United States Environmental Protection Agency

ProUCL Version 5.2.0 Technical Guide

Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations

Publication #
Release Date
www.epa.gov

ProUCL Version 5.2.0 Technical Guide

Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations

Prepared for:

Felicia Barnett, Director
Office of Research and Development (ORD)
Center for Environmental Solutions and Emergency Response (CESER)
Technical Support Coordination Division (TSCD)
Site Characterization and Monitoring Technical Support Center (SCMTSC)
U.S. Environmental Protection Agency
61 Forsyth Street, Atlanta, GA 30303

Version 5.2.0 prepared by:
Neptune and Company, Inc.
1435 Garrison Street, Suite 201
Lakewood, CO 80215

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

NOTICE

The United States Environmental Protection Agency (U.S. EPA) through its Office of Research and Development (ORD) funded and managed the research described in ProUCL Technical Guide and methods incorporated in the ProUCL software. It has been peer reviewed by the U.S. EPA and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation by the U.S. EPA for use.

- Versions of the ProUCL software up to version ProUCL 5.1 have been developed by Lockheed Martin, IS&GS - CIVIL under the Science, Engineering, Response and Analytical contract with the U.S. EPA. Improvements included in version 5.2 were made by Neptune and Company, Inc. under the ProUCL and Statistical Support for Site Characterization and Monitor Technical Support Center (SCMTSC) contract with the U.S. EPA and is made available through the U.S. EPA Technical Support Center (TSC) in Atlanta, Georgia (GA).
- 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.

With respect to ProUCL distributed software and documentation, neither the U.S. 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.

ProUCL software is a statistical software package providing statistical methods described in various U.S. EPA guidance documents. ProUCL does not describe U.S. EPA policies and should not be considered to represent U.S. EPA policies.

Software Requirements

ProUCL 5.2 has been developed in the Microsoft .NET Framework 4.7.2 using the C# programming language and will run on Windows 10 that has this framework pre-installed. The downloadable .NET Framework 4.7.2 files can also be obtained from the following websites:

https://dotnet.microsoft.com/download/dotnet-framework/net472

Installation Instructions when Downloading ProUCL 5.2 from the EPA Web Site

- Download the file SETUP.EXE from the EPA Web site and save to a temporary location. Note: You can delete this file when the installation is complete.
- Right click on PROUCL5_2.zip and select "Extract All...". A dialog window will open that allows you to browse to the location you want ProUCL to be installed. Once you choose a location, click "Extract". This will create a directory named ProUCL 5.2 and add all the necessary files; create two subdirectories, adding sample data to one and documentation to the other.
- To run the program, use Windows Explorer to locate the ProUCL application file, and Double click on it, or use the RUN command from the start menu to locate the ProUCL.exe file, and run ProUCL.exe.
- To uninstall the program, use Windows Explorer to locate and delete the ProUCL folder.

Caution: If you have previous versions of the ProUCL, which were installed on your computer, you should remove or rename the directory in which earlier ProUCL versions are currently located.

Creating a Shortcut for ProUCL 5.2 on Desktop or Pin to Taskbar

- To create a shortcut of the ProUCL program on your desktop, go to your ProUCL directory and right click on the executable program and send it to desktop. A ProUCL icon will be displayed on your desktop. This shortcut will point to the ProUCL directory consisting of all files required to execute ProUCL 5.2.
- To pin ProUCL to Taskbar, open ProUCL and then right click ProUCL icon displayed on Taskbar and click Pin to Taskbar option.

Caution: Because all files in your ProUCL directory are needed to execute the ProUCL software, you need to generate a shortcut using the process described above. Simply dragging the ProUCL executable file from Window Explorer onto your desktop will not work successfully (an error message will appear) as all files needed to run the software are not available on your desktop. Your shortcut should point to the directory path with all required ProUCL files.

ProUCL 5.2

Software ProUCL version 5.2.0 (ProUCL 5.2), its earlier versions: ProUCL version 3.00.01, 4.00.02, 4.00.04, 4.00.05, 4.1.00, 4.1.01, and ProUCL 5.0.00, 5.1.002 and associated Facts Sheet, User Guides and Technical Guides (e.g., EPA 2010b, 2010c, 2013a, 2013b) can be downloaded from the following EPA website:

https://www.epa.gov/land-research/proucl-software

Recordings of ProUCL webinars offered in 2020 can be downloaded from:

ProUCL Utilization 2020: Part 1: ProUCL A to Z
https://clu-in.org/conf/tio/ProUCLAtoZ1/

ProUCL Utilization 2020: Part 2: Trend Analysis
https://clu-in.org/conf/tio/ProUCLAtoZ2/

ProUCL Utilization 2020: Part 3: Background Level Calculations
https://clu-in.org/conf/tio/ProUCLAtoZ3/

Relevant literature used in the development of various ProUCL versions can be downloaded from:
https://www.epa.gov/land-research/proucl-software

Contact Information for all Versions of ProUCL

Since 1999, the ProUCL software has been developed under the direction of the Technical Support Center (TSC). As of November 2007, the direction of the TSC is transferred from Brian Schumacher to Felicia Barnett. Therefore, any comments or questions concerning all versions of ProUCL software should be addressed to:

Felicia Barnett, Director
ORD Site Characterization and Monitoring Technical Support Center (SCMTSC)
Superfund and Technology Liaison, Region 4
U.S. Environmental Protection Agency
61 Forsyth Street SW, Atlanta, GA 30303-8960

barnett.felicia@epa.gov
(404)562-8659
Fax: (404) 562-8439

EXECUTIVE SUMMARY

ProUCL is software package for commonly used environmental statistics. It was initially developed as a research tool for U.S. EPA scientists and researchers of the Technical Support Center (TSC) and ORD-National Exposure Research Laboratory (NERL), Las Vegas. The intent was to provide a tool for basic statistical calculations that are applicable to site characterization and remediation. As a response to user feedback some additional statistical needs of the environmental projects of the U.S. EPA were addressed. Over the years ProUCL software has been upgraded and enhanced to include more graphical tools and statistical methods described in many EPA guidance documents.

Methods incorporated in ProUCL cover many common environmental situations and allow environmental practitioners with limited knowledge of statistics to perform calculations to estimate DQO based sample size, establish background levels, compare background and site sample data sets for site evaluation and risk assessment, and perform basic trend analysis. Some methods for analysis of data sets with nondetect values are built in this software. Statistical modules are organized as drop-down menus to allow users easy access to statistical methods and tests.

However, as any software, ProUCL has limitations. Software does not include advanced statistical methods applicable to very skewed data sets or biased sampling designs and does not include geostatistical methods. ProUCL also lacks capabilities to perform simulations or automation of repeating tasks. Therefore, environmental practitioners are strongly encouraged to seek advice from environmental statisticians on planning of environmental studies and choosing applicable statistical methods for sampling design used in the project.

Several improvements have been made to the decision logic for the recommendation of UCLs for version 5.2. The reliance on goodness of fit tests to select appropriate UCLs is reduced. The Chebyshev UCL is no longer recommended, and the H UCL is only recommended in cases of very large sample sizes when there is high confidence that the assumption of lognormality is met to a good approximation. In some cases, data may be too skewed or not numerous enough to determine an appropriate UCL. Version 5.2 does not provide a recommendation in these cases but encourages the user to: verify that the data were collected randomly (rather than through biased sampling, such as hot spot delineation sampling or best professional judgment sampling); consider site knowledge that may explain why the data may be skewed (such as small areas of high concentrations), and to contact a statistician if ProUCL cannot provide a recommendation.

Another improvement of ProUCL 5.2 is that libraries and developer tools (Microsoft .NET, Spread.NET (previously FarPoint), ChartFX, and Visual Studio) were updated to the latest available version. These tools have all had one or more version releases since 2016 when version ProUCL 5.1 was released.

In parallel with ProUCL improvements released as version 5.2, the ProUCL User guide and Technical guide were updated as well. The User Guide was reorganized to be better aligned with the software layout. Sections are now organized in the same order as ProUCL software drop-down menus. The last chapter of User Guide provides some limited guidance on the use of statistical methods incorporated in ProUCL software. Technical guide was updated to include the description and justification for decision logic improvements incorporated in version 5.2.

ProUCL has been verified against, and is agreement with, the results obtained by using other software packages including Minitab, SAS®, and CRAN R packages. Statistical methods incorporated in ProUCL have also been tested and verified extensively by the developers, researchers, scientists, and users. Software is continuously improved to address findings and observations of hundreds of users with different levels of statistical background spanning from environmental practitioners to professional statisticians performing analysis on thousands of environmental data sets.

ProUCL is available for free at the U.S. EPA Site Characterization and Monitoring Technical Support Center (SCMTSC) web site. SCMTSC also provides some user support. This may include answering questions related to the use of ProUCL software and technical support to EPA superfund project managers or technical staff.

ACRONYMS and ABBREVIATIONS

ACL	Alternative compliance or concentration limit
A-D, AD	Anderson-Darling test
AL	Action limit
AOC	Area(s) of concern
ANOVA	Analysis of variance
A_0	Not to exceed compliance limit or specified action level
BC	Box-Cox transformation
BCA	Bias-corrected accelerated bootstrap method
BD	Binomial distribution
BISS	Background Incremental Sample Simulator
BTV	Background threshold value
CC, cc	Confidence coefficient
CERCLA	Comprehensive Environmental Recovery, Compensation, and Liability Act
CL	Compliance limit
CLT	Central Limit Theorem
COPC	Contaminant/constituent of potential concern
C_s	Cleanup standards
CSM	Conceptual site model
Df	Degrees of freedom
DL	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 NDs replaced by 1/2 of the respective detection limits
DOE	Department of Energy
DQOs	Data quality objectives
DU	Decision unit
EA	Exposure area
EDF	Empirical distribution function
EM	Expectation maximization
EPA	United States Environmental Protection Agency
EPC	Exposure point concentration
GA	Georgia
GB	Gigabyte
GHz	Gigahertz
GROS	Gamma ROS
GOF, G.O.F.	Goodness-of-fit
GUI	Graphical user interface
GW	Groundwater
H_A	Alternative hypothesis
H_0	Null hypothesis
H-UCL	UCL based upon Land's H-statistic
i.i.d.	Independently and identically distributed
ISM	Incremental sampling methodology
ITRC	Interstate Technology & Regulatory Council

k, K	Positive integer representing future or next k observations
K	Shape parameter of a gamma distribution
K, k	Number of nondetects in a data set
k hat	MLE of the shape parameter of a gamma distribution
k star	Biased corrected MLE of the shape parameter of a gamma distribution
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 critical value
KM (z)	UCL based upon Kaplan-Meier estimates using critical value of a standard normal distribution
K-M, KM	Kaplan-Meier
K-S, KS	Kolmogorov-Smirnov
K-W	Kruskal Wallis
LCL	Lower confidence limit
LN, *ln*	Lognormal distribution
LCL	Lower confidence limit of mean
LPL	Lower prediction limit
LROS	LogROS; robust ROS
LTL	Lower tolerance limit
LSL	Lower simultaneous limit
M,m	Applied to incremental sampling: number in increments in an ISM sample
MARSSIM	Multi-Agency Radiation Survey and Site Investigation Manual

MCL	Maximum concentration limit, maximum compliance limit
MDD	Minimum detectable difference
MDL	Method detection limit
MK, M-K	Mann-Kendall
ML	Maximum likelihood
MLE	Maximum likelihood estimate
N	Number of observations/measurements in a sample
N	Number of observations/measurements in a population
MVUE	Minimum variance unbiased estimate
MW	Monitoring well
NARPM	National Association of Remedial Project Managers
ND, nd, Nd	Nondetect
NERL	National Exposure Research Laboratory
NRC	Nuclear Regulatory Commission
OKG	Orthogonalized Kettenring Gnanadesikan
OLS	Ordinary least squares
ORD	Office of Research and Development
OSRTI	Office of Superfund Remediation and Technology Innovation
OU	Operating unit
PCA	Principal component analysis
PDF, pdf	Probability density function
.pdf	Files in Portable Document Format

PRG	Preliminary remediation goals
PROP	Proposed influence function
p-values	Probability-values
QA	Quality assurance
QC	Quality
Q-Q	Quantile-quantile
R,r	Applied to incremental sampling: number of replicates of ISM samples
RAGS	Risk Assessment Guidance for Superfund
RCRA	Resource Conservation and Recovery Act
RL	Reporting limit
RMLE	Restricted maximum likelihood estimate
ROS	Regression on order statistics
RPM	Remedial Project Manager
RSD	Relative standard deviation
RV	Random variable
S	Substantial difference
SCMTSC	Site Characterization and Monitoring Technical Support Center
SD, Sd, sd	Standard deviation
SND	Standard Normal Distribution
SNV	Standard Normal Variate
SE	Standard error
SSL	Soil screening levels

SQL	Sample quantitation limit
SU	Sampling unit
S-W, SW	Shapiro-Wilk
T-S	Theil-Sen
TSC	Technical Support Center
TW, T-W	Tarone-Ware
UCL	Upper confidence limit
UCL95	95% upper confidence limit
UPL	Upper prediction limit
UPL95	95% upper prediction limit
U.S. EPA, EPA	United States Environmental Protection Agency
UTL	Upper tolerance limit
UTL95-95	95% upper tolerance limit with 95% coverage
USGS	U.S. Geological Survey
USL	Upper simultaneous limit
vs.	Versus
WMW	Wilcoxon-Mann-Whitney
WRS	Wilcoxon Rank Sum
WSR	Wilcoxon Signed Rank
X_p	p^{th} percentile of a distribution
<	Less than
>	Greater than

≥	Greater than or equal to
≤	Less than or equal to
Δ	Greek letter denoting the width of the gray region associated with hypothesis testing
Σ	Greek letter representing the summation of several mathematical quantities, numbers
%	Percent
α	Type I error rate
β	Type II error rate
Θ	Scale parameter of the gamma distribution
Σ	Standard deviation of the log-transformed data
^	carat sign over a parameter, indicates that it represents a statistic/estimate computed using the sampled data

GLOSSARY

Anderson-Darling (A-D) test: The Anderson-Darling test assesses whether known data come from a specified distribution. In ProUCL the A-D test is used to test the null hypothesis that a sample data set, x_1, ..., x_n came from a gamma distributed population.

Background Measurements: Measurements that are not site-related or impacted by site activities. 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.

Censored Data Sets: Data sets that contain one or more observations which are nondetects.

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. It is also known as the relative standard deviation (RSD).

Confidence Coefficient (CC): 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, x_0, 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.

Critical Value: The critical value for a hypothesis test is a threshold to which the value of the test statistic is compared to determine whether or not the null hypothesis is rejected. The critical value for any hypothesis

test depends on the sample size, the significance level, α at which the test is carried out, and whether the test is one-sided or two-sided.

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. It is 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 density function (PDF). EM is typically used to compute maximum likelihood estimates given incomplete samples.

Exposure Point Concentration (EPC): The constituent 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: Values that are well-separated from the majority of the data set coming from the far/extreme tails of the data distribution.

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 constituent 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, Δ, in this guidance.

H-Statistic: Land's statistic used to compute UCL of mean of a lognormal population

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.

Kolmogorov-Smirnov (KS) test: The Kolmogorov-Smirnov test is used to decide if a data set comes from a population with a specific distribution. The Kolmogorov-Smirnov test is based on the empirical distribution function (EDF). ProUCL uses the KS test to test the null hypothesis if a data set follows a gamma distribution.

Left-censored Data Set: An observation is left-censored when it is below a certain value (detection limit) but it is unknown by how much; left-censored observations are also called nondetect (ND) observations. A data set consisting of left-censored observations is called a left-censored data set. In environmental applications trace concentrations of chemicals may indeed be present in an environmental sample (e.g., groundwater, soil, sediment) but cannot be detected and are reported as less than the detection limit of the analytical instrument or laboratory method used.

Level of Significance (α): 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 goodness-of-fit test that tests for normality of large data sets when population mean and variance are unknown.

Maximum Likelihood Estimates (MLE): 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. It is 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 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) values: Censored data values. Typically, in environmental applications, concentrations or measurements that are less than the analytical/instrument method detection limit or reporting limit.

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.

Probability - Values (*p*-value): In statistical hypothesis testing, the *p*-value associated with 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 $t_{observed}$. The null hypothesis is rejected for all levels of significance, α greater than or equal to the *p*-value.

Parameter: A parameter is an unknown or known constant associated with the distribution used to model the population.

Parametric: A term describing statistical methods that assume a probability distribution such as a normal, lognormal, or a gamma distribution.

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

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

Probability of Type II (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 (1) 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 or p^{th} Quantile: 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.

Quality Assurance (QA): 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, quality control (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/quantiles 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 is used to fill in values imputed 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: see 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 and underlying assumptions) for a wide variety of data distributions.

Resistant Estimate: A test/estimate which is not affected by outliers is called a resistant test/estimate

Sample: 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: Shapiro-Wilk test is a goodness-of-fit test that tests the null hypothesis that a sample data set, x_1, ..., x_n came from a normally distributed population.

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

Stable UCL: The UCL of a population mean is a stable UCL if it represents a number of practical merit (e.g., a realistic value which can occur at a site), which also has some physical meaning. That is, a stable UCL represents a realistic number (e.g., constituent 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*, sd, 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 is the estimate. Standard errors are needed to construct confidence intervals for the parameters of interests such as the population mean and population percentiles.

Substitution Method: The substitution method is a method for handling NDs in a data set, where the ND is replaced by a defined value such as 0, DL/2 or DL prior to statistical calculations or graphical analyses.

This method has been included starting with ProUCL 5.1 for historical comparative purposes but **is not recommended** for use. The **bias** introduced by applying the substitution method **cannot be quantified** with any certainty. ProUCL will provide a warning when this option is chosen.

Uncensored Data Set: A data set without any censored (nondetects) observations.

