-Q plot) and their selection should be avoided as nonparametric estimates of the BTVs.

It is also noted that the largest value is also used as an estimate of the BTV when a 95% UTL (especially for lognormal skewed data sets) exceeds the largest value in a data set. Moreover, when a background UTL does exceed the maximum value in the background data set, it is quite likely that the maximum value may represent an extreme value (perhaps from some impacted site area). The use of maximum value can be avoided by using appropriate upper percentiles (e.g., 95%) described earlier as estimates of background values.

A brief description of the computation of nonparametric UTLs (confidence intervals for percentiles) now follows. For details, the interested readers are referred to David and Nagaraja (2003), Conover (1980), and Hahn and Meeker (1991). The binomial distribution is used with the number of trials = the sample size, n, and the probability of success = p (the proportion of observations for which coverage is being sought). Using the cumulative binomial probabilities, a number, r: $1 \le r \le n$, is chosen such that the cumulative binomial probability:

$\sum_{i=0}^{i=r} \binom{n}{i} p^i (1-p)^{(n-i)}$ becomes as close as possible to $(1 - \alpha)$. Then, the r^{th} order statistic, $x_{(r)}$ is picked as

the $(1 - \alpha)100\%$ UTL providing coverage to $p100\%$ of the observations in the sample under study (background data set here). This algorithm has been incorporated in ProUCL for samples of sizes up to 2000. ProUCL 4.0 also prints out the order statistic, the estimate of the UTL, and the associated coverage provided (achieved) by that order statistic. For larger samples, one can use the following approximation, which has also been incorporated in ProUCL. For $n > 2000$, a large sample normal approximation to the binomial distribution can be used to obtain an upper confidence limit for the p^{th} percentile. The number, r: $1 \le r \le n$, used to compute the r^{th} order statistics (to estimate the BTV) is obtained using the following equation:

$$r = np + z_{(1-\alpha)}\sqrt{np(1-p)} + 0.5 \qquad (3\text{-}8)$$

3.4.5 Example 2: Computation of Upper Tolerance Limits

Using the background data set used earlier associated with a Superfund site, the computation of the various parametric and nonparametric upper tolerance limits have been illustrated in this example. To illustrate the differences in the values of the UTLs as a function of the coverage coefficient, p%, UTLs have been computed for four (4) different coverage coefficients, p% = 80%, 90%, 95%, and 99%.

3.4.5.1 Normal Upper Tolerance Limits

As noted earlier (Figure 3-1), aluminum concentrations follow a normal distribution. Therefore, the various normal UTLs with coverage coefficients of 80%, 90%, 95%, and 99% are listed as follows.

95% UTL (80% coverage) = $\bar{x} + K_{(24,0.95,0.80)} * s$ = 7789.1667 + 1.331 * 4263.8969 = 13387.663

95% UTL (90% coverage) = $\bar{x} + K_{(24,0.95,0.90)} * s$ = 7789.1667 + 1.853 * 4263.8969 = 15690.168

95% UTL (95% coverage) = $\bar{x} + K_{(24,0.95,0.95)} * s$ = 7789.1667 + 2.309 * 4263.8969 = 17634.505

95% UTL (99% coverage) = $\bar{x} + K_{(24,0.95,0.99)} * s$ = 7789.1667 + 3.181 * 4263.8969 = 21362.623

3.4.5.2 Lognormal Upper Tolerance Limits

As noted earlier, chromium background concentrations follow a lognormal distribution. The 95% UTLs based upon a lognormal distribution are given as follows.

95% UTL (80% coverage) = $\exp(\bar{y} + K_{(24,0.95,0.80)} * s_y)$ = exp(2.3344 + 1.331 * 5678) = 21.7583

95% UTL (90% coverage) = $\exp(\bar{y} + K_{(24,0.95,0.90)} * s_y)$ = exp(2.3344 + 1.853 * 5678) = 29.5659

95% UTL (95% coverage) = $\exp(\bar{y} + K_{(24,0.95,0.95)} * s_y)$ = exp(2.3344 + 2.309 * 5678) = 38.3039

95% UTL (99% coverage) = $\exp(\bar{y} + K_{(24,0.95,0.99)} * s_y)$ = exp(2.3344 + 3.181 * 5678) = 62.8464

3.4.5.3 Nonparametric Upper Tolerance Limits

Earlier, it was determined that the background lead concentrations do not follow any of the known distributions as incorporated in ProUCL 4.0. For lead, 95% UTLs based upon binomial cumulative probabilities and order statistics are given in Table 3-2. It should be noted that the resulting UTL might not achieve the exact specified CC of 0.95.

Following the procedure described earlier, a 95% UTL with coverage coefficient of 80% is represented by the 22nd order statistic. The resulting UTL (22nd order statistic) covers about 80% of the observations (that is 80% observations are $\leq x_{(22)}$) with a probability (confidence coefficient) of 0.967 (instead of 0.95). A 95% UTL with a coverage coefficient of 90% is represented by the 23rd order statistic. The resulting UTL (23rd order statistic) covers about 90% of the observations (that is 90% observations are $\leq x_{(23)}$) with a probability (confidence coefficient) of 0.922 (instead of 0.95). Using order statistics, the actual achieved confidence, (1 - α) is often different from the user requested confidence coefficient of 95%. ProUCL 4.0 selects the order statistic that achieves the confidence coefficient closest to the user specified confidence coefficient.

Table 3-2. Nonparametric Upper Tolerance Limits for Lead

Coverage Coefficient (p%)	Order Statistic		Achieved Confidence (1 – α)
	Order	Value	
80%	$x_{(22)}$	53.3	96.7%
90%	$x_{(23)}$	98.5	92.2%
95%	$x_{(24)}$	109	100%
99%	$x_{(24)}$	109	100%

Caution: Since nonparametric UTLs are given by order statistics, every effort should be made to make sure that the chosen order statistic to estimate the BTV does not represent an outlying observation coming from a population other than the background population.

3.5 Nonparametric Upper Limit Based Upon Interquartile Range (IQR) – IQR Upper Limit

Sometimes, an upper limit based upon the IQR of the background data set is used as an estimate of the BTV. In this chapter, we denote this limit by the IQR Upper Limit. It is very simple to compute and is briefly described below. The simple formula to compute IQR Upper Limit is:

$$\text{IQR Upper Limit} = Q3 + 1.5 * \text{IQR}. \quad (3\text{-}9)$$

Here IQR = Q3 – Q1, the interquartile range, the difference between the third (upper) quartile, Q3, and the first (lower) quartile, Q1, of the background data set. The quartiles of a data set are defined in most applied statistical books (e.g., Hoaglin, Mosteller and Tukey, 1983). The three quartiles, Q1, Q2, and Q3 of a data set divide the data set into four (4) equal parts. Note that the second quartile represents the median of the data set. Thus, 25% of the data lie at or below Q1, 50% of the data lie at or below Q2 (median), and 75% of the data lie at or below Q3; therefore, 25% of the data lie above Q3. Just like all other limits described in this chapter, individual site observations are compared with the IQR Upper Limit. Any site concentration exceeding the background level IQR Upper Limit may be considered justification to consider contamination at the site. The computation of IQR Upper Limit has also been incorporated in ProUCL 4.0 software package.

Note: *The behavior of an IQR-based limit as an estimate of a BTV is not well studied. Therefore, this limit should be used with caution to estimate the BTVs or not-to-exceed values.*

3.5.1 Example 3: IQR Upper Limit

Sometimes, in practice, a nonparametric upper limit based upon the interquartile range (IQR) of the data set under study is used as an estimate of the background threshold value. Since lead does not follow any of the parametric distributions as incorporated in ProUCL, an upper limit based upon the IQR can be used as an estimate of the BTV. This will require the use of the first quartile, Q1, and the third quartile, Q3. Here Q1 = 8.7, Q3 = 19, and

$$\text{IQR} = Q3 - Q1 = 10.3$$

An estimate of the BTV based upon IQR is given as follows.

$$\text{IQR Upper Limit} = Q3 + 1.5 * \text{IQR} = 19 + 1.5 * 10.3 = 34.45$$

3.6 Upper Prediction Limits

As mentioned before, both the 1989 RCRA document and its 1992 addendum suggest the use of upper prediction limits (UPLs) as estimates of background level threshold values. If the background and site contaminant distributions are comparable, then a typical site observation should lie below a 95% UPL based upon a background data set with probability 0.95. A site observation exceeding the background 95% UPL can be considered as providing some evidence of contamination due to site related industrial activities. Since a UPL does represent a plausible way of expressing background level contaminant concentration, a brief discussion of both parametric as well as nonparametric UTLs is presented in this section.

3.6.1 Normal Upper Prediction Limit

The sample mean, $\bar{x}$, and the standard deviation (*sd*), *s*, are computed first based upon a defensible unique (e.g., outliers, multiple populations, mixture populations not allowed) background data set without the outliers. For normally distributed data sets, an upper (1 – α)100% prediction limit is given by the following well known equation:

$$\text{UPL} = \bar{x} + t_{((1-\alpha),(n-1))} * s * \sqrt{(1+1/n)} \qquad (3\text{-}10)$$

Here $t_{((1-\alpha),(n-1))}$ is a critical value from Student's t-distribution with (*n*–1) degrees of freedom.

3.6.2 Lognormal Upper Prediction Limit

An upper (1 – α)100% lognormal UPL is similarly given by the following equation:

$$\text{UPL} = \exp(\bar{y} + t_{((1-\alpha),(n-1))} * s_y * \sqrt{(1+1/n)}) \qquad (3\text{-}11)$$

Here $t_{((1-\alpha),(n-1))}$ is a critical value from Student's t-distribution with (*n*–1) degrees of freedom.

3.6.3 Nonparametric Upper Prediction Limit

As mentioned before, the background data set under consideration should represent a single population, and should be free of outlying observations, which may represent data from impacted areas of the site. A one-sided nonparametric UPL is simple to compute and is given by the following m^{th} order statistic. One can use linear interpolation if the resulting number, m, given below does not represent a whole number (a positive integer).

$$\text{UPL} = X_{(m)}, \text{ where } m = (n + 1) * (1 - \alpha). \qquad (3\text{-}12)$$

For example, for a nonparametric data set of size 25, a 90% UPL is desired. Then *m* = (26*0.90) = 23.4. Thus, a 90% nonparametric UPL can be obtained by using the 23^{rd} and the 24^{th} ordered statistics and is given by the following equation:

$$\text{UPL} = X_{(23)} + 0.4 * (X_{(24)} - X_{(23)})$$

Similarly, if a nonparametric 95% UPL is desired, then m = 0.95 * (25 + 1) = 24.7, and a 95% UPL can be similarly obtained by using linear interpolation between the 24^{th} and 25^{th} order statistics. However, if a 99% UPL needs to be computed, then m = 0.99 * 26 = 25.74, which exceeds 25, the sample size. Therefore, for such cases, the highest order statistic (the largest value) has to be used as the 99% UPL of the background data set under study. The largest value(s) should be used with caution (as they may represent outliers) to estimate the BTVs.

3.6.4 Example 4

The same background data set as used earlier has been used in this example. All computations have been carried out using ProUCL 4.0.

3.6.4.1 Normal Upper Prediction Limit

As noted earlier, the aluminum concentrations in our example do follow a normal distribution. A 95% UPL for aluminum is given as follows.

UPL = $\bar{x} + t_{((1-\alpha),(n-1))} * s * \sqrt{(1+1/n)}$ = 7789.17 + 1.7139 * 4.263.90 * 1.02 = 15247.628

3.6.4.2 Lognormal Upper Prediction Limit

The chromium background concentrations of Example 1 follow a lognormal distribution. A 95% UPL for chromium is given by the following equation.

UPL = $\exp(\bar{y} + t_{((1-\alpha),(n-1))} * s_y * \sqrt{(1+1/n)})$ = exp(2.3344 + 1.7139 * 0.5678* 1.02) = 27.8738

3.6.4.3 Nonparametric Upper Prediction Limit

A nonparametric UPL can be computed using the following equation:

$$\text{UPL} = X_{(m)}, \text{ where } m = (n + 1) * (1 - \alpha). \qquad (3\text{-}13)$$

For lead concentrations, with n = 24 and (1 -α) = 0.95, the corresponding 95% UPL is given by the m^{th} = 23.75thorder statistic, which can be computed using simple linear interpolation as follows:
$X_{(23.75)} = X_{(23)} + 0.75(X_{(24)} - X_{(23)}) = 98.5 + 0.75(10.5) = 106.375$.

Note: *As mentioned before, nonparametric UPLs (and also UTLs) are typically represented by higher order statistics, or by some value in between (based upon linear interpolation) the higher order statistics. If those higher order statistics represent contaminated outlying observations, then a value lying between the two contaminated observations will also be an outlier. For example for lead, if the two high observations: 98.5 and 109 are considered as outliers, then the 95% UPL = 106.375 as computed above will also represent an outlier.*

Therefore, nonparametric UTLs or UPLs should be used with caution to estimate the BTVs. Every effort should be made to identify and separate the outlying observations before computing nonparametric limits to estimate the BTVs.

For the comparison sake, the 95% UPLs for aluminum, arsenic, chromium, and lead as produced by ProUCL (irrespective of the data distribution) have been summarized in Table 3-3.

Table 3-3. 95% Upper Prediction Limits for Selected Contaminants

	Inorganic Contaminants			
Distribution	**Aluminum**	**Arsenic**	**Chromium**	**Lead**
Normal	15247.63	4.1764	24.022	69.418
Lognormal	19245.46	4.6277	27.874	59.784
non-discernable	16000.00	5.3500	31.625	106.375

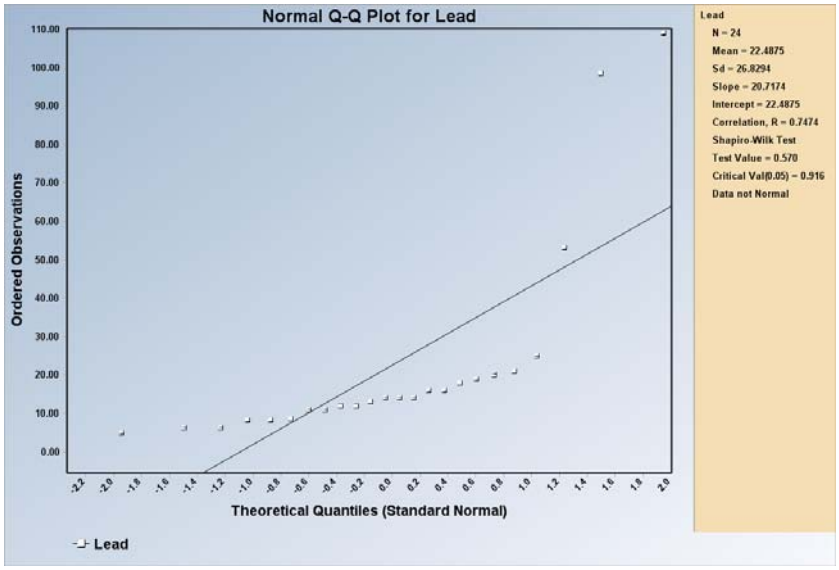

Figure 3-5. Shapiro-Wilk Normal Distribution Test for Lead

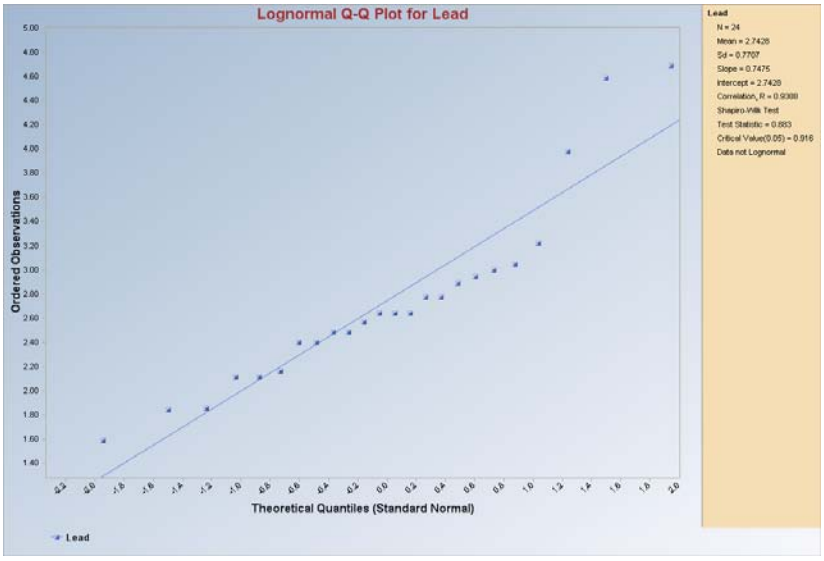

Figure 3-6. Shapiro-Wilk Lognormal Distribution Test for Lead

General Background Statistics for Full Data Sets	
User Selected Options	
From File	WorkSheet.wst
Full Precision	OFF
Confidence Coefficient	95%
Coverage	90%
Different or Future K Values	1
Number of Bootstrap Operations	2000

OS_Lead

General Statistics

Total Number of Samples	10	Number of Unique Samples	10

Raw Statistics		Log-Transformed Statistics	
Minimum	19.7	Minimum	2.981
Maximum	1940	Maximum	7.57
Second Largest	338	Second Largest	5.823
First Quantile	25.45	First Quantile	3.235
Median	41.63	Median	3.726
Third Quantile	106.2	Third Quantile	4.482
Mean	268.2	Mean	4.29
SD	595.8	SD	1.452
Coefficient of Variation	2.221		
Skewness	3.009		

Background Statistics

Normal Distribution Test		Lognormal Distribution Test	
Shapiro Wilk Test Statistic	0.479	Shapiro Wilk Test Statistic	0.812
Shapiro Wilk Critical Value	0.842	Shapiro Wilk Critical Value	0.842
Data not Normal at 5% Significance Level		Data not Lognormal at 5% Significance Level	

Assuming Normal Distribution		Assuming Lognormal Distribution	
95% UTL with 90% Coverage	1671	95% UTL with 90% Coverage	2231
95% UPL (t)	1414	95% UPL (t)	1191
90% Percentile (z)	1032	90% Percentile (z)	469.3
95% Percentile (z)	1248	95% Percentile (z)	795.4
99% Percentile (z)	1654	99% Percentile (z)	2140

Gamma Distribution Test		Data Distribution Tests
k star	0.409	Data do not follow a Discernable Distribution (0.05)
Theta Star	655.4	
nu star	8.186	

A-D Test Statistic	1.421	**Nonparametric Statistics**		
5% A-D Critical Value	0.779	90% Percentile		338
K-S Test Statistic	0.369	95% Percentile		1139
5% K-S Critical Value	0.281	99% Percentile		1780
Data not Gamma Distributed at 5% Significance Level				
Assuming Gamma Distribution		95% UTL with	90% Coverage	1940
90% Percentile	754.5	95% Percentile Bootstrap UTL with	90% Coverage	1940
95% Percentile	1106	95% BCA Bootstrap UTL with	90% Coverage	1940
99% Percentile	1986	95% UPL		1940
		95% Chebyshev UPL		2992
		Upper Threshold Limit Based upon IQR		227.3
Note: UPL (or upper percentile for gamma distributed data) represents a preferred estimate of BTV				

Figure 3-7. All Background Statistics for Lead

Chapter 4

Computation of Upper Confidence Limit of the Population Mean Based Upon Data Sets with Nondetect Observations

4.1 Introduction

Nondetects (NDs) or below detection limit (BDL) observations are inevitable in most environmental data sets. Singh, Maichle, and Lee (EPA, 2006) studied the performances of the various parametric and nonparametric UCL computation methods for data sets with NDs and multiple detection limits. This chapter represents a sub-chapter of Singh, Maichle, and Lee (EPA, 2006) summarizing their main findings and recommendations. Based upon their findings and results, some of the better (in terms of coverage probabilities) performing UCL computation methods have been incorporated in ProUCL 4.0. Specifically, the new UCL module of ProUCL 4.0 can be used to compute UCLs of the population means based upon left-censored data sets (with single and multiple detection limits) from normal, lognormal, and gamma distributions. The details of those methods can be found in Singh, Maichle, and Lee (EPA, 2006).

The estimation of mean and standard deviation (*sd)* based upon data sets with NDs have been studied by many researchers (Cohen, 1991, Gibbons and Coleman, 2001, Schneider 1986, Gilliom and Helsel, 1986, Singh and Nocerino, 2002, and Helsel, 2005). However, it is noted that not much guidance is available on how to accurately compute the various upper limits (e.g., UCLs, UPLs, UTLs) often used in environmental applications. For example, in many exposure and risk assessment applications, one needs to compute a 95% upper confidence limit (UCL) of the mean based upon data sets with BDL observations.

Also, in several background evaluation studies (e.g., estimating background threshold values (BTVs)), one needs to compute upper prediction limits (UPLs), upper percentiles, and upper tolerance limits (UTLs) using left-censored data sets. Statistical methods to compute upper limits (UPLs, UTLs) based upon data sets with BDL observations are described in Chapter 5 of this Technical document. In the recent environmental literature (e.g., Millard 2000 and USEPA-UGD 2004), the use of some ad hoc "rule-of-thumb" type methods, based upon the Student's t-statistic or Land's H-statistic has been suggested to compute the 95% UCLs, 95% UTLs, and 95% UPLs. For example, it is noted that the Student's t-statistic (on Cohen's maximum likelihood estimates) is used to compute the upper prediction limit (pages 10-26, USEPA-UGD (2004)) without justifying its use. It should be noted that the distribution of the statistic used to construct the UPL (pages 10-26, USEPA-UGD (2004)) is unknown as it is based upon maximum likelihood estimate (MLE) of the mean and standard deviation, which are very different from the traditional mean and standard deviation used in the definition of a typical Student's t-statistic. This "rule-of-thumb" method will be even harder to defend for skewed data sets with standard deviation of log-transformed data exceeding 1.0.

Let $x_1, x_2, ..., x_n$ (including NDs and detected measurement) represent n data values obtained from samples collected from an area of concern (AOC) or some other population of interest (e.g., reference area). It is assumed that k : $1 \le k \le n$, observations lie below one or more detection limit(s), and the remaining $(n - k)$ observations represent detected observations, which might come from a well-known parametric distribution such as a normal, a lognormal, or a gamma distribution, or from a population with a non-discernable distribution. Since multiple detection limits are allowed, some of the detected data may be smaller than the detection limit(s) associated with ND observations.

Note: *It should be noted that it is not easy to verify the distribution of left-censored data sets, especially when a large percentage (> 40% -50%) of observations are being censored (nondetected). In such cases, it is desirable to use nonparametric methods to compute various statistics (upper limits, hypotheses testing statistics) of interest. Several nonparametric methods to compute various upper limits based upon data sets with ND observations are available in ProUCL 4.0.*

ProUCL 4.0 also provides some ad hoc goodness-of-fit (GOF) methods to verify the data distribution based upon quantile-quantile (Q-Q) plot of the (*n - k*) detected observations supplemented with the available GOF test (e.g., as in ProUCL 3.0) statistics computed using the detected observations. This chapter also discusses the construction of graphical displays such as Q-Q plots that may be used to determine the distribution of data sets with ND observations. These methods have been incorporated in ProUCL 4.0. Additionally, using regression on order statistics (ROS) methods for normal, lognormal, or gamma distributions, one can extrapolate and estimate the NDs. This process yields a full data set of size n with *(n-k)* original detected observations, and *k* extrapolated nondetects. ProUCL 4.0 has the capability to store these data (*k* estimated and *(n-k)* original detected) in new columns (e.g., for use in other applications) generated by ProUCL 4.0. This option is available in ProUCL 4.0 for more advanced users. Some important notes and comments that apply to data sets with and without outliers are described as follows.

Note on skewness: *In ProUCL 3.0 and also in ProUCL 4.0, for skewed distributions such as a lognormal distribution (or some non-discernable distribution), skewness is measured as a function of sd, σ (or its estimate, $\hat{\sigma}$) of log-transformed data. Also, it should be noted out that sd, σ, or its estimate, $\hat{\sigma}$ of log-transformed detected data is used to get an idea about the skewness of a data set with ND observations. This information is useful to decide which UCL should be used to estimate the EPC term for data sets with and without ND observations. For data sets with ND values, output spreadsheets generated by ProUCL 4.0 do exhibit the sd, $\hat{\sigma}$, of log-transformed data based upon detected observations. For gamma distribution, skewness is a function of the shape parameter, k. Therefore, in order to assess the skewness of gamma distributed data sets, the associated output screens exhibit the MLE of the shape parameter, k based upon detected observations.*

4.2 Pre-processing a Data Set

Throughout this chapter (and in other chapters such as Chapters 2, 3, and 5), it has been implicitly assumed that the data set under consideration represents a “single” population. Specifically, it assumed that the user had pre-processed the data by performing preliminary and exploratory data analyses. This means that the user has separated out data sets from a potentially mixture sample (if any) into component sub-samples representing single individual population(s). The user may want to consult a statistician to perform this kind of population partitioning on a mixture sample. Some of exploratory graphical methods are also available in ProUCL 4.0

4.2.1 Handling of Outliers

The main objective of using a statistical procedure is to model the majority of the data representing the main dominant population, and not to accommodate a few outliers that may yield inflated and impractical results. The background values should be estimated by reliable statistics (and not distorted statistics) obtained using data sets representing the dominant background population. High outlying values contaminate the underlying left-censored or uncensored full data set from the population under study. In practice, it is the presence of a few extreme outliers that cause the rejection of normality of a data set (Barnett and Lewis (1994)).

4.2.2 Disposition of Outliers

One can argue against "not using the outliers" while estimating the various environmental parameters such as the EPC terms and BTVs. An argument can be made that the outlying observations are inevitable and can be naturally occurring (not impacted by site activities) in some environmental data sets. For an example, in groundwater applications, a few elevated values (occurring in tails of data distribution with lower probabilities) are naturally occurring, and as such may not be representing the impacted and contaminated monitoring well data values. Those data values originating from the groundwater studies may require separate investigation, and all interested parties should decide on how to deal with data sets that contain naturally occurring unimpacted outlying observations. The entire project team should decide about the appropriate disposition of such outlying observations. The project team should come to an agreement whether to treat the outlying observations separately or to include them in the computation of the various statistics of interest such as the sample mean, standard deviation, and the associated upper limits.

4.2.3 Assessing Influence of Outliers

In order to assess the influence of outliers on the various statistics (e.g., upper limits) of interest, it is suggested to compute all relevant statistics using data sets with outliers and without outliers, and compare the results. This extra step often helps to see the direct potential influence of outlier(s) on the various statistics of interest (e.g., mean, UPLs, UTLs). This in turn will help the project team to make informative decisions about the disposition of outliers. That is, the project team and experts familiar with the site should decide which of the computed statistics (with outliers or without outliers) represent better and more accurate estimate(s) of the population parameters (e.g., mean, EPC, BTV) under consideration. Since the treatment and handling of outliers is a controversial and subjective topic, it is suggested that the outliers be treated on a site-specific basis using all existing knowledge about the site and the site background (e.g., EA, area of concern (AOC), reference area) under investigation.

4.2.4 Log-Transformation Tends to Accommodate Outliers (and Contamination)

It should also be noted that often in practice, the use of a log-transformation on a data set tends to accommodate outliers and hide contaminated locations instead of revealing them. Specifically, an observation that is a potential outlier (representing a contaminated location) in the original raw scale may not seem to be an outlier in the log-scale. Once again since the cleanup and remediation decisions have to be performed using data and statistics in original scale, the use of a log-transformation should be avoided to achieve symmetry and normality. Instead, the use of nonparametric methods is preferable.

4.2.4.1 Avoid Data Transformations

It is suggested to avoid the use of transformations (Singh, Singh, Engelhardt (1997) and Singh, Singh, and Iaci (2002)) of the raw data to achieve symmetry (approximate normality). In most environmental applications, the cleanup decisions have to be made on statistics and results computed in the original space/scale as the cleanup goals need to be attained in the original scale. Therefore, statistics and results need to be back-transformed in the original scale before making any cleanup decisions. Furthermore, the parameter (hypothesis of interest) in the transformed space is not of interest to make remediation and cleanup decisions. Often, the back-transformed estimates of the parameters (from transformed space) in the original space suffer from a significant amount of transformation bias (e.g., see the results of the back-transformation of ROS estimates or MLE obtained using log-transformed data in Singh, Maichle, and Lee, 2006). Many times the transformation bias can be unacceptably large (for highly skewed data sets) and unreasonable, leading to incorrect decisions. Therefore, AVOID the use of transformations.

It is recommended to avoid the use of equation (4-2) as given below (Shaarawi, 1989) to back-transform estimates from log-scale to original scale. The question now arises – how one should back-transform results from a log-space (or any other transformed space) to the original space. Unfortunately, no defensible guidance or procedure is available in the literature to address this issue. Moreover, the back-transformation formula will change from transformation to transformation (Box-Cox (BC)-type transformations), and the bias introduced by such transformations will remain unknown. This is one of the main reasons that ProUCL 4.0 does not compute MLE estimates (or other estimates such as fully parametric estimates using ROS on logged data) using log-transformed data.

Therefore, in cases when a data set in the "raw" scale cannot be modeled by a parametric distribution, it is desirable to use nonparametric methods rather than testing or estimating some parameter(s) in the transformed space. This is especially true when dealing with data sets with nondetect observations.

4.2.4.2 Do Not Use DL/2 (t) UCL Method

The DL/2 goodness of fit tests and UCL computation methods are included for historical reasons. It is suggested that the use of DL/2 (t) method should be avoided to estimate the EPC term or other threshold values. This UCL computation method does not provide adequate coverage (for any distribution and sample size) for the population mean, even for censoring levels as low as 10%, 15%. This is contrary to the conjecture and assertion (e.g., EPA (2006)) often made that the DL/2 method can be used for lower ($\leq$ 20%) censoring levels. The coverage provided by the DL/2 (t) method deteriorates fast as the censoring intensity increases.

4.2.5 Minimum Data Requirement

If the use of appropriate data quality objectives (DQOs) (e.g., USEPA (2006)) is not possible (e.g., data might have been already collected), every effort should be made to collect a data set with about 8-10 detected observations to compute reasonably reliable estimates of EPC terms (UCLs) and BTVs (e.g., UPLs, UTLs). Whenever possible, it is desirable to collect more detected observations, especially when the percentage of NDs becomes greater than 40%, 50%, and so on. It should also be noted that the use of the minimum of 10 to 15 detected observations is desirable to compute UCLs and other upper limits based upon re-sampling bootstrap methods. These issues have also been discussed in Chapter 1 of this Technical Guide.

Most of the comments and notes described above apply to all data sets with and without nondetect observations.

4.3 Estimation of Population Mean and Variance Based Upon Left-Censored Data Sets

For left-censored data sets with NDs, the detailed description of the various methods to estimate the population mean, μ, and *sd*, σ, including the: MLE (Cohen, 1950, 1959) method, restricted MLE (RMLE) method (Person and Rootzen, 1977), expectation maximization (EM) method (Dempster, Laird, and Rubin, 1977, Gleit, 1985), EPA Delta lognormal method (EPA, 1991), Winsorization (Dixon and Tukey 1968, Gilbert, 1987) method, regression method (Newman, Dixon, and Pinder, 1989)), regression on order statistics (ROS) method (Gilliom and Helsel 1986, Helsel, 1990), and Kaplan-Meier (KM) method (1958) can be found in Singh and Nocerino (2001), and Singh, Maichle, and Lee (2006). Only two estimation methods, namely the ROS (normal, lognormal, and gamma) methods, and KM method are

described in this section. These estimation methods can be used on data sets with multiple detection limits.

The Q-Q plots and the GOF statistics based upon detected observations, and also based upon the data set consisting of detected observations and extrapolated (estimated) NDs obtained using ROS methods can be used to determine the data distribution of the left-censored data set under study. It should be noted that, it is not easy to verify the distribution of left-censored data sets, especially when a large percentage (> 40%-50%) of observations are being censored (nondetects). Therefore, it is preferable to use nonparametric methods to compute statistics of interest using data sets with ND observations.

For data sets with NDs, ProUCL 4.0 does provide some simple GOF methods based upon the ROS methods. Those simple ad hoc methods to verify the data distribution based upon the quantile-quantile (Q-Q) plot of the (*n* - *k*) detected observations supplemented with the available GOF test statistics computed using the detected observations are briefly described in the following sections.

4.3.1 Goodness-of-Fit (GOF) Tests for Data Sets with Nondetect Observations

Throughout this chapter, let $x_1, x_2, \ldots, x_n$ (including NDs and detected measurements) represent a random sample of *n* observations obtained from a population under investigation (e.g., background area, or an AOC). Out of the *n* observations, $k: 1 \le k \le n$, values are reported as nondetects lying below one or more detection limits, and the remaining $(n-k)$ observations represent the detected data values. The $(n-k)$ detected values are ordered and are denoted by $x_{(i)}\ ;\ i:=k+1, k+2,\ldots n$. The *k* nondetect observations are denoted by $x_{(ndi)\ ;}\ i:=1,2,\ldots k$. These *k* nondetect values can be predicted or extrapolated using a ROS method. The full data set of size *n* thus obtained may be used to compute the various summary statistics, and to estimate the EPC terms, BTVs, and other not-to-exceed values using methods described in Chapters 2 and 3 of this Technical guidance document. Several ROS methods have been cited and used in the literature for left-censored data sets (e.g., Helsel (2005)) with ND observations. However, it is pointed out that the use of ROS methods often yields infeasible and negative estimates of ND observations. This is especially true when outliers may be present in a data set (Singh and Nocerino, 2002). A description of ROS methods as incorporated in ProUCL 4.0 is given as follows.

It should be noted that the ROS estimation methods are parametric in nature as they involve extrapolation of the nondetects based upon certain distributional assumptions about the detected as well as nondetected observations. In this process, nondetects are imputed (extrapolated) based upon the assumed distribution (e.g., normal, lognormal, or gamma) of the detected observations. Using the menu option, “ROS Est. NDs”, ProUCL 4.0 can be used to store the extrapolated NDs along with the original detected values in additional columns generated by ProUCL 4.0. ProUCL 4.0 assigns suitable and self-explanatory titles for those generated columns.

One of the objectives here is to determine the data distributions data sets consisting of detected and nondetected observations. An ad hoc method to verify the data distribution is based upon the quantile - quantile (Q-Q) plot of the (*n* - *k*) detected observations supplemented with the available goodness-of-fit test statistics (e.g., S-W test, A-D test, and K-S test) computed using the detected observations. Methods described below (and incorporated in ProUCL 4.0) may be used on left-censored data sets with a single detection limit as well as multiple detection limits. Some of these methods such as ROS method can also handle cases when some of the detection limits (nondetects) exceed the observed detected values.

4.3.2 Regression on Order Statistics (ROS) Estimation Methods

Typically, in a ROS method, an ordinary least squares (OLS) regression model is obtained first by fitting a linear straight line to the *(n-k)* detected values, $x_{(i)}$ (perhaps after a suitable transformation) and the *(n-k)* hypothesized (e.g., normal, gamma) quantiles, $q_{(i)}$; $i:=k+1, k+2,...,n$ associated with those *(n-k)* detected observations. It should be noted that the hypothesized quantiles are obtained for all of the *n* data values by using a specified distribution such as a normal or a gamma distribution. The *(n-k)* quantiles associated with *(n-k)* detected values are denoted by $q_{(i)}$; $i:=k+1, k+2,...,n$, and the *k* quantiles associated with nondetect observations are denoted by $q_{(ndi)}$; $i:= 1, 2, \ldots, k$. The fitted linear model based upon *(n-k)* pairs is then used to predict or extrapolate the *k* nondetect observations. Obviously, in order to obtain a reliable model (slope and intercept) and extrapolated NDs, enough (at least 8-10) detected values should be available.

When there is only a single detection limit (DL) and all values lying below DL represent nondetect observations, then the quantiles corresponding to those nondetect values typically are lower than the quantiles associated with the detected observations. However, a different quantile position (percentile) computation method (e.g., see Helsel (2005)) is used to accommodate multiple detection limits and cases when some of the detected values may be smaller than some of the detection limits. Therefore, when multiple detection limits are present, the quantiles associated with some of the nondetect observations may exceed the quantiles associated with detected data values.

As mentioned before, in a ROS method, it is assumed that the *k* censored (nondetect) observations, $x_1, x_2, ..., x_k$, follow a normal (or lognormal distribution when logged data are used), gamma, or some other distribution. Before computing the *n* hypothesized quantiles, $q_{(i)}$; $i:=k+1, k+2,...,n$, and $q_{(ndi)}$; $i:= 1, 2, \ldots, k$, the plotting positions (also known as percentiles) need to be computed for all of *n* observations with *k* nondetects and *(n-k)* detected values. There are many methods available in the literature (Blom (1958), Johnson and Wichern (2002), Helsel (2005)) to compute the appropriate plotting positions (percentiles). Once the plotting positions (empirical percentiles) have been obtained, the associated quantiles are computed based upon the hypothesized distribution (e.g., normal, gamma) that needs to be tested. A few plotting position (percentile) computation methods as incorporated in ProUCL 4.0 are described in this section.

Once the *n* plotting positions (empirical probabilities, percentiles) have been computed, the *n* quantiles, $q_{(ndi)}$; $i:= 1, 2, \ldots, k$, and $q_{(i)}$; $i:=k+1, k+2,...,n$ are computed using the specified distribution (e.g., normal, gamma) corresponding to those *n* plotting positions. As mentioned before, when there are multiple detection limits, and when some of those detection limits exceed the detected values, then quantiles, $q_{(ndi)}$ corresponding to some of those nondetect values might become greater than the quantiles, $q_{(i)}$ associated with some of the detected values. The resulting quantiles when used to obtain an OLS regression often yield extrapolated nondetect values exceeding some of the detected values. Some of these issues are illustrated by examples in Chapter 6 of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b).

4.3.3 OLS Regression Line Based Upon Detected Data

An ordinary least squares (OLS) regression line is obtained using the $(n - k)$ pairs, $(q_{(i)}, x_{(i)})$; $i:= k + 1, k + 2, \ldots, n$, where $x_{(i)}$ are the (n-*k)* detected values arranged in ascending order. The *n* quantiles, $q_{(i)}$ are computed using the hypothesized distribution (e.g., normal, gamma). The OLS regression line fitted to the $(n - k)$ pairs $(q_{(i)}, x_{(i)})$; $i:= k + 1, k + 2, \ldots, n$ corresponding to the detected values is given by:

$$x_{(i)} = a + bq_{(i)};\ i := k+1, k+2, \ldots, n. \qquad (4\text{-}1)$$

It should be noted that in (4-1) above, the hypothesized quantile $q_{(i)}$ is associated with the detected value $x_{(i)}$*; i:=k+1, k+2,...n.* When ROS is used on transformed data (e.g., log-transformed data), then ordered values, $x_{(i)}$; i: = k + 1, k + 2, …, n represent ordered detected data in that transformed scale (e.g., log scale, Box-Cox (BC)-type transformation). Equation (4-1) is used to predict or extrapolate the nondetect values in original or log-transformed space. Specifically, for quantile, $q_{(ndi)}$ corresponding to the ith nondetect, the extrapolated nondetect is given by $x_{(ndi)} = a + bq_{(ndi)}$; *i:=1,2,...k.* As mentioned before, some of the predicted nondetect values thus obtained may become negative, or may exceed some of the detection limits and detected values.

4.3.4 Computation of the Plotting Positions (Percentiles) and Quantiles

Before computing the hypothesized (e.g., normal, gamma) quantiles, one has to compute the percentiles or plotting positions for the detected as well as nondetect observations. Those plotting positions are then used to compute the quantiles based upon the distribution used (e.g., normal, gamma). Several methods are available in the literature (e.g., Johnson and Wichern (2002), Helsel (2005)) that can be used to compute the *n* percentiles or *n* plotting positions.