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

Upper Simultaneous Limit (USL): The upper boundary of the largest value.

xBar: arithmetic average of computed using the sampled data values

ACKNOWLEDGEMENTS

We wish to express our gratitude and thanks to our friends and colleagues who have contributed during the development of past versions of ProUCL and to all of the many people who reviewed, tested, and gave helpful suggestions throughout the development of the ProUCL software package. We wish to especially acknowledge EPA scientists including Deana Crumbling, Nancy Rios-Jafolla, Tim Frederick, Jean Balent, Dr. Maliha Nash, Kira Lynch, and Marc Stifleman; James Durant of ATSDR, Dr. Steve Roberts of University of Florida, Dr. Elise A. Striz of the National Regulatory Commission (NRC), and Drs. Phillip Goodrum and John Samuelian of Integral Consulting Inc., as well as Dr. D. Beal of Leidos for testing and reviewing ProUCL and its associated guidance documents, and for providing helpful comments and suggestions. Finally, we want to express gratitude to statisticians and computer scientists of Neptune and Company, Inc. for the latest improvements included in ProUCL version 5.2.

Special thanks go to Dr. Anita Singh, Ms. Donna Getty and Mr. Richard Leuser of Lockheed Martin, for significant contribution to the development of ProUCL software and providing a thorough technical and editorial review of ProUCL 5.1 and also ProUCL 5.0 User Guide and Technical Guide. A special note of thanks is due to Ms. Felicia Barnett of EPA ORD Site Characterization and Monitoring Technical Support Center (SCMTSC), without whose assistance the development of the ProUCL 5.1 software and associated guidance documents would not have been possible.

Finally, we wish to dedicate the ProUCL 5.1 (and ProUCL 5.0) software package to our friend and colleague, John M. Nocerino John M. Nocerino who had contributed significantly in the development of ProUCL and Scout software packages.

Table of Contents

ProUCL 5.2 Capabilities

Assumptions: Like most statistical methods, statistical methods for computing upper limits (e.g., UCLs, UPLs, UTLs) are also based upon certain assumptions including the availability of a randomly collected data set consisting of independently and identically distributed (*i.i.d*) observations representing the population (e.g., site area, reference area) under investigation. A UCL of the mean (of a population) and BTV estimates (UPL, UTL) should be computed using a randomly collected (simple random or systematic random) data set representing a single statistical population (e.g., site population or background population). When multiple populations (e.g., background and site data mixed together) are present in a data set, the recommendation is to separate them first by using the population partitioning techniques (e.g., Singh, Singh, and Flatman 1994) prior to computing the appropriate decision statistics (e.g., 95% UCLs). Regardless of how the populations are separated, decision statistics should be computed separately for each identified population. The topic of population partitioning and the extraction of a valid site-specific background data set from a broader mixture data set potentially consisting of both onsite and offsite data are beyond the scope of ProUCL. Parametric estimation and hypotheses testing methods (e.g., t-test, UCLs, UTLs) are based upon distributional (e.g., normal distribution, gamma) assumptions. ProUCL includes GOF tests for determining if a data set follows a normal, a gamma, or a lognormal distribution.

Multiple Constituents/Variables: Environmental scientists need to evaluate many constituents in their decision-making processes including exposure and risk assessment, background evaluations, and site versus background comparisons. ProUCL can process multiple constituents/variables simultaneously in a user-friendly manner; an option not available in other freeware or commercial software packages such as NADA for R (Helsel 2013). This option is very useful when one has to process many variables/analytes and compute decision statistics (e.g., UCLs, UPLs, and UTLs) and/or test statistics (e.g., ANOVA test, trend test) for those variables/analytes.

Analysis by a Group Variable: ProUCL also has the capability of processing data by groups. A valid group column should be included in the data file. The analyses of data categorized by a group ID variable such as: 1) Surface versus (vs.) Subsurface; 2) AOC1 vs. AOC2; 3) Site vs. Background; and 4) Upgradient vs. Downgradient MWs are common in many environmental applications. ProUCL offers this option for data sets with and without nondetects. The **Group** option provides a way to perform 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 example, the same data set may consist of analytical data from multiple groups or populations representing site, background, two or more AOCs, surface soil, subsurface soil, and GW. By using this option, the graphical displays (e.g., box plots, Q-Q plots, histograms) and statistics (including computation of background statistics, UCLs, ANOVA test, trend test and OLS regression statistics) can be easily computed separately for each group in the data set.

Exploratory Graphical Displays for Uncensored and Left-Censored Data Sets: Graphical methods included in the **Graphs** module of ProUCL include: Q-Q plots (data in same column), multiple Q-Q plots (data in different columns), box plots, multiple box plots (data in different columns), histograms, and multiple histograms. These graphs can also be generated for data sets containing ND observations. Additionally, the **OLS Regression** and **Trend Analysis** module can be used to generate graphs displaying parametric OLS regression lines with confidence and prediction intervals around the regression and nonparametric Theil-

Sen trend lines. The **Trend Analysis** module can generate trend graphs for data sets without a sampling event variable, and also generates time series graphs for data sets with a sampling event (time) variable. ProUCL accepts only numerical values for the event variable. Graphical displays of a data set are useful for gaining added insight regarding a data set that may not otherwise be clear by looking at test statistics such as T-S test or MK statistics. Unlike test statistics (e.g., t-test, MK test, AD test) and decision statistics (e.g., UCL, UTL), graphical displays do not get influenced by outliers and ND observations. It is suggested that the final decisions be made based upon statistical results as well as graphical displays.

Side-by-side box plots or multiple Q-Q plots are useful to graphically compare concentrations of two or more groups (e.g., several monitoring wells). The GOF module of ProUCL generates Q-Q plots for normal, gamma, and lognormal distributions based upon uncensored as well as left-censored data sets with NDs. All relevant information such as the test statistics, critical values and probability-values (*p*-values), when available are also displayed on the GOF Q-Q plots. In addition to providing information about the data distribution, a normal Q-Q plot in the original raw scale also helps to identify multiple populations that may be present in a data set. On a Q-Q plot, observations well-separated from the majority of the data indicate presence of multiple populations. ProUCL can also be used to display box plots with horizontal lines displayed/superimposed at pre-specified compliance limits (CLs) or computed upper limits (e.g., UPL, UTL). This kind of graph provides a visual comparison of site data with compliance limits and/or BTV estimates.

ProUCL also provides a couple of classical outlier test procedures (EPA 2006b, 2009), the Dixon test and the Rosner test. The details of these outlier tests are described in Chapter 7. It is suggested that the classical outlier procedures should always be accompanied by graphical displays such as box plots.

Extreme values may represent observations coming from populations different from the dominant population represented by the majority of the data set. Outliers distort most statistics (e.g., mean, UCLs, UPLs, test statistics) of interest. Therefore, it is desirable to compute decisions statistics based upon data sets representing the main population. 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 several nonparametric statistics including bootstrap methods such as bootstrap-t and Hall's bootstrap UCLs and other nonparametric UPLs and UTLs computed using higher order statistics.

Goodness-of-Fit Tests: In addition to computing simple summary statistics for data sets with and without NDs, ProUCL includes GOF tests for normal, lognormal and gamma distributions. To test for normality (lognormality) of a data set, ProUCL includes the Lilliefors test and the extended S-W test for samples of sizes up to 2000 (Royston 1982a, 1982b). For the gamma distribution, two GOF tests: the A-D test (Anderson and Darling 1954) and K-S test (Schneider 1976, 1978) are available in ProUCL. For samples of larger sizes (e.g., with $n > 100$) and small values of the gamma shape parameter, k (e.g., $k \leq 0.1$), significant discrepancies were found in the critical values of the two gamma GOF test statistics (A-D and K-S tests) obtained using the two gamma deviate generation algorithms: Whitaker (1974) and Marsaglia and Tsang (2000). In ProUCL, for values of $k \leq 0.2$, the critical values of the two gamma GOF tests: A-D and K-S tests have been updated using the currently available more efficient gamma deviate generation algorithm due to Marsaglia and Tsang's (2000); more details about the implementation of their algorithm can be found in Kroese, Taimre, and Botev (2011). For these two GOF and values of the shape parameter, $k=0.025, 0.05, 0.1$, and 0.2, critical value tables have been updated by incorporating the newly generated

critical values for three levels of significance: 0.05, 0.1, and 0.01. The updated tables are provided in Appendix A of the ProUCL Technical Guide. It was noted that for *k=0.2*, the older (generated in 2002) and the newly generated critical values are in general agreement; therefore, critical values for *k=0.2* were not replaced in tables summarized in Appendix A.

ProUCL also generates GOF Q-Q plots for normal, lognormal, and gamma distributions displaying all relevant statistics including GOF test statistics. GOF tests for data sets with and without NDs are described in Chapters 2 and 3 of the ProUCL Technical Guide. For data sets containing NDs, it is not easy to verify the distributional assumptions correctly, especially when the data set consists of a large percentage of NDs with multiple DLs and NDs exceeding some detected values. Historically, decisions about distributions of data sets with NDs are based upon GOF test statistics computed using the data obtained: without NDs; replacing NDs by 0, DL, or DL/2; using imputed NDs based upon a ROS (e.g., lognormal ROS) method. For data sets with NDs, ProUCL can perform GOF tests using the methods listed above. ProUCL can also generate censored probability plots (Q-Q plots) which are very similar to Q-Q plots generated using detected data. Using the **Imputed NDs using ROS Methods** option of the **Stats/Sample Sizes** module of ProUCL, additional columns can be generated for storing imputed (estimated) values for NDs based upon normal ROS, gamma ROS, and lognormal ROS (also known as robust ROS) methods.

Sample Size Determination and Power Evaluation: The **Sample Sizes** module in ProUCL can be used to develop DQO-based sampling designs needed to address statistical issues associated with environmental projects. ProUCL provides user-friendly options for entering the desired/pre-specified values for decision parameters (e.g., Type I and Type II error rates) and other DQOs used to determine minimum sample sizes for statistical applications including: estimation of the mean, single and two-sample hypothesis testing approaches, and acceptance sampling for discrete items (e.g., drums containing hazardous waste). Both parametric (e.g., t-test) and nonparametric (e.g., Sign test, WRS test) sample size determination methods as described in EPA (2000, 2002c, 2006a, 2006b) guidance documents are available in ProUCL. ProUCL also has the sample size determination option for acceptance sampling of lots of discrete objects such as a lot (batch, set) of drums containing hazardous waste (e.g., RCRA applications, EPA 2002c). When the sample size for an application (e.g., verification of cleanup level) is not computed using the DQOs-based sampling design process, the **Sample Size** module can be used to assess the power of the test statistic used in retrospect. The mathematical details of the **Sample Sizes** module are given in Chapter 8 of the ProUCL Technical Guide.

Bootstrap Methods: Bootstrap methods are computer intensive nonparametric methods which can be used to compute decision statistics of interest when a data set does not follow a known distribution, or when it is difficult to analytically derive the distributions of statistics of interest. It is well-known that for moderately skewed to highly skewed data sets, UCLs based upon standard bootstrap and the percentile bootstrap methods do not perform well (e.g., Efron [1981, 1982]; Efron and Tibshirani 1993; Hall [1988,1992]; Singh, Singh, and Iaci 2002; Singh, Maichle and Lee 2006) as the interval estimates based upon these bootstrap methods fail to provide the specified coverage to the population mean (e.g., UCL95 does not provide adequate 95% coverage of population mean). For skewed data sets, Efron and Tibshirani (1993) and Hall (1988, 1992) considered other bootstrap methods such as the BCA, bootstrap-t and Hall's bootstrap methods. For skewed data sets, bootstrap-t and Hall's bootstrap (meant to adjust for skewness) methods perform better (e.g., in terms of coverage for the population mean) than the other bootstrap methods. However, it has been noted (e.g., Efron and Tibshirani 1993, Singh, Singh, and Iaci 2002) that

these two bootstrap methods tend to yield erratic and inflated UCL values (orders of magnitude higher than other UCLs) in the presence of outliers. Similar behavior of the bootstrap-t UCL and Hall's bootstrap UCL methods is observed for data sets consisting of NDs and outliers. Due to the reasons described above, whenever applicable, ProUCL provides cautionary notes and warning messages regarding the use of bootstrap-t and Halls bootstrap UCL methods.

Hypotheses Testing Approaches: ProUCL software has both single-sample (e.g., Student's t-test, sign test, proportion test, WSR test) and two-sample (Student's t-test, WMW test, Gehan test, and T-W test) parametric and nonparametric hypotheses testing approaches. Hypotheses testing approaches in ProUCL can handle both full-uncensored data sets and left-censored data sets with NDs. Most of the hypotheses tests also report associated *p*-values. For some hypotheses tests (e.g., WMW test, WSR test, proportion test), large sample *p*-values based upon the normal approximation are computed using continuity correction factors. The mathematical details of the various single-sample and two-sample hypotheses testing approaches are described in Chapter 6 the ProUCL Technical Guide

Single-Sample Tests: Parametric (Student's t-test) and nonparametric (Sign test, WSR test, tests for proportions and percentiles) hypotheses testing approaches are available in ProUCL. Single-sample hypotheses tests are used when environmental parameters such as the cleanup standard, action level, or compliance limits are known, and the objective is to compare site concentrations with those known threshold values. A t-test (or a sign test) may be used to verify the attainment of cleanup levels in an AOC after a remediation activity has taken place or a test for proportion may be used to verify if the proportion of exceedances of an action level (A_0 or a CL) by sample observations collected from an AOC (or a MW) exceeds a certain specified proportion (e.g., 1%, 5%, 10%).

The differences between these tests should be noted and understood. 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 determines 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., action level).

Two-Sample Tests: Hypotheses tests (Student's t-test, WMW test, Gehan test, T-W test) are used to perform site versus background comparisons, compare concentrations of two or more AOCs, or to compare concentrations of GW collected from MWs. As cited in the literature, some of the hypotheses testing approaches (e.g., nonparametric two-sample WMW) deal with a single detection limit scenario. When using the WMW test on a data set with multiple detection limits, all observations (detects and NDs) below the largest detection limit need to be considered as NDs (Gilbert 1987). This in turn tends to reduce the power and increase uncertainty associated with test. As mentioned before, it is always desirable to supplement the test statistics and conclusions with graphical displays such as multiple Q-Q plots and side-by-side box plots. The Gehan test or T-W test should be used in cases where multiple detection limits are present.

Note about Quantile Test: For smaller data sets, the Quantile test as described in U.S. EPA documents (U.S. EPA [1994, 2006b]; Hollander and Wolfe, 1999) is available in ProUCL 4.1(see ProUCL 4.1 Technical Guide). In the past, some users incorrectly used this test for larger data sets. Due to lack of resources, this

test has not been expanded for data sets of all sizes. Therefore, to avoid confusion and its misuse for larger data sets, the Quantile test was not included in ProUCL 5.0 and later versions.

Computation of Upper Limits including UCLs, UPLs, UTLs, and USLs: ProUCL software has parametric and nonparametric methods including bootstrap and Chebyshev inequality based methods to compute decision making statistics such as UCLs of the mean (EPA 2002a), percentiles, UPLs for future k (≥1) observations, UTLs (U.S. EPA [1992b and 2009]) and upper simultaneous limits (USLs) (Singh and Nocerino [1995, 2002]) based upon uncensored full data sets and left-censored data sets containing NDs with multiple DLs. Methods incorporated in ProUCL cover a wide range of skewed data distributions with and without NDs. In addition to normal and lognormal distributions based upper limits, ProUCL can compute parametric UCLs, percentiles, UPLs for future k (≥1) observations, UTLs, and USLs based upon gamma distributed data sets. For data sets with NDs, ProUCL has several estimation methods including the Kaplan-Meier (KM) method (1958), ROS methods (Helsel 2005) and substitution methods such as replacing NDs with the DL or DL/2 (Gilbert 1987; U.S. EPA 2006b). Substitution method and other poor performing methods (e.g., H-UCL for lognormal distribution) have been retained, as requested by U.S. EPA scientists, in ProUCL 5.0/ 5.1 /5.2 for research and comparison purposes. *One may not interpret the availability of these poor performing methods in ProUCL as recommended methods by ProUCL or by the U.S EPA for computing decision statistics.*

Computation of UCLs Based upon Uncensored Data Sets without NDs: Parametric UCL computation methods in ProUCL for uncensored data sets include: Student's t-UCL, Approximate gamma UCL (using chi-square approximation), Adjusted gamma UCL (adjusted for level significance), Land's H-UCL, and Chebyshev inequality-based UCL (using minimum variance unbiased estimates (MVUEs) of parameters of a lognormal distribution). Nonparametric UCL computation methods for data sets without NDs include: CLT-based UCL, Modified-t-statistic-based UCL (adjusted for skewness), Adjusted-CLT-based UCL (adjusted for skewness), Chebyshev inequality-based UCL (using sample mean and standard deviation), UCL based upon standard bootstrap, UCL based upon percentile bootstrap, UCL based upon BCA bootstrap, UCL based upon bootstrap-t, and UCL based upon Hall's bootstrap method. The details of UCL computation methods for uncensored data sets are summarized in Chapter 2 of the ProUCL Technical Guide.

Computations of UPLs, UTLs, and USLs Based upon Uncensored Data Sets without NDs: For uncensored data sets without NDs, ProUCL can compute parametric percentiles, UPLs for k (k≥1) future observations, UPLs for mean of k (≥1) future observations, UTLs, and USLs based upon the normal, gamma, and lognormal distributions. Nonparametric upper limits are typically based upon order statistics of a data set. Depending upon the size of the data set, the higher order statistics (maximum, second largest, third largest, and so on) are used to compute these upper limits (e.g., UTLs). Depending upon the sample size, specified CC and coverage probability, ProUCL outputs the actual CC achieved by a nonparametric UTL. The details of the parametric and nonparametric computation methods for UPLs, UTLs, and USLs are described in Chapter 3 of the ProUCL Technical Guide.

Computation of UCLs, UPLs, UTLs, and USLs Based upon Left-Censored Data Sets with NDs: For data sets with NDs, ProUCL computes UCLs, UPLs, UTLs, and USLs based upon the mean and *sd* computed using lognormal ROS (LROS, robust ROS), Gamma ROS (GROS), KM, and DL/2 substitution methods. To adjust for skewness in non-normally distributed data sets, ProUCL uses bootstrap methods and

Chebyshev inequality when computing UCLs and other limits using estimates of the mean and *sd* obtained using the methods (details in Chapters 4 and 5) listed above. ProUCL uses parametric methods on KM (and ROS) estimates, provided detected observations in the left-censored data set follow a parametric distribution. For example, if the detected data follow a gamma distribution, ProUCL uses KM estimates in gamma distribution-based equations when computing UCLs, UTLs, and other upper limits. When detected data do not follow a discernible distribution, depending upon size and skewness of detected data, ProUCL recommends the use of Kaplan-Meier (1958) estimates in bootstrap methods for computing nonparametric decision statistics (e.g., UCL95, UPL, UTL) of interest. ProUCL computes KM estimates directly using left-censored data sets without flipping data and requiring re-flipping of decision statistics. The KM method incorporated in ProUCL computes both *sd* and standard error (*SE*) of the mean. As mentioned earlier, for historical reasons and for comparison and research purposes, the DL/2 substitution method and H-UCL based upon LROS method have been retained in ProUCL 5.0/ 5.1 /5.2. The inclusion of the substitution and LROS methods in ProUCL should not be inferred as an endorsement of those methods by ProUCL software and its developers. The details of the UCL computation methods for data sets with NDs are given in Chapter 4 and the detail description of the various other upper limits: UPLs, UTLs, and USLs for data sets with NDs are given in Chapter 5 of the ProUCL Technical Guide.

Oneway ANOVA, OLS Regression and Trend Analysis: The **Oneway ANOVA** module has both classical and nonparametric K-W ANOVA tests as described in EPA guidance documents (e.g., EPA [2006b, 2009]). Oneway ANOVA is used to compare means (or medians) of multiple groups such as comparing mean concentrations of several areas of concern or performing inter-well comparisons of COPC concentrations at several MWs. The **OLS Regression** option computes the classical OLS regression line and generates graphs displaying the OLS line, confidence bands and prediction bands around the regression line. All statistics of interest including slope, intercept, and correlation coefficient are displayed on the OLS line graph. The **Trend Analysis** module has two nonparametric trend tests: the M-K trend test and T-S trend test. Using this option, one can generate trend graphs and time-series graphs displaying a T-S trend line and all other statistics of interest with associated *p*-values. In addition to slope and intercept, the T-S test in ProUCL 5.2 computes and outputs residuals based upon the computed nonparametric T-S line.

In GW monitoring applications, OLS regression, trend tests, and time series plots are often used to identify trends (e.g., upwards, downwards) in constituent concentrations of GW monitoring wells over a certain period of time (U.S. EPA 2009e). The details of Oneway ANOVA are given in Chapter 9 and OLS regression line and Trend tests methods are described in Chapter 10 of the ProUCL Technical Guide.

Note: It is pointed out that in this document, all statements made about the capabilities of ProUCL 5.0 also apply to ProUCL version 5.1 and 5.2; and to save time, many screen shots used in ProUCL 5.0 manuals have been used in ProUCL 5.2 manuals (User Guide and Technical Guide).

Recommendations and Suggestions in ProUCL: Until 2006, not much guidance was available on how to compute a UCL95 of the mean and other upper limits (e.g., UPLs and UTLs) for skewed left-censored data sets containing NDs with multiple DLs, a common occurrence in environmental data sets. For uncensored positively skewed data sets, Singh, Singh, and Iaci (2002) summarize some simulation results comparing the performances (in terms of coverage probabilities) of several UCL computation methods described in the statistical and environmental literature. They noted that the optimal choice of a decision statistic (e.g., UCL95) depends upon the sample size, data distribution and data skewness. They incorporated the results

of their findings in ProUCL 3.1 and higher versions to select the most appropriate UCL to estimate the EPC term.

For data sets with NDs, Singh, Maichle, and Lee (2006) conducted a similar simulation study to compare the performances of the various estimation methods (in terms of bias in the mean estimate); and some UCL computation methods (in terms of coverage provided by a UCL). They demonstrated that the KM estimation method performs well in terms of bias in estimates of the mean; and for skewed data sets, the t-statistic, CLT, and the percentile bootstrap method based UCLs computed using KM estimates (and ROS estimates) underestimate the population mean. From these findings summarized in Singh, Singh, and Iaci (2002) and Singh, Maichle, and Lee (2006), it is natural to state and assume the findings of the simulation studies performed on uncensored skewed data sets comparing performances of the various UCL computation methods can be extended to skewed left-censored data sets.

Like uncensored data sets without NDs, for data sets with NDs, there is no one single best UCL (and other upper limits such as UTL, UPL) which can be used to estimate an EPC (and background threshold values) for all data sets of varying sizes, distribution, and skewness. The optimal choice of a decision statistic depends upon the size, distribution, and skewness of detected observations.

For data sets with and without NDs, ProUCL computes decision statistics including UCLs, UPLs, and UTLs using several parametric and nonparametric methods covering a wide range of sample size, data variability and skewness. Using the results and findings summarized in the literature cited above, and based upon the sample size, data distribution, and data skewness, modules of ProUCL make suggestions about using the most appropriate decision statistic(s) to estimate population parameter(s) of interest (e.g., EPC). The suggestions made in ProUCL are based upon the extensive professional applied and theoretical experience of the developers in environmental statistical methods, published literature, results of simulation studies conducted by the developers of ProUCL and procedures described in many U.S. EPA guidance documents. These suggestions are made to help the users in selecting the most appropriate UCL to estimate an EPC which is routinely used in exposure assessment and risk management studies of the U.S. EPA. It should be pointed out that a typical simulation study cannot cover all data sets of various sizes and skewness from all types of distributions. For an analyte (data set) with skewness (*sd* of logged data) near the end points of the skewness intervals described in decision tables of Chapter 2 (e.g., Tables 2-9 through 2-11) of the ProUCL Technical Guide, the user/project team may select the most appropriate UCL based upon the site CSM, expert site knowledge, toxicity of the analyte, and exposure risks associated with that analyte. The project team should make the final decision regarding using or not using the suggestions/recommendations made by ProUCL. If deemed necessary, the project team may want to consult a statistician.

Even though, ProUCL software has been developed using limited government funding, it provides many statistical and graphical methods described in U.S. EPA documents for data sets with and without NDs. However, one may not compare the availability of methods in ProUCL with methods available in the commercial software packages such as SAS® and Minitab 16 or open source statistical computing software R. For example, trend tests correcting for seasonal/spatial variations and geostatistical methods are not available in the ProUCL software. For those methods, the user is referred to commercial software packages such as SAS® or open source R. As mentioned earlier, the developers of ProUCL recommended supplementing test results (e.g., two-sample test) with graphical displays (e.g., Q-Q plots, side-by-side box plots) especially when data sets contain NDs and outliers. With the inclusion of the **Oneway ANOVA**,

OLS Regression Trend and the user-friendly DQOs based **Sample Size** modules, ProUCL represents a comprehensive software package equipped with statistical methods and graphical tools needed to address many environmental sampling and statistical needs as described in the various CERCLA (U.S. EPA 1989a, 1992a, 2002a, 2002b, 2006a, 2006b), MARSSIM (U.S. EPA 2000), and RCRA (U.S. EPA 1989b, 1992b, 2002c, 2009) guidance documents.

Finally, the users of ProUCL are cautioned about the use of methods and suggestions described in some recent environmental literature. For example, many decision statistics (e.g., UCLs, UPLs, UTLs,) computed using the methods (e.g., percentile bootstrap, statistics using KM estimates and t-critical values) described in Helsel (2005, 2012) will fail to provide the desired coverage for environmental parameters of interest (mean, upper percentile) of moderately skewed to highly skewed populations and conclusions derived based upon those decisions statistics may lead to incorrect conclusions which may not be cost-effective or protective of human health and the environment.

Note: The look and feel of ProUCL 5.2 is similar to that of ProUCL 5.1 and 5.0; and they share the same names for the various modules and drop-down menus. For modules where no changes have been made in ProUCL since 2010 (e.g., Sample Sizes), screen shots as used in ProUCL 5.0 and 5.1 documents have been used in ProUCL 5.2 documents.

ProUCL 5.2 User Guide

In addition to this Technical Guide, a User Guide also accompanies the ProUCL 5.2 software, providing details of using the statistical and graphical methods incorporated in ProUCL 5.2. The User Guide provides details about the input and output operations that can be performed using ProUCL 5.2. The User guide also provides details about saving edited input files, output Excel-type spreadsheets and graphical displays generated by ProUCL 5.2.

CHAPTER 1

Guidance on the Use of Statistical Methods in ProUCL Software

Decisions based upon statistics computed using discrete data sets of small sizes (e.g., < 6) cannot be considered reliable enough to make decisions that affect human health and the environment. For example, a background data set of size < 6 is not large enough to characterize a background population, compute BTV estimates, or to perform background versus site comparisons. Several U.S. EPA guidance documents (e.g., EPA 2000, 2006a, 2006b) detail DQOs and minimum sample size requirements needed to address statistical issues associated with different environmental applications. In order to obtain reliable statistical results, an adequate amount of data should be collected using project-specified DQOs (i.e., CC, decision error rates). The **Sample Sizes** module of ProUCL computes minimum sample sizes based on DQOs specified by the user and described in many guidance documents. In some cases, it may not be possible (e.g., due to resource constraints) to collect the calculated number of samples needed to meet the project-specific DQOs. Under these circumstances one can use the **Sample Sizes** module to assess the power of the test statistic resulting from the reduced number of samples which were collected. Based upon professional experience, the developers of ProUCL 4 software and its later versions have been making some rule-of-thumb suggestions regarding minimum sample size requirements needed to perform statistical evaluations such as: estimation of environmental parameters of interest (i.e., EPCs and BTVs), comparing site data with background data or with some pre-established screening levels (e.g., action levels [ALs], compliance limits [CLs]). Those rule-of thumb suggestions are described later in Section 1.7 of this chapter. It is noted that those minimum sample requirements have been adopted by some other guidance documents including the RCRA Guidance Document (EPA 2009e).

This chapter also describes the differences between the various statistical upper limits including upper confidence limits (UCLs) of the mean, upper prediction limits (UPLs) for future observations, and upper tolerance intervals (UTLs) often used to estimate the environmental parameters of interest including EPC terms and BTVs. The use of a statistical method depends upon the environmental parameter(s) being estimated or compared. The measures of central tendency (e.g., means, medians, or their UCLs) are used to compare site mean concentrations with a cleanup standard, C_s, also 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 CLs, alternative concentration limits (ACL), or not-to-exceed values, are used when individual point-by-point observations are compared with those threshold values. Depending upon whether the environmental parameters (e.g., BTVs, not-to-exceed value, or EPC term) 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. Several upper limits, and single and two sample hypotheses testing approaches, for both full-uncensored and left-censored data sets are available in the ProUCL software package for performing the comparisons described above.

1.1 Background Data Sets

Based upon the CSM and regional and expert knowledge about the site, the project team selects background or reference areas. Depending upon the site activities and the pollutants, the background area can be site-specific or a general reference area with conditions comparable to the site before contamination due to site

related activities. An appropriate random sample of independent observations (*i.i.d*) should be collected from the background area. A defensible background data set represents a "single" environmental population possibly without any outliers.

Background data set needs to be evaluated for the presence of data caused by reporting and/or laboratory errors, and extreme values that are suspects of misrepresenting the observed population. Statistical outlier tests give probabilistic evidence for the "misfit" of extreme values. However, their drawback is that they assume normal distribution of the data without outliers. This is often not the case with environmental data, which tend to be naturally right-skewed. Therefore, statistical outlier tests available in ProUCL should only be used to identify potential suspect data points that require further investigation to gain an understanding of extreme values in the context of site processes, geology, and historical use. For example, extreme values may represent contamination from the site (hot spots). However, it is not unusual for a background to consist of different subpopulations due to the presence of varying soil types, textures, vegetation, historical use of the site, etc. It may have, therefore, have higher variability than expected in the planning process.

To obtain representative estimates for the decision-making statistics (e.g., UCLs, UPLs and UTLs), data need to be critically evaluated. Following five-step process as described in EPA QA/G-9S (2006) Data Quality Assessment: Statistical Methods for Practitioners is recommended:

- Identify extreme values that may be potential outliers;
- Apply statistical test;
- Scientifically review statistical outliers and decide on their disposition;
- Conduct data analyses with and without statistical outliers; and
- Document the entire process.

When calculating BTV, the objective is to compute background statistics based upon a data set which is representative of the background population. The occurrence of elevated outliers is not uncommon when background samples are collected from various onsite areas (e.g., large Federal Facilities). The proper disposition of outliers, to include or not include them in statistical computations, should be decided by the project team. The project team may want to compute decision statistics with and without the outliers to evaluate the influence of outliers on the decision making statistics.

A couple of classical outlier tests (Dixon and Rosner tests) are available in ProUCL. These tests assume normal distribution of the data without outliers. Therefore, a distribution of the data needs to be verified before outlier tests are applied. If the data are not normally distributed, they should be normalized by using an appropriate transformation before outlier tests are applied. It is also recommended that these classical outlier tests be supplemented with graphical displays such as a box plot and Q-Q plot. The use of exploratory graphical displays helps in determining the number of outliers potentially present in a data set.

An appropriate background data set of a reasonable size (preferably computed using the DQOs processes) is needed for the data set to be representative of background conditions and to compute upper limits (e.g., estimates of BTVs) and compare site and background data sets using hypotheses testing approaches. A background data set should have a minimum of 10 observations, however more observations is preferable.

1.2 Site Data Sets

A data set collected from a site population (e.g., AOC, exposure area [EA], DU, group of MWs) should be representative of the population under investigation. Depending upon the 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 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 the planning stages when developing sampling designs. Specifically, the availability of an adequate amount of representative data is required from each of those site sub-populations/strata defined by sample depths, soil types, and other characteristics.

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

- Point-by-point onsite comparisons with pre-established ALs or estimated BTVs. Typically, this approach is used when only a small number (e.g., < 6) of onsite observations are compared with a BTV or some other not-to-exceed value. If many onsite values need to be compared with a BTV, the recommended upper limit to use is the UTL or upper simultaneous limit (USL) to control the false positive error rate (Type I Error Rate). More details can be found in Chapter 3 of this guidance document. Alternatively, one can use hypothesis testing approaches (Chapter 6) provided enough observations (at least 10, more are preferred) are available.
- Single-sample hypotheses tests to compare site data with a pre-established cleanup standards, C_s (e.g., representing a measure of central tendency); proportion test to compare site proportion of exceedances of an AL with a pre-specified allowable proportion, P_0. These hypotheses testing approaches are used on site data when enough site observations are available. Specifically, when at least 10 (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 both types of error rates (Type 1 and Type 2) 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 a BTV (e.g., 95^{th} percentile) by onsite observations, even when the site and background populations have comparable distributions. The probabilities of these chance exceedances increase as the site sample size increases.
- Two-sample hypotheses tests to compare site data distribution with background data distribution to determine if the site concentrations are comparable to background concentrations. An adequate amount of data needs to be made available from the site as well as the background populations. It is preferable to collect at least 10 observations from each population under comparison. Note that if background data sets are re-used for multiple sites, the false positive rate may be higher than the user intends.

Table 1-1. Probability of at Least One Sample Exceeding a BTV for Various Sample Sizes

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

Notes: From a mathematical point of view, one can perform hypothesis tests on data sets consisting of only 3-4 data values; however, the reliability of the test statistics (and the conclusions derived) obtained is questionable. In these situations, it is suggested to supplement the test statistics decisions with graphical displays.

1.3 Discrete Samples or Composite Samples?

ProUCL can be used for discrete sample data sets, as well as on composite sample data sets. However, in a data set (background or site), 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.

- 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 ($\geq$ 10) site observations are available. Details of the single-sample hypothesis approaches are widely available in EPA guidance documents (MARSSIM 2000, EPA 1989a, 2006b). Several single-sample hypotheses testing procedures available in ProUCL are described in Chapter 6 of this document.
- If a two-sample hypothesis 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 approaches are used when many (e.g., at least 10) site, as well as background, observations are available. For better results with higher statistical power, the availability of more observations perhaps based upon an appropriate DQOs process (EPA 2006a) is desirable. Several two-sample hypotheses tests available in ProUCL are described in Chapter 6 of this document.

1.4 Upper Limits and Their Use

The computation and use of statistical limits depend upon their applications and the parameters (e.g., EPC term, BTVs) they are supposed to be estimating. Depending upon the objective of the study, a pre-specified cleanup standard, $C_{s,}$ can be viewed as representing: 1) an average (or median) constituent concentration, μ_0; or 2) a not-to-exceed upper threshold concentration value, A_0. These two threshold values, μ_0, and A_0, represent two significantly different parameters, and different statistical methods and limits are used to compare the site data with these two very different threshold values. Statistical limits, such as a UCL of the population mean, a 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 UTLs are often used to estimate the environmental parameters: EPC (μ_0) and a BTV (A_0). A new upper limit, USL was included in ProUCL 5.0 which may be used to estimate a BTV based upon a well-established background data set representing a single statistical population without any outliers.

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. The differences between UCLs and UPLs (or upper percentiles), and UCLs and UTLs should be clearly understood. A UCL with a 95% confidence limit (UCL95) of the mean represents an estimate of the population mean (measure of the central tendency), whereas a UPL95, a UTL95%-95% (UTL95-95), and an upper 95th percentile represent estimates of a threshold from the upper tail of the population distribution such as the 95th percentile. Here, UPL95 represents a 95% upper prediction limit, and UTL95-95 represents a 95% confidence limit of the 95th percentile. For mildly skewed to moderately skewed data sets, the numerical values of these limits tend to follow the order given as follows.

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

Example 1-1. Consider a real data set collected from a Superfund site. The data set has several inorganic COPCs, including aluminum (Al), arsenic (As), chromium (Cr), iron (Fe), lead (Pb), manganese (Mn), thallium (Tl) and vanadium (V). Iron concentrations follow a normal distribution. This data set has been used in several examples throughout the two ProUCL guidance documents (Technical Guide and User Guide), therefore it is provided as follows.