For a data set of size n, the most commonly used plotting positions for the ith observation (ordered) are given by $(i - 3/8) / (n + 1/4)$ or $(i - 1/2)/n$. Typically, these plotting positions are used when one wants to graph a Q-Q plot of full-uncensored data set or wants to use ROS method on a data set with a single detection limit, DL with all nondetect values lying below the detection limit DL. For the single DL case (with all of the nondetect observations below the DL), ProUCL 4.0 uses percentiles, $(i - 3/8) / (n + 1/4)$ for normal and lognormal distribution, and uses empirical percentiles given by $(i - 1/2)/n$ for a gamma distribution.

Once, the plotting positions have been obtained, the n normal quantiles, $q_{(i)}$ are computed using the probability statement: $P(Z \le q_{(i)}) = (i - 3/8) / (n + 1/4)$, i : = 1, 2, …, n , where Z represents a standard normal variable (SNV). The gamma quantiles are computed using the probability statement: $P(X \le q_{(i)}) = (i - 1/2) /n$, i : = 1, 2, …, n , where X represents a gamma random variable. The details of the computation of the gamma quantiles are also given in the ProUCL 3.0 Technical Guide (EPA, 2004).

In case multiple detection limits are present, perhaps with some of the detected values smaller than some of the detection limits (or nondetects exceeding some of the detected values), the plotting positions (percentiles) are computed using appropriate methods that adjust for multiple detection limits. The details of the computation of such plotting positions (percentiles), p_i; *i: =1, 2, ... ,n,* for data sets with multiple detection limits or with nondetect observations exceeding the detection limits are given in Helsel (2005). The associated hypothesized quantiles, $q_{(i)}$ are obtained by using the following probability statements:

$$P\ (Z \le q_{(i)}) = p_i;\ i := 1, 2, \ldots, n \text{ (Normal or Lognormal Distribution)}$$

$$P\ (X \le q_{(i)}) = p_i;\ i := 1, 2, \ldots, n \text{ (Gamma Distribution)}$$

4.3.5 ROS Estimates Obtained Using Only Detected Observations

For full data sets, Barnett (1976) used the intercept and the slope of the regression line to estimate the population mean and the standard deviation. Newman, Dixon, and Pinder (1989) followed a similar approach (using normal quantiles), and used the intercept and the slope of the OLS line given by (4-1) to estimate population mean, μ, and standard deviation, σ, using left-censored data sets. It is noted that these

estimates ignore all of the information contained in nondetect observations. Singh and Nocerino (2002) noted that the use of this method using only (*n - k*) detected values results in biased estimates of the mean and standard deviation. Therefore, this method is not a recommended method to compute summary statistics and various upper limits used in environmental applications.

4.3.5.1 ROS Method for Normal Distribution

Once the OLS regression model given by (4-1) has been obtained based upon the *(n-k)* detected observations and the associated hypothesized quantiles (e.g., normal, lognormal, or gamma), the *k* nondetect observations can be predicted or extrapolated using the linear model given by (4-1). The use of this process yields a full data set of *n* values with *k* imputed new values, and *(n-k)* detected original data values. One may then want to use any of the methods including the bootstrap resampling methods as described in Chapters 2 and 3 to estimate the EPC terms, BTVs, and other not-to-exceed values.

It should be pointed out that the use of ROS methods often results in infeasible estimates of nondetect observations. Specifically, the imputed nondetects become negative (e.g., when using ROS on a normal distribution), larger than the detection limits, and even larger than the detected values. This is especially true when some outlying observations may be present in the data set. The occurrence of outliers distorts all statistics including the slope and intercept of the linear OLS fit, and all extrapolated nondetect values. The use of distorted predicted nondetect observations in turn yields biased estimates of the population mean (e.g., reference area), standard deviation, and all other statistics of interest including UPLs and UTLs. *Therefore, the use of ROS method assuming a normal distribution is not recommended.*

In such situations with infeasible estimates of NDs, some subjective checks may be provided to modify the regression method: negative estimates of NDs may be replaced by DL/2, and the estimated nondetects greater than DL may be replaced by DL itself. The mean and variance are then computed using the replacement values. Singh and Nocerino (2002) studied this method in their simulation study and concluded that the modified regression method yields biased estimates of population mean and variance. Therefore, this modified ROS method for normal distribution (or for any other distribution) is not available in ProUCL 4.0.

4.3.5.2 ROS Method for Lognormal Distribution

For the ROS method on log-transformed data, the OLS model given by (4-1) is obtained using the log-transformed detected data and the corresponding normal quantiles. Let *Org* stand for the data in the original unit and *Ln* stand for the data in the natural log-transformed unit. Using the OLS linear model on log-transformed detected observations, the nondetects in transformed log-units are obtained by extrapolation corresponding to the *k* normal quantiles, $q_{(ndi)}$ associated with ND observations.

Once the *k* nondetects have been estimated, the sample mean and *sd* can be computed using the back-transformation formula (Shaarawi, 1989) given by equation (4-2) below. This method is also known as fully parametric method (Helsel, 2005). The mean, $\hat{\mu}_{Ln}$, and *sd,* $\hat{\sigma}_{Ln}$, are computed in log-scale using a full data set obtained by combining the (*n - k*) detected log-transformed data values and the *k* extrapolated nondetect (in log scale) values. Assuming, lognormality, El-Shaarawi (1989) estimated μ and σ by back-transformation using the following equations as one of the several ways of computing these estimates. Note that these estimates suffer from significant amount of transformation bias as can be seen in examples discussed in Singh, Maichle, and Lee (2006). The estimates given by equation (4-2) are neither unbiased nor have the minimum variance (Gilbert (1987)). Therefore, it is recommended to avoid

the use of this version of ROS method on log-transformed data (also called the fully parametric ROS method, Helsel, 2005) to compute UCL95 and various other statistics.

$$\hat{\mu}_{Org} = \exp(\hat{\mu}_{Ln} + \hat{\sigma}^2_{Ln}/2), \text{ and } \hat{\sigma}^2_{Org} = \hat{\mu}^2_{Org}(\exp(\hat{\sigma}^2_{Ln}) - 1) \quad (4\text{-}2)$$

Note: *As mentioned before, it is recommended to avoid the use of equation (4-2) as given in Shaarawi (1989) to back-transform estimates from log-scale to original scale. Often, the back-transformed estimates of the parameters (from transformed space) in the original space suffer from a significant amount of transformation bias (e.g., see the results of back-transformation of ROS estimates or MLEs obtained using log-transformed data in Singh, Maichle, and Lee, 2006). Many times the transformation bias can be unacceptably large (for highly skewed data sets) and unreasonable, leading to incorrect decisions.*

It is also noted that the same formula given by equation (4-2) above is used on other parametric estimates (e.g., MLE, EM, RMLE, robust MLE) obtained in log-scale (assuming a lognormal distribution) to back-transform them in the original scale. Therefore, the MLE and all other parametric estimates obtained using a lognormal distribution suffer from significant amount of back-transformation bias (e.g., Singh, Maichle, and Lee (EPA, 2006)). The use of such estimates obtained using some kind of back-transformation is not recommended. Instead, the use of nonparametric methods (e.g., KM and bootstrap methods) on data sets in the original scale is preferred to compute estimates of environmental parameters including BTVs and EPC terms.

This is one of the main reasons that ProUCL 4.0 does not compute MLE estimates (or other estimates such as fully parametric estimates using ROS method or EM method on logged data) using log-transformed data, as those estimates do require back-transformation based upon equations given by (4-2) above. Therefore, in cases when a data set in the "raw" scale cannot be modeled by a parametric distribution, it is desirable to use nonparametric methods rather than testing or estimating some parameter(s) in the transformed space. This is especially true when dealing with data sets with nondetect observations.

4.3.5.3 Robust ROS Method on Log-Transformed Data

Robust ROS method is also performed on log-transformed data as described above. In this robust ROS method (Helsel, 2005), nondetect observations are first extrapolated (predicted) in log scale based upon a linear ROS model fitted to the log-transformed detected values and normal quantiles. The estimated nondetects are transformed back in the original scale by simple exponentiation. It should be noted that, the process of using the ROS method based upon a lognormal distribution and imputing NDs by exponentiation will not yield negative estimates for nondetect values. However, this process still may yield some infeasible nondetects with some of the NDs exceeding their respective detection or reporting limits, and the estimated NDs exceeding the detected values. In any case, this process yields a full data set of size n. Any of the methods as described in Chapters 2 and 3 (and available in ProUCL 4.0) can be used to compute estimates of EPC terms, BTVs, not-to-exceed, and all other statistics of interest.

Using the "ROS Est. with NDs" Option of ProUCL 4.0, one can save and store the extrapolated NDs along with the original detected values in additional columns generated by ProUCL 4.0. ProUCL 4.0 assigns suitable self-explanatory titles for those generated columns.

4.3.5.4 Gamma ROS Method

Many positively skewed data sets follow a lognormal as well as a gamma distribution. Singh, Singh, and Iaci (2002) noted that gamma distributions are better suited to model positively skewed environmental data sets. For full data sets, it is observed that the use of a gamma distribution results in reliable and stable 95% *UCL* values. It is also noted that in order to use a gamma distribution, there is no need to transform the data and back-transform the resulting statistics. However, one has to estimate the gamma shape and scale parameters before computing the gamma quantiles and estimating (predicting) the NDs using a gamma distribution for detected data values. This process may have some effect on the adequacy and accuracy of the estimated gamma quantiles (based upon estimated value of the shape parameter, *k*), and consequently on the accuracy of the extrapolated NDs. Just like all other distributions, outliers, when present, can distort all statistics including slope, intercept, extrapolated NDs, mean, *sd*, UCL95, UPLs, and percentiles.

If a left-censored data set follows a gamma distribution (can be verified using goodness-of-fit tests in ProUCL using any of the two empirical distribution function (EDF) tests – the K-S test or the A-D test on detected data); then, those NDs can be extrapolated using the regression model based upon (*n* - *k*) pairs given by: ((*n* - *k*) gamma quantiles, ordered (*n* - *k*) detected observations). As in normally distributed left-censored data sets, for gamma distributed left-censored data sets, one can fit a linear regression model to (*n* - *k*) pairs: ((*n* - *k*) gamma quantiles, (*n* - *k*) ordered detected values) as given by formula (4-1). In these *(n-k)* pairs, the gamma quantile, $q_{(i)}$ is associated with the i^{th} detected observations (arranged in ascending order).

Specifically, let $x_{nd1}, x_{nd2}, \ldots, x_{ndk}, x_{k+1}, x_{k+2}, \ldots, x_n$, be a random sample (with *k* ND observations and *(n-k)* detected values) of size n from the gamma distribution, $G(k,\theta)$. Note that some of the ND values may be greater than the detected values. It is pointed out once again that in the data spreadsheet for ProUCL 4.0, all NDs, or <DL values are entered as the respective DL value (multiple DLs are allowed); and a column of 0 (associated with a nondetect) and 1 (associated with a detected value) value is assigned to each such contaminant (variable) with ND values. The n plotting positions, p_i; i:= 1,2,…,n are computed for each observation (detected and nondetected) using the methods (for single detection limit and multiple detection limit cases) as described earlier in this chapter. One should not get confused with *k*, the shape gamma parameter, which is different from *k*, the number of NDs as used above.

Let $x_{(k+1)} \le x_{(k+2)} \le \ldots \le x_{(n)}$ represent the ordered detected values (some of the DLs may be lying in between these *(n-k)* ordered detected values). In order to compute *n* gamma quantiles associated with the n plotting positions (percentiles, empirical probabilities), one needs to estimate the gamma parameters, *k* and θ. For left-censored data sets, these parameters are estimated based upon *(n-k)* detected values. Obviously enough detected data (e.g., at least 8-10) values are needed to compute reliable estimates of *k* and θ. Let $\hat{k}$ and $\hat{\theta}$ represent the maximum likelihood estimates (*MLEs*) of *k* and θ respectively. For details of the computation of *MLEs* of *k* and θ, refer to Singh, Singh, and Iaci (2002). Just like all other ROS methods, in order to be able to compute reliable estimates of the nondetects and the resulting upper limits, enough detected observations (> 8-10) should be available.

The Q-Q plot for a gamma distribution is obtained by plotting the scatter plot of pairs $(x_{0i}, x_{(i)})$ $i := k+1, 2, \ldots, n$, where *k*=number of nondetects. The *n* quantiles, x_{0i}, are given by the equation, $x_{0i} = z_{0i}\hat{\theta}/2;\ i := 1, 2, \ldots, n$, where the quantiles z_{0i} (already ordered) are obtained by using the inverse chi-square distribution given as follows.

$$\int_0^{z_{0i}} f(\chi^2_{2\hat{k}})d\chi^2_{2\hat{k}} = (i-1/2)/n; \quad i := 1, 2, \ldots, n \quad \text{(Single DL Case)} \tag{4-3}$$

$$\int_0^{z_{0i}} f(\chi^2_{2\hat{k}})d\chi^2_{2\hat{k}} = p_i; \quad i := 1, 2, \ldots, n \quad \text{(Multiple DL Case)} \tag{4-4}$$

In the above equation, $\chi^2_{2\hat{k}}$ represents a chi-square random variable with $2\hat{k}$ degrees of freedom (*df*), and p_i are the plotting positions (percentiles) obtained using the process described in Singh, Maichle, and Lee (2006) and Helsel (2005). The process of computing plotting positions, p_i , *i:=1,2,...,n,* for left-censored data sets with multiple detection limits has been incorporated in ProUCL 4.0. The program, PPCHI2 (Algorithm AS91) from Best and Roberts (1975) has been used to compute the inverse chi-square percentage points, z_{0i}, as given by the above equation.

Using a linear regression model given by equation (4-1) fitted to the (*n* - *k*) pairs, (gamma quantiles associated with detected data, detected data), one can extrapolate the *k* NDs contained in the data set. This will yield a full data set of size *n* = *k* + (*n* - *k*).

It is noted that in order to use gamma ROS method, one has to estimate the gamma shape and scale parameters before computing the gamma quantiles and estimating (predicting) the NDs using a gamma distribution for detected data values. This process may have some effect on the adequacy and accuracy of the estimated gamma quantiles (based upon estimated value of the shape parameter, *k*), and consequently on the accuracy of the extrapolated NDs. Obviously, in order to obtain reasonably reliable estimates of gamma parameters, enough (e.g., > 8-10, more are desirable) detected data should be made available.

Note: *On data sets obtained using a ROS method, any of the available methods (including bootstrap methods) for full data sets may be used to compute estimates of BTVs and not-to-exceed values, and all other statistics of interest. These ROS methods are specifically useful when the BTVs and not-to-exceed values are to be estimated by percentiles. However, when the BTVs need to be estimated using UPLs and UTLs for left-censored data sets, one may want to use direct bootstrap methods on left-censored data sets.*

The fitted OLS linear model as given by (4-1) based upon (n-k) pairs is used to predict or extrapolate the k nondetect observations. Therefore, in order to obtain a reliable model (slope and intercept) and extrapolated NDs, enough (at least 8-10) detected values should be available.

It is important to point out that in practice, it is not reasonable or justifiable to assume or expect that the NDs should follow the same distribution as assumed for the detected values. This is especially true when multiple detection limits may be present, and the % of NDs may be quite high (e.g., > 40%). Additionally, ROS methods often yield infeasible predicted NDs such as negative values, or greater than the respective detection limits.

It is also noted that in order to use a gamma distribution, there is no need to transform the data and back-transform the resulting statistics. However, one has to estimate the gamma shape and scale parameters before computing the gamma quantiles and estimating (predicting) the NDs using a gamma distribution for detected data values. This process may have some effect on the adequacy and accuracy of the estimated gamma quantiles (based upon estimated value of the shape parameter, k), and consequently on the accuracy of the extrapolated NDs. Just like all other distributions, outliers, when present, can distort all statistics including slope, intercept, extrapolated NDs, mean, sd, UCL95, UPLs, and percentiles.

4.4 Saving Extrapolated Nondetect Values Using ROS Est. with NDs Option of ProUCL 4.0

Using this option, nondetects are imputed (extrapolated) based upon the assumed distribution (e.g., normal, lognormal, or gamma) of the detected observations. Using the menu "ROS Est. NDs," ProUCL 4.0 can be used to save and store the extrapolated NDs along with the original detected values in additional columns generated by ProUCL 4.0. ProUCL 4.0 assigns suitable self-explanatory titles for those generated columns. This is available in ProUCL 4.0 for advanced users who want to use full data sets consisting of detected and extrapolated nondetect observations for other applications (e.g., Analysis of Variance, Principal Component Analysis).

4.5 Influence of Outliers on ROS methods

This section briefly discusses the influence of potential outliers on the various statistics of interest including the UCLs, UPLs, and UTLs. Singh and Nocerino (2002) demonstrated that, the classical MLE methods and the various ROS approaches (on raw or log-transformed data) do not perform well in the presence of outliers. The estimates obtained using the classical methods in the original scale or in log-transformed scale get distorted by outliers. This results in distorted estimates of intercept (population mean) and slope (*sd*), which gives rise to infeasible extrapolated nondetects. For example, the estimated nondetects can become negative (when dealing with raw data), larger than the detection limit (DL), and even larger than some of the observed values (e.g., $x_{(k)}$). The use of such extrapolated NDs results in biased estimates of the population mean and *sd*. Conclusions derived using distorted statistics, UCLs, and UPLs can be incorrect and misleading. In these situations, subjective checks may be provided to modify the regression method: negative estimates of NDs may be replaced by DL/2, and the estimated nondetects greater than DL may be replaced by DL itself. The mean and variance are then computed using these replacement values. Singh and Nocerino (2002) considered this method in their simulation study, and concluded that the modified regression method also yields biased estimates of population mean and variance.

4.6 Nonparametric Kaplan-Meier (KM) Estimation Method

The Kaplan-Meier (KM) (1958) estimation method, also known as the product limit estimate (PLE), is based upon a statistical distribution function estimate, like the sample distribution function, except that this method adjusts for censoring. The KM method is quite popular in survival analysis (dealing with right-censored data such as dealing with terminally ill patients) and various other medical applications. Some practitioners (e.g., Helsel 2005) are also recommending the use of KM method when dealing with left-censored environmental data sets. For data sets with ND observations with multiple detection limits, the KM estimation method has been incorporated in ProUCL 4.0. A brief description of the KM method to estimate the population mean and standard deviation, and standard error (SE) of the mean for left-censored data sets is described in this section. For details, refer to Kaplan-Meier (1958) and the report by prepared by Bechtel Jacobs Company for DOE (2000). It should be noted that the KM method has an added advantage as it can be used on data sets with multiple detection limits.

Let $x_1, x_2, ..., x_n$ (detection limits or actual measurement) represent n data values obtained from samples collected from an area of concern (AOC), and let $x'_1 < x'_2, \ldots < x'_n$ denote the n' distinct values at which detects are observed. That is, n' ($\leq n$) represents distinct observed values in the collected data set of size n. For $j = 1, \ldots, n'$, let m_j denote the number of detects at x'_j and let n_j denote the number of xi $\leq x'_j$. Also, let $x(1)$ denote the smallest x_i. Then

$$\tilde{\tilde{F}}(x) = 1, \qquad x \le x'_{n'}$$

$$\tilde{\tilde{F}}(x) = \prod_{j \ \ such \ \ that \ \ x'_j > x} \frac{n_j - m_j}{n_j}, \qquad x'_1 \le x \le x'_{n'}$$

$$\tilde{\tilde{F}}(x) = \tilde{\tilde{F}}(x'_1), \qquad x(1) \le x \le x'_1$$

$$\tilde{\tilde{F}}(x) = 0 \text{ or undefined}, \qquad 0 \le x \le x(1)$$

Note that in the last equality statement of $\tilde{\tilde{F}}(x)$ above, $\tilde{\tilde{F}}(x) = 0$ when $x(1)$ is a detect, and is undefined when $x(1)$ is a nondetect. The estimation of the population mean using the KM method is described as follows.

$$\hat{\mu} = \sum_{i=1}^{n'} x'_i [\tilde{\tilde{F}}(x'_i) - \tilde{\tilde{F}}(x'_{i-1})], \text{ with } x_0 = 0 \qquad (4\text{-}5)$$

Using the PLE (or KM) method as described above, an estimate of the standard error (SE) of the mean can be obtained. This is given by the following equation.

$$\hat{\sigma}^2_{SE} = \frac{n-k}{n-k-1} \sum_{i=1}^{n'-1} a_i^2 \frac{m_{i+1}}{n_{i+1}(n_{i+1} - m_{i+1})}, \qquad (4\text{-}6)$$

where k = number of observation below the detection limit and

$$a_i = \sum_{j=1}^{i} (x'_{j+1} - x'_j) \tilde{\tilde{F}}(x'_j), \ i: =1, 2, \ldots, n'\text{-}1. \qquad (4\text{-}7)$$

As mentioned before, some researchers (e.g., Helsel 2005) have suggested that this method perhaps is the most appropriate to compute the sample mean, SE, and a UCL95 for left-censored data sets. Helsel (2005) also suggested using the percentile bootstrap method on the KM estimate of the mean to compute a 95% UCL of the population mean.

Using the KM estimates of the mean and the SE of the mean, some investigators have also suggested using a normal distribution-based cut-off value (Helsel, 2005) or a Student's t-distribution-based cut-off value (Bechtel, 2002) to compute a 95% UCL of the mean. Specifically, using a t cut-off value, a 95% UCL of the mean based upon the KM estimates is given by the following equation.

$$\text{UCL95} = \hat{\mu} + t_{0.95,(n-1)} \sqrt{\hat{\sigma}^2_{SE}} \qquad (4\text{-}8)$$

ProUCL 4.0 computes a 95% UCL of the mean based upon the PLE method (KM method) using: 1) the normal approximation based upon standard normal critical values, z_α; 2) several of the bootstrap methods, including the percentile bootstrap method and the bias-corrected accelerated (BCA) bootstrap method, and 3) the Chebyshev inequality. It is noted that the approximate KM-UCL95 based upon the normal

approximation does not provide adequate coverage to the mean of non-normal skewed populations. For details of the findings and results of the simulation experiments, refer to Singh, Maichle, and Lee (2006).

4.7 Bootstrap UCL Computation Methods for Left-Censored Data Sets

The use of bootstrap methods has become popular with easy access to fast personal computers. For full-uncensored data sets, repeated samples of size n are drawn with replacement (that is each x_i has the same probability = $1/n$ of being selected in each of the N bootstrap replications) from the given set of observations (as described in Chapter 2). The process is repeated a large number of times, N (e.g., 1000-2000), and each time an estimate, $\hat{\theta}$ of θ (the mean, here) is computed. The estimates thus obtained are used to compute an estimate of the standard error of the estimate, $\hat{\theta}$. Just as for the full-uncensored data sets without any NDs, for censored data sets also, the bootstrap resamples are obtained with replacement. However, an indicator variable, I (0 = detected value, and 1 = nondetected value), is tagged to each observation in a bootstrap sample. This process is described as follows.

Singh, Maichle, and Lee (EPA, 2006) studied the performances (in terms of coverage probabilities) of four bootstrap methods (e.g., see Efron and Tibshirani 1993) to compute UCL95 for data sets with BDL observations. The four bootstrap methods include standard bootstrap method, bootstrap t method, percentile bootstrap method, and the bias-corrected accelerated (BCA) bootstrap method (Efron and Tibshirani, 1993, and Many, 1997). These methods are specifically useful when the exact distributions of the statistics used or needed (e.g., Cohen's MLE, RMLE) are not known; or the critical values are not available; the test statistics and their distribution depend upon the unknown number of nondetects, k that might be present in a data set. For left-censored data sets, the exact (and even approximate) distribution of the test statistics and the associated critical values needed to compute the various upper limits (UCL, UPL, and UTL) are not known (Kroll, 1996). Some authors (Helsel, 2005) have suggested the use of bootstrap methods on full data sets obtained using extrapolated NDs based upon ROS methods, especially the robust ROS method. One can also use bootstrap resampling methods directly on left-censored data sets. Some bootstrap methods (as incorporated in ProUCL 4.0) to compute upper limits based upon left-censored data sets are briefly discussed in this section. The details can be found in Singh, Maichle, and Lee (EPA, 2006).

4.7.1 Bootstrapping Data Sets with Nondetect Observations

Formally, let $x_{nd1}, x_{nd2}, \ldots, x_{ndk}, x_{k+1}, x_{k+2}, \ldots, x_n$ be a random sample of size n from a population (e.g., AOC, or background area) with an unknown parameter θ such as the mean, μ, or the p^{th} upper percentile, x_p, that needs to be estimated (e.g., by the corresponding sample percentile, or by a $(1 - \alpha)*100\%$ UCL for the p^{th} upper percentile = UTL) based upon the sampled data set with ND observations. As before, the sample is left-censored with k nondetect observations, and $(n - k)$ detected data values. Let $\hat{\theta}$ be an estimate of θ, which is a function of k nondetect and $(n - k)$ detected observations. For example, the parameter, θ, could be the population mean, μ, and a reasonable choice for the estimate, $\hat{\theta}$, might be the MLE, robust ROS, gamma ROS, or KM estimate (as discussed earlier) of the population mean. If the parameter, θ, represents the p^{th} upper percentile, then the estimate, $\hat{\theta}$, may represent the p^{th} sample percentile, $\hat{x}_p$, based upon a full data set obtained using one of the ROS method described above.

The bootstrap procedure on a censored data set is similar to the general bootstrap technique used on uncensored data sets. The only difference is that an indicator variable, I (taking only two values: 0 and 1), is assigned to each observation (detected or nondetected) when dealing with left-censored data sets (e.g.,

see Efron, 1981 and Barber and Jennison, 1999). The indicator variables, $I_{j:}\ j:=1,2,...,n$ is associated with the detection status of the sampled observations, x_j ; $j: = 1, 2,..., n$. Just like simple bootstrap samples, a large number, N (e.g., 1000, 2000) of two-dimensional bootstrap resamples, $(x_{iJ}, I_{iJ}), j:= j: = 1, 2,..., N$, and $i: = 1, 2,..., n$, of size n are drawn with replacement. The indicator variable, I, takes on a value = 1 when a detected value is selected and $I = 0$ if a nondetected value is selected in a bootstrap re-sample. The two-dimensional bootstrap process keeps track of the detection status of each observation in a bootstrap re-sample. In this setting, the detection limits are fixed, and the number of nondetects may vary from bootstrap sample to bootstrap sample. There may be k_1 nondetects in the first bootstrap sample, k_2 nondetects in the second sample, ..., and k_N nondetects in the N^{th} bootstrap sample. Since the sampling is conducted with replacement, the number of nondetects, k_i, i: = 1, 2, ..., N, can take any value from 0 to n inclusive. This is typical of a Type I left-censoring bootstrap process. On each of the N bootstrap resample, one can use any of the nondetect estimation methods (e.g., KM, MLE, ROS) to compute the statistics of interest (e.g., mean, *sd*, upper limits) using the methods as described in Singh, Maichle, and Lee (EPA, 2006).

It should be noted there is a positive chance that all (or most) observations in a bootstrap resample are equal. This is specifically true, when one is dealing with small data sets. In order to avoid such situations (with all values in a bootstrap sample to be the same), it desirable to have at least 10-15 (preferably more) detected observations in a left-censored data set. It should also be pointed out that it is not advisable and desirable to compute statistics based upon a bootstrap resample consisting of only a few detected values such as < 4-5. Therefore, it is suggested that all bootstrap resamples consisting of fewer than 4-5 detected values should not be used in the computation of the various statistics of interest (e.g., summary statistics, upper limits). Bootstrap procedures as incorporated in ProUCL 4.0 use this convention; that is bootstrap resamples with less than 4-5 detected values have not been included (in such cases additional resamples are drawn) in the estimation process.

Let $\hat{\theta}$ be an estimate of θ based upon the original left-censored data set of size n. For an example, if the parameter, θ, represents the population mean, then a reasonable choice for the estimate, $\hat{\theta}$, can be Cohen's MLE mean, ROS mean, or Kaplan-Meier (KM) mean. Similarly, calculate the standard deviation (*sd*) using one of these methods for left-censored data sets. The following two steps are common to all of the bootstrap methods incorporated in ProUCL 4.0

Step 1. Let $(x_{i1}, x_{i2}, ..., x_{in})$ represent the i^{th} bootstrap resample of size n with replacement from the original left-censored data set $(x_1, x_2, ..., x_n)$. Note that an indicator variable (as mentioned above) is tagged along with each data value, taking values 1 (if a detected value is chosen) and 0 (if a nondetect is chosen in the resample). Compute an estimate of the mean (e.g., MLE, KM, and ROS) using the i^{th} bootstrap resample, i: = 1, 2, ..., N.

Step 2. Repeat Step 1 independently N times (e.g., $N = 2000$), each time calculating new estimates (e.g., KM estimates) of population mean. Denote these estimates (e.g., MLE, KM means, and ROS means) by $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N$. The bootstrap estimate of the population mean is given by the arithmetic mean, $\bar{x}_B$, of the N estimates $\bar{x}_i$ (N MLEs or N KM means). The bootstrap estimate of the standard error is given by:

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{x}_i - \bar{x}_B)^2}\ . \qquad (4\text{-}9)$$

In general, a bootstrap estimate of θ may be denoted by $\overline{\theta}_B$ (instead of $\overline{x}_B$). The estimate, $\overline{\theta}_B$ is the arithmetic mean of the *N* bootstrap estimates (e.g., KM mean, or MLE mean) given by $\hat{\theta}_i$, *i:=1,2,…N*. It is pointed out that the N bootstrap estimates are computed in the similar way as the original estimate, $\hat{\theta}$ (e.g., KM or MLE mean) of the parameter, θ. Note that if the estimate, $\hat{\theta}$ represents the KM estimate of, θ, then $\hat{\theta}_i$ (denoted by $\overline{x}_i$ in the above paragraph) also represents the KM mean based upon the i[th] bootstrap resample. The difference, $\overline{\theta}_B - \hat{\theta}$, provides an estimate of the bias of the estimate, $\hat{\theta}$. After these two steps, a bootstrap procedure (BCA, bootstrap t) is similar to a conventional bootstrap procedure used on a full data set as described earlier in Chapter 2, and in ProUCL 3.0 Technical Guide (2004).

A brief description of the various bootstrap UCL computation methods for left-censored data sets is given in the following subsections.

4.7.1.1 UCL of Mean Based Upon Standard Bootstrap Method

Once the desired number of bootstrap samples, have been obtained following the two steps described above, a UCL of mean based upon the standard bootstrap method can be computed as follows. The standard bootstrap confidence interval is derived from the following pivotal quantity, t:

$$t = \frac{\hat{\theta} - \theta}{\hat{\sigma}_B}. \qquad (4\text{-}10)$$

A (1 – α)100% standard bootstrap UCL for θ is given as follows:

$$\text{UCL} = \hat{\theta} + z_{\alpha}\hat{\sigma}_B \qquad (4\text{-}11)$$

Here z_α is the upper α^{th} critical value (quantile) of the standard normal distribution (SND). It is observed that the standard bootstrap method does not adequately adjust for skewness, and the UCL given by the above equation fails to provide the specified (1–α)100% coverage to the population mean of skewed (e.g., lognormal and gamma) data distributions.

4.7.1.2 UCL of Mean Based Upon Bootstrap t Method

This main process is similar to the bootstrap t method as described in Chapter 2 for full-uncensored data sets without NDs. A (1–α)100% UCL of the mean based upon the bootstrap t method is given as follows.

$$\text{UCL} = \overline{x} - t_{(\alpha N)}\frac{s_x}{\sqrt{n}} \qquad (4\text{-}12)$$

It should be noted that the mean and *sd* used in the above equation represent estimates (e.g., KM estimates, MLE estimates) obtained using data set with ND observations. Similarly, the t-cutoff value used in the above equation is used using the pivotal t-values based upon KM estimates or some other estimates obtained using data sets with NDs. Typically, for skewed data sets (e.g., gamma, lognormal), the 95% UCL based upon the bootstrap t method performs better than the 95% UCLs based upon the simple percentile and the BCA percentile methods. However, it should be pointed out that the bootstrap t

method sometimes results in unstable and erratic UCL values, especially in the presence of outliers (Efron and Tibshirani (1993)) or when the data set may appear to look skewed (perhaps due to the presence of contaminated observations). Therefore, the bootstrap t method should be used with caution. In case this method results in erratic unstable UCL values, the use of an appropriate Chebyshev inequality-based UCL is recommended. Additional suggestions on this topic are described in Chapter 2.

4.7.1.3 Percentile Bootstrap Method

In Chapter 2, one can find the description of the percentile bootstrap method. For left-censored data sets, sample means are computed for each bootstrap samples using a specified method (e.g., MLE, KM, ROS), which are arranged in ascending order. The 95% UCL of the mean is the 95th percentile and is given by:

$$95\%\ Percentile - UCL = 95^{th}\%\ \bar{x}_i\ ;\ i{:}\ = 1, 2, ..., N \tag{4-13}$$

For example, when N = 1000, a simple 95% percentile-UCL is given by the 950th ordered mean value given by $\bar{x}_{(950)}$. It is observed that for skewed (lognormal and gamma) data sets, the BCA bootstrap method performs (described below) slightly better (in terms of coverage probability) than the simple percentile method.

4.7.1.4 Bias-Corrected Accelerated (BCA) Percentile Bootstrap Procedure

It is observed (Singh, Maichle, and Lee, 2006) that for skewed data sets, the BCA method does represent a slight improvement (in terms of coverage probability) over the simple percentile method. However, for moderately skewed to highly skewed data sets with the *sd* of log-transformed data >1, this improvement is not adequate enough and yields UCLs with a coverage probability lower than the specified coverage of 0.95. The BCA upper confidence limit of the intended (1-α) coverage for a selected estimation method (e.g., MLE, KM, Chebyshev, and so on) is given by the following equation:

$$(1-\alpha)100\%\ UCL_{PROC} = BCA - UCL = \bar{x}_{PROC}^{\alpha_2} \tag{4-14}$$

Here $\bar{x}_{PROC}^{\alpha_2}$ is the $\alpha_2 100^{th}$ percentile of the distribution of statistics given by $\bar{x}_{PROC}$; $i{:}\ = 1, 2, ..., N$, and PROC is one of the many (e.g., MLE, KM, DL/2, Chebyshev, and so on) mean estimation methods included in this simulation study.

Here α_2 is given by the following probability statement:

$$\alpha_2 = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{\alpha}(\hat{z}_0 + z^{(1-\alpha)})}\right] \tag{4-15}$$

$\Phi(Z)$ is the standard normal cumulative distribution function and $z^{(1-\alpha)}$ is the $100(1-\alpha)^{th}$ percentile of a standard normal distribution. Also, $\hat{z}_0$ (bias correction) and $\hat{\alpha}$ (acceleration factor) are given as follows.

$$\hat{z}_0 = \Phi^{-1}\left[\frac{\#(\bar{x}_{PROC,i} < \bar{x}_{PROC})}{N}\right],\ i{:}\ = 1, 2, ..., N \tag{4-16}$$

$\Phi^{-1}(x)$ is the inverse function of a standard normal cumulative distribution function, e.g., $\Phi^{-1}(0.95) = 1.645$ and # $\hat{\alpha}$ is the acceleration factor and is given by the following equation.

$$\hat{\alpha} = \frac{\sum(\bar{x}_{PROC} - \bar{x}_{-i,PROC})^3}{6[\sum(\bar{x}_{PROC} - \bar{x}_{-i,PROC})^2]^{1.5}} \quad (4\text{-}17)$$

Here the summation in the above equations is being carried from $i = 1$ to $i = n$, the sample size. $\bar{x}_{PROC}$ and $\bar{x}_{-i,PROC}$ are respectively the PROC mean (e.g., KM mean) based upon all n observations, and PROC mean of (n-1) observations without the i^{th} observation, i: = 1, 2, ..., n.

4.8 Additional UCL Computation Methods for Left-Censored Data Sets

Some other UCL computation methods as incorporated in ProUCL 4.0 are briefly described in this section. The details can be found in Singh, Maichle, and Lee (2006).

4.8.1 Ad hoc UCL95 Computation Method Based Upon Student's t-distribution

Several authors (e.g., Helsel 2005 and Millard 2000) and documents (e.g., USEPA-UGD, 2004) suggest the use of Student's t-statistic to compute a UCL95 for left-censored data sets. That is, using an appropriate method (e.g., Cohen's MLE (CMLE), unbiased MLE (UMLE)) to estimate the mean and *sd*, it has been suggested to use Student's t-statistic to compute a UCL95. One such UCL95 based upon Cohen's MLE method is given as follows:

$$\text{UCL95} = \hat{\mu}_{MLE} + t_{0.95,(n-1)}\sqrt{(\hat{\sigma}^2_{MLE}/n)} \quad (4\text{-}18)$$

Similar UCL95 equations have been suggested for the RMLE, UMLE, and EM methods. It is noticed that for normally distributed left-censored data sets (with low censoring intensities, such as lower than 20%), the UCL95 based upon CMLE(t) (or MLE (t)) ad hoc method does provide about 95% coverage to the population mean. The coverage by this method decreases slowly as the censoring intensity (percentage of NDs) increases.

4.8.2 (1 – α)100% UCL Based Upon Chebyshev Theorem Using Sample Mean and Sd

The Chebyshev-Type inequality (as used in Chapter 2) can also be used to compute a UCL95 of mean for left-censored data sets. The two-sided Chebyshev theorem (Dudewicz and Misra (1988)) states that given a random variable, X, with finite mean and standard deviation, μ_1 and σ_1, we have

$$P(-k\sigma_1 \le X - \mu_1 \le k\sigma_1) \ge 1 - 1/k^2.$$

This inequality can be used to compute a UCL of mean based upon data sets with nondetect observations. A *(1 – α)100% UCL* of population mean, μ_1, can be obtained by:

$$\text{UCL} = \bar{x} + \sqrt{((1/\alpha)-1)}s_x/\sqrt{n}. \quad (4\text{-}20)$$

In the above UCL equation, the sample mean and standard deviation are computed using one of the many estimation methods for data sets with ND observations. To obtain such UCLs, instead of using the sample

mean and *sd* (or *SE* of mean), one uses the sample mean and *sd* (or SE) obtained using the various MLE, EM, and KM methods for left-censored data sets as described above. It is noted that the UCL95 based upon Chebyshev inequality (with KM estimates) yields a reasonable but conservative UCL of the mean. This is especially true for highly skewed data sets from lognormal and gamma distributions.

4.8.3 UCL95 Based Upon Tiku's Method (for symmetrical censoring)

For symmetrical Type II censoring, Tiku (1971) suggested the use of a t-distribution with (*n* - *k*-1) degrees of freedom. This method can be used on any of the MLE methods (e.g., MLE, RMLE, and UMLE). In practice, this method is also used for Type 1 censoring. This is just an approximation method. Due to symmetrical censoring, MLE estimates of the mean and standard deviation are independent (Schneider 1986), and Student's t statistic may be used to construct a UCL of the mean.

A *(1 – α)100% UCL* of the mean, as proposed by Tiku (1971), is given as follows:

$$UCL_{Tk} = \hat{\mu}_{MLE} + t_{\alpha,(n-k-1)}\sqrt{V(\hat{\mu}_{MLE})} \tag{4-21}$$

This can also be written as follows:

$$UCL_{Tk} = \hat{\mu}_{MLE} + t_{\alpha,(n-k-1)}\hat{\sigma}_{MLE}\, Gam_{11}\Big/\sqrt{(n+k+1)} \tag{4-22}$$

Here Gam_{11} is computed using the Fisher's information matrix (I), (e.g., see Schneider, 1986). Tiku's approximate method is simple and performs better (in terms of coverage probabilities) than Schneider's approximation (Singh, Maichle, and Lee, 2006).

4.9 Comments on the Use of Substitution Methods and Minimum Sample Size Requirements

It is noted that even though, most of the statistics used to estimate the EPC terms and BTV estimates could be computed based upon data sets even with 3-4 detected values; the use of those statistics is not recommended due to high level of uncertainty associated with them. It is re-emphasized that when only a few (e.g., <4-6) observations are detected, it is not appropriate (due to high uncertainty associated with such estimates) to use substitution methods (e.g., DL/2 method), and replace the rest (majority) of the nondetect data by their respective DL/2 values (or some other proxy value), and compute an estimate of the EPC term or BTV based upon the resulting data set (Singh, Maichle, and Lee (2006)).

In case, only a small number (e.g., < 4-6) of detected observations are available, it is desirable to collect more samples and data (from the area under investigation), so that enough detected observations will be available to compute reasonably reliable estimates the various environmental parameters including BTVs and EPC terms. If it is not possible to collect more samples (as often is the case), it is suggested to use professional judgment and available historical information (e.g., from similar sites) to estimate the BTVs rather than using statistical methods on the fabricated data sets obtained using some substitution method (e.g., DL/2). In such cases, BTVs and EPC terms may be estimated on case-by-case basis perhaps based upon the site knowledge and experts' opinion.

4.9.1 Use of Ad hoc Estimation Methods on Case-by-Case Basis

- For data sets, with high percentage on nondetects (e.g., > 80%, > 90%, ...), the EPC term or BTV may be estimated using simple ad hoc methods rather than estimating them based upon mean and *sd* statistics obtained using the fabricated (e.g., 0, DL/2, DL) values. For an example, when all or most of the background data values are reported as nondetects, the BTV or a not-to-exceed value should also be reported as a nondetect value (or some pre-specified action level), perhaps by the maximum RL or maximum RL/2.

- The median or mode (instead of using mean and *sd* based upon fabricated data) of the data with majority of the ND values may also be used to estimate EPC terms or BTVs.

- Also, when only a few detected values (< 4-6) are available in a data set of larger size (e.g., > 15-20), one may again use the maximum RL, or maximum RL/2 as an estimate of the BTV or some pre-specified action level.

- When only a few detected values (e.g., < 4-6) are available in data sets of smaller sizes (e.g., < 8-10), one may use the maximum detected data value or the second largest detected value to estimate the BTV.

- The uncertainty associated with all such estimates listed above will be high; and statistical properties such as bias, accuracy, and precision of such estimates would remain unknown.

- Statistics is the science of estimating population values (e.g., EPC, BTV) based upon sampled data values representing a much smaller but representative fraction of the population under study. In order to be able to use defensible statistical methods and computing reliable estimates, it is desirable to have adequate amount of data. Specifically, when a data set associated with a COPC consists of at least 8-10 detected (more are desirable) observations, one can use one of the recommended methods as incorporated in ProUCL 4.0. The details of those methods are given in Singh, Maichle, and Lee (EPA, 2006).

- Thus, it is recommended to avoid using statistical methods (to estimate the BTVs and other environmental parameters) on data sets with less than 4-5 detected values (more are desirable). If possible, it is desirable to collect more samples (data) with detected values. Statistics computed based upon small data sets or on data sets with only a few detected values (e.g., < 4-6) cannot be considered reliable enough to make important remediation and cleanup decisions potentially affecting the human health and the environment.

4.10 Summary and Recommendations

The following observations and recommendations as incorporated in ProUCL 4.0 are based upon the results and findings of the study and simulation experiments as summarized in Singh, Maichle, and Lee (EPA, 2006).

4.10.1 General Observations and Comments

- It is not easy to verify (perform goodness-of-fit) the distribution of a left-censored data set. Therefore, emphasis is given on the use of nonparametric upper limit (e.g., UCLs, UPLs, and UTLs) computation methods.

- Most of the parametric MLE methods (e.g., MLE, RMLE, and EM) assume that there is only one detection limit. But in practice, a left-censored data set often has multiple detection limits. For such methods, the KM method or one of the ROS methods as incorporated in ProUCL 4.0 may be used.

- For reliable and accurate results, the user should ensure that the data set under study represents a single statistical population (e.g., background reference area, or an AOC) and not a mixture population (e.g., clean and polluted site areas).

- It is recommended to identify all potential outliers and investigate them separately. Decisions about the disposition of outliers should be made by all interested members of the project team. Several references are available in the literature to properly identify outliers (Rousseeuw and Leroy (1987) and Singh and Nocerino (1995)), and to partition a mixture sample into component sub-samples (Singh, Singh, and Flatman (1994)). A new chapter describing the population partitioning methods has been included in the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b), currently under revision by the NERL, Technical Support Center, EPA, Las Vegas.

- Avoid the use of transformations (to achieve symmetry) while computing the upper limits for various environmental applications. It is easier and safer to interpret the results computed in the original scale. Moreover, the results and statistics computed in the original scale do not suffer from transformations bias.

- Specifically, avoid the use of a lognormal model even when the data appear to be lognormally distributed. Its use often results in incorrect and unrealistic statistics of no practical merit (Singh, Singh, and Iaci, 2002).

- The parameter in the transformed space may not be of interest to make cleanup decisions. The cleanup and remediation decision are often made in original raw scale; therefore, the statistics (e.g., UCL95) computed in transformed space need to be back-transformed in the original scale. It is not clear to a typical user how to back-transform results in log-scale or any other scale obtained using a Box-Cox (BC)-type transformation to original raw scale. Moreover, transformed results suffer from significant amount of transformation bias.

- It is recommended to avoid the use of equation (4-2) as described in Shaarawi (1989) to back-transform estimates from the log-scale to original scale. The question now arises - how one should back-transform results from a log-space (or any other transformed space) to the original space. Unfortunately, no defensible guidance is available in the environmental literature to address this question. Moreover, the back-transformation formula will change from transformation to transformation (BC-type transformations), and the bias introduced by such transformations will remain unknown. This is one of the

main reasons that ProUCL 4.0 does not compute MLE estimates (or other estimates such as fully parametric estimates using ROS on logged data) using log-transformed data.

- Therefore, in cases when a data set in the "raw" scale cannot be modeled by a parametric distribution, it is desirable to use nonparametric methods rather than testing or estimating some parameter in the transformed space.

- For the various parametric (gamma and lognormal) and nonparametric skewed distributions, the performance (in terms of coverage percentage for the population mean) of Robust ROS method followed by percentile or BCA bootstrap methods is not better than the KM (Chebyshev) and KM (BCA) UCL methods. It is also observed that, for left-censor data sets of all sizes and various censoring levels, the robust ROS UCL (both percentile bootstrap and BCA bootstrap methods) fail to provide adequate coverage for the population mean of highly skewed distributions.

- On page (78) of Helsel (2005), the use of the robust ROS MLE method (Kroll, C.N. and J.R. Stedinger, 1996) has been suggested to compute summary statistics. In this hybrid method, MLEs are computed using log-transformed data. Using the regression model as given by equation (4-1), the MLEs of mean (used as intercept) and *sd* (used as slope) in the log-scale are used to extrapolate the NDs in the log-scale. Just like in Helsel's robust ROS method, all of the NDs are transformed back in the original scale by exponentiation. This results in a full data set in the original scale. One may then compute the mean and *sd* using the full data set. The estimates thus obtained are called robust ROS ML estimates (Helsel (2005), and Kroll and Stedinger, 1996). However, the performance of such a hybrid estimation method is not well known. Moreover, for higher censoring levels, the MLE methods sometimes behave in an unstable manner.

    - It should be noted that the performance of this hybrid method is unknown.
    - It is not known why this method is called a robust method.
    - The stability of the MLEs obtained using the log-transformed data is doubtful especially for higher censoring levels.
    - The BCA and (%) UCLs based upon this method will fail to provide the adequate coverage for the population mean for moderately skewed to highly skewed data sets.

- The maximum censoring level considered in the present simulation study is 70%. For data sets having a larger % of nondetects (e.g., 80%, 90%, or 99% nondetects), statistical estimates may not be reliable. Decisions about the use of an appropriate method should be made by the risk assessors and regulatory personnel on a site-specific basis. The use of nonparametric methods based upon the hypothesis testing for proportion of exceedances is recommended in such cases (USEPA (2006); Helsel (2005)) with % censoring exceeding 70%-80%. This method, the single sample hypothesis testing for proportions, is also available in ProUCL 4.0.

- Do Not Use DL/2 (t) UCL method. This UCL computation method does not provide adequate coverage (for any distribution and sample size) for the population mean, even for censoring levels as low as 10%, 15%. This is contrary to the conjecture and assertion (e.g., EPA (2006)) often made that the DL/2 method can be used for lower ($\leq$ 20%) censoring levels. The coverage provided by the DL/2 (t) method deteriorates fast as the

censoring intensity increases. The DL/2 (t) method is not recommended by the authors or developers of this text and ProUCL 4.0.

- The KM method is a preferred method as it can handle multiple detection limits. Moreover, the various nonparametric UCL95 methods (KM (BCA), KM (z), KM (%), KM (t)) based upon the KM estimates provide good coverages for the population mean.

- For a symmetric distribution (approximate normality), several UCL95 methods provide good coverage (~95%) for the population mean, including the Winsorization mean, Cohen's MLE (t), Cohen's MLE (Tiku), KM (z), KM (t), KM (%) and KM (BCA). Specific recommendations for the various distributions considered in this report are described as follows.

4.10.2 Recommended UCL95 Methods for Normal (Approximate Normal) Distribution

- For normal and approximately normal (e.g., symmetric) distributions, the most appropriate UCL95 computation methods are the KM (t) or KM (%) methods. For symmetric distributions, both of these methods perform equally well on left-censored data sets for all censoring levels and sample sizes.

4.10.3 Recommended UCL95 Methods for Gamma Distribution

- For highly skewed gamma distributions, G(k, θ), with a shape parameter, $k \leq 1$:
  - Use the nonparametric KM (Chebyshev) UCL95 method for censoring levels < 30%,
  - Use the nonparametric KM (BCA) UCL95 method for censoring levels in the interval [30%, 50%),
  - Use the nonparametric KM (t) UCL95 method for censoring levels ≥ 50%.
- For moderately skewed gamma distributions, G (k, θ), with shape parameter, $1< k \leq 2$:
  - For censoring level ≤ 10%, use the KM (Chebyshev) UCL95 method,
  - For higher censoring levels (10%, 25%, ...), use the KM (BCA) UCL95 method,
  - For censoring levels in [25%, 40%), use the KM (%) UCL95 method,
  - For censoring levels ≥ 40%, use the KM (t) UCL95 method.
- For mildly skewed gamma distributions, G (k, θ), with $k > 2$:
  - Use the KM (BCA) UCL95 method for lower censoring levels (≤ 20%)
  - For censoring levels in the interval (20%, 40%), use the KM (%) UCL95,
  - For censoring ≥ 40%, use the KM (t) UCL95 computation method.

4.10.4 Recommended UCL95 Methods for Lognormal Distribution

- For mildly skewed data sets with $\hat{\sigma} \leq 1$:
    - For censoring levels (≤ 20%) and sample of sizes less than 50-70, use the KM (Chebyshev) UCL95,
    - For censoring levels (≤ 20%) and samples of sizes greater than 50-70, use the KM (BCA) UCL95,
    - For censoring levels in the interval (20%, 40%) and all sample sizes, use the KM (BCA) UCL95,
    - For censoring level ≥ 40%, use the KM (%) or KM (t) UCL95 method.
- For data sets with $\hat{\sigma}$ in the interval (1, 1.5]:
    - For censoring levels ≤ 50% and samples of sizes < 40, use 97.5% the KM (Chebyshev) UCL,
    - For censoring levels ≤ 50%, samples of size ≥ 40, use 95% KM (Chebyshev) UCL,
    - For censoring levels > 50%, use the KM (BCA) UCL95 for samples of all sizes.
- For highly skewed data sets with $\hat{\sigma}$ in the interval (1.5, 2]:
    - For sample sizes < 40, and censoring levels <50%, use 99% KM (Chebyshev) UCL,
    - For sample sizes ≥ 40 and censoring levels < 50%, use 97.5% KM (Chebyshev) UCL,
    - For samples of sizes < 40-50 and censoring levels ≥ 50%, use the 97.5% KM (Chebyshev) UCL,
    - For samples of sizes ≥ 40-50, censoring ≥ 50%, use 95% KM (Chebyshev) UCL.
- Use a similar pattern for more highly skewed data sets with $\hat{\sigma} > 2.0, 3.0$:
    - For extremely highly skewed data sets, an appropriate estimate of the EPC term (in terms of adequate coverage) is given by a UCL based upon Chebyshev inequality and KM estimates. The confidence coefficient to be used depends upon the skewness. For highly skewed data sets, a higher (e.g., > 95%) confidence coefficient may have to be used to estimate the EPC.
    - As the skewness increases, the confidence coefficient also increases.
    - For such highly skewed distributions (with $\hat{\sigma} > 2.0, 3.0$), for lower sample sizes (e.g., < 50-60), one may simply use a 99% KM (Chebyshev) UCL to estimate the

population mean, EPC term, and other relevant threshold values. For sample sizes greater than 60, one may use a 97.5% KM (Chebyshev) UCL as an estimate of the population mean or mass.

- For sample sizes greater than 60, one may use a 97.5% KM (Chebyshev) UCL as an estimate of the population mean or mass.

4.10.5 Recommended UCL95 Methods for Non-Discernable Distributions

- For symmetric or approximately symmetric nonparametric data distributions, one may use the same UCL computation methods as for the data sets coming from a normal or an approximate normal or symmetric population.

- It is noted that most of the recommended UCL computation methods for a lognormal distribution do not assume the lognormality of the data set. Therefore, those UCL computation methods can be used on skewed nonparametric data sets that do not follow any of the well-known parametric distributions.

Chapter 5

Estimating Background Threshold Values and Establishing Site-Specific Background Concentrations Using Data Sets with Nondetect (ND) Observations

5.1 Introduction

As discussed in Chapter 4, nondetects (NDs) or below detection limit (BDL) observations are inevitable in most environmental data sets. Environmental scientists often encounter trace level concentrations of contaminants of potential concern (COPC) when evaluating sampled analytical results. Those low level analytical results cannot be measured accurately, and therefore are typically reported as less than one or more detection limit (DL) values. However, the practitioners (e.g., environmental scientists) do need to obtain reliable estimates (point estimates and interval estimates) of the population parameters such as the population mean and population percentiles. Parametric and nonparametric methods to compute interval estimates of population mean (e.g., UCLs) for data sets with NDs are described in Chapter 4. Specifically, Chapter 4 of this Technical Guide discusses in detail the various GOF tests, methods to estimate population mean and *sd*, and parametric and nonparametric UCL computation methods based upon data sets with nondetect observations. All of those GOF tests and methods to estimate the population mean and *sd* as described in Chapter 4 are also applicable to this Chapter 5. Chapter 3 of the Technical Guide discusses various parametric and nonparametric methods to compute upper limits (UPLs, UTLs, and upper percentiles) that can be used to estimate the background threshold values (BTVs) and other not-to-exceed values. This chapter discusses parametric and nonparametric methods as incorporated in ProUCL 4.0 that can be used to compute upper limits including UPLs, UTLs, and upper percentiles often used to estimate the BTVs based upon data sets with ND observations.

Even though several methods are available (as described in Chapter 4) to estimate population mean and *sd* based upon data sets with nondetects, it is noted that not much guidance is available in the statistical literature on how to compute appropriate UPLs and UTLs based upon left-censored data sets of varying degree of skewness. The estimation methods (e.g., MLE, ROS, KM, bootstrap) to estimate the population mean and standard deviation as incorporated in ProUCL 4.0 are described in Chapter 4. Using the results and findings of the study performed by Singh, Maichle, and Lee (2006), both parametric and distribution-free (nonparametric) methods and limits used to estimate the BTVs and not-to-exceed values (as incorporated in ProUCL 4.0) based upon left-censored data sets are briefly described in this chapter. More details of the computations of upper limits can be found in the newly added Chapter 6 of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b). It is noted that ProUCL 4.0 is perhaps the first software package, which has several statistical methods including the KM and bootstrap methods that can be used to estimate the BTVs and not-to-exceed values based upon left-censored data sets with ND observations.

As mentioned in earlier chapters, it is not easy to reliably perform GOF tests on left-censored data sets, especially when the percentage of nondetects is quite high (e.g., > 40%) and the number of detected observations is small (e.g., < 8-10). Therefore, just as in Chapter 4, in this chapter also emphasis is given to the computation of upper limits (UPLs, UTLs, upper percentiles) based upon distribution-free nonparametric methods such as the Kaplan-Meier (KM) method, Chebyshev inequality, and other computer intensive bootstrap resampling methods. It should be noted that, in practice the Chebyshev inequality often yields conservative estimates (providing higher coverage than the specified 90% or 95% coverage) of the BTVs and other not-to-exceed values. Therefore, instead of using a 95% UPL based

upon Chebyshev inequality, it is suggested to use Chebyshev UPL with a lower confidence coefficient (e.g., 85%, 90%) as an estimate of BTV and not-to-exceed value.

Note: *Unless, the data are highly skewed, it is preferred to use KM estimates and the associated upper limits to estimate the BTVs and not-to-exceed values. Just as in Chapter 3, no specific recommendations have been made regarding the most appropriate upper limit(s) that may be used to estimate the BTVs and not-to-exceed values. However, the author of this chapter prefers to use UPLs as estimates BTVs and other not-to-exceed values.*

Once again, throughout this chapter, it is assumed that the user is dealing with a data set collected from a single background population; has pre-processed the data set, and has identified all of the potential outliers (if any) and multiple populations. In order to obtain meaningful, reliable, and practical results, the procedures described in this chapter should be used on data sets that represent "single" (e.g., a background area), and "not mixture" populations. These statements are made to familiarize the user with the underlying assumptions required by the various statistical methods, including the estimation methods based upon left-censored data sets. Some simple classical outlier tests as incorporated in ProUCL 4.0 are discussed in Chapter 7 of this Technical guidance document.

Note: *It should be noted that the mathematical UPL and UTL computation formulae and methods as described in this chapter (and incorporated in ProUCL 4.0) could be used on any left-censored data set with or without the outliers. The user should keep this in mind that the estimates of the BTVs based upon data sets with potential outliers or mixture populations may not be reliable.*

Background evaluation and comparison studies often require the computation of UPLs and UTLs based upon left-censored data sets. Recent environmental literature (e.g., Millard (2002) and USEPA-UGD (2004)) lists and cites the use of ad hoc "rule-of-thumb" type methods based upon the Student's t-statistic or Land's H-statistic to compute 95% UCLs, 95% UPLs, and 95% UTLs. For example, it is noted that the Student's t-statistic (e.g., on Cohen's maximum likelihood estimates) is used to compute an upper prediction limit (pages 10-26, USEPA-UGD (2004)). However, it is noted that the distribution of the t-type statistic based upon the MLE or KM estimates of mean and standard deviation used to construct a UPL (pages 10-26, USEPA-UGD, 2004) is not known. The MLEs or KM estimates of population mean and standard deviation (*sd*) based upon left-censored data sets are very different from the traditional mean and *sd* used in the definition of a typical Student's t-statistic. The Student's t-statistic-based "rule-of-thumb" methods to compute UCLs, UPLs and UTLs are difficult to defend for moderately skewed to highly skewed data sets with standard deviation of the log-transformed data exceeding 0.75-1.0.

5.2 Underlying Assumptions

Pre-processing of data to identify potential outliers and multiple populations (if any) should be conducted to obtain accurate and reliable estimates of the environmental parameters such as BTVs and exposure point concentration (EPC) terms. The user may want use informal graphical displays (e.g., quantile-quantile plots, histograms) and formal population partitioning methods (e.g., see Singh, Singh, and Flatman (1994)) to identify multiple populations (if any). A simple iterative population partitioning method is presented in the newly added Chapter 7 of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b). The estimation and upper limits computation methods as described in this chapter should be used on data sets coming from a "single" statistical population such as a single contaminated or remediated area of the site, an unimpacted clean background or reference population. The sampled data set should represent a random sample from the

area under study such as a background area (BA). This means that the data set should be representative of the entire background population of interest under study.

A few outlying observations (e.g., representing contaminated locations) in a full-uncensored data set or in a left-censored data set often distort all statistics including the maximum likelihood estimates (MLEs) of mean and standard deviation (*sd*); Kaplan-Meier (KM) estimates; and regression on order statistics (ROS) estimates (e.g., slope and intercept, and extrapolated NDs) both in raw as well as log-scale (Singh and Nocerino, 2002). The use of such distorted estimates of mean, *sd*, and of ND values (e.g., using ROS) will yield distorted estimates (e.g., UPLs, UTLs, upper percentiles) of BTVs and not-to-exceed values. Therefore, instead of computing distorted estimates for the entire area of interest (e.g., BA, AOC), it is desirable to identify potential outliers and study them separately.

Identified environmental outliers perhaps represented by their spatial locations, time period, laboratory, or analytical methods require further and perhaps separate investigation. Reliable and defensible statistics can be obtained based upon the majority of a data set representing the main body of the dominant population (e.g., reference area, BA) under study. Instead of computing distorted statistics and estimates by including a few low probability outlying observations, it is desirable that those low probability extreme high outliers coming from the upper tail of the data distribution should be treated separately.

Note: *The main objective of using a statistical procedure is to model the majority of the data representing the main dominant population, and not to accommodate a few low probability outliers that may yield inflated and impractical results. The background thresholds should be estimated by reliable statistics (and not distorted statistics) obtained using data sets representing the dominant background population. High outlying values contaminate the underlying left-censored or uncensored full data set from the population under study. In practice, it is the presence of a few extreme outliers that cause the rejection of normality of a data set (Barnett and Lewis (1994)).*

As mentioned before, statistics (e.g., mean, UCL, UPL) based upon data sets with outliers would yield distorted and inflated estimates for entire population (e.g., mean, upper threshold value, upper limits), which in turn may result in incorrect conclusions and decisions. In such cases, it is suggested to compute all relevant statistics using data sets with outliers and without outliers, and compare the results. This extra step often helps to see the direct potential influence of outlier(s) on the various statistics of interest (e.g., mean, UPLs). This in turn will help the project team to make informative decisions about the disposition of outliers. That is, the project team and experts familiar with the site should decide which of the computed statistics (with outliers or without outliers) represent better and more accurate estimate(s) of the population parameters (e.g., mean, EPC, BTV) under consideration. Since the treatment and handling of outliers is a controversial and subjective topic, it is suggested that the outliers be treated on a site-specific basis using all existing knowledge about the site and the site background (e.g., EA, area of concern (AOC), reference area) under investigation.

5.3 Identification of High (in Upper Tail) Outliers for Left-Censored Data Sets

As well known, outliers when present in uncensored or censored (with NDs) data sets distort all statistics of interest. Decisions based upon distorted (inflated) estimates of the BTVs and not-to-exceed values can be hazardous to human health and the environment. It is, therefore important that outliers be identified before computing the estimates of BTVs and not-to-exceed values. Therefore, the user should make a sincere effort to identify all potential outliers using effective, robust and resistant statistical methods (e.g., Rousseeuw and Leroy (1987), Barnett and Lewis (1994), Singh (1993), and Singh and Nocerino (1995)) before proceeding with the estimation of population mean, standard deviation, UCLs, UPLs, UTLs, and

other summary statistics based upon left-censored data sets. Some robust and resistant estimation methods for left-censored data sets are given in Singh and Nocerino (1995). The detailed discussion of those robust and resistant estimation methods is beyond the scope of ProUCL 4.0. A couple of classical outlier tests (e.g., Dixon test and Rosner test) are available in ProUCL 4.0. These tests can be used on data sets with and without nondetect observation.

Additionally, one can always (it is recommended) use graphical displays such as Q-Q plots and box plots to visually identify high outliers in a left-censored data set. It should be pointed out that in many environmental applications; it is the identification of high outliers (perhaps representing contaminated locations and hot spots) that is important. The occurrence of nondetect (less than values) observations and other low values is quite common in environmental data sets, especially when the data are collected from a background or a reference area.

5.3.1 Outlier Testing Procedures for Data Sets with NDs

For the purpose of the identification of high outliers, one may replace the nondetect values by their respective detection limits or half of the detection limits, or may just ignore them (especially when the number of detected values is large such as exceeding 8-10) from any of the outlier test (e.g., Rosner test, Dixon test) computation including the graphical displays such as Q-Q plots. These simple outlier testing procedures (ignoring NDs, and replacing them by DL/2) are available in ProUCL for data sets with ND values. Except for the identification of high outlying observations, the outlier testing statistics (computed with NDs or without NDs) are not used in any of the estimation and decision making process. Therefore, for the purpose of testing for high outliers, it does not matter how the nondetect observations are treated. After outliers have been determined (this should also be verified using graphical displays), the project team and experts familiar with the site should make the final decision about the proper disposition of outliers.

5.3.2 Q-Q Plots for Data Sets with Nondetect Observations

ProUCL 4.0 can be used to obtain Q-Q plots using only the detected data values, or based upon full data sets obtained using some proxy values (e.g., DL, DL/2) for nondetect observations. ProUCL 4.0 also generates Q-Q plots based upon full data sets obtained using estimated NDs based upon ROS methods (e.g., normal, lognormal, gamma). The details of those ROS methods are described in Chapter 4. Thus outlier test statistics (e.g., Rosner test) as discussed in Chapter 7, and the visual displays shown by Q-Q plots can identify high outlying observations that may be present in an environmental data set. It is suggested to use both graphical displays as well outlier test statistics to test and confirm the presence of outlier(s) in an environmental data set. Some examples illustrating the influence of outliers on estimates of BTVs are also considered in later sections of this chapter.

Note*: NERL-EPA, Las Vegas is currently upgrading Scout software package (EPA, 1999) that will provide state-of-the-art robust and resistant multiple outlier identification procedures for both univariate as well as multivariate data sets.*

5.4 Estimating BTVs and Not-to-Exceed Values Based Upon Left-Censored Data Sets

This section describes some parametric and nonparametric methods to compute upper limits (e.g., UPLs, UTLs, upper percentiles) that may be used to estimate the BTVs and other not-to-exceed levels based upon data sets with nondetect observations. As described in Chapter 3, sample statistics such as upper

percentiles, UPLs, and UTLs are often used as estimates of BTVs, compliance limits, or not-to-exceed values. It is noted that for left-censored data sets, there are no preferred and recommended methods available in the literature that can be used to estimate the BTVs and not-to-exceed values. Some ad hoc computation methods based upon the Student's t-statistic (assuming Student's t-distribution without theoretical justification) or normal Z-scores have been cited or mentioned in the literature (Helsel (2005), Millard (2002), USEPA-UGD (2004)) to compute percentiles, UPLs, and UTLs based upon statistics (e.g., mean, *sd*) obtained using MLE method, KM method, or ROS methods (California's Ocean Plan (2005)).

However, it should be noted that the t-type statistic used to derive such limits (e.g., equation (5-1) given below) does not really follow the Student's t-distribution as assumed in the references listed above. These methods to compute upper limits (e.g., UPL, UTL, and percentiles) based upon Student's t-distribution, non-central t-distribution, or normal Z-scores are described in this chapter for historical reasons; and the use of such methods in not recommended due to lack of theoretical justification. These methods may yield reasonable upper limits (e.g., with proper coverage) for mildly skewed data sets with *sd* of log-transformed data less than 0.75-1.0. Singh, Maichle, and Lee (EPA 2006) demonstrated that the use of Student's t-statistic on moderately to highly skewed left-censored data sets yields UCL95 with coverage probabilities much lower than the specified confidence coefficient, 0.95. Moreover, based upon the results and conclusions summarized in Singh, Maichle, and Lee (2006), it is anticipated that the performance of Student's t-type upper limits (e.g., UPLs, UTLs) based upon MLE estimates for moderately skewed to highly skewed data sets would be less than acceptable.

Therefore, instead of making incorrect assumptions or using parametric methods without theoretical justification, the use of nonparametric methods is preferred to estimate BTVs. Several of those methods including bootstrap methods for left-censored data sets (Chapter 4) are available in ProUCL 4.0. Methods as incorporated in ProUCL 4.0 to compute upper limits (UPLs, UTLs, and upper percentiles) based upon data sets with ND observations are described in the following sections.

Note: *It should also be noted that the BTV estimation methods for full data sets as described in Chapter 3 and available in ProUCL 4.0 may also be used on generated data sets obtained using ROS methods on data sets with nondetect observations. Moreover, all other comments about the use of substitution methods, disposition of outliers, and minimum sample size requirements as described in Chapter 4 also apply to BTV estimation methods for data sets with nondetect observations as described in this chapter.*

5.4.1 Computing Upper Prediction Limits (UPLs) for Left-Censored Data Sets

This section describes some parametric and nonparametric methods to compute UPLs based upon left-censored data sets.

5.4.1.1 UPLs Based Upon Student's t-type Statistic

As noted before, some documents (e.g., Helsel (2005), Millard (2002), and Unified Guidance Document (UGD)-EPA (2004)) list or mention the use of Student's t-statistic as one of the potential methods to compute UCL95, UPL95, and UTLs. Specifically, KM estimates (Helsel (2005)) or Cohen's MLEs of mean and standard deviation are used to compute Student's t-statistic-based UPL as an estimate of the BTV. For an example, a $(1 - \alpha)100\%$ UPL for a future (or next, or from a different population) observation (observation not belonging to the collected background sample under study) based upon MLE estimates is given by the following equation.

$$\text{UPL} = \hat{\mu}_{MLE} + t_{((1-\alpha),(n-1))}\sqrt{\hat{\sigma}^2_{MLE}(1+1/n)} \quad (5\text{-}1)$$

Here $t_{((1-\alpha),(n-1))}$ is the critical value of Student's t-distribution with *(n–1)* degrees of freedom (*df*). Similar UPL equations for a single future observation and also for the next (or future) $k \geq 1$ observations (as described in Chapter 3) can be developed for other estimation methods including the KM and ROS methods. If the distributions of the site data and the background data are comparable, then a new (next) observation coming from the site population (e.g., site) should lie at or below the UPL95 probability 0.95. It is noted that, just like UCLs, the UPLs based upon Student's t-distribution might fail to provide the specified (e.g., 0.95) coverage, especially when data are moderately skewed to highly skewed with *sd* of log-transformed data > 1.0.

5.4.1.2 UPL Based Upon the Chebyshev Inequality

The Chebyshev inequality can also be used to obtain a reasonably conservative but stable estimate of the BTV, and is given as follows.

$$UPL = \bar{x} + [\sqrt{((1/\alpha)-1)*(1+1/n)}]s_x \quad (5\text{-}2)$$

The mean $\bar{x}$ and standard deviation, s_x used in the above equation are computed using one of the estimation methods (e.g., KM, MLE, ROS method) on the left-censored data set. Since this Chebyshev method does not require any distributional assumptions about the data set under study, this is a nonparametric method. It should be noted that just like the Chebyshev UCL, UPL based upon Chebyshev inequality often yields much higher estimate of not-to-exceed values and BTVs than various other methods as described in this chapter. This is especially true when skewness is mild (e.g., *sd* of log-transformed data is low < 0.75-1.0), and the sample size is large (e.g., > 30). The user is advised to use professional judgment before using this method to compute a UPL. Specifically, instead of using a 95% UPL based upon Chebyshev inequality, it is suggested to use Chebyshev UPL with a lower confidence coefficient (e.g., 85%, 90%) as an estimate of BTV and not-to-exceed value. ProUCL 4.0 can compute these limits for any level of confidence coefficient.

5.4.1.3 UPLs Based Upon ROS Methods

As described earlier, ROS methods first predict *k* nondetect values using an OLS linear regression model (Chapter 4). This results in a full data set of size *n*. For ROS methods (normal, gamma, lognormal), ProUCL 4.0 generates additional columns consisting of *(n-k)* detected values, and *k* predicted values of the *k* nondetect observations for each variable selected by the user. Once, the nondetect observations have been estimated, an ***experienced*** user may use any of the available parametric and nonparametric BTV and not-to-exceed value estimation methods for full data sets (without NDs) as described in Chapter 3 and incorporated in ProUCL 4.0. Those estimation methods are not repeated here. The user of this method is assumed to know the behavior of the various ROS methods as incorporated in ProUCL 4.0. Specifically, it is expected that the user knows how the presence of outliers can yield distorted and infeasible estimates of ND observations.

Note: *It is noted that a linear model (regression line) can be obtained even when only two (2) detected observations are available. Therefore, the methods as discussed here and also incorporated in ProUCL 4.0 can be used on data sets with 2 or more detected observations. Obviously, in order to be able to*

compute reliable estimates of nondetects and to compute defensible upper limits, enough detected observations should be made available.

*5.4.2 Computing Upper p*100% Percentiles for Left-Censored Data Sets*

This section briefly describes some parametric and nonparametric methods to compute upper percentiles based upon left-censored data sets.

5.4.2.1 Upper Percentiles Based Upon Standard Normal Z-Scores

The use of standard normal Z-scores (e.g., Helsel, 2005) has also been listed as one of the potential method to estimate upper percentiles using KM estimates or MLE estimates based upon left-censored data sets. The p^{th} percentile based upon KM estimates (as incorporated in ProUCL 4.0) is given as follows.

$$\hat{x}_p = \hat{\mu}_{KM} + z_p \sqrt{\hat{\sigma}^2_{KM}} \qquad (5\text{-}3)$$

Here z_p is the p*100th percentile of a standard normal, N (0, 1) distribution, which means that the area (under the standard normal curve) to the left of z_p is p. If the distributions of the site data and the background data are comparable and similar (meaning no contaminated observations from the site), then an observation coming from a population (e.g., site) similar (comparable) to that of the background population should lie at or below the p*100% percentile, with probability p. The 95th normal percentile given by the above equation (for p = 0.95) represents one of the many estimates of the BTVs.

5.4.2.2 Upper Percentiles Based Upon ROS Methods

As noted in Chapter 4, all ROS methods first predict *k* nondetect values using an OLS linear regression model (Chapter 4) assuming a specified distribution of detected and nondetected observations. This process results in a full data set of size *n* consisting of *k* extrapolated NDs and *(n-k)* detected values. For ROS methods (normal, gamma, lognormal), ProUCL 4.0 generates additional columns consisting of the *(n-k)* detected values, and *k* predicted values of the *k* nondetect observations for each variable selected by the user. Once, the nondetect observations have been estimated, an experienced user may use any of the parametric or nonparametric percentile computation methods (e.g., gamma percentiles) for full data sets as described in Chapter 3 and incorporated in ProUCL 4.0.

5.4.3 Computing Upper Tolerance Limits (UTLs) for Left-Censored Data Sets

This section briefly describes some parametric and nonparametric methods to compute UTLs based upon left-censored data sets.

5.4.3.1 UTLs Based Upon K Critical Values Obtained Using a Non-Central t-Distribution

Just like Student's t-statistic-based UPL, the use of the following equation has been mentioned to compute an upper $(1 - \alpha)100\%$ tolerance limit with tolerance or coverage coefficient = p (that is providing coverage to at least p*100% of observations):

$$\text{UTL} = \hat{\mu}_{MLE} + K * \sqrt{\hat{\sigma}^2_{MLE}} \qquad (5\text{-}4)$$

Here K = K (n, α, p) is the tolerance factor and depends upon the sample size, n, confidence coefficient (*CC*) = (1- α), and the coverage proportion = p. The K critical values are based upon non-central t-distribution, and have been tabulated extensively in the statistical literature (e.g., Hahn and Meeker (1991)). It should be pointed out that for left-censored data sets, the distribution of the MLEs (or KM estimates) based-statistic used to develop a UTL is unknown, and hence the use of non-central t-distribution-based K-cut off values has no theoretical backup and justification. The use of similar UTLs based upon other estimation methods including the KM and ROS methods has also been suggested. These UTLs are described here for historical reasons. The use of such UTLs without theoretical justification is not recommended.

5.4.3.2 UTLs Based Upon ROS Methods

The ROS methods as incorporated in ProUCL 4.0, first predict *k* nondetect values using an OLS linear regression model (Chapter 4). This process yields a full data set of size *n* with *(n-k)* original detected values, and *k* new predicted (extrapolated) values. For ROS methods (normal, gamma, lognormal), ProUCL 4.0 generates additional columns consisting of *(n-k)* detected values, and predicted values of the *k* nondetect observations for each variable selected by the user. Once, the nondetect observations have been estimated, an experienced user may use any of the UTL computation methods for full data sets as described in Chapter 3 and incorporated in ProUCL 4.0. Those estimation methods are not repeated here. An example illustrating the use of ROS methods to estimate BTVs is described as follows.

5.5 Estimating BTVs Using Nonparametric Methods Based Upon Higher Order Statistics

It is noted that for full data sets without any discernable distribution, nonparametric UTLs and UPLs are computed using higher order statistics. Therefore, when the data set consists of enough detected observations, and if some of those detected data are larger than all of the NDs and the detection limits, the UTLs, UPLs, and upper percentiles can be estimated by simple nonparametric methods as described in Chapter 3. That is, just as in full data sets, nonparametric UPLs, UTLs, and percentiles for left-censored data sets are also estimated by upper ordered statistics such as the largest or the second largest value, and so on. This nonparametric approach to compute UTLs and UPLs is also available in ProUCL 4.0.

Since, nonparametric UTLs, UPLs, and upper percentiles are represented by higher order statistics (or by some value in between higher order statistic obtained using linear interpolation) in a data set, therefore, every effort should be made to make sure that those higher order statistics do not represent contaminating outlying observations potentially coming from population(s) other than the background population under study. Nonparametric UTLs, UPLs, and percentiles should be used with caution to estimate the BTVs and compliance limits. Every effort should be made to identify and separate the outlying observations before computing order statistics-based upper limits to estimate the BTVs.

5.5.1 Using Ad hoc Estimation Methods on Case-by-Case Basis

This topic has also been discussed in Chapter 4. In order to be able to compute reasonable and reliable estimates using statistical methods, it is desirable to have enough detected values. Statistical methods are often based upon certain assumptions such as randomness, independence, and data distributions that are hard to verify and justify when dealing with small data sets having only a few (e.g., <4-6) detected values. Therefore, instead of using statistical methods on such small data sets, it is desirable to collect more samples (preferably using DQOs) from the area under investigation. If collection of more data is not feasible (e.g., due to lack of resources), it is desirable to use professional judgment, historical information,

information from similar sites, and all expert site knowledge to estimate the parameters of interest such as the EPC term, BTV, or a compliance limit. For an example when the number of detected values is less than 4-6, or when the majority of the data are NDs (e.g., > 80%, > 90%), then the compliance limit or the BTV based upon such data sets also represents a nondetect. In such cases, the BTVs or compliance limits may be determined using information from similar sites or by action levels determined by state or federal agencies.

5.5.2 Example 1

This data set of size 54 was also considered in Chapter 5 of the revised background guidance document. In this example, some data values have been treated as nondetect values with DL= 0.02. The example is used here to illustrate how one can identify high outliers in data sets consisting of some ND values in the lower tail of the data distribution. Typically, one is interested in identifying outliers in the right tail of the data distributions, the same outlier identification procedures (e.g., Rosner test) as used on full data sets can be used to identify outliers in data sets with ND values. The outlier procedures for data sets with and without outliers (as available in ProUCL 4.0) are described in Chapter 7. In the present example, the data set has at least one obvious outlier (= 130,000) as can be seen in Figure 5-1. From Figure 5-2, it is observed that 19,000 may also be an outlier, which got masked (in Figure 5-1) by the occurrence of the outlier, 130,000. Outlier, 130,000 was removed from the data set, and Rosner test was performed to identify more outliers that may be present in the data set. The Rosner test identified at least two more (with 2 suspected outliers) outlying observations as described in Table 5-1.