Table 1-2. Example Data Set from Superfund Site.

Aluminum	Arsenic	Chromium	Iron	Lead	Manganese	Thallium	Vanadium
6280	1.3	8.7	4600	16	39	0.0835	12
3830	1.2	8.1	4330	6.4	30	0.068	8.4
3900	2	11	13000	4.9	10	0.155	11
5130	1.2	5.1	4300	8.3	92	0.0665	9
9310	3.2	12	11300	18	530	0.071	22
15300	5.9	20	18700	14	140	0.427	32
9730	2.3	12	10000	12	440	0.352	19

Aluminum	Arsenic	Chromium	Iron	Lead	Manganese	Thallium	Vanadium
7840	1.9	11	8900	8.7	130	0.228	17
10400	2.9	13	12400	11	120	0.068	21
16200	3.7	20	18200	12	70	0.456	32
6350	1.8	9.8	7340	14	60	0.067	15
10700	2.3	14	10900	14	110	0.0695	21
15400	2.4	17	14400	19	340	0.07	28
12500	2.2	15	11800	21	85	0.214	25
2850	1.1	8.4	4090	16	41	0.0665	8
9040	3.7	14	15300	25	66	0.4355	24
2700	1.1	4.5	6030	20	21	0.0675	11
1710	1	3	3060	11	8.6	0.066	7.2
3430	1.5	4	4470	6.3	19	0.067	8.1
6790	2.6	11	9230	13	140	0.068	16
11600	2.4	16.4		98.5	72.5	0.13	
4110	1.1	7.6		53.3	27.2	0.068	
7230	2.1	35.5		109	118	0.095	
4610	0.66	6.1		8.3	22.5	0.07	

Several upper limits for iron are summarized as follows, and it be seen that they follow the order (in magnitude) as described above.

Table 1-3. 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

For highly skewed data sets, these limits may not follow the 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% and the largest sample value. 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 and the sample size is large (e.g., > 50, 100).

There is a great deal of confusion about the appropriate use of these upper limits. A brief discussion about the differences between the applications and uses of the statistical limits described above is provided as follows.

- A UCL represents an average value that is compared with a threshold value also representing an average value (pre-established or estimated), such as a mean C_s. For example, a site 95% UCL exceeding a C_s, may lead to the conclusion that the cleanup standard, C_s has not been attained by the average site area concentration. It should also be noted that UCLs of means are typically computed from the site data set.
- 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 a site average or a site median with a specified or pre-established cleanup standard (single-sample hypothesis), or with the background population average or median (two-sample hypothesis).
- A UPL, an upper percentile, or a 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 point-by-point onsite observations are compared with those limits. A site observation exceeding a background UTL may lead to the conclusion that the constituent is present at the site at levels greater than the background concentrations level.
- When enough (e.g., at least 10) site observations are available, it is preferable 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.

It is re-emphasized that only averages should be compared with averages or UCLs, and individual site observations should be compared with UPLs, upper percentiles, UTLs, or USLs. For 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.

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

The point-by-point observation comparison method is used when a small number (e.g., < 6) of site observations are compared with pre-established or estimated BTVs, screening levels, or preliminary remediation goals (PRGs). Typically, a single exceedance of the BTV by an onsite (or a monitoring well) observation may be considered an indication of the presence 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) at the site location (or a monitoring well) exhibiting constituent concentrations in excess of the BTV. If all collocated sample observations (or all sample observations collected during the same time period) from the same site location (or well) exceed the BTV or PRG, then it may be concluded that the location (well) requires further investigation (e.g., continuing treatment and monitoring) and possibly cleanup.

When BTV constituent concentrations are not known or pre-established, one has to collect a background data set of an appropriate size that can be considered representative of the site background. Statistical upper limits are computed using the background data set thus obtained, which are used as estimates of BTVs. To compute reasonably reliable estimates of BTVs, a minimum of 10 background observations should be collected, perhaps using an appropriate DQOs process as described in EPA (2000, 2006a). Several statistical limits listed above are used to estimate BTVs based upon a defensible (free of outliers, representing the background population) background data set of an adequate size.

The point-by-point comparison method is also useful when quick turnaround comparisons are required in real time. Specifically, when decisions have to be made in real time by a sampling/screening crew, or when only a few site samples are available, then individual point-by-point site concentrations are compared either with pre-established cleanup goals or with estimated BTVs. The sampling crew can use these comparisons to: 1) screen and identify the COPCs, 2) identify the potentially polluted site AOCs, or 3) continue or stop remediation or excavation at an onsite area of concern.

If a larger number of samples (e.g., >10) are available from the AOC, then the use of hypotheses testing approaches (both single-sample and a two-sample) is preferred. The use of hypothesis testing approaches tends to control the error rates more tightly and efficiently than the individual point-by-point site comparisons.

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 constituent concentrations of two (e.g., site versus background) or more populations (e.g., MWs).

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

When pre-established BTVs are used such as the U.S. Geological Survey (USGS) background values (Shacklette and Boerngen 1984), or thresholds obtained from similar sites, there is no need to establish or collect a background 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 DQOs) or decide (depending upon resources) 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 AOCs. As mentioned earlier, when the number of available site samples is < 6, one might perform point-by-point site observation comparisons with a BTV; and when enough site observations (at least 10) are available, it is desirable to use single-sample hypothesis testing approaches. Depending upon the parameter (μ_0, A_0),

represented by the known threshold value, one can use single-sample hypotheses tests for population mean or median (t-test, sign test), or use single-sample tests for proportions and percentiles. The details of the single-sample hypotheses testing approaches can be found in EPA (2006b) guidance document and in Chapter 6 of this document.

One-Sample t-Test: This test is used to compare the site mean, μ, with some specified cleanup standard, C_s, where the C_s represents an average threshold value, μ_0. The Student's t-test (or a UCL of the mean) is used (assuming normality of site data set or when sample size is large, such as larger than 30, 50) to verify 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 ND observations, provided the detection limits of all NDs fall below the specified threshold value, C_s. These tests are used to compare the site location (e.g., median, mean) with some specified 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 constituent concentration distribution rather than the mean threshold value, μ_0, then a test for proportion or a test for percentile (equivalently 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.

1.6.2 Two-Sample Hypotheses (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, two-sample hypothesis testing approaches are 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 from each of the two populations. Site and background data requirements (e.g., based upon DQOs) for performing two-sample hypothesis test approaches are described in EPA (2000, 2002b, 2006a, 2006b) and also in Chapter 6 of this Technical Guide. 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) in sample size determination. That is, a larger number (>15-20) of representative background (and site) samples should be collected from larger background (and site) areas; every effort should be made to collect as many samples as determined by the DQOs-based sample sizes.

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. Specifically, in environmental applications, two-sample hypotheses testing approaches are used to compare average or median constituent concentrations of two or more populations. To derive reliable conclusions with higher statistical power based upon hypothesis testing approaches, an adequate amount of data (e.g., minimum of 10 samples) should be collected from all of the populations under investigation.

The two-sample hypotheses testing approaches incorporated in ProUCL are listed as follows:

- Student t-test (with equal and unequal variances) – Parametric test assumes normality
- Wilcoxon-Mann-Whitney (WMW) test – Nonparametric test handles data with NDs with one DL - assumes two populations have comparable shapes and variability
- Gehan test – Nonparametric test handles data sets with NDs and multiple DLs - assumes comparable shapes and variability
- Tarone-Ware (T-W) test – Nonparametric test handles data sets with NDs and multiple DLs - assumes comparable shapes and variability

The Gehan and T-W tests are meant to be used on left-censored data sets with multiple DLs. For best results, the samples collected from the two (or more) populations should all be of the same type obtained using similar analytical methods and apparatus; the collected site and background samples should all be 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: ProUCL has been developed using limited government funding. ProUCL is equipped with statistical and graphical methods needed to address many environmental sampling and statistical issues as described in the various CERCLA, MARSSIM, and RCRA documents cited earlier. However, one may not compare the availability of methods in ProUCL with methods incorporated in commercial software packages such as SAS® and Minitab 16 or open source software for statistical computing R. Not all methods available in the statistical literature are available in ProUCL.

1.7 Minimum Sample Size Requirements and Power Evaluations

Due to resource limitations, it is not be possible (nor needed) to sample the entire population (e.g., background area, site area, AOCs, EAs) under study. Statistics is used to draw inference(s) about the populations (clean, dirty) and their known or unknown statistical parameters (e.g., mean, variance, upper threshold values) based upon much smaller data sets (samples) collected from those populations. To determine and establish BTVs and site specific screening levels, defensible data set(s) of appropriate size(s) representing the background population (e.g., site-specific, general reference area, or historical data) need to be collected. 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 area and boundaries based upon all current and intended future uses, and the objectives of data collection. 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 constituent concentrations are comparable; or if the site concentrations exceed the background threshold concentration level; or if an adequate amount of remediation approaching the BTV or some cleanup level has been performed at polluted site AOCs.

To perform statistical tests and compute upper limits, determine the number of samples that need to be collected from the populations (e.g., site and background) under investigation using appropriate DQOs processes (EPA 2000, 2006a, 2006b). ProUCL has the **Sample Sizes** module which can be used to develop DQOs based sampling designs needed to address statistical issues associated with polluted sites projects. ProUCL provides user-friendly options to enter the desired/pre-specified values of decision parameters (e.g., Type I and Type II error rates) to determine minimum sample sizes for the selected statistical applications including: estimation of mean, single and two-sample hypothesis testing approaches, and acceptance sampling. Sample size determination methods are available for the sampling of continuous characteristics (e.g., lead or Radium 226), as well as for attributes (e.g., proportion of occurrences exceeding a specified threshold). Both parametric (e.g., t-tests) and nonparametric (e.g., Sign test, test for proportions, WRS test) sample size determination methods are available in ProUCL 5.2 and in its earlier versions (e.g., ProUCL 4.1). ProUCL also has sample size determination methods for acceptance sampling of lots of discrete objects such as a batch of drums containing hazardous waste (e.g., RCRA applications, U.S. EPA 2002c).

However, due to budgetary or logistical constraints, it may not be possible to collect the same number of samples as determined by applying a DQO process. For example, the data might have already been collected (as often is the case) without using a DQO process, or due to resource constraints, it may not have been possible to collect as many samples as determined by using a DQO-based sample size formula. In practice, the project team and the decision makers tend not to collect enough background samples. It is suggested to collect at least 10 background observations before using statistical methods to perform background evaluations based upon data collected using discrete samples. The minimum sample size recommendations described here are useful when resources are limited, and it may not be possible to collect as many background and site samples as computed using DQOs based sample size determination formulae. In case data are collected without using a DQO process, the **Sample Sizes** module can be used to assess the power of the test statistic in retrospect. Specifically, one can use the standard deviation of the computed test statistic (EPA 2006b) and compute the sample size needed to meet the desired DQOs. If the computed sample size is greater than the size of the data set used, the project team may want to collect additional samples to meet the desired DQOs.

Note: From a mathematical point of view, the statistical methods incorporated in ProUCL and described in this guidance document for estimating EPC terms and BTVs, and comparing site versus background concentrations can be performed on small site and background data sets (e.g., of sizes as small as 3). However, those statistics may not be considered representative and reliable enough to make important cleanup and remediation decisions which will potentially impact human health and the environment. ProUCL provides messages when the number of detects is <4-5, and suggests collecting at least 8-10 observations. Based upon professional judgment, as a rule-of-thumb, ProUCL guidance documents recommend collecting a minimum of 10 observations when data sets of a size determined by a DQOs process (EPA 2006) cannot be collected. This however, should not be interpreted as the general recommendation and every effort should be made to collect DQOs based number of samples. Some recent guidance documents (e.g., EPA 2009e) have also adopted this rule-of-thumb and suggest collecting a minimum of about 8-10 samples in the circumstance that data cannot be collected using a DQO-based process. However, the project team needs to make these determinations based upon their comfort level and knowledge of site conditions.

To allow users to compute decision statistics using data from ISM (ITRC 2012 and ITRC 2020) samples, ProUCL will compute decision statistics (e.g., UCLs, UPLs, UTLs) and conduct hypothesis tests based upon samples of sizes as small as 3. Note that if discrete samples are used, the sample size for any statistical computation should be at least 8. The user is referred to the ITRC ISM-2 Technical Regulatory Guide (2020) for additional information on ISM considerations; however, note that ITRC (2012, 2020) may recommend the Chebyshev UCL, which has been shown to grossly overestimate the mean. Refer to Section 2.4.7 for discussion of the Chebyshev UCL.

1.7.1 Why a Data Set of Minimum Size, n = 10?

Typically, the computation of parametric upper limits (UPL, UTL, UCL) depends upon three values: the sample mean, sample variability (standard deviation) and a critical value. A critical value depends upon sample size, data distribution, and confidence level. For samples of small size (< 10), the critical values are large and unstable, and upper limits (e.g., UTLs, UCLs) based upon a data set with fewer than 10 observations are mainly driven by those critical values. The differences in the corresponding critical values tend to stabilize when the sample size becomes larger than 10 (see tables below, where degrees of freedom [*df*] = sample size - 1). This is one of the reasons ProUCL guidance documents suggest a minimum data set size of 10 when the number of observations determined from sample-size calculations based upon EPA DQO process exceed the logistical/financial/temporal/constraints of a project. For samples of sizes 2-11, 95% critical values used to compute upper limits (UCLs, UPLs, UTLs, and USLs) based upon a normal distribution are summarized in the subsequent tables. In general, a similar pattern is followed for critical values used in the computation of upper limits based upon other distributions.

For the normal distribution, Student's t-critical values are used to compute UCLs and UPLs which are summarized as follows.

Table 1-4. Crititical Values of t-Statistic

df= sample size-1= *(n-1)*

	Upper-tail probability *p*				
df	.10	.05	.025	.02	.01
1	3.078	6.314	12.71	15.89	31.82
2	1.886	2.920	4.303	4.849	6.965
3	1.638	2.353	3.182	3.482	4.541
4	1.533	2.132	2.776	2.999	3.747
5	1.476	2.015	2.571	2.757	3.365
6	1.440	1.943	2.447	2.612	3.143
7	1.415	1.895	2.365	2.517	2.998
8	1.397	1.860	2.306	2.449	2.896
9	1.383	1.833	2.262	2.398	2.821
10	1.372	1.812	2.228	2.359	2.764

One can see that once the sample size starts exceeding 9-10 (*df* = 8, 9), the difference between the critical values starts stabilizing. For example, for upper tail probability (= level of significance) of 0.05, the difference between critical values for *df* = 9 and *df* =10 is only 0.021, where as the difference between critical values for *df*= 4 and 5 is 0.117; similar patterns are noted for other levels of significance. For the normal distribution, critical values used to compute UTL90-95, UTL95-95, USL90, and USL95 are

described as follows. One can see that once the sample size starts exceeding 9-10, the difference between the critical values starts decreasing significantly.

Table 1-5. UTLs Computed Using the t-Statistic (for Normally Distributed Data).

n	UTL90-95	UTL95-95	USL90	USL95
3	6.155	7.656	1.148	1.153
4	4.162	5.144	1.425	1.462
5	3.407	4.203	1.602	1.671
6	3.006	3.708	1.729	1.822
7	2.755	3.399	1.828	1.938
8	2.582	3.187	1.909	2.032
9	2.454	3.031	1.977	2.11
10	2.355	2.911	2.036	2.176
11	2.275	2.815	2.088	2.234

Note: Nonparametric upper limits (UPLs, UTLs, and USLs) are computed using higher order statistics of a data set. To achieve the desired confidence coefficient, samples of sizes much greater than 10 are required. For details, refer to Chapter 3. It should be noted that critical values of USLs are significantly lower than critical values for UTLs. Critical values associated with UTLs decrease as the sample size increases. Since, as the sample size increases the maximum of the data set also increases, and critical values associated with USLs increase with the sample size.

1.7.2 Sample Sizes for Bootstrap Methods

Several nonparametric methods including bootstrap methods for computing UCL, UTL, and other limits for both full-uncensored data sets and left-censored data sets with NDs are available in ProUCL. Bootstrap resampling methods are useful when not too few (e.g., < 15-20) and not too many (e.g., > 500-1000) 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 are drawn with replacement from the same data set. Therefore, 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 15-20. Also, it is not necessary 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., > 500) 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.

Note: Rules-of-thumb about minimum sample size requirements described in this section are based upon professional experience of the developers. ProUCL software is not a policy software. It is recommended that the users/project teams/agencies make determinations about the minimum number of observations and minimum number of detects that should be present in a data set before using a statistical method.

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 common in environmental applications. ProUCL offers this option for data sets with and without NDs. The Group Option provides a tool for performing separate statistical tests and for generating separate graphical displays for each member/category of the group (samples from different populations) that may be present in a data set. The graphical displays (e.g., box plots, quantile-quantile plots) and statistics (e.g., background statistics, UCLs, hypotheses tests) of interest can be computed separately for each group by using this option. Moreover, using the Group Option, graphical methods can display multiple graphs (e.g., Q-Q plots) on the same graph providing graphical comparison of multiple groups.

It should be pointed out that it is the user's responsibility to provide an adequate amount of data to perform the group operations. For example, if the user desires to produce a graphical Q-Q plot (e.g., using only detected data) with regression lines displayed, then there should be at least two detected data values (to compute slope, intercept, *sd*) in the data set. Similarly if the graphs are desired for each group specified by the group ID variable, there should be at least two observations in each group specified by the group variable. When ProUCL data requirements are not met, ProUCL does not perform any computations, and generates a warning message (colored orange) in the lower Log Panel of the output screen of ProUCL.

1.9 Statistical Analyses for Many Constituents/Variables

ProUCL software can process multiple analytes/variables simultaneously in a user-friendly manner This option is useful when one has to process multiple variables and compute decision statistics (e.g., UCLs, UPLs, and UTLs) and test statistics (e.g., ANOVA test, trend test) for multiple variables. It is the user's responsibility to make sure that each selected variable has an adequate amount of data so that ProUCL can perform the selected statistical method correctly. ProUCL displays warning messages when a selected variable does not have enough data needed to perform the selected statistical method.