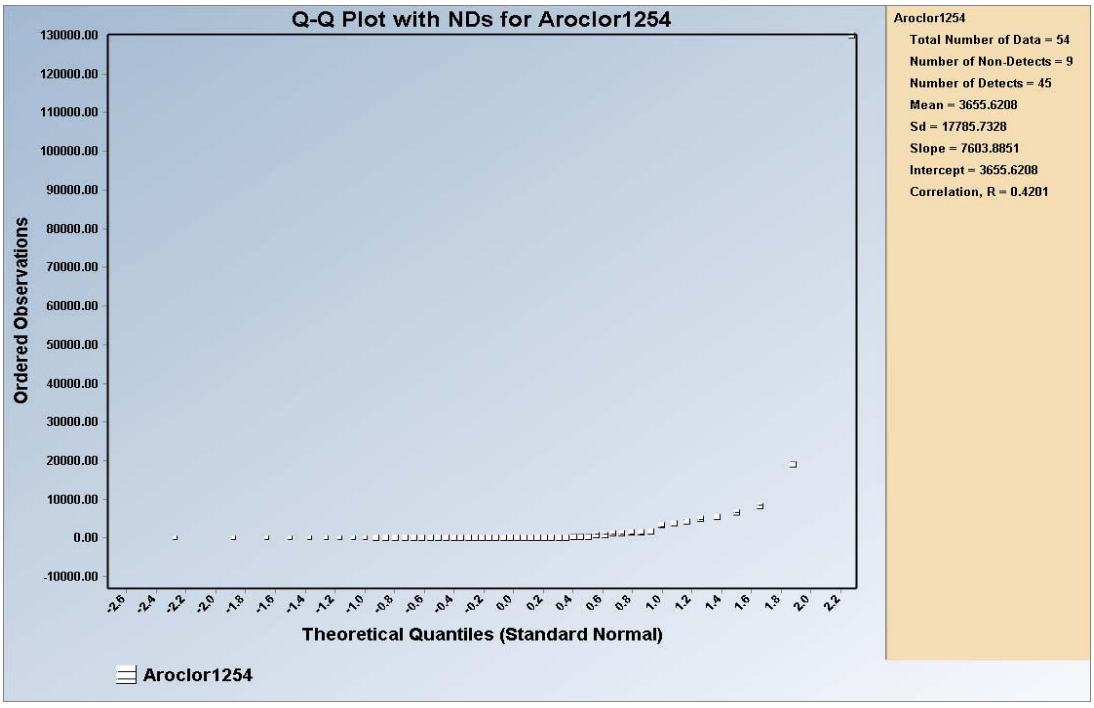

Figure 5-1. Q-Q Plot of Aroclor Data Set

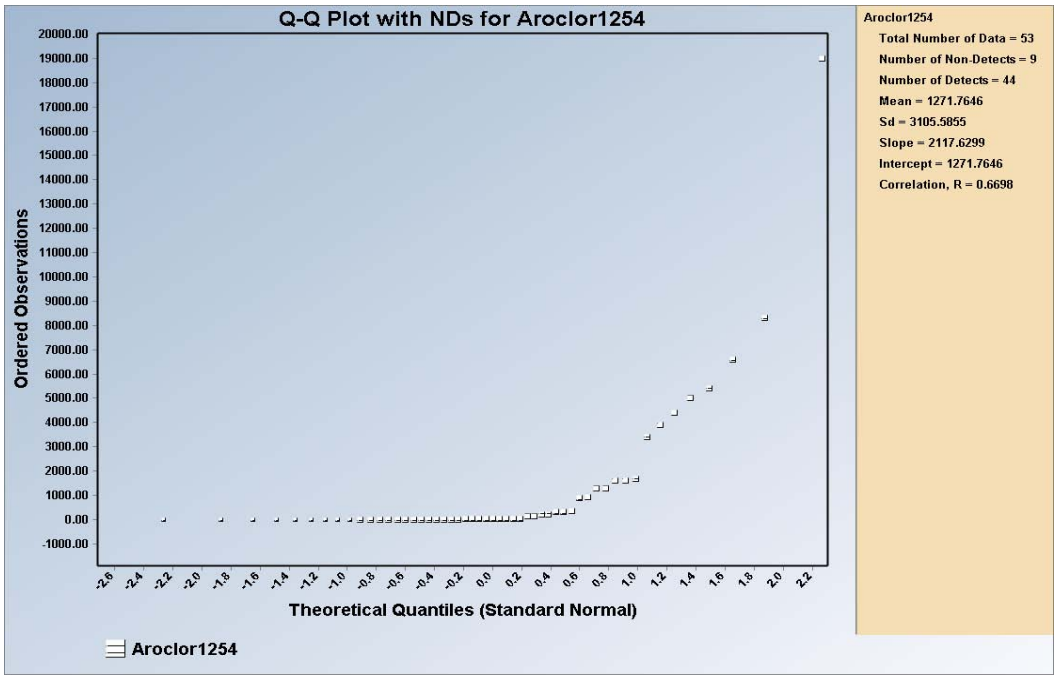

Figure 5-2. Q-Q Plot of Aroclor Data Set without Outlier 130,000

Below are shown the results of the Rosner's test performed on the data set without using nondetect (ND) observations (as one is interested to identify high outlying observations potentially representing hot spots). The two suspected outliers are 19,000 and 8,300. The test values (5.27 and 3.59) are considerably higher than the critical values at 5% significance level.

Table 5-1. Rosner's Outlier Test for Aroclor1254 (Excluding NDs and Outlier 130,000)

Rosner's Outlier Test for Aroclor_Without_NonDetects

Number of data: 44

Number of suspected outliers: 2

#	Mean	sd	Potential outlier	Test value	Critical value (5%)	Critical value (1%)
1	1531.88	3316.59	19000.00	5.27	3.08	3.43
2	1125.65	1998.60	8300.00	3.59	3.07	3.41

For 5% significance level, there are 2 Potential Outliers

Therefore, Potential Statistical Outliers are

19000.00, 8300.00

For 1% Significance Level, there are 2 Potential Outliers

Therefore, Potential Statistical Outliers are

19000.00, 8300.00

In order to illustrate the influence of the two highest outliers (19,000 and 8,300) on the computation of background statistics, the background statistics (for data sets with NDs) as described in this chapter (and in ProUCL 4.0) have been computed using data set with outliers (Table 5-2) and without the two outliers shown in Table 5-3.

Table 5-2. Background Statistics for Aroclor

General Background Statistics for Data Sets with Non-Detects			
User Selected Options			
From File	D:\Narain\ProUCL 4.0\Data\Aroclor 1254.wst		
Full Precision	OFF		
Confidence Coefficient	95%		
Coverage	90%		
Different or Future K Values	1		
Number of Bootstrap Operations	2000		
Aroclor1254			
General Statistics			
Number of Valid Samples	53	Number of Detected Data	44
Number of Unique Samples	41	Number of Non-Detect Data	9
		Percent Non-Detects	16.98%
Raw Statistics		**Log-transformed Statistics**	
Minimum Detected	0.21	Minimum Detected	-1.561
Maximum Detected	19000	Maximum Detected	9.852
Mean of Detected	1532	Mean of Detected	4.474
SD of Detected	3355	SD of Detected	3.163
Minimum Non-Detect	0.02	Minimum Non-Detect	-3.912
Maximum Non-Detect	0.2	Maximum Non-Detect	-1.609
Data with Multiple Detection Limits		**Single Detection Limit Scenario**	
Note: Data have multiple DLs - Use of KM Method is recommended		Number treated as Non-Detect with Single DL	9
For all methods (except KM, DL/2, and ROS Methods),		Number treated as Detected with Single DL	44
Observations < Largest ND are treated as NDs		Single DL Non-Detect Percentage	16.98%
Background Statistics			
Normal Distribution Test with Detected Values Only		**Lognormal Distribution Test with Detected Values Only**	
Lilliefors Test Statistic	0.526	Lilliefors Test Statistic	0.948
5% Lilliefors Critical Value	0.944	5% Lilliefors Critical Value	0.944
Data not Normal at 5% Significance Level		**Data appear Lognormal at 5% Significance Level**	

Data not Normal at 5% Significance Level		Data appear Lognormal at 5% Significance Level	
Assuming Normal Distribution		**Assuming Lognormal Distribution**	
DL/2 Substitution Method		DL/2 Substitution Method	
Mean	1272	Mean (Log Scale)	3.08
SD	3106	SD (Log Scale)	4.257
95% UTL 90% Coverage	6328	95% UTL 90% Coverage	22261
95% UPL (t)	6521	95% UPL (t)	29043
90% Percentile (z)	5252	90% Percentile (z)	5095
95% Percentile (z)	6380	95% Percentile (z)	23922
99% Percentile (z)	8496	99% Percentile (z)	435301
Maximum Likelihood Estimate(MLE) Method		Log ROS Method	
Mean	850.4	Mean in Original Scale	1272
SD	3470	SD in Original Scale	3106
95% UTL with 90% Coverage	6500	95% UTL with 90% Coverage	16104
		95% BCA UTL with 90% Coverage	6240
		95% Bootstrap (%) UTL with 90% Coverage	6240
95% UPL (t)	6717	95% UPL (t)	20589
90% Percentile (z)	5298	90% Percentile (z)	4124
95% Percentile (z)	6559	95% Percentile (z)	17211
99% Percentile (z)	8924	99% Percentile (z)	251088
Gamma Distribution Test with Detected Values Only		**Data Distribution Test with Detected Values Only**	
k star (bias corrected)	0.247	Data appear Lognormal at 5% Significance Level	
Theta Star	6208		
nu star	21.72		
A-D Test Statistic	1.01	**Nonparametric Statistics**	
5% A-D Critical Value	0.884	Kaplan-Meier (KM) Method	
K-S Test Statistic	0.169	Mean	1272
5% K-S Critical Value	0.146	SD	3076
Data not Gamma Distributed at 5% Significance Level		SE of Mean	427.4
		95% KM UTL with 90% Coverage	6280
Assuming Gamma Distribution		95% KM Chebyshev UPL	14806
Gamma ROS Statistics with extrapolated Data		95% KM UPL (t)	6472
Mean	1272	90% Percentile (z)	5214
Median	41	95% Percentile (z)	6332
SD	3106	99% Percentile (z)	8428
k star	0.121		
Theta star	10526		
Nu star	12.81		
95% Percentile of Chisquare (2k)	1.378		
90% Percentile	3610		
95% Percentile	7251		
99% Percentile	18344		
Note: UPL (or upper percentile for gamma distributed data) represents a preferred estimate of BTV			
For an Example: KM-UPL may be used when multiple detection limits are present			
Note: DL/2 is not a recommended method.			

Table 5-3. Background Statistics for Aroclor Data Set without Outliers 19,000 and 8,300

General Background Statistics for Data Sets with Non-Detects			
User Selected Options			
From File	D:\Narain\ProUCL 4.0\Data\Aroclor 1254.wst		
Full Precision	OFF		
Confidence Coefficient	95%		
Coverage	90%		
Different or Future K Values	1		
Number of Bootstrap Operations	2000		
Aroclor1254			
General Statistics			
Number of Valid Samples	51	Number of Detected Data	42
Number of Unique Samples	39	Number of Non-Detect Data	9
		Percent Non-Detects	17.65%
Raw Statistics		**Log-transformed Statistics**	
Minimum Detected	0.21	Minimum Detected	-1.561
Maximum Detected	6600	Maximum Detected	8.795
Mean of Detected	954.8	Mean of Detected	4.238
SD of Detected	1704	SD of Detected	3.037
Minimum Non-Detect	0.02	Minimum Non-Detect	-3.912
Maximum Non-Detect	0.2	Maximum Non-Detect	-1.609
Data with Multiple Detection Limits		**Single Detection Limit Scenario**	
Note: Data have multiple DLs - Use of KM Method is recommended		Number treated as Non-Detect with Single DL	9
For all methods (except KM, DL/2, and ROS Methods),		Number treated as Detected with Single DL	42
Observations < Largest ND are treated as NDs		Single DL Non-Detect Percentage	17.65%
Background Statistics			
Normal Distribution Test with Detected Values Only		**Lognormal Distribution Test with Detected Values Only**	
Lilliefors Test Statistic	0.611	Lilliefors Test Statistic	0.888
5% Lilliefors Critical Value	0.942	5% Lilliefors Critical Value	0.942
Data not Normal at 5% Significance Level		**Data not Lognormal at 5% Significance Level**	

Assuming Normal Distribution		Assuming Lognormal Distribution	
DL/2 Substitution Method		DL/2 Substitution Method	
Mean	786.3	Mean (Log Scale)	2.831
SD	1586	SD (Log Scale)	4.143
95% UTL 90% Coverage	3381	95% UTL 90% Coverage	14868
95% UPL (t)	3471	95% UPL (t)	18792
90% Percentile (z)	2819	90% Percentile (z)	3428
95% Percentile (z)	3396	95% Percentile (z)	15441
99% Percentile (z)	4477	99% Percentile (z)	259875
Maximum Likelihood Estimate(MLE) Method		Log ROS Method	
Mean	567.4	Mean in Original Scale	786.4
SD	1790	SD in Original Scale	1586
95% UTL with 90% Coverage	3496	95% UTL with 90% Coverage	10472
		95% BCA UTL with 90% Coverage	4890
		95% Bootstrap (%) UTL with 90% Coverage	4940
95% UPL (t)	3597	95% UPL (t)	12976
90% Percentile (z)	2862	90% Percentile (z)	2733
95% Percentile (z)	3512	95% Percentile (z)	10841
99% Percentile (z)	4732	99% Percentile (z)	143726
Gamma Distribution Test with Detected Values Only		**Data Distribution Test with Detected Values Only**	
k star (bias corrected)	0.264	Data do not follow a Discernable Distribution (0.05)	
Theta Star	3610		
nu star	22.21		
A-D Test Statistic	1.002	**Nonparametric Statistics**	
5% A-D Critical Value	0.874	Kaplan-Meier (KM) Method	
K-S Test Statistic	0.173	Mean	786.4
5% K-S Critical Value	0.149	SD	1571
Data not Gamma Distributed at 5% Significance Level		SE of Mean	222.6
		95% KM UTL with 90% Coverage	3356
Assuming Gamma Distribution		95% KM Chebyshev UPL	7700
Gamma ROS Statistics with extrapolated Data		95% KM UPL (t)	3444
Mean	786.3	90% Percentile (z)	2799
Median	35	95% Percentile (z)	3370
SD	1586	99% Percentile (z)	4440
k star	0.123		
Theta star	6408		
Nu star	12.52		
95% Percentile of Chisquare (2k)	1.396		
90% Percentile	2241		
95% Percentile	4474		
99% Percentile	11251		
Note: UPL (or upper percentile for gamma distributed data) represents a preferred estimate of BTV			
For an Example: KM-UPL may be used when multiple detection limits are present			
Note: DL/2 is not a recommended method.			

A quick comparison of statistics as summarized in Table 5-2 and Table 5-3 reveals that the inclusion of two outliers distorted all parametric and nonparametric estimates. For an example, for data set with outliers, a 95% UPL based upon KM estimates is 33,426, where as the corresponding 95% KM UPL without the two outliers is 4088. The project team should decide which of the two statistics represents a more accurate, reasonable, and realistic estimate of the background threshold value.

Furthermore, in order to illustrate the influence of outliers on various other nonparametric limits based upon ranks (e.g., UPLs, UTLs), those nonparametric upper limits (as in ProUCL 4.0) on data sets with (Table 5-4) and without (Table 5-5) outliers are summarized in the following two tables.

Table 5-4. Nonparametric Background Statistics for the Data Sets with the Outliers

	Nonparametric Background Statistics for Data Sets with Non-Detects
User Selected Options	
From File	D:\Narain\ProUCL 4.0\Data\Aroclor 1254.wst
Full Precision	OFF
Confidence Coefficient	95%
Coverage	90%
Different or Future K Values	1

Aroclor1254

Total Number of Data	53
Number of Non-Detect Data	9
Number of Detected Data	44
Minimum Detected	0.21
Maximum Detected	8300
Percent Non-Detects	16.98%
Minimum Non-detect	0.02
Maximum Non-detect	0.2
Mean of Detected Data	1143
SD of Detected Data	2002
Mean of Log-Transformed Detected Data	4.422
SD of Log-Transformed Detected Data	3.09

Data do not follow a Discernable Distribution (0.05)

Nonparametric Background Statistics

95% UTL with 90% Coverage

Order Statistic	51
Achieved CC	0.974
UTL	5400
Largest Non-detect at Order	9

95% UPL

95% UPL	5760

Kaplan-Meier (KM) Method

Mean	949.1
SD	1854
Standard Error of Mean	257.6
95% UTL 90% Coverage	3967
95% KM Chebyshev UPL	9105
95% KM UPL (t)	4083
90% KM Percentile (z)	3325
95% KM Percentile (z)	3998
99% KM Percentile (z)	5261

Note: UPL (or upper percentile for gamma distributed data)
represents a preferred estimate of BTV. For an Example:
KM-UPL may be used when multiple detection limits are present

Table 5-5. Nonparametric Background Statistics for the Data Sets without the Outliers

Nonparametric Background Statistics for Data Sets with Non-Detects	
User Selected Options	
From File	D:\Narain\ProUCL 4.0\Data\Aroclor 1254.wst
Full Precision	OFF
Confidence Coefficient	95%
Coverage	90%
Different or Future K Values	1

Aroclor1254

Total Number of Data	51
Number of Non-Detect Data	9
Number of Detected Data	42
Minimum Detected	0.21
Maximum Detected	6600
Percent Non-Detects	17.65%
Minimum Non-detect	0.02
Maximum Non-detect	0.2
Mean of Detected Data	954.8
SD of Detected Data	1704
Mean of Log-Transformed Detected Data	4.238
SD of Log-Transformed Detected Data	3.037

Data do not follow a Discernable Distribution (0.05)

Nonparametric Background Statistics

95% UTL with 90% Coverage

Order Statistic	49
Achieved CC	0.969
UTL	5000
Largest Non-detect at Order	9

95% UPL

95% UPL	5160

Kaplan-Meier (KM) Method

Mean	786.4
SD	1571
Standard Error of Mean	222.6
95% UTL 90% Coverage	3356
95% KM Chebyshev UPL	7700
95% KM UPL (t)	3444
90% KM Percentile (z)	2799
95% KM Percentile (z)	3370
99% KM Percentile (z)	4440

Note: UPL (or upper percentile for gamma distributed data) represents a preferred estimate of BTV. For an Example: KM-UPL may be used when multiple detection limits are present

5.6 Minimum Sample Size Requirements

These issues have been discussed throughout this guide and included here for convenience. Estimates of the various population (e.g., background area) parameters (e.g., not-to-exceed value) are obtained based upon sampled data collected from the population of interest. Determination of the adequate sample size

needed to obtain reliable estimates is perhaps the most important decision that one has to make when dealing with data sets with BDL observations. For data sets with and without NDs, some guidance about the minimum sample size requirements has been described in earlier chapters (e.g., Chapter 1, Chapter 4). For convenience, some of those recommendations and suggestions for data sets with NDs have been summarized below.

It is noted that even though, most of the statistics (including fitting a linear regression line) used to estimate the EPC terms and BTV estimates could be computed based upon data sets even with 2 detected values; the use of those statistics is not recommended due to high level of uncertainty associated with them. It is re-emphasized that when only a few (e.g., < 4-6) observations are detected, it is not appropriate (due to high uncertainty associated with such estimates) to replace the rest (majority) of the nondetect data by their respective DL/2 values (or some other proxy value), and compute an estimate of the EPC term or BTV based upon the resulting data set (Singh, Maichle, and Lee (2006)).

In case, only a small number (e.g., < 4-6) of detected observations are available, it is desirable to collect more samples and data (from the area under investigation), so that enough detected observations will be available to compute reasonably reliable estimates the various environmental parameters including BTVs and EPC terms. If it is not possible to collect more samples (as often is the case), it is suggested to use professional judgment and available historical information (e.g., from similar sites) to estimate the BTVs rather than using a statistical method on a fabricated data set obtained using some substitution method (e.g., DL/2). In such cases, BTVs and EPC terms may be estimated on case-by-case basis perhaps based upon the site knowledge (and information from similar sites) and experts' opinions.

For data sets consisting of mostly of nondetects, the EPC term or BTV may be estimated using simple ad hoc methods. Specifically, when all of the background sample data values are reported as nondetects, the BTV or a not-to-exceed value should also be reported as a nondetect value (or some pre-specified action level), perhaps by the maximum RL or maximum RL/2.

Also, when only a few detected values (< 4-6) are available in a data set of larger size (e.g., > 15-20), one may again use the maximum RL, or maximum RL/2 as an estimate of the BTV or some pre-specified action level. When only a few detected values (e.g., < 4-6) are available in data sets of smaller sizes (e.g., < 8-10), one may use the maximum detected data value or the second largest detected value to estimate the BTV.

Note: *The uncertainty associated with all such estimates listed above will be high; and statistical properties such as bias, accuracy, and precision of such estimates would remain unknown. It is suggested to avoid using statistical methods (to estimate the BTVs and other environmental parameters) on data sets with less than 4-6 detected values (more are preferred). If possible, it is desirable to collect more samples (data) with detected values. Statistics computed based upon small data sets or on data sets with only a few detected values (e.g., < 4-6) cannot be considered reliable enough to make important remediation and cleanup decisions potentially affecting the human health and the environment.*

5.7 Additional Suggestions and Comments

- *It is noted that a linear model (regression line) can be obtained even when only two (2) detected observations are available. Therefore, the methods as discussed here and also incorporated in ProUCL 4.0 can be used on data sets with two or more detected observations. Obviously, in order to be able to compute reliable estimates of nondetects*

and to compute defensible upper limits, enough detected observations should be made available.

- *If the use of appropriate data quality objectives (e.g., USEPA (2006)) is not possible to collect enough observations, every effort should be made to obtain a sample with about 8-10 detected observations. ProUCL 4.0 prints out a message when estimates (of mean, sd, BTVs) are obtained using data sets with not many detected (e.g., < 8-10) observations. For accurate and reliable results and estimates, whenever possible, at least 8-10 (more are desirable) detected observations should be made available, especially when the percentage of NDs becomes greater than 40%, 50%, and so on. This is especially true when one wants to compute the upper limits (e.g., UPLs, UTLs, and UCLs) based upon bootstrap resampling methods.*

- *Also, in order to be able to use bootstrap resampling methods, it is desirable that the data set has at least five unique observations. Otherwise, the bootstrap procedures may result is resamples with all identical values.*

- *It is important to point out that in practice, it is not reasonable or justifiable to assume or expect that the NDs should follow the same distribution as assumed for the detected values. This is especially true when multiple detection limits may be present, and the % of NDs may be quite high (e.g., > 40%). Additionally, ROS methods often yield infeasible predicted NDs such as negative values, or greater than the respective detection limits.*

Chapter 6

Single and Two-Sample Hypotheses Testing Approaches as Incorporated in ProUCL 4.0

Both single sample and two-sample hypotheses testing approaches are used to make cleanup decisions at polluted sites, and also to compare contaminant concentrations of two (e.g., site versus background) or more (several monitoring wells (MWs)) populations. This chapter briefly discusses some guidance on when to use single sample hypothesis test and when to use two-sample hypotheses approaches. These issues were also discussed in Chapter 1 of this Technical Guide. This chapter presents brief description of mathematical formulations of various parametric and nonparametric hypotheses testing approaches as available in ProUCL 4.0. ProUCL 4.0 also provides hypotheses testing approaches for data sets with nondetect (ND) observations. The details of those approaches can also be found in EPA (2006).

Note: *As discussed in Chapter 1, it should be pointed out that while developing ProUCL 4.0, emphasis is given to the practical applicability of the estimation and hypotheses testing methods as incorporated in ProUCL 4.0. It should also be noted that ProUCL 4.0 provides many graphical and statistical methods often used in the various statistical applications. ProUCL 4.0 assumes that the user (project team) has collected appropriate amount of good quality data, perhaps based upon DQOs. Consequently, ProUCL 4.0 does not deal with statistical methods that may be used to compute sample sizes based upon performance measures and DQO processes (EPA, 2006). Those sample size determination methods have been discussed in detail in other EPA documents (EPA, 1997, EPA 2006). Moreover, some freeware packages such as VSP (2005) and DataQUEST (EPA, 1997) are also available that may be used to compute DQOs-based sample sizes from the populations under study. However, as mentioned before, some practical guidance on the minimum sample size requirements to be able to use methods as available in ProUCL 4.0 has been provided in Chapter 1. Similar statements and suggestions have been made in some other chapters (e.g., Chapters 4 and 5) of this Technical Guide.*

6.1 When to Use Single Sample Hypotheses Approaches

When pre-established BTVs and not-to-exceed values are used, such as the USGS background values (Shacklette and Boerngen (1984)), thresholds obtained from similar sites, pre-established thresholds and not-to-exceed values, PRGs, or RBCs, there is no need to extract, establish, or collect a background or reference data set. When the BTVs and cleanup standards are known, one-sample hypotheses are used to compare site data (provided enough site data are available) with known and pre-established threshold values. It is suggested that the project team determine (e.g., preferably using appropriate DQOs) or decide (depending upon resources) about the number of site observations that should be collected and compared with the "pre-established" standards before coming to a conclusion about the status (clean or polluted) of the site area (e.g., RU, AOC) under investigation. When the number of available detected site samples is less than 4-6, one might perform point-by-point site observation comparisons with a BTV; and when enough detected site observations (> 8 to 10, more are preferable) are available, it is desirable to use single sample hypothesis testing approaches. Some of these issues have also been discussed in Chapter 1 of this Technical Guide.

Depending upon the parameter (e.g., the average value, μ_0, or a not-to-exceed value, A_0), represented by the known threshold value, one can use single sample hypothesis tests for population mean (t-test, sign

test) or single sample tests for proportions and percentiles. Several single sample tests listed as follows are available in ProUCL 4.0.

One-Sample t-Test: This test is used to compare the site mean, μ, with some specified cleanup standard, C_s *(or C)*, where the cleanup standard, C_s or C represents a specified value of the true average threshold value, μ. The Student's t- test (or a UCL of mean) is often used (assuming normality of site data, or when site sample size is large such as larger than 30, 50, 100) to determine the attainment of cleanup levels at a polluted site, perhaps after some remediation activities performed at the site. Note that the large sample size requirement (30, 50, 100, ...) depends upon the data skewness. Specifically, as skewness increases (measured in terms of *sd* of log-transformed data), the large sample size requirement also increases to be able to use central limit theorem (CLT).

One-Sample Sign Test or Wilcoxon Signed Rank (WSR) Test: These tests are nonparametric tests and can also handle nondetect observations provided all nondetects (e.g., associated detection limits) fall below the specified threshold value, C_s(C). These tests are used to compare the site location (e.g., median, mean) with some specified cleanup standard, C_s or C representing the similar location measure.

One-Sample Proportion Test or Percentile Test: When a specified cleanup standard, $A_{0,}$ such as a PRG, BTV, compliance limit, or a not-to-exceed value represents an upper threshold value of a contaminant concentration distribution (e.g., not-to-exceed value, compliance limit) rather than the mean threshold value, μ, of the contaminant concentration distribution, then:

A test for proportion or a test for percentile (or equivalently a UTL 95%-95%, UTL 95%-90%) might be used to compare the site proportion, p of exceedances of the action level, A_0 by site observations to some pre-specified allowable proportion, P_0 of exceedances of A_0 by site observations (perhaps after some remediation activity). It is noted that the proportion test can also handle NDs provided all NDs are below the action level, A_0. A test for single sample proportion has been incorporated in ProUCL 4.0.

As always, it is desirable to use a sampling plan based upon a DQO process to collect appropriate amount of detected data. In any case, in order to obtain reasonably reliable estimates and compute reliable test statistics, an adequate amount of representative site data (at least 8 to 10 detected observations) should be made available to perform single sample hypotheses tests listed above.

Note: *As mentioned before, in case only a few (e.g., < 4 to 6) detected site observations are available, instead of using hypotheses testing approaches, point-by-point site concentrations may be compared with the specified action level, A_0. It should be noted that individual point-by-point observations are not compared with the average cleanup or threshold level, C or C_s.*

6.2 When to Use Two-Sample Hypotheses Testing Approaches

When BTVs, not-to-exceed values, and other cleanup standards are not available, then site data are compared directly with the background data. In such cases, a two-sample hypothesis testing approach can be used to perform site versus background comparisons provided enough data are available from each of the two populations. Note that this approach can be used to compare concentrations of any two populations including two different site areas or two different monitoring wells (MWs). In order to use and perform a two-sample hypothesis testing approach, enough data of good quality should be available (collected) from each of the two populations under investigation. Site and background data requirements (e.g., based upon DQOs) to perform two-sample hypothesis test approaches are described in EPA (2006),

Breckenridge and Crockett (1995), and the VSP (2005) software package. Some minimum sample size requirements are provided in Chapter 1 of this Technical guidance document.

While collecting site and background data, for better representation of populations under investigation, one may also want to account for the size of the background area (and site area for site samples) into sample size determinations. That is, a larger number (>10 to 15) of representative background (or site) samples should be collected from larger background (or site) areas. As mentioned before, every effort should be made to collect as many samples as determined using DQO processes as described in EPA documents (2006).

The two-sample (or more) hypotheses approaches are used when the site parameters (e.g., mean, shape, distribution) are being compared with the background parameters (e.g., mean, shape, distribution). The two-sample hypotheses testing approach is also used when the cleanup standards or screening levels are not known *a priori*, and they need to be estimated based upon a data set from a background or reference population. Specifically, two-sample hypotheses testing approaches are used to compare: 1) the average (also medians or upper tails) contaminant concentrations of two or more populations such as the background population and the potentially contaminated site areas, or 2) the proportions of site and background observations exceeding a pre-established compliance limit, A_0. In order to derive reliable conclusions with higher statistical power based upon hypothesis testing approaches, enough data (e.g., minimum of 8 to 10 samples) should be available from all of the populations under investigation.

It is desirable and recommended to always supplement statistical methods and test statistics with graphical displays, such as the double Q-Q plots, or side-by-side multiple box plots, as available in ProUCL 4.0. Several parametric and nonparametric two-sample hypotheses testing approaches, including Student's t-test, the Wilcoxon-Mann-Whitney (WMW) test, Gehan's test, and the quantile test are included in ProUCL 4.0. Some details of those methods are described in this chapter. It should be noted that the WMW, Gehan, and quantile tests can also be used on data sets with NDs. Gehan's test is specifically meant to be used on data sets with multiple detection limits. It is suggested that for best and reliable conclusions, both WMW test and quantile tests should be used in parallel on the same data set. The details of these two tests with examples are given in EPA (1994, 2006).

The samples collected from the two (or more) populations should all be of the same type obtained using similar analytical methods and apparatus. In other words, the collected site and background samples should be all discrete or all composite (obtained using the same design and pattern), and be collected from the same medium (soil) at similar depths (e.g., all surface samples or all subsurface samples) and time (e.g., during the same quarter in groundwater applications) using comparable (preferably same) analytical methods. Good sample collection methods and sampling strategies are given in EPA (1996, 2003) guidance documents.

Note: *As mentioned before, it is noted that ProUCL 4.0 does not deal with DQOs and sample size determinations for the various statistical applications. There are other free software packages (e.g., VSP, 2005, and DataQUEST, 1997) available that may be used to compute DQO-based sample sizes for the populations under investigations. In ProUCL 4.0, emphasis is given to the practical applicability of statistical and graphical methods that may be used to make sense of data collected from the various environmental applications and studies. Specifically, it is assumed that the user has collected adequate amount of data of good quality (perhaps based upon appropriate DQOs) to be able to use statistical methods as available in ProUCL 4.0.*

In order to familiarize the users with the statistical terminology used in all hypotheses testing approaches as incorporated in ProUCL 4.0, a brief discussion of the various terms used is described next. A brief discussion of the error rates and associated sample sizes is also included in the following. Detailed descriptions of these terminologies can be found in EPA (2002, 2006).

6.3 Statistical Terminology Used in Hypotheses Testing Approaches

The first step in developing a hypothesis test is to transform the problem into statistical terminology by developing a *null hypothesis*, H_0, and an *alternative hypothesis*, H_A. These hypotheses tests result in two alternative decisions (acceptance of the null hypothesis or the rejection of the null hypothesis) that a hypothesis test statistic (e.g., t-statistic, WMW test statistic) will evaluate and determine. In this section, these terminologies (e.g., error rates, hypotheses statements, Form 1, Form2, two sided tests) are explained in terms of two-sample hypotheses testing approaches often used to compare background and site parameters and data distributions. Similar terminology applies to all parametric and nonparametric hypotheses testing approaches include the single sample and two-sample hypotheses testing approaches, as incorporated in ProUCL 4.0.

In two-sample comparisons such as background area versus potentially impacted site area comparisons, the parameter of interest is symbolized by the Greek letter, *delta* (Δ), the amount by which the mean of the distribution of concentrations in potentially impacted areas exceeds the mean of the background distribution. Delta is an unknown parameter, and statistical tests may be used to evaluate hypotheses relating to its possible values. The statistical tests are designed to reject or not reject hypotheses about Δ based on test statistics computed based upon sampled data collected from the background and the impacted site area(s).

The action level for background comparisons is the largest value of the "difference in means" that is acceptable to the decision maker. In this guidance, the action level for the difference (site mean - background mean) in means is defined as a substantial difference, S, which may be zero or a positive value based on the risk assessment, an applicable regulation, a screening level, or guidance. In some cases, the largest acceptable value for the difference in means may be S = 0. Typically, the value for "S" is determined on a case-by-case basis and the analyte under comparison. Some discussion on the selection of the substantial difference, S, is given in Appendix A of the *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b). It should be noted that whenever applicable, ProUCL 4.0 provides an option to select a suitable value of S while testing for Form 2 null hypotheses defined below. If the user is not very sure about the selection for a value of the substantial difference, S, the user may perform the test for more than one value of S. If deemed necessary, the user may perform sensitivity analyses by using ProUCL 4.0 for several values of S.

Estimates of Δ are obtained by measuring the differences in contaminant concentrations in potentially impacted areas and in background areas. For example, one estimate of the mean concentration in potentially impacted areas is the simple arithmetic average of measurements from those site areas. An estimate of the mean background concentration is similarly calculated. An estimate of the difference in population means, Δ, is obtained by subtracting the mean background concentration from the mean concentration in potentially impacted areas. In most cases of interest, the estimate of Δ will be a positive number. If there is little or no contamination on the site, then the estimate for Δ may be near zero or slightly negative. Note that the estimated value for Δ calculated by using this simple procedure (or by some more complicated procedure) only represents an estimate of the true value of Δ. Hence, decisions based on any estimated value for Δ may be incorrect due to uncertainty and variability associate with its estimated value.

Adopting hypothesis tests and an appropriate data quality objective (DQO) approach (EPA, 2006) can control the probability of making decision errors. However, incorrect use of hypothesis tests can lead to erratic decisions. Each type of hypothesis test is based on a set of assumptions that should be verified to confirm proper use of the test. Procedures and assumptions for determining the selection and proper use of parametric tests (such as a t-test), and nonparametric test (e.g., WMW test) are provided in (EPA, 2006). It is pointed out that some goodness-of-fit (GOF) tests and graphical procedures are available in ProUCL 4.0 to verify the distributional assumptions.

Hypothesis testing is a quantitative method to determine whether or not a specific statement (called the null hypothesis) concerning Δ can be rejected. Decisions concerning the true value of Δ reduce to a choice between "yes" or "no." When viewed in this way, two types of incorrect decisions, or decision errors, may occur:

Incorrectly deciding the answer is "yes" when the true answer is "no"; and
Incorrectly deciding the answer is "no" when the true answer is "yes."

While the possibility of decision errors can never be completely eliminated, it can be controlled. To control decision errors, it is necessary to control the uncertainty in the estimate of Δ. It should be noted that the uncertainty typically arises from three sources:

1. Sampling error;

2. Measurement error; and

3. Natural site variability.

The decision maker has some control of the first two sources of uncertainty. For example, a larger number of samples may lead to fewer decision errors because the probability of a decision error decreases as the number of samples increases. Use of more precise measurement techniques or duplicate measurements can reduce measurement error, thus minimizing the likelihood of a decision error. The third source of uncertainty is more difficult to control. However, the site variability may also be controlled to some extent if more rigorous sampling procedures are used to collect samples. Some good sample collection methods and sampling strategies (e.g., based upon Gy sampling theory) are given in EPA (1996, 2003) guidance documents.

Natural variability arises from the uneven distribution of chemical concentrations on the site and in background areas. Natural variability is measured by the true standard deviation, σ of the concentration distribution. A large value of σ indicates that a large number of measurements will be needed to achieve a desired limit on decision errors. Since variability is usually higher in impacted areas of the site than in background locations, data collected from the site is used to estimate σ. An estimate for σ frequently is obtained from historical data, if available. Estimates of variability reported elsewhere at similar sites with similar contamination problems may also be used. If an estimate of the mean concentration in contaminated site areas is available, then the coefficient of variation (*CV*) observed at other sites may be multiplied by the mean to obtain an estimate the standard deviation (*sd*). If no acceptable historical source for an estimate of σ is available, then it may be necessary to conduct a small-scale pilot survey on site using 20 or more random samples to estimate σ. Due to the small sample size of the pilot, it is advisable to use an 80 or 90 percent upper confidence limit for the estimate of σ rather than an unbiased estimate to

avoid underestimating the true variability. A crude estimate of σ is obtained by dividing the anticipated range (maximum - minimum) by number, 6.

The hypothesis testing process provides a formal procedure to quantify the decision maker's acceptable limits for decision errors. The decision maker's limits on decision errors are used to establish performance goals for data collection that reduce the chance of making decision errors of both types. The gray region is a range of possible values of Δ where the consequences of making a decision error (of both type) are relatively minor. Specifically, a reasonable statistical test should have a low probability of reflecting a substantial positive (> 0) difference, S, when the site and background distributions are comparable (false positive), but has a high probability of reflecting a substantial difference when the concentrations in potentially impacted areas greatly (significantly) exceed the background concentrations. In the gray region between these two extremes, the statistical test has relatively poor performance. When the test procedure is applied to a site with a true mean concentration in the gray region, the test may indicate that the site exceeds background, or may indicate that the site does not exceed background, depending on random fluctuations in the sample.

It is necessary to specify a gray region for the test because the decision may be "too close to call" due to uncertainty in the estimate of Δ. This may occur when the difference in means is small compared to the minimum detectable difference (MDD) for the test. The minimum detectable difference (MDD) is the smallest difference in means that the statistical test can resolve. The MDD depends on sample-to-sample variability, the number of samples, and the power of the statistical test. In the gray region, the uncertainty in the measurement of Δ is larger than the difference between Δ and the action level, so it may not be possible for the test to yield a correct decision with a high probability. One step in the hypothesis test procedure is to assign upper bounds on the decision error rates for values of Δ above and below the gray region. These bounds limit the probability of occurrence of decision errors.

Typically, the gray region depends upon the type of hypothesis test that is selected by the decision maker. In general, the gray region for Δ is to the right of the origin (Δ = 0) and bounded from above by the substantial difference (Δ = S). Some guidance on specifying a gray region for the test is available in Chapter 6 of *Guidance for the Data Quality Objectives Process* (EPA, 1994). The size of the gray region may also depend on specific regulatory requirements or policy decisions that may not be addressed in the DQO guidance.

The width of the gray region is called the "minimum detectable difference" for the statistical test, indicating that differences smaller than the MDD cannot be detected reliably by the test. If a test is used to determine if concentrations of the potentially impacted areas exceed background concentrations by more than S units, then it is necessary to ensure that the MDD for the test is less than S. In the planning stage, this requirement is met by designing a sampling plan with sufficient power to detect differences as small as S.

Note: *If possible, it is suggested that if data were collected without the benefit of a sampling plan (e.g., with specified decision errors),then the practitioners or the project team may want to*

Table 6-1. Required Sample Sizes for Selected Values of σ (α = β = 0.10 and MDD = 50 mg/kg)

σ (mg/kg)	MDD/σ	n	N
25	2	3.7	5
50	1	13.55	16
75	0.67	29.97	35
100	0.5	52.97	62
125	0.4	82.53	96
150	0.33	118.66	138
175	0.29	161.36	188
200	0.25	210.63	245

perform retrospective power calculation of the test before making a decision about the acceptance or rejection of the null hypothesis under consideration.

It is also noted that the use of graphical displays (e.g., side-by-side box plots, multiple Q-Q plots) provide added insight and information about the comparability of data sets that may not be possible to observe and assess based upon simple test statistics such as t-test or Gehan's test statistic.

In the planning stage, the absolute size of the MDD is of less importance than the ratio of the MDD to the natural variability of the contaminant concentrations in the potentially impacted area. This ratio is termed the "relative difference" and defined as MDD/σ, where σ is an estimate of the standard deviation of the distribution of concentrations on the site. The relative difference expresses the power of resolution of the statistical test in units of uncertainty. Relative differences much less than one standard deviation (MDD/σ < 1) are more difficult to resolve unless a larger number of measurements are available. Relative differences of more than three standard deviations (MDD/σ > 3) are easier to resolve. As a general rule, values of MDD/σ near 1 will result in acceptable sample sizes. The required number of samples may increase dramatically when MDD/σ is much smaller than one. Conversely, designs with MDD/σ larger than three may be inefficient. If MDD/σ is greater than three, then additional measurement precision is available at a minimal cost by reducing the width of the gray region. The cost of the data collection plan should be examined quantitatively for a range of possible values of the MDD before selecting a final value. A tradeoff exists between cost (number of samples required) and benefit (better power of resolution of the test).

The number of measurements required to achieve the specified decision error rates has an inverse relationship with the value of MDD/σ. An example of this inverse relationship is demonstrated in Table 6-1 for hypothetical values of $\alpha = \beta = 0.10$ and MDD = 50 mg/kg. Sample sizes may be obtained using the approximate formula given in EPA (2006), written here as:

$$n = (0.25)\, z^2_{1-\alpha} + 2\,(z_{1-\alpha} + z_{1-\beta})^2\, \sigma^2 / (MDD)^2,$$

Here z_p is the p^{th} percentile of the standard normal distribution. Note the inverse-squared dependence of n on MDD/σ. The smaller values of α and β (leading to larger values for the z terms) magnify the strength of this inverse relationship. A recommended sample size of N = (1.16)*n is tabulated for a variety of σ values in the table. Note the dramatic increase in the sample size as the value of MDD/σ is lowered from 1 to 0.25.

Table 6-2. Achievable Values of $\alpha = \beta$ for Selected Values of N with MDD/ σ = 0.5

N	n	$z_{1-\alpha}$	$\alpha = \beta$
10	8.62	0.517	0.303
15	12.93	0.633	0.263
20	17.24	0.731	0.232
25	21.55	0.817	0.207
30	25.86	0.896	0.185
40	34.48	1.034	0.151
50	43.1	1.156	0.124
60	51.72	1.266	0.103
70	60.34	1.368	0.086
100	86.21	1.635	0.051
150	129.31	2.002	0.023
200	172.41	2.312	0.01

Letting $\alpha = \beta$, we can solve for $z_{1-\alpha} = z_{1-\beta}$:

$$z^2_{1-\alpha} = n / [0.25 + 8\,\sigma^2 / (MDD)^2\,].$$

For any fixed value of MDD/σ, the decision error rate, α, is a function of n:

$$\alpha = 1 - \Phi[z_{1-\alpha}(n)],$$

Here Φ is the cumulative normal distribution function. Achievable values of α (and β) for selected sample sizes with a hypothetical value of MDD/σ = 0.5 are shown in Table 6-2.

A tradeoff analysis should begin with the analysis of the choice MDD = S, where S is a substantial difference. Note that a choice of MDD > S would lead to a sample size that does not have sufficient power to distinguish a difference between the site and background means as small as S. Hence, the minimum acceptable number of samples for the decision is obtained when MDD = S. If S/σ is less than one, then this indicates that MDD/σ is also less than one, and a relatively large number of samples will be required to make the decision. If S/σ exceeds three, then a reasonably small number of samples are required for this minimally acceptable test design. More information about the choice of MDD and S can be found in EPA (2002).

Two forms (Form 1 and Form 2) of the statistical hypothesis test are useful for many environmental applications. In addition to testing the two forms of hypotheses testing, ProUCL 4.0 can also be used to perform two-sided hypotheses tests. In this section, these forms of hypotheses are described for two-sample hypotheses tests often used to compare site concentrations with background concentrations. The null hypothesis in the first form (Form 1) of the test states that there is *no statistically significant difference* between the means of the concentration distributions measured at the site and in the selected background areas. The null hypothesis in the second form (Form 2) of the test is that the concentrations of the impacted site *exceed the background concentrations by a substantial difference, S*. Both forms of the null hypothesis are described next.

6.3.1 Test Form 1

The null hypothesis for background comparisons, "the concentrations in potentially impacted sites areas do not exceed (or are less than equal) background concentrations," is formulated for the express purpose of being rejected.

- *The null hypothesis (H_0)*: The mean contaminant concentration in samples collected from potentially impacted areas is less than or equal to the mean concentration in samples collected from background areas (H_0: $\Delta \leq 0$).

- *The alternative hypothesis (H_A)*. The mean contaminant concentration in samples from potentially impacted areas is greater than the mean in background areas (H_A: $\Delta > 0$).

When using this form of hypothesis test, the collected data should provide statistically significant evidence that the null hypothesis is false leading to the conclusion that the site mean does exceed background mean concentration. Otherwise, the null hypothesis cannot be rejected based on the available data, and the concentrations found in the potentially impacted areas are considered equivalent and comparable those of the background.

Two serious problems arise when using the Background Tests Form 1. One type of problem arises when there is a very large amount of data. In this case, the MDD for the test will be very small, and the test may reject the null hypothesis when there is only a very small difference between the site and background mean concentrations. If the site exceeds background by only a small amount, then there is a very high probability that the null hypothesis will be rejected if a sufficiently large number of samples were taken. This case can be avoided by selecting the Background Test Form 2, which incorporates an acceptable level for the difference between the site and the background concentrations (see Section 6.3.2).

A second type of problem may arise in the use of the Background Test Form 1 when insufficient data are available. This may occur, for example, when the onsite or background variability was underestimated in the design phase. An estimated value for σ is used during the preliminary phase of the DQO planning process to determine the required number of samples. When the samples are actually collected, σ can then be re-estimated, and the power of the analysis should be re-evaluated. If the variance estimate used in the planning stage was too low, the statistical test is unlikely to reject the null hypothesis due to the lack of sufficient power.

Hence, when using the Background Test Form 1, if possible (if resources are available), it is desirable to conduct a *retrospective power analysis* to ensure that the power of the test was adequate to detect a site area with a mean contamination that exceeds the background by more than the MDD. A procedure to perform this is to re-compute the required sample size using the sample variance in place of the estimated variance that was used to determine the required sample size in the planning phase. If the actual sample size is greater than this post-calculated size, then it is likely that the test has adequate power. Alternatively, if the retrospective analysis indicates that adequate power was not obtained, it may be necessary to collect more samples. Hence, if large uncertainties exist concerning the variability of the contaminant concentration in potentially impacted areas, the Background Test Form 1 may lead to inconclusive results. Therefore, the sample size should exceed the minimum number of samples required to give the test sufficient power.

Note: *In cases, when it is not possible to collect enough samples, one should also use graphical displays to compare site concentrations with background concentrations to gain additional information about the populations under comparison. It should be noted that a minimum of at least 8-10 observations should be made available to include and use graphical displays in the decision making processes regarding site versus background comparisons.*

6.3.2 Test Form 2

In the Background Test Form 2, the null hypothesis is stated as, "the concentration in potentially impacted areas exceeds background concentration."

- *The null hypothesis (H_0):* The mean contaminant concentration in potentially impacted areas exceeds the background mean concentration by more than S units. Symbolically, the null hypothesis is written as H_0: $\Delta \geq S$.

- *The alternative hypothesis (*H_A*):* The mean contaminant concentration in potentially impacted areas does not exceed the background mean concentration by more than S (H_A: $\Delta < S$).

Here, S is the background investigation level. Although there is no explicit use of the quantity S in the hypothesis statement used in the Background Test Form 1, an estimate of S is important for determining an upper limit for the MDD for the Background Test Form 1. The background investigation level, S, is determined on a case-by-case basis by the project team, EPA, and other stakeholders. ProUCL 4.0 can also be used to determine a value for S (by performing sensitivity analysis to determine a value of S).

6.3.3 Selecting a Test Form

The test forms described above are commonly used in background versus site comparison evaluations. Therefore, these test forms are also known as Background Test Form 1 and Background Test Form 2

(EPA, 2002). Background Test Form 1 uses a conservative investigation level of $\Delta = 0$, but relaxes the burden of proof by selecting the null hypothesis that the contaminant concentrations in potentially impacted areas are not statistically greater than the background concentrations. Background Test Form 2 requires a stricter burden of proof, but relaxes the investigation level from 0 to S.

6.3.4 Errors Tests and Confidence Levels

A key step in developing a sampling and analysis plan is to establish the level of precision required of the data (EPA (1994)). Whether a null hypothesis is rejected or not depends on the results of the sampling. Due to the uncertainties that result from sampling variation, decisions made using hypotheses tests will be subject to errors, also known as decision errors. Decisions should be made about the width of the gray region and the degree of decision errors that is acceptable. There are two ways to err when analyzing data (Table 6-3).

- *Type I Error*: Based on the observed data, the test may reject the null hypothesis when in fact the null hypothesis is true (a false positive or equivalently a false rejection). This is a *Type I error*. The probability of making a Type I error is often denoted by α (*alpha*); and

- *Type II Error*: On the other hand, the test may fail to reject the null hypothesis when the null hypothesis is in fact false (a false negative or equivalently a false acceptance). This is called *Type II error*. The probability of making a Type II error is denoted by β (*beta*).

Table 6-3. Hypothesis Testing: Type I and Type II Errors

Decision Based on Sample Data	**Actual Site Condition**	
	H_0 is True	**H_0 is not true**
H_0 is not rejected	Correct Decision: $(1 - \alpha)$	Type II Error: False Negative (β)
H_0 is rejected	Type I Error: False Positive (α)	Correct Decision: $(1 - \beta)$

The *acceptable level of decision error* associated with hypothesis testing is defined by two key parameters – *confidence level* and *power*. These parameters are related to the two error probabilities, α and β.

- *Confidence level* $100(1 - \alpha)\%$: As the confidence level is lowered (or alternatively, as α is increased), the likelihood of committing a Type I error increases.

- *Power* $100(1 - \beta)\%$: As the power is lowered (or alternatively, as β is increased), the likelihood of committing a Type II error increases.

Although a range of values can be selected for these two parameters, as the demand for precision increases, the number of samples and the associated cost (e.g., sampling cost) will generally also increase. The cost of sampling is often an important determining factor in selecting the acceptable level of decision errors. However, unwarranted cost reduction at the sampling stage may incur greater costs later in terms of increased threats to human health and the environment, or unnecessary cleanup at a site area under investigation. The number of samples, and hence the cost of sampling, can be reduced but at the expense of a higher possibility of making decision errors that may result in the need for additional sampling, unnecessary remediation, or increased risk to the human health and the environment.

The selection of appropriate levels for decision errors and the resulting number of samples is a critical component of the DQO process that should concern all stakeholders. Because there is an inherent tradeoff between the probabilities of committing Type I or Type II error, a simultaneous reduction in both types of errors can only occur by increasing the number of samples. If the probability of committing a false positive error is reduced by increasing the level of confidence associated with the test (in other words, by decreasing,α), the probability of committing a false negative is increased because the power of the test is reduced (increasing β).

Typically, the following values for error probabilities are selected as the minimum recommended performance measures (EPA, 1990 and EPA, 2002).

- For the Background Test Form 1, the confidence level should be at least 80% ($\alpha = 0.20$) and the power should be at least 90% ($\beta = 0.10$).

- For the Background Test Form 2, the confidence level should be at least 90% ($\alpha = 0.10$) and the power should be at least 80% ($\beta = 0.20$).

It should be noted that when using the Background Test Form 1, a Type I error (false positive) is less serious than a Type II error (false negative). This approach favors the protection of human health and the environment. To ensure that there is a low probability of committing Type II errors, a Test Form 1 statistical test should have adequate power at the right edge of the gray region.

When the Background Test Form 2 is used, a Type II error is preferable to committing a Type I error. This approach favors the protection of human health and the environment. The choice of the hypotheses used in the Background Test Form 2 is designed to be protective of human health and the environment by requiring that the data contain evidence of no substantial contamination.

6.4 Parametric Hypotheses Tests

Parametric statistical tests assume that the data have a known distributional form. They may also assume that the data are statistically independent or that there are no spatial trends in the data.
Many statistical tests and models are only appropriate for data that follow a particular distribution. Statistical tests that rely on knowledge of the form of the population distribution for the data are known as *parametric* tests, because the test is usually phrased in terms of the parameters of the distribution assumed for the data. The most important distribution for tests involving environmental data is the normal distribution. Till recent past, the use of a lognormal distribution has been quite common in many environmental applications. It is well known that the use of a lognormal distribution often results in unstable and impractical estimates of the EPC terms (e.g., Singh, Singh, and Iaci, 2002, and Singh, Maichle, and Lee, 2006) and t-test statistic used on log-transformed data. Some of these issues will be illustrated later in this chapter.

Goodness-of-fit tests for data distribution (such as the Shapiro-Wilk test for normality) often fail if there are insufficient data, if the data contain multiple populations, or if there is a high proportion of nondetects in the collected data set. Tests for normality lack statistical power for small sample sizes. In this context, a sample consisting of less than 20 observations (EPA, 2002) may be considered a small sample. However, in practice, many times it is not possible and feasible (due to resource constraints) to collect data sets of sizes greater than 8-10. This is especially true for background data sets. Furthermore, the decision makers often do not want to collect many background samples, and they want to make cleanup decisions based upon data sets of sizes even smaller than 8. As discussed in Chapter 1, it again is suggested to avoid

deriving cleanup decisions (with potential effect on the human health and the environment) based upon data sets of small sizes such as smaller than 8.

Note: *Statistics computed based upon smaller data sets of sizes 4 to 5 cannot be considered reliable enough to derive important decisions affecting the human health and the environment. Every effort should be made by all parties involved including the project team and the decision makers to provide enough resources and budget so that adequate amount of data (perhaps based upon DQOs) can be collected from the various areas under investigation.*

6.5 Nonparametric Hypotheses Tests

Statistical tests that do not assume a specific statistical form for the population distribution(s) are called distribution-free or *nonparametric* statistical tests. Nonparametric tests have good test performance for a wide variety of distributions, and their performances are not unduly affected by the outlying observations. Nonparametric tests can be used for normal or non-normal data sets. In two-sample comparisons, if one or both of the data sets fail to meet the test for normality, or if the data sets appear to come from different distributions, then nonparametric tests may be used to perform site versus background comparisons. Several two-sample nonparametric hypotheses tests (e.g., the WMW test, the Gehan test, the quantile test) are available in ProUCL 4.0.

The relative performances of different testing procedures may be assessed by comparing, *p-values* associated with those tests. The *p-value* of a statistical test is defined as the smallest value of α (level of significance, Type I error) for which the null hypothesis would be rejected for the given set of observations. The *p-value* of a test is sometimes called the critical level or the significance level of the test.

Performance of statistical tests is also compared based on their *robustness*. Robustness means that the test has good performance for a wide variety of data distributions, and that its performance is not significantly affected by the occurrence of outliers. It should be pointed out that not all nonparametric methods are robust and resistant to outliers. Specifically, nonparametric upper limits used to estimate BTVs and not-to-exceed values can get affected and misrepresented by outliers. This issue has been discussed in Chapter 3 of this Technical Guide. In addition, nonparametric tests used to compare population means and medians generally are unaffected by a reasonable number of nondetect values. There are other circumstances and situations that should be considered:

- If a parametric test for comparing means is applied to data from a non-normal population and the sample size is large, then the parametric test may work well provided that the data sets are not heavily skewed. For heavily skewed data sets, the sample size requirement associated with central limit theorem (CLT) can become quite large such as larger than 100 (Singh, Singh, and Iaci, 2002, Singh and Singh, 2003). For moderately skewed (as defined in Chapter 4) data sets, the CLT ensures that parametric tests for the mean will work because parametric tests for the mean are robust to deviations from normal distributions as long as the sample size is large. Unfortunately, the answer to the question of how large is large enough depends on the nature of the particular distribution. Unless the population distribution is very peculiar (e.g., highly skewed), one may choose a parametric test for comparing means when there are at least 25-30 data points in each group.

- If a nonparametric test for comparing means is applied to data from a normal population and the sample size is large, then the nonparametric test will work well. In this case, the p-values tend to be a little too large, but the discrepancy is small. In other words, non-parametric tests for comparing means are only slightly less powerful than parametric tests with large samples.

- If a parametric test is applied to data from a non-normal population and the sample size is small (for example, less than 20 data points), then the p-value may be inaccurate because the central limit theory does not apply in this case.

- If a nonparametric test is applied to data from a non-normal population and the sample size is small, then the p-values tend to be too high. In other words, nonparametric tests may lack statistical power with small samples.

In conclusion, large data sets do not present any problems. In this case, the nonparametric tests are powerful and the parametric tests are robust. However, small data sets are challenging. In this case, the nonparametric tests are not powerful, and the parametric tests are not robust.

Note: *It is re-stated that there is no substitute for visual graphical displays of data sets. The users should always supplement their findings and conclusions by using graphical displays for visual comparisons of two or more data sets. ProUCL 4.0 offers both side-by-side box plots and multiple Q-Q plots that can be used to graphically compare two or more populations.*

Having discussed the various terminologies associated with hypotheses testing approaches, some parametric and nonparametric single sample hypotheses testing approaches as incorporated in ProUCL 4.0 are described in the next section. Some of these approaches can also handle data sets with ND observations.

6.6 Single Sample Hypotheses Testing Approaches

This section briefly describes the mathematical formulation of parametric and nonparametric single sample hypotheses testing approaches as incorporated in ProUCL 4.0. Some details of the single sample hypotheses tests can be found in EPA (2006).

6.6.1 The One-Sample t-Test

The one-sample t-test is a parametric test used for testing a difference between a population (site area, AOC) mean and a fixed pre-established threshold also representing a mean concentration level. Some minimum sample size requirements associated with this test have been described in Chapter 1. However, it is suggested that every effort should be made to collect adequate amount of data, perhaps using an appropriate DQO process. The collected sample should represent a random sample properly representing the site area or an area of concern (AOC) under investigation.

6.6.1.1 Limitations and Robustness

- This test is not robust in the presence of outliers.
- Does not give reliable results in the presence of less than values. It is suggested not to use this test when dealing with data sets with NDs. Some other nonparametric tests described as follows may be used in case NDs are present in the samples data set.

- It may yield reliable results when performed on mildly or moderately skewed data sets. Note that skewness is discussed in Chapters 3 and 4.
- Its use should be avoided when data are highly skewed, even when the data set is of a large size such as 100.

6.6.1.2 Directions for the One-Sample t-Test

Even though, the user can perform this test using ProUCL4.0, a brief description of the t-test is given in this section. Let $X_1, X_2, \ldots, X_n$ represent a random sample of size, n, from the site area (an AOC, EA) under investigations. The following directions are written keeping in mind that the user may be manually performing the test. It is pointed out ProUCL 4.0 computes all statistics and prints out the conclusion based upon the data set.

STEP 1: Specify an average cleanup goal/level, μ_0 (or C, C_s), and state the following null hypotheses (as available in ProUCL 4.0):

Form 1: H_0: site $\mu \leq \mu_0$ vs. alternative H_1: site $\mu > \mu_0$
Form 2: H_0: site $\mu \geq \mu_0$ vs. alternative H_1: site $\mu < \mu_0$
Two Sided: H_0: site $\mu = \mu_0$ vs. alternative H_1: site $\mu \neq \mu_0$.
Form 2 with substantial difference, S: H_0: site $\mu \geq \mu_0 + S$ vs. alternative H_0: site $\mu < \mu_0 + S$, here $S \geq 0$.

STEP 2: Calculate the test statistic:

$$t_0 = \frac{\bar{X} - \mu_0 - S}{\frac{sd}{\sqrt{n}}} \qquad (6\text{-}1)$$

Note: *In the above equation, S is assumed to be equal to "0", except for Form 2 with substantial difference.*

STEP 3: Use a t-table (ProUCL 4.0 computes them) to find the critical values of t-statistic.

Conclusion:
Form 1: If $t_0 > t_{n-1,\, 1-\alpha}$, then reject the null hypothesis that the site population mean is less than the cleanup level.
Form 2: If $t_0 < -t_{n-1,\, 1-\alpha}$, then reject the null hypothesis that the site population mean exceeds the cleanup level.
Two Sided: If $|t_0| > t_{n-1,\, 1-\alpha/2}$, then reject the null hypothesis that the site population mean is same as the cleanup level.
Form 2 with substantial difference, S: If $t_0 < -t_{n-1,\, 1-\alpha}$, then reject the null hypothesis that the site population mean is more than the cleanup level, μ_0 + the substantial difference, S.

P-values

A p-value is the smallest value for which the null hypothesis is rejected in favor of the alternative hypotheses. Thus, based upon the given data, the null hypothesis is rejected for all values of α (the level of significance) greater than or equal to the p-value. The details of computing a p-value for t- test can be found in any statistical text book such as Daniel (1995). ProUCL 4.0 computes p-values for t-test

associated with each form of null hypothesis. If the computed p-value is smaller than the specified value of, α, the conclusion is to reject the null hypothesis based upon the collected data set used in the various computations.

6.6.2 The One-Sample Test for Proportions

The one-sample test for proportions represents a test for a difference between the population proportion, *P*, and a fixed desired threshold proportion, P_0. Based upon the sampled data set and sample proportion, *p*, of exceedances of an action level, A_0, by the *n* site observations; the objective is to determine if the population proportion (of exceedances of a threshold value, A_0) exceeds the pre-specified proportion level, P_0. It is noted that this proportion test is equivalent to a sign test (described as follows), when $P_0 = 0.5$.

6.6.2.1 Limitations and Robustness

- Normal approximation is applicable when both (nP_0) and $n\ (1- P_0)$ are at least 5.
- For smaller data sets, ProUCL 4.0 uses exact binomial distribution (e.g., Conover, 1999) to compute the critical values when the above statement is not true.
- The proportion test may also be used on data sets with ND observations (also available in ProUCL 4.0), provided all NDs are smaller than the action level, A_0.

6.6.2.2 Directions for the One-Sample Test for Proportions

Let $X_1, X_2, \ldots, X_n$ represent a random sample of size, *n*, from the site (AOC, EA) under investigation. Let A_0 represents a compliance limit or a threshold level to be met by site data (perhaps after some remediation activity). It is expected (e.g., after remediation) that the proportion of site observations exceeding the compliance limit, A_0, is smaller than the specified proportion, P_0.

Let B = # of site values in the sample exceeding the compliance limit, A_0. A typical observed sample value of B (based upon a data set) is denoted by b. It is noted that B follows a binomial distribution (BD) ~ $B(n, P)$ with n as the sample size, and P being the unknown population proportion. Under the null hypothesis, the variable B follows ~ $B(n, P_0)$.

Note*: The sample proportion, p, is computed by comparing the n site observations with the action level, A_0. Specifically, sample proportion p =B/n = (# of site values in the sample > A_0)/n*

STEP 1: Specify a proportion threshold value, P_0, and state the following null hypotheses:

Form 1: $H_0: P \leq P_0$ *vs.* $H_1: P > P_0$
Form 2: $H_0: P \geq P_0$ *vs.* $H_1: P < P_0$
Two Sided: $H_0: P = P_0$ *vs.* $H_1: P \neq P_0$

STEP 2: Calculate the test statistic:

$$z_0 = \frac{p + c - P_0}{\sqrt{P_0(1-P_0)/n}} \qquad (6\text{-}2)$$

$$\text{where } c = \begin{cases} \frac{-0.5}{n} if, p > P_0 \\ \frac{0.5}{n} if, p < P_0 \end{cases} \text{ and } p = \frac{x(\#\text{ of site values} > A_0)}{n}$$

Here *c* is the continuity correction factor to be able to use normal approximation.

Use of Large Sample Normal Approximation

STEP 3: Typically, one should use BD (as described above) to perform this test. However, when both (nP_0) and $n(1-P_0)$ are at least 5, a normal, z-table (automatically computed by ProUCL 4.0) approximation may be used to compute the critical values and p-values.

STEP 4: Conclusion – Given for approximate test based upon normal approximation:

Form 1: If $z_0 > z_{1-\alpha}$, then reject the null hypothesis that the population proportion, P of exceedances of action level, A_0 is less than the specified proportion, P_0.
Form 2: If $z_0 < -z_{1-\alpha}$, then reject the null hypothesis that the population proportion, P is more than the specified proportion, P_0.
Two Sided: If $|z_0| > z_{1-\alpha/2}$, then reject the null hypothesis that the population proportion, P is the same as the specified proportion, P_0.

P-Values Based Upon a Normal Approximation

As mentioned before, a p-value is the smallest value for which the null hypothesis is rejected in favor of the alternative hypotheses. Thus, based upon the given data, the null hypothesis is rejected for all values of α (the level of significance) greater than or equal to the p-value. The details of computing a p-value for proportion test based upon large sample normal approximation can be found in any statistical text book such as Daniel (1995), and also in EPA (2006). ProUCL 4.0 computes large sample p-values for proportion test associated with each form of null hypothesis. If the computed p-value is smaller than the specified value of, α, the conclusion is to reject the null hypothesis based upon the collected data set used in the various computations.

Note: *ProUCL 4.0 also performs the proportion test based upon the exact binomial distribution when the sample size is small, and one may not be able to use the normal approximation as described above. ProUCL 4.0 checks for the availability of appropriate amount of data, and performs the tests using a normal approximation or the exact binomial distribution accordingly.*

Use of the Exact Binomial Distribution for Smaller Samples

STEP 1: When the sample size is small (e.g., < 30), either (nP_0), or $n(1-P_0)$ is less than 5, one should use the exact BD to perform this test. ProUCL 4.0 automatically performs this test based upon BD, when the above conditions are not satisfied. In such cases, ProUCL 4.0 computes the critical values and p-

values based upon the BD and its cumulative distribution function (CDF). The probability statements concerning the computation of p-values can be found in Conover (1999).

STEP 2: Conclusion – Based Upon Exact Binomial Distribution:

Form 1: Large values of B cause the rejection of the null hypothesis. Therefore, reject the null hypothesis, when $B \geq b$. Here b is obtained using the binomial cumulative probabilities based upon a BD (n, P_0). The critical value, b (associated with α) is given by the probability statement: $Prob(B \geq b) = \alpha$, or equivalently, $P(B < b) = (1 - \alpha)$. As mentioned before, since B is a discrete binomial random variable, the level, α may not be exactly achieved by the critical value, b.

Form 2: For this form, small values of B will cause the rejection of the null hypothesis. Therefore, reject the null hypothesis, when $B \leq b$. Here b is obtained using the binomial cumulative probabilities based upon BD (n, P_0). The critical value, b (associated with α) is given by the probability statement: $P(B \leq b) = \alpha$. As mentioned before, since B is a discrete binomial random variable, the level, α may not be exactly achieved by the critical value, b.

Two Sided Alternative: Here the critical or the rejection region for the null hypothesis is made of two areas, one in the right tail (of area ~ α_2) and the other in the left tail (with area ~ α_1), so that the combined area of the two tails is approximately, $\alpha = \alpha_1 + \alpha_2$. That is for this hypothesis form, both small values and large values of B will cause the rejection of the null hypothesis. Therefore, reject the null hypothesis, when $B \leq b_1$ or $B > b_2$. Typically α_1 and α_2 are roughly equal, and in ProUCL 4.0, both are chosen to be equal = $\alpha/2$. Thus, b_1 and b_2 are given by the statements: $P(B \leq b_1) \sim \alpha/2$, and $P(B > b_2) \sim \alpha/2$. Since B is a discrete binomial random variable, the level, α may not be exactly achieved by the critical values, b_1 and b_2.

P-Values Based Upon Binomial Distribution as Incorporated in ProUCL 4.0

As mentioned before, a p-value is the smallest value for which the null hypothesis is rejected in favor of the alternative hypothesis. Thus, based upon the given data, the null hypothesis is rejected for all values of α (the level of significance) greater than or equal to the p-value. Note, for discrete distributions such a BD, the exact level of significance, α cannot be achieved. The probability statements for computing a p-value for proportion test based upon BD can be found in Conover (1999). Using the BD, ProUCL 4.0 computes p-values for proportion test associated with each form of null hypothesis. If the computed p-value is smaller than the specified value of, α, the conclusion is to reject the null hypothesis based upon the collected data set used in the various computations. It is noted that there are some variations in the literature about the computation of p-values for proportion test based upon the exact BD. Therefore, the p-value computation procedure as incorporated in ProUCL 4.0 is described as follows.

Let b be the calculated value of the binomial random variable, B under the null hypothesis. ProUCL 4.0 computes the p-values using the following statements:

Form 1: p-value = Prob ($B \geq b$)
Form 2: p-value = Prob ($B \leq b$)

Two sided Alternative:

For $b > (n-b)$
P-value = 2* Prob ($B \leq b$)
For $b \leq (n-b)$

P-value = 2*Prob (B≥b)

Some single sample distribution-free tests based upon the ranks of the data as incorporated in ProUCL 4.0 are described as follows. These tests can also be used on data sets with NDs, provided the NDs are smaller than the cleanup C_s or C. In this Technical Guide, both C_s and C have been used to represent cleanup goals or cleanup levels.

6.6.3 The Sign Test

The sign test is used to detect a difference between the population median and a fixed cleanup goal, *C*. This test makes no distributional assumptions like the t-test. The sign test is used when the data are not symmetric and the sample size is small (EPA, 2006).

6.6.3.1 Limitations and Robustness

- This test can handle nondetects, provided all NDs are smaller than the cleanup limit, *C*.
- Compared to one-sample t-test, and the Wilcoxon Signed Rank (WSR) test described below, the sign test has less power.

6.6.3.2 Sign Test in the Presence of Nondetects

- A principal requirement when applying the sign test is that the cleanup limit/goal, *C* should be greater than the greatest less-than value; all NDs should be smaller than *C*.

6.6.3.3 Directions for the Sign Test

Let $X_1, X_2, \ldots, X_n$ represent a random sample of size *n* collected from a site area under investigation. As before, let *C* represent the cleanup level. In the following, the substantial difference, S is ≥ 0. It should be noted that the substantial difference, S ≥ 0 is used only with Form 2 hypothesis with substantial difference.

STEP 1: Let $\tilde{\mu}_X$ be the site population median.
State the following null and the alternative hypotheses:

Form 1: H_0: $\tilde{\mu}_X \le C$ *vs.* H_1: $\tilde{\mu}_X > C$

Form 2: H_0: $\tilde{\mu}_X \ge C$ *vs.* H_1: $\tilde{\mu}_X < C$

Two Sided: H_0: $\tilde{\mu}_X = C$ *vs.* H_1: $\tilde{\mu}_X \ne C$

Form 2 with substantial difference, *S*: H_o: $\tilde{\mu}_X \ge C + S$ *vs.* H_1: $\tilde{\mu}_X < C + S$

STEP 2: Calculate the deviations $d_i = x_i - C$. If some of the $d_i = 0$, then reduce the sample size until all the remaining $d_i > 0$. This means that all observations tied at C are ignored from the computation.
Compute the binomial random variable, *B* representing the number of $d_i > 0$,
i: = 1,2,...,n. Note that under the null hypothesis, the BD random variable B follows a BD (n, ½).
Thus, one can use the exact BD to compute the critical values and p-values associated with this test.

STEP 3: Use the test statistic, *B* with exact binomial distribution (BD) for $n \le 40$ (programmed in ProUCL 4.0).

For $n > 40$, one may use the approximate normal test statistic given by,

$$z_0 = \frac{B - \frac{n}{2} - S}{\sqrt{\frac{n}{4}}}. \qquad (6\text{-}3)$$

Note*: As before, the substantial difference, S =0, except for Form 2 hypotheses with substantial difference.*

STEP 4: For $n \le 40$, use the BD table as in EPA, 2006 (critical values automatically computed by ProUCL 4.0) to calculate the critical values, and for $n > 40$, use the normal approximation and the associated normal z critical values.

STEP 5: Conclusion when $n \le 40$ (following EPA 2006):

Form 1: If $B \ge B_{UPPER}(n, 2\alpha)$, then reject the null hypothesis that the population median is less than the cleanup level, *C*.
Form 2: If $B \le B_{UPPER}(n, 2\alpha)$, then reject the null hypothesis that the population median is more than the cleanup level.
Two Sided: If $B \ge B_{UPPER}(n, \alpha)$ or $B \le B_{UPPER}(n, \alpha) - 1$, then reject the null hypothesis that the population median is comparable to the cleanup level, *C*.
Form 2 with substantial difference, *S*: If $B \le B_{UPPER}(n, 2\alpha)$, then reject the null hypothesis that the population median is more than the cleanup level, *C* + substantial difference, *S*.

Note that ProUCL 4.0 automatically calculates the critical values and p-values based upon the BD (n, ½) for both small samples and large samples.

Conclusion: When $n > 40$ – Large Sample Approximation:

Form 1: If $z_0 > z_{1-\alpha}$, then reject the null hypothesis that the population median is less than the cleanup level, *C*.
Form 2: If $z_0 < -z_{1-\alpha}$, then reject the null hypothesis that the population median is more than the cleanup level, *C*.
Two Sided: If $|z_0| > z_{1-\alpha/2}$, then reject the null hypothesis that the population median is comparable to the cleanup level, *C*.
Form 2 with substantial difference, S: If $z_0 < -z_{1-\alpha}$, then reject the null hypothesis that the population median is more than the cleanup level, *C* + substantial difference, *S*.

P-Values for One-Sample Sign Test

ProUCL 4.0 automatically calculates the critical values and p-values based upon: the BD (n, ½) for small data sets; and normal approximation for larger data sets as described above.

6.6.4 The Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank (WSR) test is used for the testing the difference between the location parameter (mean or median) of a population and a fixed cleanup threshold level such as C, C_s also representing a location value such as mean concentration.

6.6.4.1 Limitations and Robustness

- For symmetric distributions, typically, the Wilcoxon Signed Rank test has more power than the sign test.
- It may give incorrect results in the presence of many tied values.
- The presence of different detection limits makes this test less powerful.
- For large samples (n > 50) and the normality assumption of the mean (due to *CLT*), the one-sample t-test is more powerful than the Wilcoxon Signed Rank test.

6.6.4.2 Wilcoxon Signed Rank (WSR) Test in the Presence of Nondetects

- When all the data have the same detection limits, replacement of censored data by a surrogate value of DL/2 has been recommended (EPA (2006)).
- In the presence of multiple detection limits, all observations at the highest detection limit may be censored (just as in WMW test) – this may result in some loss of power.

6.6.4.3 Directions for the Wilcoxon Signed Rank Test

Let $X_1, X_2, \ldots, X_n$ represent the n data points from site under investigation, and C represent the Cleanup Level.

STEP 1: State the following null hypotheses:

Form 1: H_0: *Site location* $\leq C$ *vs.* H_1: *Site location* $> C$
Form 2: H_0: *Site location* $\geq C$ *vs.* H_1: *Site location* $< C$
Two Sided: H_0: *Site location* $= C$ *vs.* H_1: *Site location* $\neq C$
Form 2 with substantial difference, S: H_0: *Site location* $\geq C + S$ *vs.* H_1: *Site location* $< C + S$, here $S \geq 0$.

STEP 2: Calculate the deviations $d_i = x_i - C$. If $d_i = 0$, then reduce the sample size until $d_i > 0$.

STEP 3: Rank the absolute deviations, $|d_i|$, from smallest to the largest. Assign an average rank to the tied observations.

STEP 4: Let R_i be the signed rank of $|d_i|$, where the sign of R_i is determined by the sign of d_i.

STEP 5: Test Statistic Calculations:

For $n \leq 20$, compute $T^+ = \sum_{\{i:R_i>0\}} R_i$, where T^+ is the sum of the positive signed ranks.

For $n > 20$, approximate test is given by

$$z_0 = \frac{T^+ - n(n+1)/4}{\sqrt{\text{var}(T^+)}} \qquad (6\text{-}4)$$

where $\text{var}(T^+)$ is given by $\text{var}(T^+) = \frac{n(n+1)(2n+1)}{24} - \frac{1}{48}\sum_{j=1}^{g} t_j\left(t_j{}^2 - 1\right)$ with g = number of tied groups.

STEP 6: Conclusion when $n \le 20$:

Form 1: If $T^+ \ge \frac{n(n+1)}{2}$, then reject the null hypothesis that the location parameter is less than the cleanup level, C.
Form 2: If $T^+ \le w_\alpha$, then reject the null hypothesis that the location parameter is more than the cleanup level, C.
Two Sided: If $T^+ \ge \frac{n(n+1)}{2} - w_{\alpha/2}$ or $T^+ \le w_\alpha$, then reject the null hypothesis that the location parameter is comparable to the action level, C.
Form 2 with substantial difference, *S*: If $T^+ \le w_\alpha$, then reject the null hypothesis that the location parameter is more than the cleanup level, C + the substantial difference, S.

Conclusion when $n > 20$:

Form 1: If $z_0 > z_{1-\alpha}$, then reject the null hypothesis that the location parameter is less than the cleanup level, *C*.
Form 2: If $z_0 < -z_{1-\alpha}$, then reject the null hypothesis that the location parameter is more than the cleanup level, *C*.
Two Sided: If $|z_0| > z_{1-\alpha/2}$, then reject the null hypothesis that the location parameter is comparable to the cleanup level, *C*.
Form 2 with substantial difference, S: If $z_0 < -z_{1-\alpha}$, then reject the null hypothesis that the location parameter is more than the cleanup level, *C* + the substantial difference, S.

Note: *The critical values, w_α as given in EPA (2006) have been programmed in ProUCL 4.0. The details of computation of p-values for small samples (based upon a BD) and larges samples (based upon a normal distribution) as incorporated in ProUCL 4.0 are given in EPA (2006). For small data sets, ProUCL 4.0 used tables of critical values as given in EPA (2006).*

6.7 Two-Sample Hypotheses Testing Approaches

The use of parametric and nonparametric two-sample hypotheses testing approaches is quite common in many environmental applications including site versus background comparison studies. Several of those approaches for data sets with and without ND observations have been incorporated in ProUCL 4.0. Additionally some graphical methods (box plots and Q-Q plots) for data sets with and without NDs are also available in ProUCL 4.0 to visually compare two or more populations. Some of the parametric and nonparametric methods are described in the following subsections. Additional details can be found in EPA (1994, 1997, and 2006).

Note: *It is suggested to always supplement the conclusions derived using test statistics (e.g., t-test, WMW test, Gehan test) with graphical displays such as multiple Q-Q plots and side-by-side box plots. Graphical displays of data sets often provide useful information about the behavior of the populations (and their parameters) under investigation that may not be obtained and understood by simple test statistics such as t-test, Gehan test, and even GOF test statistics. Therefore, one should always use graphical displays before deriving any conclusions about the data distributions, or equality of two data distributions.*

Student's two-sample t-test is used to compare the means of the two populations such as the potentially impacted site area and a background or reference area. Two cases arise: 1) the variations (dispersion) of the two populations are the same, and 2) the dispersions of the two populations are not the same. Generally, a t-test is robust and not sensitive to small deviations from the assumptions of normality.

6.7.1 Student's Two-Sample t-Test (Equal Variances)

6.7.1.1 Assumptions and Their Verification

- $X_1, X_2, \ldots, X_n$ represent systematic site samples and $Y_1, Y_2, \ldots, Y_m$ represent systematic background samples that are drawn at random from those independent populations. The validity of the random sampling and independence assumptions should be confirmed by reviewing the procedures used to select the sampling points (EPA, 2006).

- The sample means $\overline{X}$ (site) and $\overline{Y}$ (background) are approximately normally distributed provided the underlying data distributions are not heavily skewed (Singh, Singh, and Iaci, 2002). If both *m* and *n* are large (and data are mildly to moderately skewed), one may make this assumption without further verification. If the data are heavily skewed (skewness discussed in Chapters 3 and 4), the use of nonparametric tests such as WMW test and quantile test is preferable. Normality or approximate normality of data sets should be checked by using GOF tests as incorporated in ProUCL 4.0 and described in Chapter 3 of this Technical Guide.

6.7.1.2 Limitations and Robustness

- The two-sample t-test with equal variances is robust to violations of the assumption of normality. However, if the investigator has tested and rejected normality or equality of variances, then nonparametric procedures such as the Wilcoxon-Mann-Whitney (WMW) may be applied.