1.10 Use of Maximum Detected Value as Estimates of Upper Limits

Some practitioners 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 1992a). Also, many times in practice, the BTVs and not-to-exceed values are estimated by the maximum detected value (e.g., nonparametric UTLs, USLs).

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

BTVs and not-to-exceed values represent upper threshold values from 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 largest or the second largest detected value. A nonparametric UPL, UTL, and USL are often estimated by higher order statistics such as the maximum value or the second largest value (EPA 1992b, 2009, Hahn and Meeker 1991). The use of higher order statistics to estimate the UTLs depends upon the sample size. For data sets of size: 1) 59 to 92 observations, a nonparametric UTL95-95 is given by the maximum detected value; 2) 93 to 123 observations, a nonparametric UTL95-95 is given by the second

largest maximum detected value; and 3) 124 to 152 observations, a UTL95-95 is given by the third largest detected value in the sample, and so on.

1.10.2 Use of Maximum Detected Value to Estimate EPC Terms

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 value. Specifically, the EPA (1992a) document 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) exceeds the maximum value in a data set. ProUCL computes 95% UCLs of the mean using several methods based upon normal, gamma, lognormal, and non-discernible distributions. In the past, a lognormal distribution was used as the default distribution to model positively skewed environmental data sets. Additionally, only two methods were used to estimate the EPC term based upon: 1) normal distribution and Student's t-statistic, and 2) lognormal distribution and Land's H-statistic (Land 1971, 1975). The use of the H-statistic often yields unstable and impractically large UCL95 of the mean (Singh, Singh, and Engelhardt 1997; Singh, Singh, and Iaci 2002). For highly skewed data sets of smaller sizes (< 30, < 50), H-UCL often exceeds the maximum detected value. Since the use of a lognormal distribution has been quite common (suggested as a default model in the risk assessment guidance for Superfund [RAGS] document [EPA 1992a]), the exceedance of the maximum value by an H-UCL95 is frequent for many skewed data sets of smaller sizes (e.g., < 30, < 50). 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 an impacted location. Obviously, it is not desirable to use an observation potentially representing an impacted location to estimate the EPC for an AOC. The EPC term represents the average exposure contracted by an individual over an EA during a long period of time; 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. As can be seen in figures described in Appendix B, for data sets of small sizes (e.g., < 10-20), the Max Test (U.S. EPA 1996)does not provide the specified 95% coverage to the population mean, and for larger data sets it overestimates the EPC term, which may lead to unnecessary further remediation.

Several methods, some of which are described in EPA (2002a) and other EPA documents, are available in versions of ProUCL (i.e., ProUCL 3.00.02 [EPA 2004], ProUCL 4.0 [U.S. EPA 2007], ProUCL 4.00.05 [EPA 2009c, 2010], ProUCL 4.1 [EPA 2011]) for estimating the EPC terms. For data sets with NDs, ProUCL 5.0 and newer versions has some new UCL (and other limits) computation methods which were not available in earlier versions of ProUCL. It is unlikely that the UCLs based upon those methods will exceed the maximum detected value, unless some outliers are present in the data set.

1.10.2.1 Alternative UCL95 Computations

ProUCL 5.2 displays a warning message when the suggested 95% UCL (e.g., Hall's or bootstrap-t UCL) of the mean exceeds the detected maximum concentration. When a 95% UCL does exceed the maximum observed value, an alternative UCL computation method may be used. The choice of alternative UCL will

depend on the particular data set and may require professional judgement. Practitioners are encouraged to contact a statistician for guidance.

Notes: Using the maximum observed value to estimate the EPC term representing the average exposure contracted by an individual over an EA is not recommended. For the sake of interested users, ProUCL displays a warning message when the recommended 95% UCL (e.g., Hall's bootstrap UCL) of the mean exceeds the observed maximum concentration. Note that ProUCL no longer recommends the Chebyshev UCL.

1.11 Samples with Nondetect Observations

ND observations are inevitable in most environmental data sets. Singh, Maichle, and Lee (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 ND observations. They concluded that the UCLs obtained using the substitution methods, including the replacement of NDs by DL/2; do not perform well even when the percentage of ND observations is low, such as less than 5% to 10%. They recommended avoiding the use of substitution methods for computing UCL95 based upon data sets with ND observations.

1.11.1 Avoid the Use of the DL/2 Substitution Method to Compute UCL95

Based upon the results of the report by Singh, Maichle, and Lee (2006), it is recommended to avoid the use of the DL/2 substitution method when performing a GOF test, and when computing the summary statistics and various other limits (e.g., UCL, UPL, UTLs) often used to estimate the EPC terms and BTVs. Until recently, the substitution method has been the most commonly used method for computing various statistics of interest for data sets which include NDs. The main reason for this has been the lack of the availability of the other rigorous methods and associated software programs that can be used to estimate the various environmental parameters of interest. Today, several methods (e.g., using KM estimates) with better performance, including bootstrap methods, are available for computing the upper limits of interest. Several of those parametric and nonparametric methods are available in ProUCL 4.0 and higher versions. The DL/2 method is included in ProUCL for historical reasons as it had been the most commonly used and recommended method until recently (EPA 2006b). EPA scientists and several reviewers of the ProUCL software had suggested and requested the inclusion of the DL/2 substitution method in ProUCL for comparison and research purposes.

Notes: Even though the DL/2 substitution method has been incorporated in ProUCL, its use is **not recommended** due to its poor performance. The DL/2 substitution method has been retained in ProUCL 5.2 for historical and comparison purposes. NERL-EPA, Las Vegas strongly recommends avoiding the use of this method even when the percentage of NDs is as low as 5% to 10%.

1.11.2 ProUCL Does Not Distinguish between Detection Limits, Reporting limits, or Method Detection Limits

ProUCL 5.2 (and all previous versions) does not make distinctions between method detection limits (MDLs), adjusted MDLs, sample quantitation limits (SQLs), reporting limits (RLs), or DLs. Multiple DLs (or RLs) in ProUCL mean different values of the detection limits. It is user's responsibility to understand

the differences between these limits and use appropriate values (e.g., DLs) for nondetect values below which the laboratory cannot reliably detect/measure the presence of the analyte in collected samples (e.g., soil samples). A data set consisting of values less than the DLs (or MDLs, RLs) is considered a left-censored data set. ProUCL uses statistical methods available in the statistical literature for left-censored data sets for computing statistics of interest including mean, sd, UCL, and estimates of BTVs.

The user determines which qualifiers (e.g., J, U, UJ) will be considered as nondetects. Typically, all values with U or UJ qualifiers are considered as nondetect values. It is the user's responsibility to enter a value which can be used to represent a ND value. For NDs, the user enters the associated DLs or RLs (and not zeros or half of the detection limits). An indicator column/variable, D_x taking a value, 0, for all nondetects and a value, 1, for all detects is assigned to each variable, x, with NDs. It is the user's responsibility to supply the numerical values for NDs (should be entered as reported DLs) not qualifiers (e.g., J, U, B, UJ). For example, for thallium with nondetect values, the user creates an associated column labeled as D_thallium to tell the software that the data set will have nondetect values. This column, D_thallium consists of only zeros (0) and ones (1); zeros are used for all values reported as NDs and ones are used for all values reported as detects.

1.12 Samples with Low Frequency of Detection

When all of the sampled values are reported as NDs, the EPC term and other statistical limits should also be reported as a ND value, perhaps by the maximum RL or the maximum RL/2. The project team will need to make this determination. Statistics (e.g., UCL95) based upon only a few detected values (e.g., < 4) cannot be considered reliable enough to estimate EPCs which can have a potential impact on human health and the environment. When the number of detected values is small, it is preferable to use ad hoc methods rather than using statistical methods to compute EPCs and other upper limits. Specifically, for data sets consisting of < 4 detects and for small data sets (e.g., size < 10) with low detection frequency (e.g., < 10%), the project team and the decision makers should decide, on a site-specific basis, how to estimate the average exposure (EPC) for the constituent and area under consideration. For 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, when most (e.g., > 95%) of the observations for a constituent lie below the DLs, the sample median or the sample mode (rather than the sample average) may be used as an estimate of the EPC. Note that when the majority of the data are NDs, the median and the mode may also be represented by a ND value. The uncertainty associated with such estimates will be high. 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.

1.13 Some Other Applications of Methods in ProUCL 5.2

In addition to performing background versus site comparisons for CERCLA and RCRA sites, performing trend evaluations based upon time-series data sets, and estimating EPCs in exposure and risk evaluation studies, the statistical methods in ProUCL can be used to address other issues dealing with environmental investigations that are conducted at Superfund or RCRA sites.

1.13.1 Identification of COPCs

Risk assessors and remedial project managers (RPMs) often use screening levels or BTVs to identify COPCs during the screening phase of a cleanup project at a contaminated site. The screening for COPCs is performed prior to any characterization and remediation activities that are conducted at the site. This comparison is performed to screen out those constituents 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 constituents may be eliminated from all future site investigations, and risk assessment and risk management studies.

To identify the COPCs, point-by-point site observations are compared with some pre-established soil 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 onsite. The project team should decide the type of site samples (discrete or composite) and the number of site observations that should be collected and compared with the screening levels or the BTVs. In case BTVs or screening levels are not known, the availability of a defensible site-specific background or reference data set of reasonable size (e.g., at least 10) is required for computing reliable and representative estimates of BTVs and screening levels. The constituents with concentrations exceeding the respective screening values or BTVs may be considered COPCs, whereas constituents with concentrations (e.g., in all collected samples) lower than the screening values or BTVs may be omitted from all future evaluations.

1.13.2 Identification of Non-Compliance Monitoring Wells

In MW compliance assessment applications, individual (often discrete) constituent concentrations from a MW are compared with some pre-established limits such as an ACL or a maximum concentration limit (MCL). An exceedance of the MCL or the BTV (e.g., estimated by a UTL95-95 or a UPL95) by a MW concentration may be considered an indication of contamination in that MW. For 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 constituents in the original sample and re-sample(s) exceed the MCL or BTV, then that MW may require further scrutiny, perhaps triggering remediation activates. If the concentration data from a MW for 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.

1.13.3 Verification of the Attainment of Cleanup Standards, Cs

Hypothesis testing approaches are used to verify the attainment of the cleanup standard, C_s, at site AOCs after conducting remediation and cleanup at those site AOCs (EPA 1989a, 1994). In order to assess the attainment of cleanup levels, a representative data set of adequate size perhaps obtained using the DQO process (or a minimum of 10 observations should be collected) needs to be made available from the remediated/excavated areas of the site under investigation. The sample size should also account for the size of the remediated site areas: meaning that larger site areas should be sampled more (with more observations) to obtain a representative sample of the remediated areas under investigation. Typically, the null hypothesis of interest is H_0: Site Mean, $\mu_s \geq C_s$ versus the alternative hypothesis, H_1: Site Mean, $\mu_s < C_s$, where the cleanup standard, C_s, is known *a priori*.

1.13.4 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 are compared with a pre-established or estimated BTV. Exceedances of the BTV by site observations may represent impacted locations with elevated concentrations (hot spots).

1.14 Some General Issues, Suggestions and Recommendations made by ProUCL

Some general issues regarding the handling of multiple DLs by ProUCL and recommendations made about various substitution and ROS methods for data sets with NDs are described in the following sections.

1.14.1 Handling of Field Duplicates

ProUCL does not pre-process field duplicates. The project team determines how field duplicates will be handled and pre-processes the data accordingly. For an example, if the project team decides to use average values for field duplicates, then averages need to be computed and field duplicates need to be replaced by their respective average values. It is the user's responsibility to feed in appropriate values (e.g., averages, maximum) for field duplicates. The user is advised to refer to the appropriate EPA guidance documents related to collection and use of field duplicates for more information.

1.14.2 ProUCL Recommendation about ROS Method and Substitution (DL/2) Method

For data sets with NDs, ProUCL can compute point estimates of population mean and standard deviation using the KM and ROS methods (and also using the DL/2 substitution method). The substitution method has been retained in ProUCL for historical and research purposes. ProUCL uses Chebyshev inequality, bootstrap methods, and normal, gamma, and lognormal distribution-based equations on KM (or ROS) estimates to compute upper limits (e.g., UCLs, UTLs). The simulation study conducted by Singh, Maichle and Lee (2006) demonstrated that the KM method yields accurate estimates of the population mean. They also demonstrated that for moderately skewed to highly skewed data sets, UCLs based upon KM estimates and BCA bootstrap (mild skewness), KM estimates and Chebyshev inequality (moderate to high skewness), and KM estimates and bootstrap-t method (moderate to high skewness) yield better (in terms of coverage probability) estimates of EPCs than other UCL methods based upon the Student's t-statistic on KM estimates, percentile bootstrap method on KM or ROS estimates.

1.14.3 Unhandled Exceptions and Crashes in ProUCL

A typical statistical software, especially developed under limited resources may not be able to accommodate data sets with all kinds of deficiencies such as all missing values for a variable, or all nondetect values for a variable. An inappropriate/insufficient data set can occur in various forms and not all of them can be addressed in a scientific program like ProUCL. Specifically, from a programming point of view, it can be quite burdensome on the programmer to address all potential deficiencies that can occur in a data set. ProUCL addresses many data deficiencies and produces warming messages. All data deficiencies causing unhandled exceptions which were identified by users have been addressed in ProUCL. However, when ProUCL yields an unhandled exception or crashes, it is highly likely that there is something wrong with the

data set; the user is advised to review the input data set to make sure that the data set follows ProUCL data and formatting requirements.

1.14.4 95% UCL (UCL95) Computed by ProUCL and NADA Packages in R and for Minitab

The fundamental assumption when computing a UCL95 (by any software) of mean is that the data set used represents a single statistical population. Simulation results used to make the suggestions regarding the selection of an appropriate UCL95 do not cover data sets representing multiple populations (with varying means and standard deviations). Typically, a mixture data set representing multiple populations cannot be modeled by a known probability distribution (e.g., normal, gamma, ...) Since the suggestions made by ProUCL are based upon simulation experiments, they may not cover all "Real-World" data sets, especially highly skewed nonparametric data sets. For such data sets, it is recommended that the project team seek advice from a qualified statistician. In some cases, the project team may want to make decisions on a case-by-case basis using their expert knowledge about the Site.

It is noted that NADA packages developed by Practical Stats in R and for Minitab (e.g., Helsel, 2012]) compute 95% UCLs using simple z-statistic, t-statistic, and some bootstrap methods (e.g., standard or percentile bootstrap). For moderately skewed to highly skewed data sets, these simple 95% UCLs fail to provide the desired 95% coverage (confidence) to the population mean. Commercial software packages (e.g., SAS and Minitab, NADA) compute 95% UCLs of mean without automatically performing goodness-of-fit (GOF) tests and do not make any suggestions/recommendations about the use of a UCL. The users are expected to make their own selection of a UCL95 to estimate EPC terms. It is suggested that even for well-behaved (e.g., without outliers representing a single population) data sets, one should not use ProUCL (and other software packages) as a black box tool. One should use graphical displays (to identify potential patterns present in the data set), perform GOF tests and outlier tests before computing decision making statistics (e.g., UCL, UPL, and UTL). Once the project team has made a decision about the disposition (include or not include) of identified outliers, one computes decision statistics based upon GOF test results. For complex data sets (e.g., with outliers, multiple populations, negative values, and/or nondetects), it is advised to use expert advice of qualified statisticians.

For skewed to highly skewed nonparametric data sets, ProUCL computes and displays 95% UCLs applying bootstrap-t and Hall's bootstrap methods and using the Chebyshev inequality. UCLs based upon bootstrap-t and Hall's bootstrap methods tend to get distorted, resulting in elevated values, when outliers are present in the data set. Also, UCLs based upon all bootstrap methods become unreliable when negative values are present in a data set. When the data set is determined not to follow a normal, gamma, or lognormal distribution, the Student's *t*-UCL is recommended. In the absence of reasonable assumptions about the underlying distribution of the data, it is reasonable to use the CLT for UCL computations. After all, the UCL is an estimate of the mean, and the mean tends toward a normal distribution (Section 2.5.1).

1.15 The Unofficial User Guide to ProUCL4 (Helsel and Gilroy 2012)

Several ProUCL 4.1 users sent inquiries about the validity of the comments made about the ProUCL software in the Unofficial User Guide to ProUCL4 (Helsel and Gilroy, 2012) and in the Practical Stats webinar, "ProUCL v4: The Unofficial User Guide," presented by Dr. Helsel on October 15, 2012 (Helsel 2012a). Their inquiries led us to review comments made about the ProUCL4 software and its associated

guidance documents (EPA 2007, 2009a, 2009b, 2010c, 2010d, and 2011) in the “The Unofficial Users Guide to ProUCL4” and in the webinar, "ProUCL v4: The Unofficial User Guide". These two documents collectively are referred to as the Unofficial ProUCLv4 User Guide in this ProUCL document. The pdf document describing the material presented in the Practical Stats Webinar (Helsel 2012a) was downloaded from the http://www.practicalstats.com website.

In the "ProUCL v4: The Unofficial User Guide", comments have been made about the software and its guidance documents, therefore, it is appropriate to address those comments in the present ProUCL guidance document. It is necessary to provide the detailed response to assure that: 1) rigorous statistical methods are used to compute decision making statistics; and 2) the methods incorporated in ProUCL software are not misrepresented and misinterpreted. Some general responses and comments about the material presented in the webinar and in the Unofficial User Guide to ProUCLv4 are described as follows. Specific comments and responses are also considered in the respective chapters of ProUCL guidance documents.

Note: It is noted that the Kindle version of "ProUCL v4: Unofficial User Guide" is no longer available on Amazon. Several incorrect theoretical statements and statements misrepresenting ProUCL 4 were made in that Unofficial User Guide; therefore, a brief response to some of those statements has been retained in ProUCL guidance documents.