- This test is not robust to outliers because sample means and standard deviations are sensitive to outliers. As mentioned before, it is suggested not to use a t-test on log-transformed data sets. This issue has been discussed in Chapter 3 of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites (EPA, 2002b).*

- The test assumes the equality of variances of the background and the potentially impacted data sets. Therefore, if the two variances are not equal and normality assumptions of the means are valid, then Satterthwaite's t-test (available in ProUCL 4.0) should be applied. However, if the variances are not equal and the normality of the means is not applicable (due to small samples, or skewed data), then nonparametric WMW test should be applied.

- In the presence of less than values, it is suggested to use WMW test or Gehan test. Sometimes, users tend to use a t-test on data sets obtained by replacing all less than values by surrogate values, such as respective DL/2 values, or DL values.

- **Note:** *As mentioned many times before, it is suggested to avoid the use of such proxy and substitution methods (e.g., DL/2) to compute t-test statistics.*

6.7.1.3 Guidance on Implementing the Student's Two-Sample t-Test

The number of site (*n*) and background (*m*) measurements required to conduct the two-sample t-test should be calculated based upon appropriate DQO procedures (EPA, 2006). In case, it is not possible to use DQOs, or to collect as many samples as determined using DQO-based sample size determination formulae, one should follow the minimum sample size requirements as described in Chapter 1 of this Technical Guide.

ProUCL 4.0 provides a test to verify the equality of two dispersions. ProUCL 4.0 automatically performs this test for dispersions before using one of the two t-tests incorporated in ProUCL 4.0. If some measurements appear to be unusually large compared to the remainder of the measurements in the data set, then a test for outliers (Chapter 7) should be conducted. Some of the identified outliers may represent contaminated locations needing further investigation. Once any identified outliers have been investigated for being mistakes or errors and, if necessary, discarded, the site and background data sets should be re-tested for normality using both probability plots and formal GOF tests.

The project team should decide the proper disposition of outliers. In practice, it is advantageous to carry out the tests on data sets with and without the outliers. This extra step helps the users to assess and determine the influence of outliers on the various test statistics and the resulting conclusions. This process also helps the users in making appropriate decisions about the proper disposition (include or exclude from the data analyses) of outliers. Many times, the outliers represent contaminated site locations requiring separate and additional investigation.

6.7.1.4 Directions for the Student's Two-Sample t-Test

Let $X_1, X_2, \ldots, X_n$ represent systematic and random site samples and $Y_1, Y_2, \ldots, Y_m$ represent systematic and random background samples drawn from independent populations.

STEP 1: State the following null and the alternative hypotheses:

Form 1: H_0: $\mu_X - \mu_Y \le 0$ *vs.* H_1: $\mu_X - \mu_Y > 0$

Form 2: H_0: $\mu_X - \mu_Y \ge 0$ *vs.* H_1: $\mu_X - \mu_Y < 0$

Two Sided: H_0: $\mu_X - \mu_Y = 0$ *vs.* H_1: $\mu_X - \mu_Y \ne 0$

Form 2 with substantial difference, S: H_0: $\mu_X - \mu_Y \ge S$ *vs.* H_1: $\mu_X - \mu_Y < S$

STEP 2: Calculate the sample mean $\overline{X}$ and the sample variances S_X^2 for the site data and compute the sample mean $\overline{Y}$ and the sample variance S_Y^2 for the background data.

STEP 3: Determine if the variances of the two populations are equal. If the variances of the two populations are not equal, use the Satterthwaite's t-test. Calculate,

$$s_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{(m-1)+(n-1)}} \tag{6-5}$$

$$\text{and} \quad t_0 = \frac{(\overline{X} - \overline{Y}) - S}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \tag{6-6}$$

STEP 4: Use ProUCL 4.0 to find the critical value $t_{1-\alpha}$ such that $100(1-\alpha)$ % of the t-distribution with $(m + n - 2)$ degrees of freedom (df) is below $t_{1-\alpha}$.

STEP 5: Conclusion:

Form 1: If $t_0 > t_{m+n-2,\,1-\alpha}$, then reject the null hypothesis that the site population mean is less than or equal to the background population mean.
Form 2: If $t_0 < -t_{m+n-2,\,1-\alpha}$, then reject the null hypothesis that the site population mean is greater than or equal to the background population mean.
Two Sided: If $|t_0| > t_{m+n-2,\,1-\alpha/2}$, then reject the null hypothesis that the site population mean comparable to the background population mean.
Form 2 with substantial difference, S: If $t_0 < -t_{m+n-2,\,1-\alpha}$, then reject the null hypothesis that the site population mean is greater than or equal to the background population mean + the substantial difference, S.

6.7.2 The Satterthwaite Two-Sample t-Test (Unequal Variances)

Satterthwaite's t-test should be used to compare two population means when the variances of the two populations are not equal. It requires the same assumptions as the two-sample t-test (described above) except for the assumption of equal variance.

6.7.2.1 Limitations and Robustness

- In the presence of less than values, replacement by a surrogate value such as the detection limit or one-half of the detection limit gives biased results. Its use should be avoided. Instead the use of nonparametric tests is suggested.

- In cases where the assumptions of normality of means are violated, the use of nonparametric tests is preferred.

- Wilcoxon-Mann-Whitney (WMW) and quantile tests are recommended when data sets have less than values. Moreover, if the less-than values have multiple detection limits (for example, < 10, < 15, etc.), then the Gehan test should be used in place of the WMW test.

6.7.2.2 Directions for the Satterthwaite Two-Sample t-Test

Let $X_1, X_2, \ldots, X_n$ represent systematic and random site samples and $Y_1, Y_2, \ldots, Y_m$ represent systematic and random background samples drawn from independent populations.

STEP 1: State the following null and the alternative hypotheses:

Form 1: H_0: $\mu_X - \mu_Y \leq 0$ *vs.* H_1: $\mu_X - \mu_Y > 0$

Form 2: H_0: $\mu_X - \mu_Y \geq 0$ *vs.* H_1: $\mu_X - \mu_Y < 0$

Two Sided: H_0: $\mu_X - \mu_Y = 0$ *vs.* H_1: $\mu_X - \mu_Y \neq 0$

Form 2 with substantial difference, S: H_0: $\mu_X - \mu_Y \geq S$ *vs.* H_1: $\mu_X - \mu_Y < S$

STEP 2: Calculate the sample mean $\overline{X}$ and the sample variances S_X^2 for the site data and compute the sample mean $\overline{Y}$ and the sample variance S_Y^2 for the background data.

STEP 3: Use F test as described below (also in ProUCL 4.0) to determine if the variances of the two populations are equal.

$$t_0 = \frac{\overline{X} - \overline{Y} - S}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}} \qquad (6\text{-}7)$$

Here S = 0, except when used in Form 2 hypothesis with substantial difference, S ≥ 0.

STEP 4: Use a t-table (ProUCL 4.0 computes it) to find the critical value $t_{1-\alpha}$ such that $100(1-\alpha)\%$ of the t-distribution with f degrees of freedom is below $t_{1-\alpha}$, where the Satterthwaite's Approximation for df is given by:

$$df = \frac{\left[\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right]^2}{\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}} \qquad (6\text{-}8)$$

STEP 5: Conclusion:

Form 1: If $t_0 > t_{df,\, 1-\alpha}$, then reject the null hypothesis that the site population mean is less than or equal to the background population mean.

Form 2: If $t_0 < -t_{df,\, 1-\alpha}$, then reject the null hypothesis that the site population mean is greater than or equal to the background population mean.

Two Sided: If $|t_0| > t_{df,\, 1-\alpha/2}$, then reject the null hypothesis that the site population mean is comparable to the background population mean.

Form 2 with substantial difference, S: If $t_0 < -t_{df, 1-\alpha}$, then reject the null hypothesis that the site population mean is greater than or equal to the background population mean + the substantial difference, S.

P-Values for Two-Sample t-Test

A p-value is the smallest value for which the null hypothesis is rejected in favor of the alternative hypotheses. Thus, based upon the given data, the null hypothesis is rejected for all values of α (the level of significance) greater than or equal to the p-value. The details of computing a p-value for two-sample t-test for comparing two means are given in EPA (2006). ProUCL 4.0 directly computes (based upon an appropriate t-distribution) p-values for two-sample t-tests associated with each form of null hypothesis. If the computed p-value is smaller than the specified value of, α, the conclusion is to reject the null hypothesis based upon the collected data set used in the various computations.

6.8 Tests for Equality of Dispersions

This section describes a test (also available in ProUCL 4.0) that verifies the assumption of the equality of two variances. This assumption is needed to perform a simple two-sample Student's t-test as described above.

6.8.1 The F-Test for the Equality of Two-Variance

An F-test may be used to test whether the true underlying variances of two populations are equal. Usually the F-test is employed as a preliminary test, before conducting the two-sample t-test for the equality of two means. The assumptions underlying the F-test are that the two-samples are independent random samples from two underlying normal populations. The F-test for equality of variances is highly sensitive to departures from normality.

6.8.1.1 Directions for the F-Test

Let $X_1, X_2, \ldots, X_n$ represent the n data points from site and $Y_1, Y_2, \ldots, Y_m$ represent the m data points from background. To perform an F-test, proceed as follows.

STEP 1: Calculate the sample variances s_X^2 (for the X's) and s_Y^2 (for the Y's)

STEP 2: Calculate the variance ratios $F_X = s_X^2/s_Y^2$ and $F_Y = s_Y^2/s_X^2$. Let F equal the larger of these two values. If $F = F_x$, then let k = n - 1 and q = m - 1. If $F = F_y$, then let k = m - 1 and q = n – 1.

STEP 3: Using a table of F distribution (ProUCL 4.0 computes them), find a cutoff, $U = f_{1-\alpha/2}(k, q)$ associated with F distribution with k and q degrees of freedom for some significance level, α. If F calculated as above > U, conclude that the variances of the two populations are not the same.

P-Values for Two-Sample Dispersion Test for Equality of Variances

As mentioned before, a p-value is the smallest value for which the null hypothesis is rejected in favor of the alternative hypotheses. Thus, based upon the given data, the null hypothesis is rejected for all values of α (the level of significance) greater than or equal to the p-value. The details of computing a p-value for a two-sample F-test to compare equality of two variances (of two normal populations) are given in EPA (2006). ProUCL 4.0 computes p-values for the two-sample F-test based upon an appropriate F-

distribution. If the computed p-value is smaller than the specified value of, α, the conclusion is to reject the null hypothesis based upon the collected data set used in the various computations.

6.9 Nonparametric Tests

The statistical tests discussed in the previous section rely on the assumption of normality of the population (e.g., background and site) distributions. When the data do not follow the assumed distribution, use of parametric statistical tests may lead to inaccurate conclusions. Additionally, if the data sets contain outliers or nondetect values, an additional level of uncertainty is faced when conducting parametric tests. Since most environmental data sets do contain outliers and nondetect values, it is unlikely that the current widespread use of parametric tests is justified, given that these tests may be adversely affected by outliers and by assumptions made for handling nondetect values. Several nonparametric tests have been incorporated in ProUCL 4.0 that can be used on data sets with ND observations with single and multiple detection limits.

As mentioned earlier, tests that do not require specific mathematical form for the underlying distribution are called distribution-free or nonparametric statistical tests. The property of *robustness* is the main advantage of nonparametric statistical tests. Nonparametric tests have good test performances for a wide variety of distributions, and those performances are not unduly affected by outliers. This is especially true when the statistics and tests are not based upon higher order statistics such as nonparametric UPLs and UTLs. An example illustrating these issues has been considered in Chapter 5. It was shown that the nonparametric background statistics do get distorted by outliers. This statement is true for data sets with or without ND observations.

Nonparametric tests can be used on normal as well as for non-normal data sets. If one or both of the data sets fail to meet the normality test, or if the data sets appear to come from different types of populations, then the use of nonparametric tests is preferable. *It should be noted that parametric tests are more powerful provided the underlying assumptions associated with those tests are satisfied.* Nonparametric tests compare the shape and location of the two distributions instead of a statistical parameter such as mean. Nonparametric tests are preferred methods whenever data set consists of ND observations, especially when the percentage of nondetects becomes very high.

Note: *Once again, it is suggested to avoid the use of DL/2 method (e.g., replacing NDs by respective DL/2 values or some other substitution values) to perform hypotheses testing approaches. ProUCL 4.0 provides alternative methods (e.g., WMW test, Gehan test) that may be used when data sets consist of ND observations.*

6.9.1 The Wilcoxon-Mann-Whitney (WMW) Test

The Mann-Whitney (or Wilcoxon-Mann-Whitney) test (Bain and Engelhardt, 1991) is a nonparametric test used for determining whether a difference exists between the site and the background population distributions. The WMW test tests whether or not measurements (location, central) from one population consistently tend to be larger (or smaller) than those from the other population based upon the assumption that the dispersion of the two distributions are roughly the same. This test determines which distribution is higher by comparing the relative ranks of the two data sets when the data from both sources are sorted into a single list. One assumes that any difference between the background and site concentration distributions is due to a shift in location (mean, median) of the site concentrations to higher values (due to the presence of contamination in addition to the background).

6.9.1.1 Advantages and Disadvantages

The WMW test has three advantages for background comparisons:

- The two data sets are not required to be from a known type of distribution. The WMW test does not assume that the data are normally distributed, although a normal distribution approximation is used to determine the critical value of the test for large sample sizes.

Note: *It is suggested again to avoid the use of t-test on log-transformed data sets to compare the mean concentrations of two populations. It should be noted that the equality of means in log-scale does not imply the equality of means in the original raw scale. The cleanup decisions often are made based upon concentrations in threshold levels in the original scale.*

- WMW tests allows for nondetect measurements to be present in both data sets. Specifically, the WMW test can handle a moderate number of nondetect values in either or both data sets by treating them as ties. In practice, the WMW test may be used with up to 40 percent or more nondetect measurements in either the background or the site data. If more than 40 percent of the data from either the background or the site are nondetect values, the WMW test should not be used. The use of Gehan test is preferable in those situations.

- It is robust with respect to outliers because the analysis is conducted in terms of the ranks of the data. This limits the influence of outliers because a given data point can be no more extreme than the first or the last rank.

- The WMW test should not be used if more than 40% of the site or background data sets are less-than values. The measurement laboratories should be instructed to report actual measurements for all soil samples, whenever possible, even if the reported measurements are negative. Although negative concentrations cannot occur in nature, negative *measurements* can occur, due to measurement uncertainties, when the true concentrations are very close to zero.

- The WMW test does not place enough weight on the larger site and background measurements. This means, a WMW may lead to the conclusion that two populations are comparable even when the observations in the right tail of one distribution (e.g., site) are significantly larger than the right tail observations of other population (e.g., background). The quantile test (EPA, 1994) is used to compare upper tails of the two distributions. WMW test uses and considers *all* measurements, rather than focusing on larger measurements as is done by the quantile test. The quantile test is available in ProUCL 4.0. It is suggested that both the quantile test and the WMW test be used on the same data sets.

- The WMW test may be applied to either null hypothesis in the two forms of background tests as discussed throughout this chapter. In all forms, the null hypothesis is assumed to be true unless the evidence in the data indicates that it should be rejected in favor of the alternative hypothesis.

6.9.1.2 WMW Test in the Presence of Nondetects

- If there are *t* nondetect values, then they are considered as "ties" and are assigned the average rank for this group. Their average rank is the average of the first *t* integers, $\frac{t+1}{2}$.
- Note that if there are no NDs, the process is the same and ranks are assigned individually to all detected observations.
- If more than one detection limit was in use, then all of the observations below the largest detection limit should be treated as nondetects. This of course will result in some loss of power associated with WMW test. Alternatively, the Gehan test may be performed.

6.9.1.3 WMW Test Assumptions and Their Verification

The underlying assumptions of the WMW test are:

- The soil sample measurements obtained from the site and background areas are independent (not correlated). This assumption requires: 1) that an appropriate probability-based sampling design strategy be used to determine (identify) the sampling locations of the soil samples for collection, and 2) those soil sampling locations are spaced far enough apart that a spatial correlation among concentrations at different locations is not likely to be present.
- The underlying probability distribution of the measurements from a site area is similar to (e.g., including shape and spread) the probability distribution of measurements collected from a background or reference area. Under the alternative hypothesis (Form 2), the distribution of site data may be shifted to higher concentrations than the distribution for the background area. The assumption of equal variances of the two regions (site and background) should also be evaluated using descriptive statistics and graphical displays such as side-by-side box plots and histograms.

6.9.1.4 Directions for the WMW Test when the Number of Site and Background Measurements is small ($n \leq 20$ and $m \leq 20$)

Let $X_1, X_2, \ldots, X_n$ represent the n data points from the site population and $Y_1, Y_2, \ldots, Y_m$ represent the m data points from the background population.

STEP 1: Let $\tilde{\mu}_X$ represent the site population median and $\tilde{\mu}_Y$ represent the background population median. State the following null and the alternative hypotheses:

Form 1: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y \leq 0$ vs. H_1: $\tilde{\mu}_X - \tilde{\mu}_Y > 0$

Form 2: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y \geq 0$ vs. H_1: $\tilde{\mu}_X - \tilde{\mu}_Y < 0$

Two Sided: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y = 0$ vs. H_1: $\tilde{\mu}_X - \tilde{\mu}_Y \neq 0$

Form 2 with substantial difference, S: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y \geq S$ $\tilde{\mu}_X - \tilde{\mu}_Y < S$

STEP 2: List and rank the pooled set of N = n + m site and background measurements from smallest to largest, keeping track of which measurements came from the site and which came from the background area. Assign the rank of 1 to the smallest value among the pooled data, the rank of 2 to the second smallest value among the pooled data, and so forth.

- If a few measurements are tied (identical in value), then assign the average of the ranks that would otherwise be assigned to the tied observations. If several measurement values have ties, then average the ranks separately for each of those measurement values.

- If a few less-than values occur (say, < 10%), and if all such values are less than the smallest detected measurement in the pooled data set, then handle the less-than values as tied at an arbitrary value less than the smallest detected measurement. Assign the average of the ranks that would otherwise be assigned to these tied less-than values (the same procedure as for tied detected measurements).

- If between 10% and 40% of the pooled data set are less-than values, and all are less than the smallest detected measurement, then use the approximate WMW test procedure given for large data sets (e.g., $n > 20$ and $m > 20$), even if n and m are less than 20.

Note: *The procedure is for the case where m and n are both of size 20 or larger. That procedure will provide only an approximate test if it is used when n and m are both smaller than 20. In that case, decisions of whether or not the null hypothesis is rejected should not be made until additional information is obtained by taking more samples and using a more sensitive measurement method. It is suggested to use graphical displays before deriving any conclusions about the equality of two data distributions.*

STEP 3: Calculate the sum of the ranks of the *site* measurements. Denote this sum by W_s and then calculate T as follows:

$$T = W_S - \frac{n(n+1)}{2} - S \qquad (6\text{-}9)$$

Note: *The test statistic, T is often called the MW test statistic. In this Guide and in ProUCL, this statistic is called the WMW test statistic (Bain and Engelhardt, 1991). Note the difference between the definitions of T and W_s . Some software packages such as MINITAB uses test statistic W_s (known as Wilcoxon Rank Sum (WRS) statistic) to perform this nonparametric test. Obviously the critical values for W_s and T are different. However, critical values for one test can be obtained from the critical values of the other test by using the relationship given by the above equation. These two tests (WRS test and WMW test) are equivalent tests, and conclusions derived by using these test statistics are equivalent.*

STEP 4: For specific values of n, m, and α, find an appropriate Wilcoxon-Mann-Whitney critical value, w_α, from the table as given in EPA (2006) and also in Daniel (1995).

STEP 5: Conclusion:

Form 1: If $T \geq nm - w_\alpha$, then reject the null hypothesis that the site population median is less than or equal to the background population median.
Form 2: If $T \leq w_\alpha$, then reject the null hypothesis that the site population median is greater than or equal to the background population median.

Two Sided: If $T \geq nm - w_{\alpha/2}$ or $T \leq w_{\alpha/2,}$ then reject the null hypothesis that the site population median (location) is similar to that of the background population median (location).
Form 2 with substantial difference, S: If $T \leq w_{\alpha,}$ then reject the null hypothesis that the site population median is greater than or equal to the background population median + the substantial difference, S.

Note: *S takes a positive value only for this form of the hypothesis with substantial difference, in all other forms of the null hypothesis, S = 0.*

P-Values for Two-Sample WMW Test – For Small Sample

A p-value is the smallest value for which the null hypothesis is rejected in favor of the alternative hypotheses. Thus, based upon the given data, the null hypothesis is rejected for all values of α (the level of significance) greater than or equal to the p-value. The details of computing a p-value for a two-sample nonparametric WMW test for comparing two means are given in EPA (2006). For small samples, ProUCL 4.0 computes only approximate (as computed for large samples) p-values for WMW test. If the computed p-value is smaller than the specified value of, α, the conclusion is to reject the null hypothesis based upon the collected data set used in the various computations.

6.9.1.5 Directions for the WMW Test when the Number of Site and Background Measurements is Large (*n > 20* and *m > 20*)

Let $X_1, X_2, \ldots, X_n$ represent the *n* data points from the site population and $Y_1, Y_2, \ldots, Y_m$ represent the *m* data points from the background population.

STEP 1: Let $\tilde{\mu}_X$ represent the site and $\tilde{\mu}_Y$ represent the background population medians (means). State the following null and the alternative hypotheses:

Form 1: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y \leq 0$ *vs.* H_1: $\tilde{\mu}_X - \tilde{\mu}_Y > 0$

Form 2: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y \geq 0$ *vs.* H_1: $\tilde{\mu}_X - \tilde{\mu}_Y < 0$

Two Sided: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y = 0$ *vs.* H_1: $\tilde{\mu}_X - \tilde{\mu}_Y \neq 0$

Form 2 with substantial difference, *S*: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y \geq S$ *vs.* $\tilde{\mu}_X - \tilde{\mu}_Y < S$

STEP 2: List and rank the pooled set of n + m site and background measurements from smallest to largest, keeping track of which measurements came from the site and which came from the background area. Assign the rank of 1 to the smallest value among the pooled data, the rank of 2 to the second smallest value among the pooled data, and so forth.

- If < 40% of the measurements in the pooled data is tied (identical in value), then assign the average of the ranks that would otherwise be assigned to the tied observations. If several measurement values exist for which ties occur, then average the ranks separately for each of those measurement values.

- If < 40% of the pooled data is less-than values and if all such values are less than the smallest detected measurement in the pooled data set, then handle those less-than values as being tied at an arbitrary value less than the smallest detected measurement. Assign the

average of the ranks that would otherwise be assigned to this group of tied values (the same procedure as for detected measurements that are tied).

- NOTE: The total number of tied detected measurements and tied less-than values should not exceed 40% of the total number of measurements.

- If more than 40% of the pooled data are less-than values, then the Gehan test should be used.

STEP 3: Calculate the sum of the ranks of the site measurements. Denote this sum by W_s.

STEP 4: Calculate

$$Z_0 = \frac{W_s - \frac{n(n+m+1)}{2} - S}{\sqrt{\frac{nm(n+m+1)}{12}}} \qquad (6\text{-}10)$$

STEP 5: Use the z-table to get the critical values $z_{1-\alpha}$, where $z_{1-\alpha}$ is the $100(1-\alpha)$ percentile of the standard normal distribution.

STEP 6: Conclusion for Large Sample Approximations:

Form 1: If $Z_0 > z_{1-\alpha}$, then reject the null hypothesis that the site population mean/median is less than or equal to the background population mean/median.
Form 2: If $Z_0 < -z_{1-\alpha}$, then reject the null hypothesis that the site population mean is greater than or equal to the background population mean.
Two Sided: If $|Z_0| > z_{1-\alpha/2}$, then reject the null hypothesis that the site population mean is same as the background population mean.
Form 2 with substantial difference, S: If $Z_0 < -z_{1-\alpha}$, then reject the null hypothesis that the site population mean is greater than or equal to the background population location + the substantial difference, S.

P-Values for Two-Sample WMW Test – For Large Sample

A p-value is the smallest value for which the null hypothesis is rejected in favor of the alternative hypotheses. Thus, based upon the given data, the null hypothesis is rejected for all values of α (the level of significance) greater than or equal to the p-value. The details of computing a large sample p-value for a two-sample nonparametric WMW test for comparing two means/medians are given in EPA (2006). ProUCL 4.0 directly computes (based upon normal distribution) p-values large samples WMW test for each form of null hypothesis. If the computed p-value is smaller than the specified value of, α, the conclusion is to reject the null hypothesis based upon the collected data set used in the various computations.

Note: *As suggested in the literature (EPA, 1994), both the WMW test and the quantile test should be used on the same data set to compare the two data distributions. Typically, the WMW test is used to compare the measures of locations (central tendencies assuming the equality of dispersions and shape of the distributions), whereas a quantile test is useful to determine if the observations in the upper tail of the site data set are larger than those found in the upper tail of the background data set. A typical WMW test simply compares the measures of location (central tendencies) of the two distributions (populations). A*

WMW test cannot compare the upper tails of two data distributions. Specifically, a WMW test may lead to the conclusion of the similarity/equality of the two data distributions even when the upper tail of one data distribution is much larger (shifted to the right) than the upper tail of the other data distribution. In order to detect the differences and shifts in the tails of two data distributions, one should use quantile test described as follows.

6.9.2 The Gehan Test

The Gehan test is one of several nonparametric tests that has been proposed to test for the differences between two sites when the data sets have multiple censoring points and detection limits. Among these tests, Palachek *et al.* (1993) indicate they selected the Gehan test primarily because it was the easiest to explain, because the several methods generally behave comparably, and because the Gehan test reduces to the WRS test, a relatively well-known test to environmental professionals. Palachek *et al.* (1993) used their computer code to conduct Gehan tests on data from the Rocky Flats Environmental Technology Site near Denver, CO. They recommend using the Gehan test rather than a more complicated procedure involving replacement of nondetects by a value such as one-half of the detection limit, testing for distribution shape and variance, and then conducting appropriate t- tests or the WMW test. The Gehan test as described here is available in ProUCL 4.0.

6.9.2.1 Limitations and Robustness

The Gehan test can be used when the background or site data sets contain multiple less-than values with different detection limits.

- The Gehan test is somewhat tedious to compute by hand. The use of a compute program is desirable.
- If the censoring mechanisms are different for the site and background data sets, then the test results may be an indication of this difference in censoring mechanisms rather than an indication that the null hypothesis is rejected.

Note: *The Gehan test is used when many ND observations or multiple DLs may be present in the two data sets. Therefore, the conclusions derived using this test may not be reliable when dealing with samples of sizes smaller than 10. Furthermore, it has been suggested throughout this guide to have a minimum of 8-10 observations (from each of the population) to use hypotheses testing approaches, as decisions derived based upon smaller data sets may not be reliable enough to draw important decisions about the human health and the environment. Therefore, this test (as included in ProUCL 4.0) is described here for data sets of sizes ≥ 10. The test described as follows is based upon the normal approximation of Gehan's statistic.*

6.9.2.2 Directions for the Gehan Test when *m* ≥ *10* and *n* ≥ *10*.

Let $X_1, X_2, \ldots, X_n$ represent data points from the site population $Y_1, Y_2, \ldots, Y_m$ represent background data from the background population. For data sets of sizes greater than or equal to 10, a test based upon normal approximations is described in the following.

STEP 1: Let $\tilde{\mu}_X$ represent the site and $\tilde{\mu}_Y$ represent the background population medians. State the following null and the alternative hypotheses:

Form 1: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y \le 0$ *vs.* H_1: $\tilde{\mu}_X - \tilde{\mu}_Y > 0$

Form 2: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y \ge 0$ *vs.* H_1: $\tilde{\mu}_X - \tilde{\mu}_Y < 0$

Two Sided: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y = 0$ *vs.* H_1: $\tilde{\mu}_X - \tilde{\mu}_Y \ne 0$

Form 2 with substantial difference, S: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y \ge S$ *vs.* H_1: $\tilde{\mu}_X - \tilde{\mu}_Y < S$

STEP 2: List the combined m background and n site measurements, including the less-than values, from smallest to largest, where the total number of combined samples is $N = m + n$. The less-than symbol (<) is ignored when listing the N data from smallest to largest.

STEP 3: Determine the N ranks, $R_1, R_2, \ldots, R_n$, for the N ordered data values using the method described in the example given below.

STEP 4: Compute the N scores, $a(R_1), a(R_2), \ldots, a(R_n)$, using the formula $a(R_i) = 2R_i - N - 1$, where i is successively set equal to $1, 2, \ldots, N$.

STEP 5: Compute the Gehan statistic, G, as follows:

$$G = \frac{\sum_{i=1}^{N} h_i a(R_i)}{\left[mn \sum_{i=1}^{N} \frac{[a(R_i)]^2}{N(N-1)} \right]^{\frac{1}{2}}} \qquad (6\text{-}11)$$

where $\begin{cases} h_i = 1 \\ h_i = 0 \end{cases}$

$h_i = 1$ if the i^{th} datum is from the site population
$h_i = 0$ if the i^{th} datum is from the background population
$N = n + m$
$a(R_i) = 2R_i - N - 1$, as indicated above.

STEP 6: Use the normal z-table to get the critical values.

STEP 7: Conclusion based upon approximate normal distribution of G statistic:

Form 1: If $G \ge z_{1-\alpha}$, then reject the null hypothesis that the site population median is less than or equal to the background population median.

Form 2: If $G \le -z_{1-\alpha}$, then reject the null hypothesis that the site population median is greater than or equal to the background population median.

Two Sided: If $|G| \ge z_{1-\alpha/2}$, then reject the null hypothesis that the site population median is same as the background population median.

Form 2 with substantial difference, S: If $G \leq - z_{1-\alpha}$, then reject the null hypothesis that the site population median is greater than or equal to the background population median and the substantial difference.

P-Values for Two-Sample Gehan Test

A p-value is the smallest value for which the null hypothesis is rejected in favor of the alternative hypotheses. Thus, based upon the given data, the null hypothesis is rejected for all values of α, the level of significance, greater than or equal to the p-value. For Gehan's test as described above, the p-values are computed using normal approximation for Gehan's G statistic. The p-values can be computed using the simple procedure as used for computing large sample p-values for a two-sample nonparametric WMW test. ProUCL 4.0 directly computes (based upon normal distribution) p-values for Gehan test for each form of null hypothesis. If the computed p-value is smaller than the specified value of, α, the conclusion is to reject the null hypothesis based upon the collected data set used in the various computations.

6.9.3 The Quantile Test

The quantile test (EPA, 1994) is a nonparametric test and is useful to detect a shift to the right in the right-tails of the site and background distributions. The quantile test when used in parallel with WMW test provides the user with stronger evidence to make decisions on whether the site has attained remediation (or background) levels or not (or if site and background concentrations are comparable). It is noted that the critical values for the quantile test are available for Background Form 1 hypothesis. Therefore, for quantile test, only Form 1 of hypothesis testing is available in ProUCL 4.0. Specifically, the null hypothesis is: Concentration in the cleanup unit (AOC) is comparable to that of the background area. A partial formulation of the quantile test (as given in EPA, 1994) is described as follows.

6.9.3.1 Limitations and Robustness

- It may give unreliable results if less than values are present in the largest detected observations.
- If the quantile test does not declare that the chemical is a COPC, then a Wilcoxon-Mann-Whitney test should be performed to ascertain the results.
- Since the test focuses on the right tails, presence of large outliers will bias results.
- It does not require any distributional assumption.
- As the test focuses on the right tail of the site and background distributions, it can have more power to detect differences than the Gehan, Wilcoxon-Mann-Whitney or the two-sample t-tests.
- Relatively simple to conduct.

6.9.3.2 Quantile Test in the Presence of Nondetects

- A principal requirement when applying the quantile test on censored data sets is to discard all less-than values present in the largest *r* detected observations.

6.9.3.3 Directions for the Quantile Test

Let $X_1, X_2, \ldots, X_n$ represent the n data points from the site population and $Y_1, Y_2, \ldots, Y_m$ represent the m data points from the background population.

STEP 1: State the following null and the alternative hypotheses:

Form 1: H_0: $\tilde{\mu}_X - \tilde{\mu}_Y \le 0$ vs. H_1: $\tilde{\mu}_X - \tilde{\mu}_Y > 0$

STEP 2: From the tables (EPA (1994)) look up the values of *r*, *k*, and α corresponding to *m* and *n*. The formula for computing the actual α is given below.

$$\alpha = \frac{\sum_{i=k}^{r} \binom{m+n-r}{n-i}\binom{r}{i}}{\binom{m+n}{n}} \tag{6-12}$$

STEP 3: If the actual α and the desired α (0.05, 0.01, etc.) do not agree, then increase or decrease *r* or *k* by 1 unit and again compute the actual α. Use the new *r* and *k* value in STEP 3.

Note: *As r and k are discrete in nature, getting the same value of α from the formula and from the table may not be achievable.*

STEP 4: Order the pooled data set from smallest to largest. In the presence of ties, increase *r* to include all ties in the *r* largest observations; e.g., when *r = 4* and *k = 4*, then counting down from the r^{th} largest observation if there are two ties within *r = 4* and two ties not in *r*, increase *r* by 2 and subsequently increase *k* by 2. The modified *r* and *k* values to be used will be now 6.

STEP 5: If the *m* and *n* are not multiples of 5, then the closest values of *m* and *n* are entered in the table to get the values of *r*, *k* and α. The formula in Step 1 is then used to find the value of the actual α with the original *m* and *n* values. If the actual α value is close enough to the desired α value, the quantile test is conducted with *r* and *k* from the rounded *m* and *n* values. For an example, for a required α level of 0.01 and with *m* = 47 and *n* = 77, the closest tabled values are *m* = 45 and *n* = 75. For *m* = 45 and *n* = 75, tabled values are *r* = 9, *k* = 9, and α = 0.012. To check if the quantile test can be performed, compute the actual α value as shown below:

Actual Alpha = $\alpha = \dfrac{\binom{124-9}{77-9}\binom{9}{9}}{\binom{124}{77}} = 0.012$

Since 0.012 is close enough to 0.01, it is safe to conduct the quantile test.

STEP 6: *For r < 20,* compute the probability *P* as shown in Step 1. For *r ≥ 20,* use the normal approximation and calculate *Z* as shown below:

$$z = \frac{k - 0.5 - \bar{X}}{sd} \tag{6-13}$$

where $\bar{X} = \dfrac{nr}{m+n}$ and $sd = \left[\dfrac{mnr\left(m+n+r\right)}{\left(m+n\right)^2\left(m+n-1\right)}\right]^{\frac{1}{2}}$

STEP 7: Conclusion:

When $r < 20$,
Form 1: If computed $k \geq$ *# of Site Observations in r*, then reject the null hypothesis that the site population parameter is less than or equal to the background population parameter.

When $r \geq 20$,
Form 1: If P – value < specified α, then reject the null hypothesis that the site population parameter is less than or equal to the background population parameter. The details are given in EPA (1994).

Chapter 7

Outlier Tests for Data Sets with and without Nondetect Values

Outliers are measurements that are extremely large or small relative to the rest of the data and, therefore, are suspected of misrepresenting the main dominant population from which they were collected. Outliers are inevitable in most environmental applications. Outliers may result from gross errors, such as transcription errors, data-coding errors, or measurement system problems such as instrument breakdown. However, outliers may also represent true extreme values of a distribution (for instance, hot spots) and indicate more variability in the population than was expected. Typically, outliers represent observations coming from population(s) different from the main dominant population (e.g., a background area, a site area after cleanup) under study. Typically, outliers represent low probability observations coming from the tails of the data distribution under consideration (e.g., data from an AOC). The presence of outliers in a data set distorts the computations of all classical statistics (e.g., sample mean, standard deviation, upper prediction, and upper tolerance limits, test statistics, GOF statistics, and also outlier test statistics) of interest. Some description (with examples) of the influence of outliers on the computation of various statistics, including the sample mean, sample standard deviation and several upper limits, is given in Chapters 3 and 5 of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b). The use of such distorted statistics (e.g., two-sample t-test, GOF test) may lead to incorrect cleanup decisions about the site under consideration.

Statistical tests based on parametric methods generally are more sensitive to the existence of outliers in either the site or the background data sets than are those based on nonparametric distribution-free methods. The use of nonparametric hypotheses tests for background comparisons can reduce (but not completely) the sensitivity of test results to the presence of outliers to a certain extent. Specifically, nonparametric statistics (e.g., UTLs, UPLs) based upon higher order statistics (e.g., largest, second largest) will represent outlying observations. Outliers can also lead to both Type I and Type II errors by distorting the test statistics used for hypotheses testing about the population parameters (e.g., means). They can lead to inconclusive and potentially incorrect results if the test statistics are sensitive to the outliers. This issue is illustrated below by using an example, comparing site versus background lead concentrations collected from a Superfund site.

In environmental applications, it is important to identify high outlying observations, as those high outliers may represent contaminated (or hot areas) locations of a polluted site. Typically, it is the presence of a few high outlying observations that distort the normality of a data set. That is, in practice, many data sets follow a normal distribution after the removal of potential outliers. Unfortunately such data sets with a few high outliers (often representing contaminated site areas) can also be modeled by a lognormal distribution. It is observed that in practice, the use of a log-transformation tends to hide and accommodate outlier(s) as part of the majority of the data set representing the dominant population. Statistical methods used on such contaminated data sets will yield distorted statistics and estimates (e.g., site average) for the entire site area under investigation. For an example, inclusion of a few high outliers (perhaps representing a polluted part of the site) in the computation of a UCL95 will yield a distorted estimate of the EPC for the entire site area under investigation. Decisions based upon such distorted statistics can be expensive and (or) not protective of the human health and the environment. In environmental applications the objective is to expose and identify such outlying observations as those observations often represent hot spots and (or) contaminated areas of the site, or observations possibly collected from an area other than the site background. These observations need to be identified as those requiring separate and further investigation and remediation. The project team and the decision makers involved in the project decision

should decide about the proper disposition (include or not include) of high outlying observations. Sometimes, performing statistical analyses twice on the same data set – once on using the full data set with outliers and once on data set without high outliers can help a typical user in determining the disposition of high outliers.

Statistical outlier tests give the analyst probabilistic evidence that an extreme value (potential outlier) does not "fit" with the distribution of the remainder of the data and is therefore a statistical outlier. These tests should only be used to *identify* data points that require further investigation. The classical outlier tests should be accompanied by graphical displays such as, Q-Q plots and box plots. Graphical displays provide additional insight into a data set that cannot be revealed by tests statistics (e.g., Rosner test, Dixon test, Shapiro Wilk test) alone. The statistical tests (e.g., Rosner test) alone cannot determine whether a statistical outlier should be discarded or corrected within a data set; this decision should be based on judgmental or scientific grounds. Typically, there are 5 steps involved in treating extreme values or outliers:

1. Identify extreme high values that may be potential outliers,
2. Apply a statistical test and supplement them with graphical displays,
3. Scientifically review the statistical outliers and decide on their proper disposition,
4. Conduct data analyses with and without the statistical outliers, and
5. Document the entire process.

The final disposition of outliers should be a team effort including the project team, decision makers, and experts familiar with the site conditions. As mentioned before, potential outliers may be identified through the graphical representations. Graphs, such as the box and goodness-of-fit (GOF) Q-Q plot, can be used to identify observations that are much larger or smaller than the rest of the data. If potential outliers are identified, then the next step is to apply one of the statistical tests described in the following sections.

If a data point is found to be an outlier, then the analyst may either: 1) correct the data point; 2) isolate the data point for further investigation; or 3) use the data point in all of the analyses. This decision should be based on scientific reasoning in addition to the results of the statistical test. For instance, data points containing transcription errors should be corrected, whereas data points collected while an instrument was malfunctioning may be discarded. As mentioned before, the disposition of outlier(s) should be a team effort. The project team should assess the influence (often undue influence) of outliers on the estimates (e.g., EPC, BTV) and test statistics (e.g., t-test) to be computed. It is noted that even the presence of a single outlier can distort all statistics of interest such as averages, standard deviations (*sd*), and various upper limits and test statistics. One simple way to assess the influence of outliers on the statistics to be computed is to compute the statistics with and without the high outlying observations. In other words, if a high outlier is discarded from the data set, then it is desirable that all of the statistical analysis of the data should be applied to both the full (with outliers) and the truncated data set (without high outliers) so that the influence of outliers may be assessed. For an example, if the difference between a UCL95 based upon a full data set with outliers is significantly higher (as often is the case) from the UCL95 based upon the data set without the high outlying observations, then the project team should be able to decide which of the two UCL values represents a more realistic estimate of the EPC term.