ProUCL is a freeware software package which has been developed under limited government funding to address statistical issues associated with various environmental site projects. Not all statistical methods (e.g., Levene test) described in the statistical literature have been incorporated in ProUCL. One should not compare ProUCL with commercial software packages which are expensive and not as user-friendly as the ProUCL software when addressing environmental statistical issues. The existing and some new statistical methods based upon the research conducted by ORD-NERL, EPA Las Vegas during the last couple of decades have been incorporated in ProUCL to address the statistical needs of various environmental site projects and research studies. Some of those new methods may not be available in text books, in the library of programs written in R-script, and in commercial software packages. However, those methods are described in detail in the cited published literature and also in the ProUCL Technical Guides (e.g., EPA 2007, 2009a, 2009b, 2010c, 2010d, and 2011). Even though for uncensored data sets, programs which compute gamma distribution based UCLs and UPLs are available in R Script, programs which compute a 95% UCL of mean based upon a gamma distribution on KM estimates are not as easily available.

In the Unofficial ProUCL v4 User Guide, several statements have been made about percentiles. There are several ways to compute percentiles. Percentiles computed by ProUCL may or may not be identical (don't have to be) to percentiles computed by NADA for R (Helsel 2013) or described in Helsel and Gilroy (2012). To address users' requests, ProUCL 4.1 (2011) and its higher versions compute percentiles that are comparable to the percentiles computed by Excel 2003 and higher versions.

The literature search suggests that there are a total of nine (9) known types of percentiles, i.e., 9 different methods of calculating percentiles in statistics literature (Hyndman and Fan, 1996). The R programming language (R Core Team 2012) computes percentiles using those 9 methods using the following statement in R

Quantile (x, p, type=k) where p = percentile, k = integer between 1 - 9

ProUCL computes percentiles using Type 7; Minitab 16 and SPSS compute percentiles using Type 6. It is simply a matter of choice, as there is no 'best' type to use. Many software packages use one type for calculating a percentile, and another for generating a box plot (Hyndman and Fan 1996).

An incorrect statement "*By definition, the sample mean has a 50% chance of being below the true population mean"* has been made in Helsel and Gilroy (2012) and also in Helsel (2012a). The above statement is not correct for means of skewed distributions (e.g., lognormal or gamma) commonly occurring in environmental applications. Since Helsel (2012b) prefers to use a lognormal distribution, the incorrectness of the above statement has been illustrated using a lognormal distribution. The mean and median of a lognormal distribution (details in Section 2.3.2 of Chapter 2 of this Technical Guide) are given by:

$$\text{mean} = \mu_1 = e^{(\mu + 0.5 \cdot \sigma^2)}$$

$$\text{median} = M = e^{\mu}$$

From the above equations, it is clear that the mean of a lognormal distribution is always greater than the median for all positive values of σ (*sd* of log-transformed variable). Actually the mean is greater than the p^{th} percentile when $\sigma > 2z_p$. For example, when $p = 0.80$, $z_p = 0.845$, and mean of a lognormal distribution, μ_1 exceeds $x_{0.80}$, the 80th percentile when $\sigma > 1.69$. In other words, when $\sigma > 1.69$ the lognormal mean will exceed the 80th percentile of a lognormal distribution. Here z_p represents the p^{th} percentile of the standard normal distribution (SND) with mean 0 and variance 1.

To demonstrate the incorrectness of the above statement, a small simulation study was conducted. The distribution of sample means based upon samples of size 100 were generated from lognormal distributions with $\mu = 4$, and varying skewness. The experiment was performed 10,000 times to generate the distributions of sample means. The probabilities of sample means less than the population means were computed. The following results are noted.

Table 1-6. Probabilities $P(\bar{x} < \mu)$ Computed for Lognormal Distributions with $\mu = 4$ and Varying Values of σ

Parameter	**μ=4, σ=0.5** **μ_1=61.86** **σ_1=32.97**	**μ=4, σ=1** **μ_1=90.017** **σ_1=117.997**	**μ=4,σ=1.5** **μ_1=168.17** **σ_1=489.95**	**μ=4,σ=2** **μ_1=403.43** **σ_1=2953.53**	**μ=4,σ=2.5** **μ_1=1242.65,** **σ_1=28255.23**
$p(\bar{x} < \mu_1)$	0.519	0.537	0.571	0.651	0.729
Mean	61.835	89.847	168.70	405.657	1193.67
Median	61.723	89.003	160.81	344.44	832.189

*Results are based upon 10000 Simulation Runs for Each Lognormal Distribution Considered

The probabilities summarized in the above table demonstrate that the statement about the mean made in Helsel and Gilroy (2012) is incorrect.

Graphical Methods: Graphical methods are available in ProUCL as exploratory tools which can be generated for both uncensored and left-censored data sets. Exploratory graphical methods are used to understand possible patterns present in data sets and not to compute statistics used in the decision making process. The Unofficial ProUCL Guide makes several comments about box plots and Q-Q plots incorporated in ProUCL. The Unofficial ProUCL Guide states that all graphs with NDs are incorrect. These statements are misleading and incorrect. The intent of the graphical methods in ProUCL is exploratory for the purpose of gaining information (e.g., outliers, multiple populations, data distribution, patterns, and skewness) about a data set. Based upon the data displayed (ProUCL displays a message [e.g., as a sub-title] in this regard) on those graphs, all statistics shown on those graphs generated by ProUCL are correct.

Box Plots: In statistical literature, one can find several ways to generate box plots. The practitioners may have their own preferences to use one method over the other. All box plot methods including the one in ProUCL convey the same information about the data set (outliers, mean, median, symmetry, skewness). ProUCL uses a couple of development tools such as FarPoint spread (for Excel type input and output operations) and ChartFx (for graphical displays); and ProUCL generates box plots using the built-in box plot feature in ChartFx. For all practical and exploratory purposes, box plots in ProUCL are equally good (if not better) as those available in the various commercial software packages, for examining data distribution (skewed or symmetric), identifying outliers, and comparing multiple groups (main objectives of box plots in ProUCL).

As mentioned earlier, it is a matter of choice of using percentiles/quartiles to construct a box plot. There is no 'best' method for constructing a box plot. Many software packages use one method (out of 9 as specified above) for calculating a percentile, and another for constructing a box plot (Hyndman and Fan 1996).

Q-Q plots: All Q-Q plots incorporated in ProUCL are correct and of high quality. In addition to identifying outliers, Q-Q plots are also used to assess data distributions. Multiple Q-Q plots are useful for performing point-by-point comparisons of grouped data sets, unlike box plots based upon the five-point summary statistics. ProUCL has Q-Q plots for normal, lognormal, and gamma distributions - not all of these graphical capabilities are directly available in other software packages such as NADA for R (Helsel 2013). ProUCL offers several exploratory options for generating Q-Q plots for data sets with NDs. Only detected outlying observations may require additional investigation; therefore, from an exploratory point of view, ProUCL can generate Q-Q plots excluding all NDs (and other options). Under this scenario there is no need to retain place holders (computing plotting positions used to impute NDs) as the objective is not to impute NDs. To impute NDs, ProUCL uses ROS methods (Gamma ROS and log ROS) requiring place holders; and ProUCL computes plotting positions for all detects and NDs to generate a proper regression model which is used to impute NDs. Also for comparison purposes, ProUCL can be used to generate Q-Q plots on data sets obtained by replacing NDs by their respective DLs or DL/2s. In these cases, no NDs are imputed, and there is no need to retain placeholders for NDs. On these Q-Q plots, ProUCL displays some relevant statistics which are computed based upon the data displayed on those graphs.

Helsel (2012a) states that the **Summary Statistics** module does not display KM estimates and that statistics based upon logged data are useless. Typically, estimates computed after processing the data do not represent summary statistics. Therefore, KM and ROS estimates are not displayed in the **Summary Statistics** module. These statistics are available in several other modules including the UCL and BTV modules. At the request of several users, summary statistics are computed based upon logged data. It is believed that the

mean, median, or standard deviation of logged data do provide useful information about data skewness and data variability.

To test for the equality of variances, the F-test, as incorporated in ProUCL, performs fairly well and the inclusion of the Levene's (1960) test will not add any new capability to the ProUCL software. Therefore, taking budget constraints into consideration, Levene's test has not been incorporated in the ProUCL software.

Although it makes sense to first determine if the two variances are equal or unequal, this is not a requirement to perform a t-test. The t-distribution based confidence interval or test for μ_1 - μ_2 based on the pooled sample variance does not perform better than the approximate confidence intervals based upon Satterthwaite's test. Hence testing for the equality of variances is not required to perform a two-sample t-test. The use of Welch-Satterthwaite's or Cochran's method is recommended in all situations (see Hayes 2005).

Helsel (2012a) suggests that imputed NDs should not be made available to the users. The developers of ProUCL and other researchers like to have access to imputed NDs. As a researcher, for exploratory purposes only, one may want to have access to imputed NDs to be used in exploratory advanced methods such as multivariate methods including data mining, cluster and principal component analyses. It is noted that one cannot easily perform exploratory methods on multivariate data sets with NDs. The availability of imputed NDs makes it possible for researchers and scientists to identify potential patterns present in complex multivariate data by using data mining exploratory methods on those multivariate data sets with NDs. Additional discussion on this topic is considered in Chapter 4 of this Technical Guide.

The statements summarized above should not be misinterpreted. One may not use parametric hypothesis tests such as a t-test or a classical ANOVA on data sets consisting of imputed NDs. These methods require further investigation as the decision errors associated with such methods remain unquantified. There are other methods such as the Gehan and T-W tests in ProUCL which are better suited to perform two-sample hypothesis tests using data sets with multiple detection limits.

Outliers: Helsel (2012a) and Helsel and Gilroy (2012) make several comments about outliers. The philosophy (with input from EPA scientists) of the developers of ProUCL about the outliers in environmental applications is that those outliers (unless they represent typographical errors) may potentially represent impacted (site related or otherwise) locations or monitoring wells or naturally-occuring background that differs naturally in non-impacted areas, and therefore may require further investigation.

The presence of outliers in a data set tends to destroy the normality of the data set. In other words, a data set with outliers can seldom (may be when outliers are mild, lying around the border of the central and tail parts of a normal distribution) follow a normal distribution. There are modern robust and resistant outlier identification methods (e.g., Rousseeuw and Leroy 1987; Singh and Nocerino 1995) which are better suited to identify outliers present in a data set; several of those robust outlier identification methods are available in the Scout 2008 version 1.0 (EPA 2009d) software package.

For both Rosner and Dixon tests, it is the data set (also called the main body of the data set) obtained after removing the outliers (and not the data set with outliers) that needs to follow a normal distribution (Barnett and Lewis 1994). Outliers are not known in advance. ProUCL has normal Q-Q plots which can be used to get an idea about potential outliers (or mixture populations) present in a data set. However, since a

lognormal model tends to accommodate outliers, a data set with outliers can follow a lognormal distribution; this does not imply that the outlier which may actually represent an impacted/unusual location does not exist! In environmental applications, outlier tests should be performed on raw data sets, as the cleanup decisions need to be made based upon values in the raw scale and not in log-scale or some other transformed space. More discussion about outliers can be found in Chapter 7 of this Technical Guide.

In Helsel (2012a), it is stated, "Mathematically, the lognormal is simpler and easier to interpret than the gamma (opinion)." We do agree with the opinion that the lognormal is simpler and easier to use but the log-transformation is often misunderstood and hence incorrectly used and interpreted. Numerous examples (e.g., Example 2-1 and 2-2, Chapter 2) are provided in the ProUCL guidance documents illustrating the advantages of the using a gamma distribution.

It is further stated in Helsel (2012a) that ProUCL prefers the gamma distribution because it downplays outliers as compared to the lognormal. This argument can be turned around - in other words, one can say that the lognormal is preferred by practitioners who want to inflate the effect of the outlier. Setting this argument aside, we prefer the gamma distribution as it does not transform the variable so the results are in the same scale as the collected data set. As mentioned earlier, log-transformation does appear to be simpler but problems arise when practitioners are not aware of the pitfalls (e.g., Singh and Ananda 2002; Singh, Singh, and Iaci 2002) associated with the use of lognormal distribution.

Helsel (2012a) and Helsel and Gilroy (2012) state that "lognormal and gamma are similar, so usually if one is considered possible, so is the other." This is another incorrect and misleading statement; there are significant differences in the two distributions and in their mathematical properties. Based upon the extensive experience in environmental statistics and published literature, for skewed data sets that follow both lognormal and gamma distributions, the developers favor the use of the gamma distribution over the lognormal distribution. The use of the gamma distribution based decision statistics is preferred to estimate the environmental parameters (mean, upper percentile). A lognormal model tends to hide contamination by accommodating outliers and multiple populations whereas a gamma distribution adjusts for skewness but tends not to accommodate contamination (elevated values) as can be seen in Examples 2-1 and 2-2 of Chapter 2 of this Technical Guide. The use of the lognormal distribution on a data set with outliers tends to yield inflated and distorted estimates which may not be protective of human health and the environment; this is especially true for skewed data sets of small of sizes <20-30; the sample size requirement increases with skewness.

In the context of computing a UCL95 of mean, Helsel and Gilroy (2012) and Helsel (2012a) state that GROS and LROS methods are probably never better than the KM method. It should be noted that these three estimation methods compute estimates of mean and standard deviation and not the upper limits used to estimate EPCs and BTVs. The use of the KM method does yield good estimates of the mean and standard deviation as noted by Singh, Maichle, and Lee (2006). The problem of estimating mean and standard deviation for data sets with nondetects has been studied by many researchers as described in Chapter 4 of this document. Computing good estimates of mean and *sd* based upon left-censored data sets addresses only half of the problem. The main issue is to compute decision statistics (UCL, UPL, UTL) which account for uncertainty and data skewness inherently present in environmental data sets.

Realizing that for skewed data sets, Student's t-UCL, CLT-UCL, and standard and percentile bootstrap UCLs do not provide the specified coverage to the population mean for uncensored data sets, many researchers (e.g., Johnson 1978; Chen 1995; Efron and Tibshirani 1993; Hall [1988, 1992]; Grice and Bain 1980; Singh, Singh, and Engelhardt 1997; Singh, Singh, and Iaci 2002) developed parametric (e.g., gamma) and nonparametric (e.g., bootstrap-t and Hall's bootstrap method, modified-t, and Chebyshev inequality) methods for computing confidence intervals and upper limits which adjust for data skewness. One cannot ignore the work and findings of the researchers cited above, and assume that Student's t-statistic based upper limits or percentile bootstrap method based upper limits can be used for all data sets with varying skewness and sample sizes.

Analytically, it is not feasible to compare the various estimation and UCL computation methods for skewed data sets containing ND observations. Instead, researchers use simulation experiments to learn about the distributions and performances of the various statistics (e.g., KM-t-UCL, KM-percentile bootstrap UCL, KM-bootstrap-t UCL, KM-Gamma UCL). Based upon the suggestions made in published literature and findings summarized in Singh, Maichle, and Lee (2006), it is reasonable to state and assume that the findings of the simulation studies performed on uncensored skewed data sets comparing the performances of the various UCL computation methods can be extended to skewed left-censored data sets.

Like uncensored skewed data sets, for left-censored data sets, ProUCL has several parametric and nonparametric methods to compute UCLs and other limits which adjust for data skewness. Specifically, ProUCL uses KM estimates in gamma equations; in the bootstrap-t method, and in the Chebyshev inequality to compute upper limits for left-censored skewed data sets.

Helsel (2012a) states that ProUCL 4 is based upon presuppositions. It is emphasized that ProUCL does not make any suppositions in advance. Due to the poor performance of a lognormal model, as demonstrated in the literature and illustrated via examples throughout ProUCL guidance documents, the use of a gamma distribution is preferred when a data set can be modeled by a lognormal model and a gamma model. To provide the desired coverage (as close as possible) for the population mean, in earlier versions of ProUCL (version 3.0), in lieu of H-UCL, the use of Chebyshev UCL was suggested for moderately and highly skewed data sets. In later (3.00.02 and higher) versions of ProUCL, depending upon skewness and sample size, for gamma distributed data sets, the use of the gamma distribution was suggested for computing the UCL of the mean.

Upper limits (e.g., UCLs, UPLs, UTLs) computed using the Student's t statistic and percentile bootstrap method (Helsel 2012b, NADA for R, 2013) often fail to provide the desired coverage (e.g., 95% confidence coefficient) to the parameters (mean, percentile) of most of the skewed environmental populations. It is suggested that the practitioners compute the decision making statistics (e.g., UCLs, UTLs) by taking: data distribution; data set size; and data skewness into consideration. For uncensored and left-censored data sets, several such upper limits computation methods are available in ProUCL 5.2 and its earlier versions.

Contrary to the statements made in Helsel and Gilroy (2012), ProUCL software does not favor statistics which yield higher (e.g., nonparametric Chebyshev UCL) or lower (e.g., preferring the use of a gamma distribution to using a lognormal distribution) estimates of the environmental parameters (e.g., EPC and BTVs). The main objectives of the ProUCL software funded by the U.S. EPA is to compute rigorous

decision statistics to help the decision makers and project teams in making sound decisions which are cost-effective and protective of human health and the environment.

Cautionary Note: Practitioners and scientists are cautioned about: 1) the suggestions made about the computations of upper limits described in some recent environmental literature such as the NADA books (Helsel [2005, 2012]); and 2) the misleading comments made about the ProUCL software in the training courses offered by Practical Stats during 2012 and 2013. Unfortunately, comments about ProUCL made by Practical Stats during their training courses lack professionalism and theoretical accuracy. It is noted that NADA packages in R and Minitab (2013) developed by Practical Stats do not offer methods which can be used to compute reliable or accurate decision statistics for skewed data sets. Decision statistics (e.g., UCLs, UTLs, UPLs) computed using the methods (e.g., UCLs computed using percentile bootstrap, and KM and LROS estimates and t-critical values) described in the NADA books and incorporated in NADA packages do not take data distribution and data skewness into consideration. The use of statistics suggested in NADA books and in Practical Stats training sessions often fail to provide the desired specified coverage to environmental parameters of interest for moderately skewed to highly skewed populations. Conclusions derived based upon those statistics may lead to incorrect conclusions which may not be cost-effective or protective of human health and the environment.

Page 75 (Helsel [2012]): One of the reviewers of the ProUCL 5.0 software drew our attention to the following incorrect statement made on page 75 of Helsel (2012b):

"If there is only 1 reporting limit, the result is that the mean is identical to a substitution of the reporting limit for censored observations."

An example of a left-censored data set containing ND observations with one reporting limit of 20 which illustrates this issue is described as follows.

Table 1-1. Example of a Left-Censored Data Set with a Single Reporting Limit.

Y	D_y
20	0
20	0
20	0
7	1
58	1
92	1
100	1
72	1
11	1
27	1

The mean and standard deviation based upon the KM and two substitution methods: DL/2 and DL are summarized as follows:

Kaplan-Meier (KM) Statistics	
Mean	39.4
SD	35.56

DL Substitution method (replacing censored values by the reporting limit)

Mean	42.7
SD	34.77

DL/2 Substitution method (replacing NDs by the reporting limit)

Mean	39.7
SD	37.19

The above example illustrates that the KM mean (when only 1 detection limit is present) is not actually identical to the mean estimate obtained using the substitution, DL (RL) method. The statement made in Helsel's text (and also incorrectly made in his presentations such as the one made at the U.S. EPA 2007 National Association of Regional Project Managers (NARPM) Annual Conference:

http://www.ttemidev.com/narpm2007Admin/conference/)

holds only when all observations reported as detects are greater than the single reporting limit, which is not always true for environmental data sets consisting of analytical concentrations.

1.16 Box and Whisker Plots

At the request of ProUCL users, a brief description of box plots (also known as box and whisker plots) as developed by Tukey (Hoaglin, Mosteller and Tukey 1983) is provided in this section. A box and whiskers plot represents a useful and convenient *exploratory* tool and provides a quick five point summary of a data set. In statistical literature, one can find several ways to generate box plots. The practitioners may have their own preferences to use one method over the other. Box plots are well documented in the statistical literature and description of box plots can be easily obtained by surfing the net. Therefore, the detailed description about the generation of box plots is not provided in ProUCL guidance documents. ProUCL also generates box plots for data set with NDs. Since box plots are used for exploratory purposes to identify outliers and also to compare concentrations of two or more groups, it does not really matter how NDs are displayed on those box plots. ProUCL generates box plots using detection limits and draws a horizontal line at the highest detection limit. Users can draw up to four horizontal lines at other levels (e.g., a screening level, a BTV, or an average) of their choice.

All box plot methods, including the one in ProUCL, represent five-point summary graphs including: the lowest and the highest data values, median (50^{th} percentile=second quartile, Q2), 25^{th} percentile (lower quartile, Q1), and 75^{th} percentile (upper quartile, Q3). A box and whisker plot also provides information about the degree of dispersion (interquartile range (IQR) = Q3-Q1=length/height of the box in a box plot),

the degree of skewness (suggested by the length of the whiskers) and unusual data values known as outliers. Specifically, ProUCL (and other software packages) use the following to generate a box and whisker plot.

- Q1= 25th percentile, Q2= 50th (median), and Q3 = 75th percentile
- Interquartile range= IQR = Q3-Q1 (the length/height of the box in a box plot)
- Lower whisker starts at Q1 and the upper whisker starts at Q3.
- Lower whisker extends up to the lowest observation or (Q1 - 1.5 * IQR) whichever is higher
- Upper whisker extends up to the highest observation or (Q3 + 1.5 * IQR) whichever is lower
- Horizontal bars (also known as fences) are drawn at the end of whiskers
- Guidance in statistical literature suggests that observations lying outside the fences (above the upper bar and below the lower bar) are considered potential outliers

An example box plot generated by ProUCL is shown in the following graph.

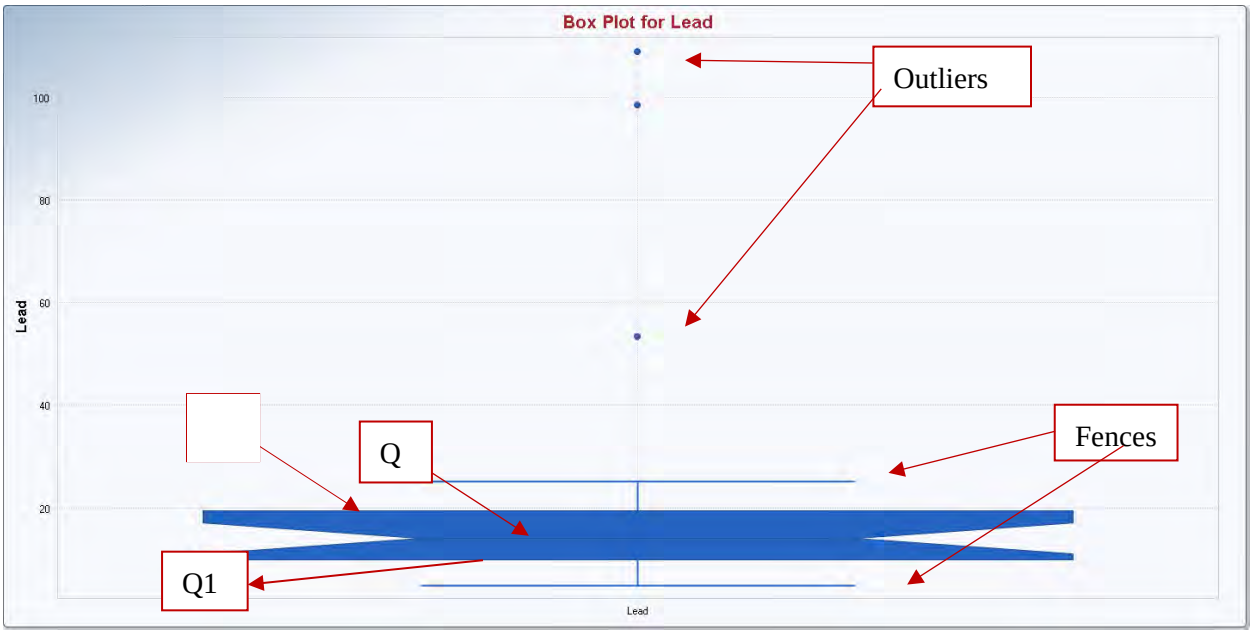

Figure 1-1. Box Plot with Fences and Outlier

It should be pointed out that the use of box plots in different scales (e.g., raw-scale and log-scale) may lead to different conclusions about outliers. Below is an example illustrating this issue.

Example 1-2. Consider an actual data set consisting of copper concentrations collected a Superfund Site. The data set is: 0.83, 0.87, 0.9, 1, 1, 2, 2, 2.18, 2.73, 5, 7, 15, 22, 46, 87.6, 92.2, 740, and 2960. Box plots using data in the raw-scale and log-scale are shown in Figures 1-2 and 1-3.

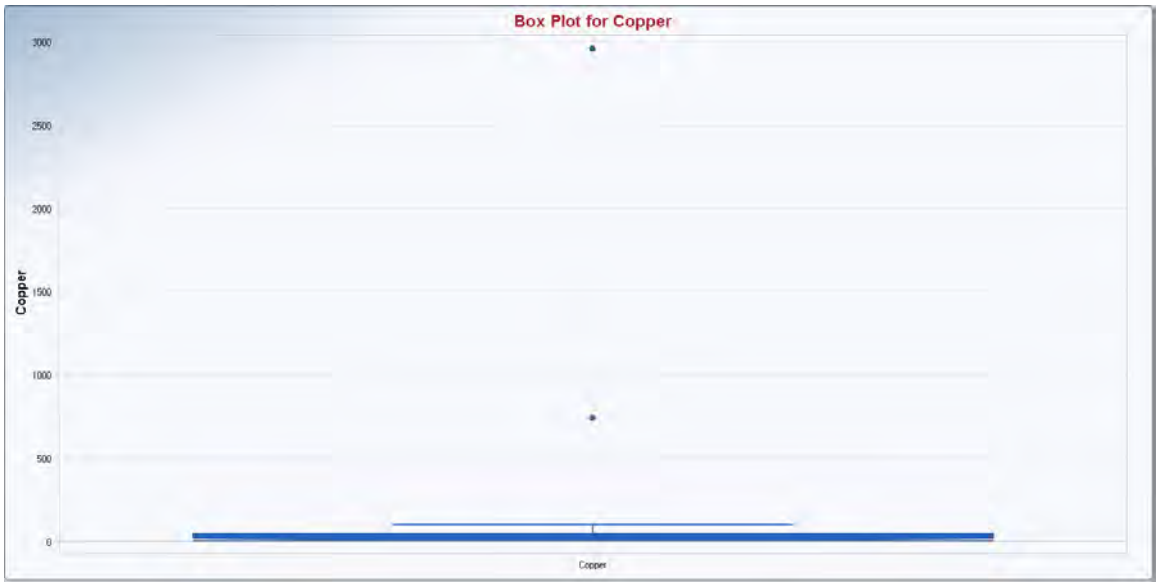

Figure 1-2. Box Plot of Raw Data in Original Scale

Based upon the last bullet point of the description of box plots described above, from Figure 1-1, it is concluded that two observations 740 and 2960 in the raw scale represent outliers.

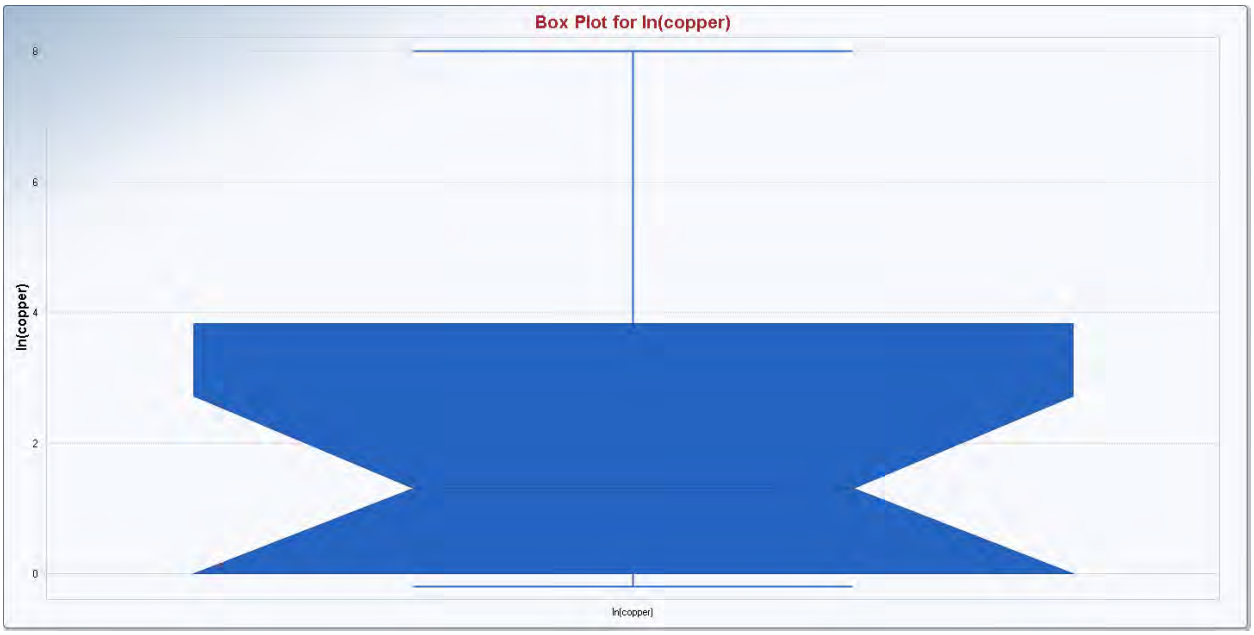

Figure 1-3. Box Plot of Data in Log-Scale

However, based upon the last bullet point about box plots, from Figure 1-3, it is concluded that two observations 740 and 2960 in the log-scale do not represent outliers. The log-transformation has accommodated the two outliers.

This is a great example demonstrating the importance of examining data distribution. Since data set is skewed, extreme values showed up as outliers in raw scale, but not in log-scale.

Note: ProUCL uses a couple of development tools such as SpreedNET and StudioFX for Excel type input and output operations and for graphical displays. ProUCL generates box plots using the built-in box plot feature in ChartFx. The programmer has no control over computing various statistics (e.g., Q1, Q2, Q3, IQR) using ChartFx. So box plots generated by ProUCL can differ slightly from box plots generated by other programs (e.g., Excel). However, for all practical and exploratory purposes, box plots in ProUCL are equally good (if not better) as available in the various commercial software packages for investigating data distribution (skewed or symmetric), identifying outliers, and comparing multiple groups (main objectives of box plots).

Precision in Box Plots: Box plots generated using ChartFx round values to the nearest integer. For increased precision of graphical displays (all graphical displays generated by ProUCL), the user can use the process described as follows.

Position your cursor on the graph and right-click, a popup menu will appear. Position the cursor on **Properties** and right-click; a windows form labeled **Properties** will appear. There are three choices at the top: **General**, **Series** and **Y-Axis**. Position the e cursor over the **Y-Axis** choice and left-click. You can change the number of decimals to increase the precision, change the step to increase or decrease the number Y-Axis values displayed and/or change the direction of the label. To show values on the plot itself, position your cursor on the graph and right-click; a pop-up menu will appear. Position the cursor on **Point Labels** and right-click. There are other options available in this pop-up menu including changing font sizes.

CHAPTER 2

Goodness-of-Fit Tests and Methods to Compute Upper Confidence Limit of Mean for Uncensored Data Sets without Nondetect Observations

2.1 Introduction

Many environmental decisions including exposure and risk assessment and management, and cleanup decisions are made based upon the mean concentrations of the contaminants/constituents of potential COPCs. To address the uncertainty associated with the sample mean, a UCL95 is used to estimate the unknown population mean, μ_1. A UCL95 is routinely used to estimate the EPC) term (EPA 1992a; EPA 2002a). A UCL95 of the mean represents that limit such that one can be 95% confident that the population mean, μ_1, will be less than that limit. From a risk point of view, a UCL95 of the mean represents a number that is considered health protective when used to compute risk and health hazards. Since, many environmental decisions are made based upon a UCL95, it is important to compute a reliable, defensible (from human health point of view) and cost-effective estimate of the EPC. To compute reliable estimates of practical merit, ProUCL software provides several parametric and nonparametric UCL computation methods covering a wide-range of skewed distributions (e.g., symmetric, mildly skewed to highly skewed) for data sets of various sizes. Based upon simulation results summarized in the literature (Singh, Singh, and Engelhardt [1997], Singh, Singh and Iaci [2002]), data distribution, data set size, and skewness, ProUCL makes suggestions on how to select an appropriate UCL95 of the mean to estimate the EPC. It should be noted that a simulation study cannot cover all possible real world data sets of various sizes and skewness following different probability distributions. This ProUCL Technical Guide provides sufficient guidance to help a user select the most appropriate UCL as an estimate of the EPC. The ProUCL software makes suggestions to help a typical user select an appropriate UCL from all the UCLs incorporated in ProUCL and those available in the statistical literature. UCL values, other than those suggested by ProUCL, may be selected based upon project personnel's experiences and project needs. The user may want to consult a statistician before selecting an appropriate UCL95.

The ITRC (2012 and 2020) regulatory documents recommend the use of a Student's t-UCL95 and Chebyshev inequality based UCL95 to estimate EPCs for ISM based soil samples collected from DUs. The Chebyshev UCL is the recommended nonparametric method for ISM data due to the fact that bootstrapping methods are typically unrealible for small sample sizes typical in ISM designs. It is also not possible to confirm samples follow a particular distribution; however, ISM samples are estimates of the average concentration and so can be assumed to follow a normal distribution in many cases. In order to facilitate the computation of ISM data-based estimates of the EPC, ProUCL5.2 /5.1/5.0 can compute a UCL95 of the mean based upon data sets of sizes as small as 3. Additionally, the UCL module of ProUCL can be used on ISM-based data sets with NDs.

However, it is advised that the users do not compute decision making statistics (e.g., UCLs, upper prediction limits [UPLs], upper tolerance limits [UTLs]) from discrete data sets consisting of less than 8-10 observations.

For uncensored data sets without ND observations, theoretical details of the Student's t- and percentile bootstrap UCL computation methods, as well as the more complicated bootstrap-t and gamma distribution methods, are described in this Chapter. One should not ignore the use of gamma distribution based UCLs (and other upper limits) just because it is easier to use a lognormal distribution. Typically, environmental data sets are positively skewed, and a default lognormal distribution (EPA 1992a) is used to model such data distributions. Additionally, an H-statistic based Land's (Land, 1971, 1975) H-UCL is then typically used to estimate the EPC. Hardin and Gilbert (1993), Singh, Singh, and Engelhardt (1997, 1999), Schultz and Griffin (1999), and Singh, Singh, and Iaci (2002) pointed out several problems associated with the use of the lognormal distribution and the H-statistic to compute UCL of the mean. For lognormal data sets with high standard deviation (*sd*), σ, of the natural log-transformed data (e.g., σ exceeding 1.0 to 1.5), the H-UCL becomes unacceptably large, exceeding the 95% and 99% data quantiles, and even the maximum observed concentration, by orders of magnitude (Singh, Singh, and Engelhardt 1997). The H-UCL is also very sensitive to a few low or a few high values. For example, the addition of a single low measurement can cause the H-UCL to increase by a large amount (Singh, Singh, and Iaci, 2002) by increasing variability. Realizing that the use of the H-statistic can result in an unreasonably large UCL, it has been recommended (EPA 1992a) that the maximum value be used as an estimate of the EPC in cases when the H-UCL exceeds the largest value in the data set. For uncensored data sets without any NDs, ProUCL makes suggestions/recommendations on how to compute an appropriate UCL95 based upon data set size, data skewness and distribution.