The practitioner should also be clear about the objective of the study. Typically, in such applications, the objective is to compute the statistics (e.g., averages, UCLs, UPLs) based upon the majority of the data set representing the dominant population. For an example, the average (estimate of population mean) should be a representative of the population based upon the majority of the data representing the dominant (main) population. A distorted estimate accommodating a few outliers tends to represent that contaminated area of the site (and not the entire dominant site area under consideration). Therefore, it is desirable that the project team decides to isolate a few high outliers and investigate their locations for the possibility of further remediation. A couple of classical outliers tests (Rosner test and Dixon test) often cited in environmental literature have been incorporated in ProUCL 4.0. These tests can be used on data sets with and without nondetect observations.

7.1 Outlier Tests for Data Sets without Nondetect Observations

A couple of classical outlier test procedures often used in environmental applications (EPA, 2006, and Gilbert, 1987) are briefly described here. It is noted that these classical tests do suffer from masking effects and may fail to identify potential outliers present in a data set. This is especially true when multiple outliers or multiple populations may be present in a data set. Such scenarios can be revealed by graphical displays, such as a Q-Q plot discussed earlier. More effective robust outlier identification procedures (Singh and Nocerino, 1995) are beyond the scope of ProUCL 4.0. Several robust estimation and outlier identification procedures are available in Scout software package (EPA, 1999), which is currently being upgraded.

7.1.1 Dixon's Test

Dixon's Extreme Value test (1953) can be used to test for statistical outliers when the sample size is less than or equal to 25. It is noted that Dixon's test considers both extreme values that are much smaller than the rest of the data (Case 1) and extreme values that are much larger than the rest of the data (Case 2). This test assumes that the data without the suspected outlier are normally distributed; therefore, it is necessary to perform a test for normality on the data without the suspected outlier before applying this test. This means that the user has to identify (guess) potential outliers that may be present in the data set. One simple way to identify and look at outliers is the use of graphical displays such as a Q-Q plot and box plot. The Dixon test often suffers from masking effects when more than one outlier may be present in the data set. If more than one outlier is suspected, the Dixon test may lead to masking where two or more outliers close in value "hide" one another. As mentioned before, the use of robust and resistant outlier procedures (Singh and Nocerino, 1995, Rousseeuw and Leroy, 1987, and Scout, 1999) is desirable. However, robust and resistant methods are beyond the scope of ProUCL 4.0. Several robust methods are available in Scout (EPA, 1999) software package, which is currently under revision and upgrade.

Even though Dixon's test finds outliers in both tails (low and high outliers) of the data distribution, it is the identification of high outlying observations (perhaps representing contamination), which is important in environmental applications. The inclusion of high outliers in a data set results in distorted statistics of interest, including estimates and test statistics. The low identified outliers (if any) may be retained in a data set to compute various statistics of interest.

7.1.1.1 Directions for the Dixon's Test

STEP 1: Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ represent the data ordered from smallest to largest. Check that the data without the suspect outlier are normally distributed. If normality fails, then apply a different outlier identification method such as a robust outlier identification procedure. *It is suggested to avoid the use of a*

transformation such as a log-transformation to achieve normality to be able to use the Dixon test. All cleanup and remediation decisions are made based upon the data set in raw scale. Therefore, outliers perhaps representing isolated contaminated locations should be identified in the original scale. As mentioned before, the use of a log-transformation tends to hide and accommodate outliers (instead of identifying them).

STEP 2: $X_{(1)}$ is a potential outlier (Case 1): Compute the test statistic, *C*, where

$$C = \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}} \text{ for } 3 \le n \le 7, \; C = \frac{X_{(3)} - X_{(1)}}{X_{(n-1)} - X_{(1)}} \text{ for } 11 \le n \le 13,$$

$$C = \frac{X_{(2)} - X_{(1)}}{X_{(n-1)} - X_{(1)}} \text{ for } 8 \le n \le 10, \; C = \frac{X_{(3)} - X_{(1)}}{X_{(n-2)} - X_{(1)}} \text{ for } 14 \le n \le 25,$$

STEP 3: If *C* exceeds the critical value for the specified significance level α, then $X_{(1)}$ is an outlier and should be further investigated.

STEP 4: $X_{(n)}$ is a potential outlier (Case 2): Compute the test statistic, *C*, where

$$C = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}} \text{ for } 3 \le n \le 7, \; C = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(2)}} \text{ for } 11 \le n \le 13,$$

$$C = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(2)}} \text{ for } 8 \le n \le 10, \; C = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(3)}} \text{ for } 14 \le n \le 25,$$

STEP 5: If *C* exceeds the critical value for the specified significance level α, then $X_{(n)}$ is an outlier and should be further investigated.

7.1.2 Rosner's Test

A parametric test developed by Rosner can be used to detect up to 10 outliers for sample sizes of 25 or more. The details of the test can be found in Gilbert (1987). This test also assumes that the data are normally distributed without the outliers; therefore, it is necessary to perform a test for normality before applying this test. Note that the test assumes that, the data without the outliers are normally distributed; therefore, the test for normality has to be performed without the suspected outliers, which need to be identified or determined first. A graphical display (Q-Q plot) can help determine to identify suspected outliers.

7.1.2.1 Directions for the Rosner's Test

To apply Rosner's test, first determine an upper limit, r_0, on the number of outliers ($r_0 \le 10$), then order the r_0 extreme values from most extreme to least extreme. Rosner's test statistic is computed using the sample mean and sample standard deviation.

STEP 1: Let $X_1, X_2, \ldots, X_n$ represent the ordered data points. By inspection, identify the maximum number of possible outliers, r_0. Check that the data are normally distributed.

STEP 2: Compute the sample mean, $\bar{x}$, and the sample standard deviation, s, for all the data. Label these values $\bar{x}^{(0)}$ and $s^{(0)}$, respectively. Determine the value that is farthest from $\bar{x}^{(0)}$ and label this observation $y^{(0)}$. Delete $y^{(0)}$ from the data and compute the sample mean, labeled $\bar{x}^{(1)}$, and the sample standard deviation, labeled $s^{(1)}$. Then determine the observation farthest from $\bar{x}^{(1)}$ and label this observation $y^{(1)}$. Delete $y^{(1)}$ and compute $\bar{x}^{(2)}$ and $s^{(2)}$. Continue this process until r_0 extreme values have been eliminated.

After carrying out the above process, the analyst should have

$[\bar{x}^{(0)}, s^{(0)}, y^{(0)}]$; $[\bar{x}^{(1)}, s^{(1)}, y^{(1)}]$; …, $[\bar{x}^{(r_0-1)}, s^{(r_0-1)}, y^{(r_0-1)}]$ where

$\bar{x}^{(i)} = \frac{1}{n-i}\sum_{j=1}^{n-i} x_j$, $s^{(i)} = \sqrt{\frac{1}{n-i}\sum_{j=1}^{n-i}(x_j - \bar{x}^{(i)})^2}$, and $y^{(i)}$ is the farthest value $\bar{x}^{(i)}$.

Note: *The above formulae for $\bar{x}^{(i)}$ and $s^{(i)}$ assume that the data have been renumbered after each outlying observation is deleted.*

STEP 3: To test if there are "*r*" outliers in the data, compute: $R_r = \frac{|\, y^{(r-1)} - \bar{x}^{(r-1)} \,|}{s^{(r-1)}}$ and compare R_r to the critical value λ_r in the tables from any statistical literature. If $R_r \geq \lambda_r$, conclude that there are *r* outliers.

First, test if there are r_0 outliers (compare R_{r_0-1} to λ_{r_0-1}). If not, then test if there are r_0 - 1 outliers (compare R_{r_0-2} to λ_{r_0-2}). If not, then test if there are r_0 - 2 outliers, and continue, until either it is determined that there are a certain number of outliers or that there are no outliers at all.

7.2 Outlier Tests for Data Sets with Nondetect Observations

For the purpose of the identification of high outliers, one may replace the nondetect values by their respective detection limits or may just ignore them (especially when the number of detected values is large such as exceeding 8-10) from any of the outlier test (e.g., Rosner test) computation, including the graphical displays such as Q-Q plots. Both of these procedures (ignoring NDs, or replacing them by DL/2) for outliers testing are available in ProUCL for data sets with ND values. Note that outlier identification procedures represent exploratory tools and are used for pre-processing of a data set to identify outliers or multiple populations that may be present in a data set. Except for the identification of high outlying observations, the outlier identification statistics (computed with NDs or without NDs) are not used in any of the estimation and decision making process. Therefore, for the purpose of the identification of high outliers, it does not matter how the nondetect observations are treated. After outliers have been identified, the project team and experts familiar with the site should make the final decision about the disposition of outliers.

Thus, nondetects and outliers are inevitable in most environmental data sets. Typically, the objective is to identify high outlying contaminating observations, the same two classical tests described above may be

used to identify outliers in data sets with nondetects observations. ProUCL 4.0 offers two options to test for outliers that may be present in the upper tail of the data distribution. The user can use Dixon test or Rosner test to test for outliers using the data set excluding all nondetect observations. Alternatively, the user may replace all nondetects by their respective DL/2 values, and use Dixon test or Rosner test on the resulting data set to test for outliers. This option should be used as an exploratory tool to identify outliers (if any). In any case, it is always desirable to use graphical displays to visually look at the outliers. ProUCL 4.0 can be used to compute Q-Q plot and box plot for data sets with nondetect observations. An example illustrating the improper influence of outliers on test statistic such as t-test statistic is considered as follows.

Example. At a Superfund site many inorganic compounds were analyzed to perform site versus background comparisons. In this example, background and site lead concentrations are being compared. This example is included here to illustrate how outliers influence the tests statistics and in turn leading to potentially incorrect inclusions. The background data set has only 6 data points (which are not enough to perform two-sample comparisons) and the site data set has 10 data points. This complete data set is included in Appendix 3 of the revised *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002b). The site data set seems to have a couple of outliers. This example illustrates how test statistics get distorted by outliers leading to potentially incorrect conclusions. Therefore, as mentioned before, it is desirable to investigate the outliers separately. The two-sample t-test results and nonparametric WMW test results for the background versus site comparisons are presented in the following tables. The parametric and nonparametric test results for full data set (with outliers) are respectively, summarized in Table 7-1 (t-test) and Table 7-2 (WMW test). Table 7-3 has outlier test results for site data set based upon Dixon test. The parametric and nonparametric tests were performed again using the site data set without the outliers. Those test results are summarized in Table 7-4 (t-test) and Table 7-5 (WMW test). All statistics have been obtained using ProUCL 4.0 software package.

Table 7-1. Parametric Site vs. Background Comparison for Lead Data with Outliers

t-Test Site vs Background Comparison for Full Data Sets without NDs

User Selected Options	
From File	WorkSheet.wst
Full Precision	OFF
Confidence Coefficient	95%
Substantial Difference (S)	0.000
Selected Null Hypothesis	Site or AOC Mean Less Than or Equal to Background Mean (Form 1)
Alternative Hypothesis	Site or AOC Mean Greater Than the Background Mean

Area of Concern Data: OS_Lead

Background Data: BG_Lead

Raw Statistics

	Site	Background
Number of Valid Samples	10	6
Number of Distinct Samples	10	6
Minimum	19.7	14.75
Maximum	1940	25.7
Mean	268.2	18.08
Median	41.63	16.55
SD	595.8	4.019
SE of Mean	188.4	1.641

Site vs Background Two-Sample t-Test

H0: Mu of Site - Mu of Background <= 0

Method	DF	t-Test Value	Critical t (0.050)	P-Value
Pooled (Equal Variance)	14	1.014	1.761	0.164
Satterthwaite (Unequal Variance)	9.0	1.328	1.833	0.108

Pooled SD 477.685

Conclusion with Alpha = 0.050

* Student t (Pooled) Test: Do Not Reject H0, Conclude Site <= Background

* Satterthwaite Test: Do Not Reject H0, Conclude Site <= Background

Test of Equality of Variances

Numerator DF	Denominator DF	F-Test Value	P-Value
9	5	21975.505	0.000

Conclusion with Alpha = 0.05

* Two variances are not equal

From Table 7-1, it is noted that, Student's t-test statistic got distorted by outliers leading to the incorrect conclusion that the mean lead concentrations of site and background populations are comparable. However, the nonparametric WMW test statistic (Table 7-2) lead to the correct conclusion that the site lead concentration levels are significantly higher than the background lead concentration levels. In order to avoid the use of distorted test statistics and deriving incorrect conclusions, it is always desirable to supplement the formal tests results with graphical displays. It is noted that a formal test statistic (e.g., t-test) alone cannot determine if the outliers are present and the conclusions derived are correct or incorrect. A quick look at the side-by-side box plot (Figure 7-1) of site and background lead data reveals that the site concentrations are significantly higher than the background concentrations. Next, Table 7-3 summarizes the results of Dixon's outlier test on site lead data set. One outlier (1940) was identified.

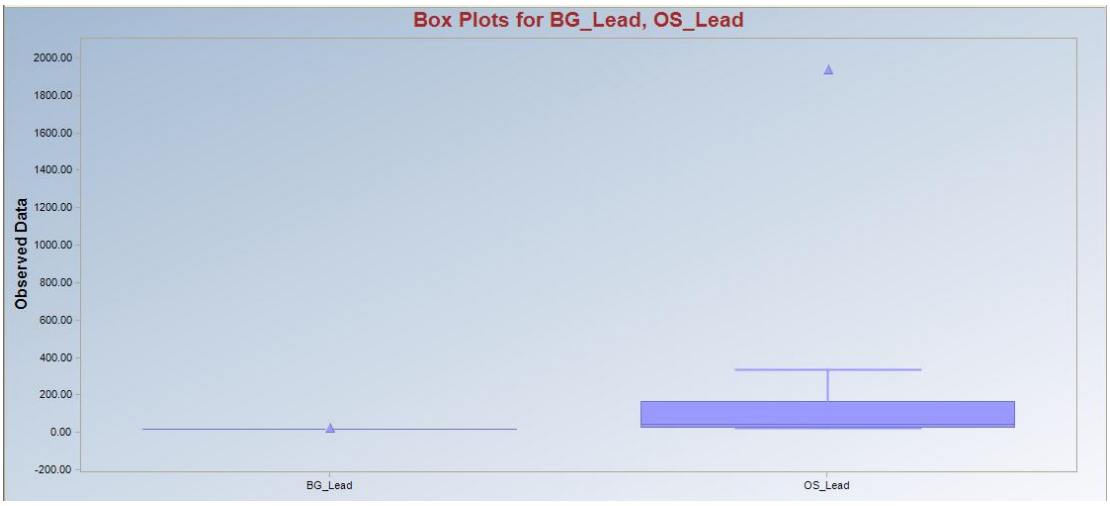

Figure 7-1. Side-by-Side Box Plots for Lead from Background and Site Areas

Table 7-2. Nonparametric Site vs. Background Comparison for Lead with Outliers

	Wilcoxon-Mann-Whitney Site vs Background Comparison Test for Full Data Sets without NDs	
User Selected Options		
From File	WorkSheet.wst	
Full Precision	OFF	
Confidence Coefficient	95%	
Substantial Difference	0.000	
Selected Null Hypothesis	Site or AOC Mean/Median Less Than or Equal to Background Mean/Median (Form 1)	
Alternative Hypothesis	Site or AOC Mean/Median Greater Than Background Mean/Median	
Area of Concern Data: OS_Lead		
Background Data: BG_Lead		
Raw Statistics		
	Site	Background
Number of Valid Samples	10	6
Number of Distinct Samples	10	6
Minimum	19.7	14.75
Maximum	1940	25.7
Mean	268.2	18.08
Median	41.63	16.55
SD	595.8	4.019
SE of Mean	188.4	1.641
Wilcoxon-Mann-Whitney (WMW) Test		
H0: Mean/Median of Site or AOC <= Mean/Median of Background		
Site Rank Sum W-Stat	113	
WMW Test U-Stat	58	
WMW Critical Value (0.050)	45	
Approximate P-Value	0.0014	
Conclusion with Alpha = 0.05		
Reject H0, Conclude Site > Background		

Table 7-3. Dixon Outlier Test Results for Site Lead Data Set

Dixon's Outlier Test for OS_Lead
Number of data = 10
10% critical value: 0.409
5% critical value: 0.477
1% critical value: 0.597
1. 1940 is a Potential Outlier (Upper Tail)
Test Statistic: 0.836
For 10% significance level, 1940 is an outlier.
For 5% significance level, 1940 is an outlier.
For 1% significance level, 1940 is an outlier.
2. 19.7 is a Potential Outlier (Lower Tail)
Test Statistic: 0.013
For 10% significance level, 19.7 is not an outlier.
For 5% significance level, 19.7 is not an outlier.
For 1% significance level, 19.7 is not an outlier.

The parametric test (Table 7-4) and nonparametric test (Table 7-5) were performed again to compare the site and background lead concentrations using site data set without the outlier, 1940. The parametric t-test (Table 7-4) still leads to the incorrect conclusion that site and background concentrations are comparable. However, the nonparametric WMW (Table 7-5) test again leads to the correct conclusion that the site lead concentrations are higher than those of the background lead concentrations. This example illustrates that when the underlying assumptions are not met (e.g., normality) or outliers are present, the parametric test can result in incorrect conclusions.

Table 7-4. Parametric Site vs. Background Comparison for Lead Data without Outlier

t-Test Site vs Background Comparison for Full Data Sets without NDs

User Selected Options	
From File	C:\Documents and Settings\Guest\Desktop\two-Sample-salford.wst
Full Precision	OFF
Confidence Coefficient	95%
Substantial Difference (S)	0.000
Selected Null Hypothesis	Site or AOC Mean Less Than or Equal to Background Mean (Form 1)
Alternative Hypothesis	Site or AOC Mean Greater Than the Background Mean

Area of Concern Data: OS_Lead

Background Data: BG_Lead

Raw Statistics

	Site	Background
Number of Valid Samples	9	6
Number of Distinct Samples	9	6
Minimum	19.7	14.75
Maximum	338	25.7
Mean	82.47	18.08
Median	38.6	16.55
SD	105.5	4.019
SE of Mean	35.18	1.641

Site vs Background Two-Sample t-Test

H0: Mu of Site - Mu of Background <= 0

		t-Test	Critical	
Method	DF	Value	t (0.050)	P-Value
Pooled (Equal Variance)	13	1.475	1.771	0.082
Satterthwaite (Unequal Variance)	8.0	1.828	1.860	0.052

Pooled SD 82.833

Conclusion with Alpha = 0.050

* Student t (Pooled) Test: Do Not Reject H0, Conclude Site <= Background

* Satterthwaite Test: Do Not Reject H0, Conclude Site <= Background

Test of Equality of Variances

Numerator DF	Denominator DF	F-Test Value	P-Value
8	5	689.687	0.000

Conclusion with Alpha = 0.05

* Two variances are not equal

Table 7-5. Nonparametric Site vs. Background Comparison for Lead without Outlier

	Wilcoxon-Mann-Whitney Site vs Background Comparison Test for Full Data Sets without NDs	
User Selected Options		
From File	C:\Documents and Settings\Guest\Desktop\two-Sample-salford.wst	
Full Precision	OFF	
Confidence Coefficient	95%	
Substantial Difference	0.000	
Selected Null Hypothesis	Site or AOC Mean/Median Less Than or Equal to Background Mean/Median (Form 1)	
Alternative Hypothesis	Site or AOC Mean/Median Greater Than Background Mean/Median	
Area of Concern Data: OS_Lead		
Background Data: BG_Lead		
Raw Statistics		
	Site	Background
Number of Valid Samples	9	6
Number of Distinct Samples	9	6
Minimum	19.7	14.75
Maximum	338	25.7
Mean	82.47	18.08
Median	38.6	16.55
SD	105.5	4.019
SE of Mean	35.18	1.641
Wilcoxon-Mann-Whitney (WMW) Test		
H0: Mean/Median of Site or AOC <= Mean/Median of Background		
Site Rank Sum W-Stat	97	
WMW Test U-Stat	52	
WMW Critical Value (0.050)	41	
Approximate P-Value	0.0019	
Conclusion with Alpha = 0.05		
Reject H0, Conclude Site > Background		

Appendix

Simulated Critical Values for Gamma GOF Tests, the Anderson-Darling Test and the Kolmogorov-Smirnov Test

Simulation Experiments

The simulation experiments performed are briefly described here. The experiments were carried out for various values of the sample size, n = 5(25)1, 30(50)5, 60(100)10, 200(500)100, and 1000. Random deviates of sample size n were generated from a gamma, (k, θ), population. Various values of k have been considered. The considered values of the shape parameter, k, are: 0.01, 0.025, 0.05, 0.1, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0, 20.0, and 50.0. These values of k cover a wide range of values of skewness, $2/\sqrt{k}$. The distributions of the Kolmogorov-Smirnov (K-S) test statistic, D, and the Anderson-Darling (A-D) test statistic, A^2, do not depend upon the scale parameter, θ , therefore, the scale parameter, θ , has been set equal to 1 in all of the simulation experiments. A typical simulation experiment can be described in the following four steps. Some details of the gamma deviate generation methods and the maximum likelihood estimation methods can be found in Singh, Singh, and Iaci (EPA, 2002).

Step 1. Generate a random sample of the specified size, n, from a gamma, G (k, 1), distribution. The algorithm as outlined in Whittaker (1974) has been used to generate the gamma deviates.

Step 2. For each generated sample, compute the MLEs of k and θ (Choi and Wette, 1969), and the K-S and the A-D test statistics (Anderson and Darling, 1954, D'Agostino and Stephens, 1986, and Schneider and Clickner, 1976) using the incomplete gamma function.

Step 3. Repeat Steps 1 and 2, 20,000 times.

Step 4. Arrange the resulting test statistics in ascending order. Compute the 80%, 90%, 95%, and 99% percentiles of the K-S test statistic and the A-D test statistic.

The resulting raw 20%, 10%, 5%, and 1% critical values for the two EDF tests are summarized in Tables 1 through 8 of this Appendix. The critical values as summarized in Tables 1-8 are in agreement (up to 3 significant digits) with all available exact or asymptotic critical values. It is also noted that the critical values for the K-S test statistic are more stable than those for the A-D test statistic. This is especially true when the shape parameter, k, is small and the sample size, n, is large. It is hoped that the availability of the critical values for the GOF tests for the gamma distribution will result in the frequent use of more practical and appropriate gamma distributions in environmental and other applications.

Table 1. Critical Values for A-D Test Statistic for Significance Level = 0.20

n\k	0.010	0.025	0.050	0.10	0.20	0.50	1.0	2.0	5.0	10.0	20.0	50.0
5	0.637	0.609	0.580	0.559	0.532	0.505	0.497	0.493	0.492	0.491	0.490	0.490
6	0.685	0.636	0.591	0.568	0.543	0.512	0.503	0.498	0.496	0.496	0.495	0.495
7	0.735	0.667	0.604	0.575	0.549	0.517	0.506	0.501	0.499	0.498	0.498	0.497
8	0.786	0.697	0.615	0.578	0.555	0.521	0.509	0.504	0.502	0.500	0.500	0.499
9	0.839	0.729	0.626	0.583	0.559	0.524	0.512	0.507	0.502	0.502	0.502	0.501
10	0.892	0.760	0.638	0.585	0.563	0.526	0.513	0.508	0.504	0.503	0.502	0.502
15	1.166	0.922	0.693	0.595	0.574	0.533	0.519	0.511	0.507	0.506	0.505	0.505
16	1.220	0.954	0.705	0.597	0.574	0.534	0.520	0.513	0.508	0.507	0.506	0.505
17	1.275	0.986	0.716	0.597	0.576	0.535	0.520	0.512	0.508	0.507	0.506	0.505
18	1.327	1.019	0.726	0.599	0.576	0.536	0.520	0.513	0.509	0.507	0.507	0.506
19	1.380	1.050	0.738	0.600	0.578	0.537	0.521	0.513	0.508	0.507	0.507	0.507
20	1.432	1.081	0.747	0.602	0.578	0.537	0.521	0.513	0.509	0.508	0.507	0.507
21	1.486	1.112	0.757	0.602	0.579	0.539	0.521	0.514	0.509	0.508	0.508	0.507
22	1.537	1.143	0.768	0.604	0.579	0.538	0.522	0.513	0.509	0.508	0.507	0.507
23	1.588	1.177	0.779	0.604	0.580	0.538	0.521	0.513	0.510	0.508	0.507	0.507
24	1.641	1.206	0.789	0.605	0.581	0.539	0.522	0.514	0.509	0.508	0.508	0.507
25	1.692	1.238	0.800	0.606	0.581	0.539	0.523	0.514	0.510	0.508	0.507	0.507
30	1.943	1.390	0.853	0.611	0.583	0.540	0.523	0.515	0.510	0.509	0.508	0.507
35	2.190	1.537	0.906	0.615	0.584	0.541	0.524	0.516	0.510	0.509	0.509	0.508
40	2.432	1.683	0.955	0.617	0.585	0.542	0.525	0.515	0.511	0.509	0.508	0.508
45	2.673	1.828	1.005	0.621	0.586	0.542	0.524	0.516	0.511	0.509	0.508	0.508
50	2.906	1.967	1.054	0.624	0.587	0.541	0.525	0.516	0.511	0.510	0.509	0.509
60	3.368	2.246	1.150	0.631	0.588	0.543	0.525	0.516	0.511	0.510	0.508	0.509
70	3.826	2.518	1.248	0.636	0.588	0.543	0.525	0.516	0.511	0.510	0.509	0.509
80	4.273	2.785	1.343	0.642	0.589	0.544	0.526	0.517	0.511	0.510	0.509	0.509
90	4.719	3.053	1.437	0.647	0.588	0.544	0.526	0.517	0.511	0.510	0.509	0.509
100	5.166	3.314	1.532	0.652	0.589	0.544	0.526	0.517	0.512	0.510	0.509	0.509
200	9.462	5.855	2.420	0.706	0.591	0.545	0.526	0.517	0.511	0.509	0.509	0.509
300	13.65	8.320	3.273	0.760	0.591	0.545	0.526	0.517	0.512	0.511	0.509	0.509
400	17.78	10.73	4.107	0.812	0.590	0.545	0.527	0.517	0.513	0.511	0.510	0.509
500	21.87	13.12	4.923	0.865	0.591	0.545	0.527	0.517	0.512	0.510	0.509	0.509
1000	42.04	24.87	8.900	1.123	0.592	0.546	0.526	0.517	0.512	0.510	0.510	0.509

Table 2. Critical Values for K-S Test Statistic for Significance Level = 0.20

n\k	0.010	0.025	0.050	0.10	0.20	0.50	1.0	2.0	5.0	10.0	20.0	50.0
5	0.349	0.341	0.332	0.328	0.323	0.313	0.307	0.304	0.303	0.302	0.301	0.301
6	0.335	0.322	0.310	0.305	0.299	0.289	0.284	0.281	0.279	0.279	0.279	0.279
7	0.321	0.306	0.292	0.285	0.279	0.270	0.265	0.262	0.261	0.260	0.260	0.260
8	0.310	0.293	0.276	0.268	0.264	0.255	0.250	0.247	0.246	0.245	0.245	0.244
9	0.301	0.283	0.264	0.255	0.251	0.242	0.237	0.235	0.233	0.232	0.232	0.232
10	0.294	0.274	0.253	0.244	0.239	0.231	0.226	0.224	0.222	0.221	0.221	0.221
15	0.266	0.243	0.216	0.203	0.199	0.192	0.188	0.185	0.184	0.183	0.183	0.183
16	0.262	0.238	0.211	0.197	0.193	0.186	0.182	0.180	0.179	0.178	0.178	0.178
17	0.259	0.234	0.206	0.191	0.188	0.181	0.177	0.175	0.174	0.173	0.173	0.173
18	0.255	0.230	0.201	0.186	0.183	0.176	0.172	0.170	0.169	0.168	0.168	0.168
19	0.252	0.227	0.197	0.182	0.179	0.172	0.168	0.166	0.165	0.164	0.164	0.164
20	0.249	0.224	0.194	0.177	0.174	0.168	0.164	0.162	0.161	0.160	0.160	0.160
21	0.246	0.220	0.190	0.173	0.170	0.164	0.160	0.158	0.157	0.157	0.157	0.156
22	0.244	0.218	0.187	0.170	0.167	0.160	0.157	0.155	0.154	0.153	0.153	0.153
23	0.241	0.215	0.184	0.166	0.163	0.157	0.154	0.152	0.150	0.150	0.150	0.150
24	0.239	0.212	0.181	0.163	0.160	0.154	0.151	0.149	0.147	0.147	0.147	0.147
25	0.237	0.210	0.178	0.160	0.157	0.151	0.148	0.146	0.145	0.144	0.144	0.144
30	0.227	0.200	0.167	0.147	0.144	0.138	0.135	0.134	0.133	0.132	0.132	0.132
35	0.220	0.192	0.158	0.137	0.134	0.129	0.126	0.124	0.123	0.123	0.123	0.123
40	0.214	0.186	0.151	0.128	0.125	0.121	0.118	0.116	0.116	0.115	0.115	0.115
45	0.208	0.180	0.145	0.121	0.119	0.114	0.112	0.110	0.109	0.109	0.109	0.109
50	0.204	0.176	0.140	0.115	0.113	0.108	0.106	0.105	0.104	0.103	0.103	0.103
60	0.197	0.168	0.132	0.106	0.103	0.099	0.097	0.096	0.095	0.095	0.095	0.095
70	0.191	0.162	0.126	0.099	0.096	0.092	0.090	0.089	0.088	0.088	0.088	0.088
80	0.187	0.158	0.121	0.093	0.090	0.086	0.084	0.083	0.083	0.082	0.082	0.082
90	0.183	0.154	0.117	0.088	0.085	0.081	0.080	0.079	0.078	0.078	0.078	0.078
100	0.180	0.150	0.113	0.084	0.080	0.077	0.076	0.075	0.074	0.074	0.074	0.074
200	0.163	0.132	0.094	0.061	0.057	0.055	0.054	0.053	0.053	0.053	0.053	0.052
300	0.155	0.125	0.086	0.051	0.047	0.045	0.044	0.044	0.043	0.043	0.043	0.043
400	0.151	0.120	0.081	0.046	0.041	0.039	0.038	0.038	0.037	0.037	0.037	0.037
500	0.148	0.117	0.077	0.042	0.036	0.035	0.034	0.034	0.034	0.033	0.033	0.033
1000	0.141	0.109	0.069	0.032	0.026	0.025	0.024	0.024	0.024	0.024	0.024	0.024

Table 3. Critical Values for A-D Test Statistic for Significance Level = 0.10

n\k	0.010	0.025	0.050	0.10	0.20	0.50	1.0	2.0	5.0	10.0	20.0	50.0
5	0.761	0.731	0.705	0.691	0.654	0.612	0.599	0.594	0.591	0.589	0.589	0.588
6	0.824	0.772	0.726	0.707	0.672	0.625	0.610	0.603	0.599	0.599	0.598	0.598
7	0.891	0.813	0.742	0.716	0.684	0.635	0.618	0.609	0.607	0.606	0.604	0.605
8	0.957	0.854	0.760	0.725	0.694	0.641	0.624	0.616	0.612	0.610	0.609	0.608
9	1.026	0.896	0.777	0.731	0.701	0.648	0.629	0.620	0.614	0.613	0.613	0.612
10	1.094	0.938	0.792	0.736	0.707	0.652	0.632	0.623	0.618	0.616	0.615	0.614
15	1.431	1.146	0.868	0.752	0.724	0.663	0.642	0.630	0.624	0.622	0.621	0.621
16	1.496	1.188	0.885	0.755	0.725	0.665	0.642	0.632	0.626	0.624	0.622	0.621
17	1.559	1.226	0.899	0.757	0.727	0.666	0.644	0.632	0.626	0.623	0.623	0.622
18	1.622	1.267	0.913	0.760	0.729	0.668	0.643	0.634	0.626	0.623	0.624	0.623
19	1.683	1.304	0.928	0.761	0.730	0.670	0.645	0.633	0.626	0.625	0.624	0.624
20	1.744	1.344	0.941	0.763	0.732	0.669	0.645	0.633	0.627	0.626	0.624	0.624
21	1.808	1.380	0.954	0.765	0.732	0.671	0.646	0.634	0.628	0.626	0.626	0.624
22	1.865	1.417	0.968	0.767	0.734	0.670	0.646	0.636	0.628	0.627	0.625	0.625
23	1.925	1.459	0.983	0.768	0.735	0.671	0.645	0.635	0.629	0.627	0.625	0.625
24	1.985	1.494	0.995	0.770	0.736	0.672	0.647	0.635	0.628	0.627	0.626	0.625
25	2.043	1.528	1.009	0.770	0.736	0.673	0.648	0.636	0.629	0.627	0.626	0.625
30	2.330	1.709	1.076	0.778	0.738	0.674	0.650	0.637	0.629	0.628	0.627	0.626
35	2.606	1.881	1.141	0.783	0.739	0.676	0.650	0.638	0.631	0.629	0.628	0.627
40	2.879	2.046	1.202	0.787	0.742	0.677	0.651	0.637	0.631	0.629	0.628	0.628
45	3.140	2.209	1.260	0.793	0.743	0.677	0.651	0.639	0.632	0.630	0.628	0.629
50	3.398	2.367	1.320	0.797	0.746	0.677	0.652	0.640	0.632	0.630	0.629	0.629
60	3.903	2.677	1.435	0.806	0.747	0.679	0.652	0.640	0.632	0.631	0.629	0.629
70	4.394	2.979	1.550	0.811	0.747	0.679	0.653	0.641	0.633	0.630	0.630	0.630
80	4.882	3.269	1.660	0.820	0.747	0.680	0.654	0.641	0.633	0.631	0.630	0.629
90	5.358	3.563	1.766	0.827	0.748	0.680	0.654	0.642	0.634	0.631	0.629	0.630
100	5.838	3.848	1.874	0.833	0.749	0.681	0.654	0.642	0.633	0.631	0.630	0.630
200	10.37	6.570	2.861	0.902	0.751	0.682	0.654	0.642	0.634	0.631	0.631	0.630
300	14.74	9.168	3.786	0.969	0.752	0.682	0.655	0.641	0.634	0.633	0.631	0.630
400	19.03	11.69	4.679	1.032	0.751	0.683	0.655	0.641	0.635	0.633	0.631	0.631
500	23.26	14.19	5.551	1.095	0.752	0.683	0.655	0.643	0.635	0.632	0.631	0.631
1000	43.96	26.32	9.737	1.399	0.752	0.684	0.655	0.643	0.635	0.632	0.631	0.630

Table 4. Critical Values for K-S Test Statistic for Significance Level = 0.10

n\k	0.010	0.025	0.050	0.10	0.20	0.50	1.0	2.0	5.0	10.0	20.0	50.0
5	0.386	0.376	0.368	0.366	0.359	0.346	0.339	0.336	0.334	0.333	0.333	0.333
6	0.369	0.357	0.345	0.339	0.332	0.319	0.313	0.310	0.307	0.307	0.307	0.307
7	0.356	0.339	0.323	0.318	0.313	0.301	0.294	0.290	0.288	0.288	0.287	0.287
8	0.344	0.327	0.308	0.301	0.296	0.284	0.278	0.274	0.272	0.271	0.271	0.271
9	0.334	0.315	0.295	0.286	0.281	0.270	0.264	0.260	0.258	0.257	0.257	0.257
10	0.326	0.305	0.282	0.273	0.268	0.257	0.251	0.248	0.246	0.245	0.245	0.245
15	0.294	0.270	0.241	0.227	0.223	0.214	0.209	0.206	0.204	0.204	0.203	0.203
16	0.290	0.265	0.236	0.221	0.217	0.208	0.203	0.200	0.198	0.198	0.197	0.197
17	0.285	0.260	0.230	0.214	0.211	0.202	0.197	0.194	0.193	0.192	0.192	0.192
18	0.281	0.256	0.225	0.209	0.205	0.197	0.192	0.189	0.188	0.187	0.187	0.187
19	0.278	0.252	0.221	0.204	0.200	0.192	0.187	0.184	0.183	0.182	0.182	0.182
20	0.274	0.248	0.217	0.199	0.196	0.187	0.183	0.180	0.179	0.178	0.178	0.178
21	0.271	0.245	0.212	0.195	0.191	0.183	0.179	0.176	0.175	0.174	0.174	0.174
22	0.268	0.241	0.209	0.190	0.187	0.179	0.175	0.172	0.171	0.170	0.170	0.170
23	0.265	0.238	0.206	0.186	0.183	0.175	0.171	0.169	0.167	0.167	0.166	0.166
24	0.262	0.235	0.202	0.183	0.180	0.172	0.168	0.165	0.164	0.163	0.163	0.163
25	0.259	0.232	0.199	0.179	0.176	0.169	0.165	0.162	0.161	0.160	0.160	0.160
30	0.248	0.221	0.187	0.165	0.162	0.155	0.151	0.149	0.147	0.147	0.147	0.147
35	0.239	0.212	0.177	0.153	0.150	0.144	0.140	0.138	0.137	0.136	0.136	0.136
40	0.232	0.204	0.168	0.144	0.141	0.135	0.132	0.130	0.128	0.128	0.128	0.128
45	0.226	0.198	0.161	0.136	0.133	0.127	0.124	0.122	0.121	0.121	0.121	0.121
50	0.221	0.192	0.156	0.130	0.127	0.121	0.118	0.116	0.115	0.115	0.115	0.115
60	0.213	0.184	0.147	0.119	0.116	0.111	0.108	0.107	0.106	0.105	0.105	0.105
70	0.206	0.177	0.139	0.111	0.107	0.103	0.100	0.099	0.098	0.098	0.097	0.097
80	0.201	0.171	0.134	0.104	0.101	0.096	0.094	0.093	0.092	0.092	0.091	0.091
90	0.196	0.166	0.129	0.099	0.095	0.091	0.089	0.088	0.087	0.086	0.086	0.086
100	0.193	0.162	0.125	0.094	0.090	0.086	0.084	0.083	0.082	0.082	0.082	0.082
200	0.172	0.141	0.102	0.069	0.064	0.062	0.060	0.059	0.059	0.058	0.058	0.058
300	0.163	0.132	0.093	0.058	0.053	0.050	0.049	0.048	0.048	0.048	0.048	0.048
400	0.157	0.126	0.087	0.051	0.046	0.044	0.043	0.042	0.042	0.042	0.041	0.041
500	0.154	0.123	0.083	0.047	0.041	0.039	0.038	0.038	0.037	0.037	0.037	0.037
1000	0.145	0.113	0.073	0.036	0.029	0.028	0.027	0.027	0.026	0.026	0.026	0.026

Table 5. Critical Values for A-D Test Statistic for Significance Level = 0.05

n\k	0.010	0.025	0.050	0.10	0.20	0.50	1.0	2.0	5.0	10.0	20.0	50.0
5	0.873	0.846	0.830	0.826	0.775	0.711	0.691	0.684	0.681	0.679	0.679	0.678
6	0.949	0.897	0.854	0.845	0.803	0.736	0.715	0.704	0.698	0.698	0.697	0.697
7	1.030	0.948	0.876	0.860	0.821	0.752	0.728	0.715	0.710	0.708	0.707	0.708
8	1.114	1.001	0.899	0.872	0.836	0.762	0.736	0.724	0.719	0.715	0.716	0.715
9	1.197	1.054	0.923	0.881	0.845	0.771	0.743	0.730	0.723	0.722	0.721	0.721
10	1.279	1.106	0.942	0.888	0.854	0.777	0.748	0.736	0.729	0.725	0.725	0.724
15	1.673	1.361	1.041	0.911	0.877	0.793	0.763	0.747	0.739	0.737	0.735	0.734
16	1.748	1.409	1.062	0.916	0.878	0.796	0.763	0.750	0.741	0.739	0.737	0.735
17	1.819	1.455	1.080	0.920	0.883	0.798	0.766	0.749	0.742	0.739	0.738	0.737
18	1.891	1.499	1.097	0.923	0.884	0.800	0.767	0.753	0.743	0.739	0.739	0.738
19	1.961	1.545	1.116	0.925	0.888	0.803	0.769	0.752	0.742	0.741	0.740	0.740
20	2.028	1.592	1.132	0.929	0.888	0.803	0.768	0.752	0.745	0.742	0.741	0.739
21	2.098	1.634	1.148	0.929	0.890	0.805	0.770	0.754	0.745	0.743	0.743	0.741
22	2.164	1.675	1.167	0.933	0.892	0.804	0.771	0.756	0.746	0.744	0.740	0.743
23	2.233	1.721	1.184	0.934	0.894	0.805	0.769	0.755	0.747	0.744	0.742	0.741
24	2.297	1.763	1.201	0.938	0.894	0.806	0.772	0.755	0.746	0.744	0.742	0.742
25	2.360	1.803	1.216	0.939	0.895	0.807	0.773	0.756	0.747	0.745	0.743	0.742
30	2.678	2.006	1.298	0.948	0.898	0.809	0.775	0.758	0.746	0.745	0.744	0.744
35	2.982	2.196	1.374	0.955	0.900	0.812	0.776	0.760	0.750	0.748	0.747	0.745
40	3.274	2.381	1.443	0.963	0.903	0.813	0.779	0.759	0.751	0.748	0.747	0.746
45	3.559	2.559	1.511	0.969	0.905	0.813	0.777	0.761	0.753	0.748	0.748	0.747
50	3.833	2.733	1.579	0.974	0.907	0.814	0.780	0.763	0.754	0.750	0.748	0.748
60	4.379	3.066	1.712	0.984	0.910	0.816	0.779	0.763	0.753	0.751	0.749	0.748
70	4.901	3.392	1.840	0.992	0.910	0.817	0.780	0.763	0.754	0.751	0.749	0.749
80	5.415	3.709	1.962	1.002	0.910	0.819	0.782	0.763	0.754	0.750	0.751	0.748
90	5.917	4.019	2.079	1.011	0.911	0.818	0.783	0.765	0.755	0.752	0.750	0.751
100	6.426	4.322	2.195	1.019	0.912	0.818	0.783	0.765	0.754	0.752	0.750	0.750
200	11.16	7.194	3.268	1.103	0.914	0.821	0.784	0.766	0.756	0.751	0.751	0.750
300	15.69	9.909	4.254	1.180	0.917	0.822	0.784	0.766	0.757	0.755	0.751	0.752
400	20.10	12.53	5.194	1.256	0.917	0.823	0.785	0.766	0.757	0.754	0.751	0.752
500	24.43	15.11	6.110	1.328	0.918	0.822	0.785	0.767	0.756	0.753	0.752	0.752
1000	45.58	27.58	10.47	1.671	0.919	0.824	0.785	0.768	0.757	0.753	0.752	0.750

Table 6. Critical Values for K-S Test Statistic for Significance Level = 0.05

n\k	0.010	0.025	0.050	0.10	0.20	0.50	1.0	2.0	5.0	10.0	20.0	50.0
5	0.419	0.409	0.401	0.398	0.388	0.372	0.364	0.360	0.358	0.358	0.357	0.357
6	0.397	0.384	0.373	0.369	0.364	0.349	0.341	0.336	0.333	0.332	0.332	0.332
7	0.385	0.369	0.353	0.348	0.342	0.327	0.320	0.315	0.313	0.312	0.311	0.311
8	0.372	0.354	0.336	0.329	0.323	0.309	0.301	0.297	0.295	0.294	0.294	0.293
9	0.362	0.342	0.321	0.312	0.307	0.294	0.287	0.282	0.280	0.279	0.279	0.279
10	0.352	0.331	0.308	0.298	0.294	0.281	0.274	0.270	0.267	0.267	0.266	0.266
15	0.318	0.293	0.264	0.249	0.245	0.234	0.228	0.224	0.222	0.222	0.221	0.221
16	0.313	0.288	0.258	0.242	0.238	0.227	0.221	0.218	0.216	0.215	0.215	0.214
17	0.308	0.283	0.252	0.235	0.231	0.221	0.215	0.212	0.210	0.209	0.209	0.208
18	0.303	0.278	0.246	0.229	0.225	0.215	0.209	0.206	0.204	0.203	0.203	0.203
19	0.299	0.273	0.242	0.224	0.220	0.210	0.204	0.201	0.199	0.199	0.198	0.198
20	0.295	0.269	0.237	0.218	0.214	0.205	0.199	0.196	0.194	0.194	0.193	0.193
21	0.291	0.265	0.232	0.213	0.210	0.200	0.195	0.192	0.190	0.189	0.189	0.189
22	0.288	0.261	0.228	0.209	0.206	0.196	0.191	0.188	0.186	0.185	0.185	0.185
23	0.285	0.258	0.225	0.205	0.201	0.192	0.187	0.184	0.182	0.182	0.181	0.181
24	0.281	0.255	0.221	0.201	0.197	0.188	0.183	0.180	0.178	0.178	0.178	0.177
25	0.279	0.252	0.218	0.197	0.193	0.184	0.180	0.177	0.175	0.175	0.174	0.174
30	0.266	0.239	0.204	0.181	0.177	0.169	0.165	0.162	0.160	0.160	0.160	0.160
35	0.256	0.228	0.193	0.168	0.165	0.157	0.153	0.151	0.149	0.149	0.148	0.148
40	0.248	0.220	0.183	0.158	0.154	0.148	0.144	0.141	0.140	0.139	0.139	0.139
45	0.241	0.212	0.176	0.150	0.146	0.139	0.136	0.133	0.132	0.132	0.132	0.131
50	0.235	0.206	0.170	0.143	0.139	0.132	0.129	0.127	0.126	0.125	0.125	0.125
60	0.226	0.196	0.159	0.131	0.127	0.121	0.118	0.116	0.115	0.115	0.114	0.114
70	0.218	0.189	0.151	0.122	0.118	0.113	0.110	0.108	0.107	0.106	0.106	0.106
80	0.212	0.182	0.145	0.114	0.111	0.105	0.103	0.101	0.100	0.100	0.099	0.099
90	0.207	0.177	0.139	0.108	0.104	0.100	0.097	0.095	0.094	0.094	0.094	0.094
100	0.203	0.173	0.135	0.103	0.099	0.095	0.092	0.091	0.090	0.089	0.089	0.089
200	0.179	0.149	0.110	0.075	0.070	0.067	0.065	0.064	0.064	0.064	0.064	0.063
300	0.169	0.138	0.099	0.063	0.058	0.055	0.054	0.053	0.052	0.052	0.052	0.052
400	0.163	0.132	0.092	0.056	0.050	0.048	0.047	0.046	0.045	0.045	0.045	0.045
500	0.159	0.127	0.087	0.051	0.045	0.043	0.042	0.041	0.041	0.040	0.040	0.040
1000	0.148	0.117	0.076	0.039	0.032	0.030	0.030	0.029	0.029	0.029	0.029	0.029

Table 7. Critical Values for A-D Test Statistic for Significance Level = 0.01

n\k	0.010	0.025	0.050	0.10	0.20	0.50	1.0	2.0	5.0	10.0	20.0	50.0
5	1.075	1.077	1.105	1.145	1.068	0.945	0.905	0.890	0.883	0.882	0.879	0.879
6	1.195	1.156	1.142	1.183	1.121	0.990	0.946	0.928	0.918	0.916	0.911	0.912
7	1.314	1.230	1.176	1.207	1.156	1.019	0.979	0.951	0.944	0.938	0.935	0.938
8	1.430	1.311	1.209	1.226	1.181	1.044	0.990	0.970	0.961	0.955	0.956	0.953
9	1.548	1.385	1.250	1.242	1.196	1.058	1.007	0.984	0.967	0.968	0.969	0.967
10	1.658	1.464	1.281	1.249	1.214	1.071	1.018	0.994	0.981	0.977	0.975	0.973
15	2.174	1.818	1.434	1.290	1.248	1.100	1.048	1.018	1.002	0.999	0.997	0.999
16	2.260	1.879	1.469	1.304	1.253	1.112	1.047	1.019	1.007	1.004	1.000	0.999
17	2.353	1.942	1.495	1.307	1.260	1.110	1.053	1.023	1.008	1.004	1.003	1.000
18	2.441	2.004	1.521	1.317	1.260	1.116	1.054	1.027	1.015	1.006	1.005	1.003
19	2.532	2.069	1.548	1.319	1.267	1.115	1.059	1.026	1.013	1.010	1.006	1.008
20	2.611	2.123	1.570	1.329	1.268	1.118	1.056	1.031	1.016	1.012	1.005	1.009
21	2.704	2.172	1.595	1.323	1.270	1.126	1.057	1.031	1.017	1.013	1.013	1.008
22	2.780	2.226	1.619	1.339	1.279	1.119	1.062	1.036	1.023	1.014	1.011	1.013
23	2.862	2.290	1.646	1.334	1.281	1.125	1.059	1.034	1.017	1.020	1.012	1.013
24	2.934	2.340	1.667	1.341	1.277	1.126	1.065	1.035	1.020	1.015	1.012	1.013
25	3.019	2.383	1.690	1.342	1.281	1.127	1.064	1.038	1.021	1.017	1.014	1.013
30	3.393	2.634	1.800	1.365	1.286	1.133	1.072	1.044	1.023	1.023	1.019	1.018
35	3.744	2.865	1.904	1.371	1.286	1.136	1.072	1.045	1.027	1.025	1.021	1.018
40	4.085	3.088	1.988	1.382	1.294	1.138	1.076	1.046	1.030	1.027	1.023	1.022
45	4.408	3.299	2.077	1.388	1.298	1.141	1.074	1.048	1.036	1.030	1.026	1.024
50	4.734	3.500	2.162	1.407	1.304	1.142	1.079	1.053	1.034	1.029	1.028	1.025
60	5.334	3.889	2.330	1.419	1.308	1.144	1.079	1.054	1.032	1.032	1.029	1.030
70	5.915	4.258	2.483	1.430	1.307	1.145	1.079	1.055	1.038	1.031	1.031	1.028
80	6.503	4.620	2.622	1.445	1.302	1.150	1.085	1.055	1.036	1.033	1.032	1.029
90	7.050	4.960	2.759	1.458	1.312	1.149	1.086	1.056	1.038	1.034	1.031	1.033
100	7.609	5.302	2.895	1.471	1.308	1.149	1.085	1.054	1.042	1.035	1.033	1.032
200	12.74	8.464	4.130	1.584	1.310	1.156	1.089	1.059	1.041	1.031	1.032	1.033
300	17.54	11.39	5.224	1.697	1.314	1.154	1.090	1.058	1.043	1.038	1.033	1.031
400	22.18	14.18	6.252	1.793	1.321	1.158	1.093	1.057	1.043	1.039	1.035	1.034
500	26.74	16.91	7.253	1.885	1.319	1.155	1.089	1.057	1.047	1.040	1.034	1.034
1000	48.73	30.00	11.94	2.296	1.325	1.157	1.092	1.060	1.043	1.035	1.036	1.031

Table 8. Critical Values for K-S Test Statistic for Significance Level = 0.01

n\k	0.010	0.025	0.050	0.10	0.20	0.50	1.0	2.0	5.0	10.0	20.0	50.0
5	0.464	0.458	0.454	0.456	0.451	0.431	0.421	0.414	0.410	0.410	0.408	0.408
6	0.453	0.441	0.431	0.431	0.423	0.402	0.391	0.385	0.382	0.381	0.380	0.380
7	0.437	0.421	0.406	0.404	0.399	0.380	0.369	0.362	0.360	0.358	0.357	0.357
8	0.423	0.407	0.388	0.384	0.379	0.360	0.349	0.344	0.340	0.339	0.339	0.338
9	0.412	0.393	0.373	0.365	0.360	0.343	0.333	0.327	0.323	0.323	0.322	0.322
10	0.402	0.382	0.358	0.350	0.345	0.328	0.318	0.312	0.309	0.308	0.308	0.307
15	0.362	0.338	0.308	0.292	0.288	0.274	0.266	0.261	0.258	0.257	0.257	0.256
16	0.356	0.331	0.301	0.284	0.280	0.266	0.258	0.253	0.251	0.250	0.249	0.249
17	0.350	0.325	0.294	0.277	0.272	0.259	0.251	0.246	0.244	0.243	0.242	0.242
18	0.345	0.319	0.288	0.270	0.265	0.252	0.245	0.240	0.237	0.236	0.236	0.236
19	0.340	0.314	0.282	0.263	0.259	0.246	0.238	0.234	0.232	0.231	0.230	0.230
20	0.335	0.309	0.276	0.257	0.253	0.240	0.233	0.228	0.226	0.225	0.225	0.225
21	0.331	0.304	0.271	0.251	0.247	0.235	0.228	0.223	0.221	0.220	0.220	0.219
22	0.327	0.300	0.267	0.246	0.242	0.230	0.223	0.219	0.216	0.216	0.215	0.215
23	0.323	0.296	0.262	0.241	0.237	0.225	0.218	0.215	0.212	0.211	0.211	0.210
24	0.318	0.292	0.258	0.236	0.232	0.221	0.214	0.210	0.208	0.207	0.207	0.206
25	0.315	0.288	0.254	0.232	0.228	0.216	0.210	0.206	0.204	0.203	0.203	0.203
30	0.300	0.272	0.237	0.213	0.209	0.199	0.193	0.189	0.187	0.186	0.186	0.185
35	0.288	0.260	0.224	0.198	0.194	0.185	0.179	0.176	0.174	0.173	0.173	0.172
40	0.278	0.249	0.213	0.187	0.182	0.173	0.168	0.165	0.163	0.162	0.162	0.162
45	0.270	0.241	0.204	0.176	0.172	0.164	0.158	0.156	0.154	0.154	0.153	0.153
50	0.263	0.233	0.196	0.168	0.164	0.156	0.151	0.148	0.146	0.146	0.146	0.145
60	0.251	0.221	0.184	0.154	0.150	0.143	0.138	0.136	0.134	0.134	0.133	0.133
70	0.242	0.212	0.174	0.144	0.139	0.132	0.128	0.126	0.124	0.124	0.124	0.124
80	0.234	0.204	0.166	0.135	0.130	0.124	0.120	0.118	0.117	0.116	0.116	0.116
90	0.228	0.198	0.159	0.128	0.123	0.117	0.114	0.111	0.110	0.110	0.109	0.110
100	0.223	0.192	0.154	0.122	0.117	0.111	0.108	0.106	0.105	0.104	0.104	0.104
200	0.194	0.163	0.124	0.089	0.083	0.079	0.077	0.075	0.074	0.074	0.074	0.074
300	0.181	0.150	0.110	0.074	0.068	0.065	0.063	0.062	0.061	0.061	0.061	0.060
400	0.173	0.142	0.102	0.066	0.059	0.056	0.054	0.053	0.053	0.053	0.053	0.053
500	0.168	0.137	0.096	0.060	0.053	0.050	0.049	0.048	0.047	0.047	0.047	0.047
1000	0.155	0.123	0.083	0.045	0.037	0.036	0.035	0.034	0.034	0.033	0.033	0.033

References

Aitchison, J. and Brown, J.A.C. 1969. *The Lognormal Distribution*, Cambridge: Cambridge University Press.

Anderson, T.W. and Darling, D. A. (1954). *Test of goodness-of-fit*. Journal of American Statistical Association, Vol. 49, 765-769.

Bain, L.J., and M. Engelhardt. 1991. *Statistical Analysis of Reliability and Life Testing Models*, Theory and Methods. 2nd Edition. Dekker, New York.

Barber, S. and Jennison, C. 1999. *Symmetric Tests and Confidence Intervals for Survival Probabilities and Quantiles of Censored Survival Data*. University of Bath, Bath, BA2 7AY, UK.

Barnett, V. 1976. *Convenient Probability Plotting Positions for the Normal Distribution*. Appl. Statist., 25, No. 1, pp. 47-50, 1976.

Barnett, V. and Lewis T. 1994. *Outliers in Statistical Data*. Third edition. John Wiley & Sons Ltd. UK.

Bechtel Jacobs Company, LLC. 2000. *Improved Methods for Calculating Concentrations used in Exposure Assessment*. Prepared for DOE. Report # BJC/OR-416.

Best, D.J. and Roberts, D.E. 1975. *The Percentage Points of the Chi-square Distribution*. Applied Statistics, 24: 385-388.

Bowman, K. O. and Shenton, L.R. 1988. *Properties of Estimators for the Gamma Distribution*, Volume 89. Marcel Dekker, Inc., New York.

Box, G.E.P. and Tiao, G.C. 1973. *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA.

Bradu, D. and Mundlak, Y. 1970. *Estimation in Lognormal Linear Models*. Journal of the American Statistical Association, 65, 198-211.

California's Ocean Plan. 2005. *Amendment of the Water Quality Control Plan for Ocean Waters Of California*. EPA, California. State Water Resources Control Board, Sacramento, California. Available at http://www.swrcb.ca.gov/plnspols/oplans/docs/draft_ffed.pdf

Chen, L. 1995. *Testing the Mean of Skewed Distributions*. Journal of the American Statistical Association, 90, 767-772.

Choi, S. C. and Wette, R. 1969. *Maximum Likelihood Estimation of the Parameters of the Gamma Distribution and Their Bias*. Technometrics, Vol. 11, 683-690.

Cochran, W. 1977. *Sampling Techniques*, New York: John Wiley.

Cohen, A. C. Jr. 1950. *Estimating the Mean and Variance of Normal Populations from Singly Truncated and Double Truncated Samples*. Ann. Math. Statist., Vol. 21, pp. 557-569.

Cohen, A. C. Jr. 1959. *Simplified Estimators for the Normal Distribution When Samples Are Singly Censored or Truncated*. Technometrics, Vol. 1, No. 3, pp. 217-237.

Cohen, A. C. Jr. 1991. *Truncated and Censored Samples*. 119, Marcel Dekker Inc. New York, NY 1991.

Colorado Water Quality Control Division (WQCD). 2003. *Determination of the Requirement to Include Water Quality Standards-Based Limits in CDPS Permits Based on Reasonable Potential. Procedural Guidance*. Colorado WQCD - Permits Unit, Denver, Colorado. Available at http://www.cdphe.state.co.us/wq/PermitsUnit/rpguide.pdf

Conover, W. J. 1980. *Practical Nonparametric Statistics*. Second Edition. John Wiley.

Conover W.J., 1999. Practical Nonparametric Statistics, 3rd Edition, John Wiley & Sons, New York.

Cressie, N. 1991. *Statistics for Spatial Data*, New York: John Wiley & Sons.

D'Agostino, R.B. and Stephens, M.A. 1986. *Goodness-of-Fit Techniques*. Marcel Dekker, Inc.

Daniel, Wayne W. 1995. *Biostatistics*. 6th Edition. John Wiley & Sons, New York.

DataQUEST: *Data Quality Evaluation Toolbox* 1997, EPA QA/G-9D, Publication EPA-600-R-96-085.

David, H.A. and Nagaraja, H.N. 2003. *Order Statistics*. Third Edition. John Wiley.

Davison, A.C. and Hinkley, D.V. 1997. *Bootstrap Methods and their Application*, Cambridge University Press.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1997. *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, Ser. B, 39, pp. 1-38.

Department of Navy (Navy). 1998. Procedural Guidance for Statistically Analyzing Environmental Background Data, Naval Facilities Engineering Command.

Department of Navy (Navy). 2003. Guidance for Environmental Background Analysis, Naval Facilities Engineering Command, Volume II: Sediment, UG-2054-ENV.

Department of Navy. 1999. *Handbook for Statistical Analysis of Environmental Background Data*. Prepared by SWDIV and EFA West of Naval Facilities Engineering Command. July 1999.

Department of Navy. 2002a. *Guidance for Environmental Background Analysis*. Volume 1 Soil. Naval Facilities Engineering Command. April 2002.

Department of Navy. 2002b. *Guidance for Environmental Background Analysis*. Volume 2 Sediment. Naval Facilities Engineering Command. May 2002.

Dixon, W.J. 1953. *Processing Data for Outliers*. *Biometrics 9: 74-89*.

Dixon, W.J. and J.W. Tukey. 1968. *Approximate Behavior of the Distribution of Winsorized-T (Trimming/ Winsorization)*. Technometrics 10: 83-98.

Dudewicz, E.D. and Misra, S.N. 1988. *Modern Mathematical Statistics*. John Wiley, New York.

Efron, B. 1981. *Censored Data and Bootstrap*. Journal of American Statistical Association, Vol. 76, pp. 312-319.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: SIAM.

Efron, B. and Tibshirani, R.J. 1993. *An Introduction to the Bootstrap*. Chapman & Hall, New York.

El-Shaarawi, A.H. 1989. *Inferences about the Mean from Censored Water Quality Data*. Water Resources Research, 25, pp. 685-690.

ExpertFit Software 2001. Averill M. Law & Associates Inc, Tucson, Arizona.

Faires, J. D., and Burden, R. L. 1993. *Numerical Methods*. PWS-Kent Publishing Company, Boston, USA.

Gibbons, R.D. and Coleman, D.E. 2001. *Statistical Methods for Detection and Quantification of Environmental Contamination*. John Wiley, NY.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring.* Van Nostrand Reinhold, New York.

Gilliom, R. H. and Helsel, D.R. 1986. *Estimations of Distributional Parameters for Censored Trace Level Water Quality Data 1. Estimation Techniques.* Water Resources Research, 22, pp. 135-146.

Gleit, A. 1985. *Estimation for Small Normal Data Sets with Detection Limits.* Environmental Science and Technology*, 19,* pp. 1206-1213, 1985.

Gogolak, C.V., G.E. Powers, and A.M. Huffert. 1998. A Nonparametric Statistical Methodology for the Design and Analysis of Final Status Decommissioning Surveys, Office of Nuclear Regulatory Research, U.S. Nuclear Regulatory Commission (NRC), NUREG-1505. June 1998 (Rev. 1).

Golden, N.H., B.A. Rattner, J.B. Cohen, D.J. Hoffman, E. Russek-Cohen, and M. A. Ottinger. 2003. *Lead Accumulation in Feathers of Nestling Black-Crowned Night Herons (Nycticorax Nycticorax) Experimentally Treated.* In the field. Environmental Toxicology and Chemistry, Vol. 22. pp. 1517-1525.

Grice, J.V., and Bain, L. J. 1980. *Inferences Concerning the Mean of the Gamma Distribution*. Journal of the American Statistical Association. Vol. 75, Number 372, 929-933.

Haas, C.H., and Scheff, P.A., 1990. *Estimation of Averages in Truncated Samples*. Environmental Science and Technology, *24*, pp. 912-919.

Hahn, J. G. and Meeker, W.Q. 1991. *Statistical Intervals. A Guide for Practitioners*. John Wiley.

Hall, P. 1988. *Theoretical comparison of bootstrap confidence intervals.* Annals of Statistics, 16, 927-953.

Hall, P. 1992. *On the Removal of Skewness by Transformation*. Journal of Royal Statistical Society, B 54, 221-228.

Hampel, F. R. 1974. *The Influence Curve and Its Role in Robust Estimation*. Journal of American Statistical Association, *69*, pp. 383-393, 1974.

Hardin, J.W. and Gilbert, R.O. 1993. *Comparing Statistical Tests for Detecting Soil Contamination Greater Than Background*. Pacific Northwest Laboratory, Battelle, Technical Report # DE 94-005498.

Helsel, D.R. 1990. *Less Than Obvious, Statistical Treatment of Data Below the Detection Limit*. ES&T Features Environmental Sci. Technol., Vol. 24, No. 12, pp. 1767-1774.

Helsel, D.R. and Hirsch, R.M. 1994. *Statistical Methods in Water Resources*. John Wiley.

Helsel, D.R. 2005. *Nondetects and Data Analysis*. Statistics for Censored Environmental Data. John Wiley and Sons, NY.

Hinton, S.W. 1993. *Δ Log-Normal Statistical Methodology Performance.* ES&T Environmental Sci. Technol., Vol. 27, No. 10, pp. 2247-2249.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. 1983. *Understanding Robust and Exploratory Data Analysis*. John Wiley, New York.

Hogg, R.V., and Craig, A.T. 1978. *Introduction to Mathematical Statistics,* New York: Macmillan Publishing Company.

Hogg, R.V. and Craig, A. 1995. *Introduction to Mathematical Statistics*; 5th edition. Macmillan.

Huber, P.J. 1981. *Robust Statistics*. John Wiley, New York.

Johnson, N.J. 1978. *Modified t-Tests and Confidence Intervals for Asymmetrical Populations*. The American Statistician, Vol. 73, 536-544.

Johnson, N.L., Kotz, S., and Balakrishnan, N. 1994. *Continuous Univariate Distributions, Vol. 1*. Second Edition. John Wiley, New York.

Johnson, R.A., and Wichern, D.W. 1988. *Applied Multivariate Statistical analysis*. Prentice Hall.

Kaplan, E.L. and Meier, O. 1958. *Nonparametric Estimation from Incomplete Observations*. Journal of the American Statistical Association, Vol. 53. 457-481.

Kleijnen, J.P.C., Kloppenburg, G.L.J., and Meeuwsen, F.L. 1986. *Testing the Mean of an Asymmetric Population: Johnson's Modified t Test Revisited*. Commun. in Statist.-Simula., 15(3), 715-731.

Kroll, C.N. and J.R. Stedinger. 1996. *Estimation of Moments and Quantiles Using Censored Data*. Water Resources, Vol. 32. pp. 1005-1012.

Land, C. E. 1971. *Confidence Intervals for Linear Functions of the Normal Mean and Variance*. Annals of Mathematical Statistics, 42, pp. 1187-1205.

Land, C. E. 1975. *Tables of Confidence Limits for Linear Functions of the Normal Mean and Variance*. In Selected Tables in Mathematical Statistics, Vol. III, American Mathematical Society, Providence, R.I., pp. 385-419.

Law, A.M. and Kelton, W.D. 2000. *Simulation Modeling and Analysis*. Third Edition. McGraw Hill.

Lehmann, E.L. and H.J.M. D'Abrera. 1998. Nonparametrics: Statistical Methods Based on Ranks, revised 1st Ed., Prentice Hall, New Jersey.

Manly, B.F.J. 1997. *Randomization, Bootstrap, and Monte Carlo Methods in Biology*. Second Edition. Chapman Hall, London.

Manly, B.F.J . 2001. *Statistics for Environmental Science and Management*. Chapman & Hall/CRC.

Millard, S.P. 2002. *EnvironmentalStats for S-PLUS*. Second Edition. Springer.

Natrella, M.G. 1963. *Experimental Statistics*. National Bureau of Standards, Hand Book No. 91, U.S. Government Printing Office, Washington, DC.
New York.

Newman, M. C., Dixon, P.M., and Pinder, J.E. 1990. *Estimating Mean and Variance for Environmental Samples with Below Detection Limit Observations*. Water Resources Bulletin, Vol. 25, No. 4, pp. 905-916.

Palachek, A.D., D.R. Weier, T.R. Gatliffe, D.M. Splett, and D.K. Sullivan. 1993. Statistical Methodology for Determining Contaminants of Concern by Comparison of Background and Site Data with Applications to Operable Unit 2, SA-93-010, Internal Report, Statistical Applications, EG&G Rocky Flats Inc., Rocky Flats Plant, Golden, CO.

Perrson, T., and Rootzen, H. 1977. *Simple and Highly Efficient Estimators for A Type I Censored Normal Sample*. Biometrika, 64, pp. 123-128.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. 1990. *Numerical Recipes in C, The Art of Scientific Computing*. Cambridge University Press. Cambridge, MA.

ProUCL 3.0. 2004. *A Statistical Software*. National Exposure Research Lab, EPA, Las Vegas Nevada, October 2004.

Rousseeuw, P.J. and Leroy, A.M. 1987. *Robust Regression and Outlier Detection*. John Wiley.

RPcalc 2.0 2005. *Reasonable Potential Analysis Calculator*, EPA, California. State Water Resources Control Board, Sacramento, California. Available at http://www.swrcb.ca.gov/plnspols/oplans/.

Saw, J. G. 1961. *The Bias for the Maximum Likelihood Estimates of Location and Scale Parameters Given A Type II Censored Normal Sample*. Biometrika, 48, pp. 448-451.

Schneider, B. E. 1978. *Kolmogorov-Smirnov Test Statistic for the Gamma Distribution with Unknown Parameters,* Dissertation, Department of Statistics, Temple University, Philadelphia, PA.

Schneider, B.E. and Clickner, R.P. 1976. *On the Distribution of the Kolmogorov-Smirnov Statistic for the Gamma Distribution with Unknown Parameters*. Mimeo Series Number 36, Department of Statistics, School of Business Administration, Temple University, Philadelphia, PA.

Schneider, H. 1986. *Truncated and Censored Samples from Normal Populations.* Vol. 70, Marcel Dekker Inc., New York, 1986.

Schulz, T. W. and Griffin, S. 1999. *Estimating Risk Assessment Exposure Point Concentrations when Data are Not Normal or Lognormal*. Risk Analysis, Vol. 19, No. 4.

Scout: *A Data Analysis Program,* Technology Support Project. EPA, NERL-LV, Las Vegas, NV 89193-3478.

She, N. 1997. *Analyzing Censored Water Quality Data Using a Non-Parametric Approach.* Journal of the American Water Resources Association 33, pp. 615-624.

Shumway, A.H., Azari, A.S., Johnson, P. 1989. *Estimating Mean Concentrations Under Transformation for Environmental Data with Detection Limits*. Technometrics, Vol. 31, No. 3, pp. 347-356.

Shumway, R.H., R.S. Azari, and M. Kayhanian. 2002. *Statistical Approaches to Estimating Mean Water Quality Concentrations with Detection Limits*. Environmental Science and Technology, Vol. 36. pp. 3345-3353.

SimCensor 2005: *A Program Developed to Perform Simulation Studies as Summarized in This Report.*

Singh, A. 1993. *Omnibus Robust Procedures for Assessment of Multivariate Normality and Detection of Multivariate Outliers*. In Multivariate Environmental Statistics, Patil G.P. and Rao, C.R., Editors, pp. 445-488. Elsevier Science Publishers.

Singh, A., Singh, A.K., and G. Flatman 1994. *Estimation of Background Levels of Contaminants.* Math Geology, Vol. 26, No, 3, 361-388.

Singh, A. and Nocerino, J.M. 1995. *Robust Procedures for the Identification of Multiple Outliers*. Handbook of Environmental Chemistry, Statistical Methods, Vol. 2.G, pp. 229-277. Springer Verlag, Germany.

Singh, A.K., Singh, A., and Engelhardt, M. 1997. *The Lognormal Distribution in Environmental Applications.* Technology Support Center Issue Paper, 182CMB97. EPA/600/R-97/006, December 1997.

Singh, A. K., Singh, A., and Engelhardt, M. 1999. *Some Practical Aspects of Sample Size and Power Computations for Estimating the Mean of Positively Skewed Distributions in Environmental Applications.* EPA/600/S-99/006, November 1999.

Singh, A. and Nocerino, J.M. 2002. *Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations.* Vol. 60, pp. 69-86.

Singh, A, Singh, A.K., and Iaci, R.J. 2002. *Estimation of the Exposure Point Concentration Term Using a Gamma Distribution*. EPA/600/R-02/084. October 2002.

Singh, A., Singh, A. K., and Iaci, R. J. 2002b. *Estimation of the Exposure Point Concentration Term Using a Gamma Distribution*. EPA/600/R-02/084.

Singh. A., Singh, A. K., Engelhardt, M., and Nocerino, J.M. 2002a. *On the Computation of the Upper Confidence Limit of the Mean of Contaminant Data Distributions*. Under EPA Review.

Singh, A. and Singh, A.K. 2003. *Estimation of the Exposure Point Concentration Term (95% UCL) Using Bias-Corrected Accelerated (BCA) Bootstrap Method and Several Other methods for Normal, Lognormal, and Gamma Distributions*. Draft EPA Internal Report.

Singh, A. 2004. *Computation of an Upper Confidence Limit (UCL) of the Unknown Population Mean Using ProUCL Version 3.0*. Part I. Download from: www.epa.gov/nerlesd1/tsc/issue.htm

Singh, A., Maichle, R.W. and S. Lee 2006. On the Computation of a 95% Upper Confidence Limit of the Unknown Population Mean Based Upon Data Sets with Below Detection Limit Observations. EPA 2006, EPA/600/R-06/022.

Staudte, R.G. and Sheather, S.J. 1990. *Robust Estimation and Testing*. John Wiley, 1990.

Stephens, M. A. 1970. *Use of Kolmogorov-Smirnov, Cramer-von Mises and Related Statistics Without Extensive Tables*. Journal of Royal Statistical Society, B 32, 115-122.

Sutton, C.D. 1993. *Computer-Intensive Methods for Tests About the Mean of an Asymmetrical Distribution*. Journal of American Statistical Society, Vol. 88, No. 423, 802-810.

Thom, H.C. S. 1968. *Direct and Inverse Tables of the Gamma Distribution*. Silver Spring, MD; Environmental Data Service.

Tiku, M.L. 1967. *Estimating Mean and Standard Deviation from A Censored Normal Sample*. Biometrika 54. pp. 155-165.

Tiku, M.L. 1971. *Estimating the Mean and Standard Deviation from Two Censored Normal Samples*. Biometrika 58. pp. 241-243.

U.S. Environmental Protection Agency (EPA). 1989. *Methods for Evaluating the Attainment of Cleanup Standards, Vol. 1, Soils and Solid Media*. Publication EPA 230/2-89/042.

U.S. Environmental Protection Agency (EPA). 1989. *Statistical Analysis of Ground-water Monitoring Data at RCRA Facilities*. Interim Final Guidance. Washington, DC: Office of Solid Waste. April 1989.

U.S. Environmental Protection Agency (EPA). 1989a. Risk Assessment Guidance for Superfund Vol. I, Human Health Evaluation Manual (Part A). Office of Emergency and Remedial Response, Washington, DC. EPA 540-1-89-002. Hereafter referred to as "RAGS."

U.S. Environmental Protection Agency (EPA). 1989b. Statistical Methods for Evaluating the Attainment of Cleanup Standards, EPA 230/02-89-042, Washington, DC.

U.S. Environmental Protection Agency (EPA). 1990. *Guidance for Data Usability in Risk Assessment: Interim Final*, October 1990. EPA 540-G-90-008, PB91-921208, Washington, DC.

U.S. Environmental Protection Agency (EPA). 1991. *Technical Support Document for Water Quality Based Toxics Control*. Office of Water Enforcement and Permits, Washington DC. March 1991.

U.S. Environmental Protection Agency (EPA). 1991. *A Guide: Methods for Evaluating the Attainment of Cleanup Standards for Soils and Solid Media*. Publication EPA/540/R95/128.

U.S. Environmental Protection Agency (EPA). 1992. *Supplemental Guidance to RAGS: Calculating the Concentration Term*. Publication EPA 9285.7-081, May 1992.

U.S. Environmental Protection Agency (EPA). 1992. *Statistical Analysis of Ground-water Monitoring Data at RCRA Facilities*. Addendum to Interim Final Guidance. Washington DC: Office of Solid Waste. July 1992.

U.S. Environmental Protection Agency (EPA). 1993. SW-846, Test Methods for Evaluating Solid Waste, Physical/Chemical Methods. Third Edition. Available at SW-846 on-line.

U.S. Environmental Protection Agency (EPA). 1994. Statistical Methods for Evaluating the Attainment of Cleanup Standards, EPA 230-R-94-004, Washington, DC.

U.S. Environmental Protection Agency (EPA). 1994a. Guidance for the Data Quality Objectives Process, EPA QA/G-4, EPA 600-R-96-065. Washington DC.

U.S. Environmental Protection Agency (EPA). 1994b. The Data Quality Objectives Decision Error Feasibility Trials (DEFT) Software (EPA QA/G-4D), EPA/600/R-96/056, Office of Research and Development, Washington, DC.

U.S. Environmental Protection Agency (EPA). 1994c. *Revised Interim Soil Lead Guidance for CERCLA Sites and RCRA Corrective Action Facilities*. OSWER Directive 9355.4-12.

U.S. Environmental Protection Agency (EPA). 1996. *A Guide: Soil Screening Guidance: Technical Background Document*. Second Edition, Publication 9355.4-04FS.

U.S. Environmental Protection Agency (EPA). 2000a. *Guidance for Choosing a Sampling Design for Environmental Data Collection*, EPA QA/G5S, U.S.EPA, Office of Environmental Information, Peer Review Draft, Aug. 2000.

U.S. Environmental Protection Agency (EPA). 2000b. *TRW Recommendations for Sampling and Analysis of Soil at Lead (Pb) Sites*. Office of Emergency and Remedial Response, Washington, DC. EPA 540-F-00-010, OSWER 9285.7-38.

U.S. Environmental Protection Agency (EPA), U.S. Nuclear Regulatory Commission, *et al.* 2000c. *Multi-Agency Radiation Survey and Site Investigation Manual (MARSSIM). Revision 1. EPA 402-R-97-016*. Available at http://www.epa.gov/radiation/marssim/ or from http://bookstore.gpo.gov/index.html (GPO Stock Number for Revision 1 is 052-020-00814-1).

U.S. Environmental Protection Agency (EPA). 2001. Requirements for Quality Assurance Project Plans, EPA QA/R-5. http://www.epa.gov/quality/qapps.html

U.S. Environmental Protection Agency (EPA). 2002a. *Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites*. OSWER 9285.6-10. December 2002.

U.S. Environmental Protection Agency (EPA). 2002b. *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites*. EPA 540-R-01-003-OSWER 9285.7-41. September, 2002.

U.S. Environmental Protection Agency (EPA). 2004. *ProUCL Version 3, A Statistical Software*. National Exposure Research Lab, EPA, Las Vegas Nevada, October 2004. The software ProUCL 3.0 can be freely downloaded from the U.S. EPA Web site at: http://www.epa.gov/nerlesd1/tsc/tsc.htm

U.S. Environmental Protection Agency (EPA). 2004. *ProUCL Version 3.1, A Statistical Software*, National Exposure Research Lab, EPA, Las Vegas Nevada, October 2004.

U.S. Environmental Protection Agency (EPA). 2004. *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities. Unified Guidance Document (UGD)*. Volumes I, II, and III. Office of Solid Waste. September, 2004.

U.S. Environmental Protection Agency (EPA). 2006. *Data Quality Assessment: Statistical Methods for Practitioners*, EPA QA/G-9S. EPA/240/B-06/003. Office of Environmental Information, Washington, DC. Download from: http://www.epa.gov/quality/qs-docs/g9s-final.pdf

UNCENSOR 5.1.2003. *A Statistical Program for Left-censored Data Sets*. University of Georgia. Savannah River Ecology Laboratory.

VSP: Visual Sample Plan. 2005 PNNL-15247, Pacific Northwest National Laboratory, Richland, Washington.

Whittaker, J. 1974. *Generating Gamma and Beta Random Variables with Non-integral Shape Parameters*. Applied Statistics, 23, No. 2, 210-214.

Wong, A. 1993. *A Note on Inference for the Mean Parameter of the Gamma Distribution*. Statistics Probability Letters, Vol. 17, 61-66.

United States
Environmental Protection
Agency

Office of Research
and Development (8101R)
Washington, DC 20460

Official Business
Penalty for Private Use
$300

EPA/600/R-07/041
April 2007
www.epa.gov

Please make all necessary changes on the below label, detach or copy, and return to the address in the upper left-hand corner.

If you do not wish to receive these reports CHECK HERE ☐; detach, or copy this cover, and return to the address in the upper left-hand corner.

PRESORTED STANDARD
POSTAGE & FEES PAID
EPA
PERMIT No. G-35