In practice, many skewed data sets follow a lognormal as well as a gamma distribution. Singh, Singh, and Iaci (2002) observed that UCLs based upon a gamma distribution yield reliable and stable values of practical merit. It is, therefore, desirable to test whether an environmental data set follows a gamma distribution. A gamma distribution based UCL95 of the mean provides approximately 95% coverage to the population mean, $\mu_1 = k\theta$ of a gamma distribution, G (k, θ) with k and θ respectively representing the shape and scale parameters. For data sets following a gamma distribution with shape parameter, $k > 1$, the EPC should be estimated using an adjusted gamma (when $n<50$) or approximate gamma (when $n\geq 50$) UCL95 of the mean. For highly skewed gamma distributed data sets with values of the shape parameter, $k \leq 1.0$, a 95% UCL may be computed using the bootstrap-t-method or Hall's bootstrap method when the sample size, n, is smaller, such as <15 to 20. For larger sample sizes with $n> 20$, a UCL of the mean may be computed using the adjusted or approximate gamma UCL (Singh, Singh, and Iaci 2002) computation method. Based upon professional judgment and practical experience of the authors, some of these suggestions have been modified. Specifically, in earlier versions ProUCL, the cutoff value for the shape parameter, k was 0.1 which has been changed to 1.0 in this version.

Unlike the percentile bootstrap method, bootstrap-t and Hall's bootstrap methods (Efron and Tibshirani, 1993) account for data skewness and their use is recommended on skewed data sets when computing UCLs of the mean. However, the 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. If the results of these two methods are much higher than the rest of the UCL computation methods, it could be an indication of erratic behavior of these two bootstrap UCL computation methods. ProUCL prints out a warning message whenever the use of these two bootstrap methods is recommended.

ProUCL has graphical (e.g., quantile-quantile [Q-Q] plots) and formal goodness-of-fit (GOF) tests for normal, lognormal, and gamma distributions. These GOF tests are available for data sets with and without NDs. The critical values of the Anderson-Darling (A-D) test statistic and the Kolmogorov-Smirnov (K-S) test statistic to test for gamma distributions were generated using Monte Carlo simulation experiments (Singh, Singh, and Iaci 2002). Those critical values have been incorporated in ProUCL software and are tabulated in Appendix A for various levels of significance.

ProUCL computes summary statistics for raw, as well as, log-transformed data sets with and without ND observations. In this Technical Guide and in ProUCL software, log-transformation (*log*) stands for the natural logarithm (*ln, LN*) or log to the base *e*. For uncensored data sets, mathematical algorithms and formulae used in ProUCL to compute the various UCLs are summarized in this chapter. ProUCL also computes the maximum likelihood estimates (MLEs*)* and the minimum variance unbiased estimates (MVUEs) of the population parameters of normal, lognormal, and gamma distributions. Nonparametric UCL computation methods in ProUCL include: central limit theorem (CLT), adjusted-CLT, modified Student's t (adjusts for skewness) Chebyshev inequality, and bootstrap methods. Moreover, it is noted that UCLs based upon the standard bootstrap and the percentile bootstrap methods do not perform well by not providing the specified coverage of the mean for skewed data sets.

Note on Computing Lower Confidence Limits (LCLs) of Mean: For some environmental projects an LCL of the unknown population mean is needed to achieve the project DQOs. At present, ProUCL does not directly compute LCLs of mean. However, for data sets with and without nondetects, excluding the bootstrap methods, gamma distribution, and H-statistic based LCLs of mean, the same critical value (e.g., normal z value, Chebyshev critical value, or t-critical value) can be used to compute a LCL of mean as used in the computation of the UCL of the mean. Specifically, to compute a LCL, the '+' sign used in the computation of the corresponding UCL needs to be replaced by the '-' sign in the equation used to compute that UCL (excluding gamma, lognormal H-statistic, and bootstrap methods). For specific details, the user may want to consult a statistician. For data sets *without nondetect* observations, the user may want to use the Scout 2008 software package (EPA 2009d, 2010) to directly compute the various parametric and nonparametric LCLs of mean.

2.2 Goodness-of-Fit (GOF) Tests

Let x_1, x_2, ..., x_n be a representative random sample (e.g., representing lead concentrations) from the underlying population (e.g., site areas under investigation) 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 chapter, 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 $Y = \log_e(X)$.

Three data distributions have been considered in ProUCL 5.2 (and in older versions). These include the normal, lognormal, and the gamma distributions. Shapiro-Wilk, for $n \leq 2000$, and Lilliefors (1967) test statistics are used to test for normality or lognormality of a data set. Lilliefors test (along with graphical Q-Q plot) seems to perform fairly well for samples of size 50 and higher. In ProUCL 5.2, updated critical values of Lilliefors test developed by Moling and Abdi (2007) and provided in the Encyclopedia of Measurement and Statistics have been used. The empirical distribution function (EDF) based methods, the K-S and A-D tests, are used to test for a gamma distribution. Extensive critical values for these two test statistics have been obtained via Monte Carlo simulation experiments (Singh, Singh, and Iaci 2002). For interested users, those critical values are given in Appendix A for various levels of significance. In addition to these formal tests, the informal histogram and Q-Q plots (also called probability plots) are also available for visual inspection of the data distributions (Looney and Gulledge 1985). Q-Q plots also provide useful information about the presence of potential outliers and multiple populations in a data set. A brief description of the GOF tests follows.

No matter which normality test is used, it may fail to detect the actual non-normality of the population distribution if the sample size is small, $n<20$ and with large sample sizes, $n>50$ or so, a small deviation from normality will lead to rejection of the normality hypothesis. The modified K-S test known as Lilliefors test is suggested for checking the normality assumption when the mean and *sd* of population distribution is not known.

2.2.1 Test Normality and Lognormality of a Data Set

ProUCL tests for normality and lognormality of a data set using three different methods described below. The program tests normality or lognormality at three different levels of significance, 0.01, 0.05, and 0.1 (or confidence levels: 0.99, 0.95, and 0.90). For normal distributions, ProUCL outputs approximate probability values (*p*-values) for the S-W GOF test. The details of those methods can be found in the cited references.

2.2.1.1 Normal Quantile-quantile (Q-Q) Plot

A normal Q-Q represents a graphical method to test for approximate normality or lognormality of a data set (Hoaglin, Mosteller, and Tukey 1983; Singh 1993; Looney and Gulledge, 1985). A linear pattern displayed by the majority of the data suggests approximate normality or lognormality (when performed on log-transformed data) of the data set. For example, a high value, 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, on this graphical display, observations well-separated from the linear pattern displayed by the majority of data may represent outlying observations not belonging to the dominant population, whose distribution one is assessing based upon a data set. Apparent jumps and breaks in the Q-Q plot may suggest the presence of multiple populations. The correlation of the Q-Q plot for such a data set may still be high but that does not signify that the data set follows a normal distribution.

Notes: Graphical displays provide added insight into a data set which might not be apparent based upon statistics such as S-W statistic or a correlation coefficient. The correlation coefficient of a Q-Q plot with curves, jumps and breaks can be high, which does not necessarily imply that the data follow a normal or

lognormal distribution. AGOF test of a data set should always be judged based upon a formal (e.g., S-W statistic) as well as informal graphical displays. The normal Q-Q plot is used as an exploratory tool to evaluate data distribution and potentially identify multiple populations. On all Q-Q plots, ProUCL displays relevant statistics including: mean, *sd,* GOF test statistic, associated critical value, *p*-value (when available), and the correlation coefficient.

There is no substitute for graphical displays of data sets. Graphical displays provide added insight about data sets and do not get distorted by outliers and/or mixture populations. The final conclusion regarding the data distribution should be based upon the formal GOF tests as wells as Q-Q plots. This statement is true for all GOF tests: normal, lognormal, and gamma distributions.

2.2.1.2 Shapiro-Wilk (S-W) Test

The S-W test is a powerful test used to test the normality or lognormality of a data set. ProUCL performs this test for samples of size up to 2000 (Royston 1982a, 1982b). For sample sizes ≤ 50, in addition to a test statistic and critical value, an approximate *p*-value associated with S-W test is also displayed. For samples of size >50, only the test statistics and approximate *p*-values are displayed (the critical value is not displayed). For the Shapiro-Wilk test, the test statistic must be below the critical value to be significant. Based upon the selected level of significance and the computed test statistic, ProUCL informs the user if the data set is normally (or lognormally) distributed. This information should be used to compute an appropriate UCL of the mean.

2.2.1.3 Lilliefors Test

This test is useful for data sets of larger size (Lilliefors 1967; Dudewicz and Misra 1988; Conover 1999). This test is a slight modification of the Kolmogorov-Smirnov (K-S) test and is more powerful than a one-sample K-S (with the estimated population mean and *sd*). In version 5.2 of ProUCL, critical values of Lilliefors test developed by Moling and Abdi and provided in the Encyclopedia of Measurement and Statistics (Salkind, N. Editor 2007) have been used and incorporated in the program. The critical values as described in Salkind (2007) are used for *n* up to 50, and for values of *n*>50 approximate critical values are computed using the following formula:

Critical Values = *Factor*/*f(n)*; where $f(n) = \frac{0.83+n}{\sqrt{n}} - 0.01$.

The Factor used in the above equation depends upon the level of significance, α; Factor values are 0.741, 0.819, 0.895, and 1.035 for α = 0.20, 0.1, 0.05, and 0.01 respectively.

Based upon the selected level of significance and the computed test statistic, ProUCL informs the user if the data set is normally or lognormally distributed. For the Lilliefors test, the test statistic must be above the critical value to be significant This information should be used to compute an appropriate UCL of the mean. The program outputs the relevant statistics on the Q-Q plot of data.

For convenience, normality, lognormality, or gamma distribution test results for a built-in level of significance of 0.05 are displayed on the UCL and background statistics output sheets. This helps the user in selecting the most appropriate UCL to estimate the EPC. It should be pointed out that sometimes, the

two GOF tests may lead to different conclusions. In such situations, ProUCL displays a message that data are approximately normally or lognormally distributed. It is suggested that the user makes a decision based upon the information provided by the associated Q-Q plot and the values of the GOF test statistics.

In an effort to streamline the decision process for computing upper limits (e.g., UCL95), some changes were made in the decision logic applied in ProUCL for suggesting/recommending UCL values. Specifically, ProUCL makes decisions about the data distribution based upon both the Lilliefors and S-W GOF test statistics for normal and lognormal distributions and both the A-D and K-S GOF test statistics for the gamma distribution. When a data set passes one of the two GOF tests for a distribution, ProUCL outputs a statement that the data set follows that approximate distribution and suggests using appropriate decision statistic(s). Specifically, when only one of the two GOF statistic leads to the conclusion that data are normal, lognormal or gamma, ProUCL outputs the conclusion that the data set follows that approximate distribution and all suggestions provided by ProUCL regarding the use of parametric or nonparametric decision statistics are made based upon this conclusion. As a result, UCLs suggested by ProUCL 5.2 can differ from the UCLs suggested by earlier versions of ProUCL. ProUCL 5.2 also contains changes to the significance levels for Lillefors and Shapiro-Wilk tests ($\alpha = 0.01$ for normality and $\alpha = 0.10$ for lognormality). Refer to Section 2.5.1 for a discussion of these changes.

Note: When dealing with a small data set, n <50, and Lilliefors test suggests that data are normal and the S-W test suggests that data are not normal, ProUCL will suggest that the data set follows an approximate normal distribution. However, for smaller data sets, Lilliefors test results may not be reliable, therefore the user is advised to review GOF tests for other distributions and proceed accordingly. It is emphasized, when a data set follows a distribution (e.g., distribution A) using all GOF tests, and also follows an approximate distribution (e.g., distribution B) using one of the available GOF tests, it is preferable to use distribution A over distribution B. However, when distribution A is a highly skewed (e.g., *sd* of logged data >1.0) lognormal distribution, use the guidance provided on the ProUCL generated output.

In practice, depending upon the power associated with statistical tests, two tests (e.g., two sample t-test vs. WMW test; S-W test vs. Lilliefors test) used to address the same statistical issue (comparing two groups, assessing data distribution) can lead to different conclusions (e.g., GOF tests for normality in Example 2-4); this is especially true when dealing with data sets of smaller sizes. The power of a test can be increased by collecting more data. If this is not feasible due to resource constraints, the collective project team should determine which conclusion to use in the decision making process. It may, in these cases, be appropriate to consult a statistician.

2.2.2 Gamma Distribution

A continuous random variable, X (e.g., concentration of an analyte), 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:

$$f(x;k,\theta)=\begin{cases}\frac{1}{\theta^k\Gamma(k)}\cdot x^{k-1}e^{\frac{-x}{\theta}}, & x>0\\ 0, & \text{otherwise}\end{cases} \qquad (2\text{-}3)$$

Many positively skewed data sets follow a lognormal as well as a gamma distribution. The use of a gamma distribution tends to yield reliable and stable 95% UCL values of practical merit. 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 data sets which follow a gamma distribution, the adjusted 95% UCL of the mean based upon a gamma distribution is optimal (Bain and Engelhardt 1991) and approximately provides the specified 95% coverage of the population mean, $\mu_1 = k\theta$ (Singh, Singh, and Iaci 2002).

The GOF test statistics for a gamma distribution are based upon the EDF. The two EDF tests incorporated in ProUCL are the K-S test and the A-D test, which are described in D'Agostino and Stephens (1986) and Stephens (1970). The graphical Q-Q plot for a gamma distribution has also been incorporated in ProUCL. The critical values for the two EDF tests are not available, especially when the shape parameter, *k*, is small ($k < 1$). Therefore, the associated critical values have been computed via extensive Monte Carlo simulation experiments (Singh, Singh, and Iaci 2002). The 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. The GOF tests for a gamma distribution depend upon the MLEs of the gamma parameters, *k* and θ, which should be computed before performing the GOF tests. Information about estimation of gamma parameters, gamma GOF tests, and construction of gamma Q-Q plots is not readily available in statistical textbooks. Therefore, a detailed description of the methods for a 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, $G(k,\theta)$; and 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; details of the computation of the MLEs of *k* and θ can be found in Singh, Singh, and Iaci (2002). The Q-Q plot for a gamma distribution is obtained by plotting the scatter plot of pairs, $(x_{0i}, x_{(i)})$ $i := 1,2,\ldots,n$. The gamma 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:

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

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) described in Best and Roberts (1975) has been used to compute the inverse chi-square percentage points given by equation (2-4). All relevant statistics including the MLE of *k* are also displayed on a gamma Q-Q plot.

Like a normal Q-Q plot, a linear pattern displayed by the majority of the data on a gamma Q-Q plot suggests that the data set follows 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 an 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. The correlation coefficient of a Q-Q plot with outliers and jumps can still be high which does not signify that the data follow a gamma distribution. Therefore, a graphical Q-Q plot and other formal

GOF tests, the A-D test or K-S test, should be used on the same data set to determine the distribution of a data set.

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

Let $F(x)$ be the cumulative distribution function (CDF) of a gamma distributed random variable, X. Let $Z = F(X)$, then Z represents a uniform U(0,1) random variable (Hogg and Craig 1995). 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 (Algorithm AS 239, Shea 1988) as given in the book *Numerical Recipes in C, the Art of Scientific Computing* (Press *et al.* 1990) has been used to compute the incomplete gamma function. Arrange the resulting z_i in ascending order as

$$z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(n)}. \text{ Let } \bar{z} = (\sum_{i=1}^{n} z_i)/n \text{ be the mean of the } n,\ z_i;\ i := 1,2,\ldots,n.$$

Compute the following two statistics:

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

The K-S test statistic is given by $D = max(D^+, D^-)$; and the A-D test statistic is given as follows:

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

As mentioned before, the critical values for these two statistics, D and A^2, are not readily available. For the A-D test, only the asymptotic critical values are available in the statistical literature (D'Agostino and Stephens 1986). Some raw critical values for the K-S test are given in Schneider (1978), and Schneider and Clickner (1976). Critical values of these test statistics are computed via Monte Carlo experiments (Singh, Singh, and Iaci 2002). It is 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 simulation experiments. In order to generate critical values, random samples from gamma distributions were generated using the algorithm as given in Whittaker (1974). It is observed that the simulated critical values are in close agreement with all available published critical values.

The critical values simulated by Singh, Singh, and Iaci (2002) for the two test statistics have been incorporated in the ProUCL software 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 set follows a gamma distribution is rejected. ProUCL computes the GOF test statistics and displays them on the gamma Q-Q plot and also on the UCL and background statistics output sheets generated by ProUCL. Like all other tests, in practice these two GOF test may lead to different conclusions. In such situations, ProUCL outputs a message that the data follow an approximate gamma distribution. The user should make a decision based upon the information provided by the associated gamma Q-Q plot and the values of the GOF test statistics.

Computation of the Gamma Distribution Based Decision Statistics and Critical Values: When computing the various decision statistics (e.g., UCL and BTVs), ProUCL uses biased corrected estimates, kstar, $\hat{k}^*$ and theta star, $\hat{\theta}^*$ (described in Section 2.3.3) of the shape, k, and scale, θ, parameters of the gamma distribution. It is noted that the critical values for the two gamma GOF tests reported in the literature (D'Agostino and

Stephens 1986; Schneider and Clickner 1976; and Schneider 1978) are computed using the MLE estimates, $\hat{k}$ and $\hat{\theta}$ of the two gamma parameters, k and θ. Therefore, the critical values of A-D and K-S tests incorporated in ProUCL have also been computed using the MLE estimates: khat, $\hat{k}$ and theta hat, $\hat{\theta}$ of the two gamma parameters, k and θ.

Updated Critical Values of Gamma GOF Test Statistics (New in ProUCL 5.0): For values of the gamma distribution shape parameter, $k \leq 0.1$, critical values of the two gamma GOF tests, A-D and K-S tests, have been updated in ProUCL 5.0 and higher versions. Critical values incorporated in earlier versions were simulated using the gamma deviate generation algorithm (Whittaker 1974) available at the time and with the source code described in the book *Numerical Recipes in C, the Art of Scientific Computing* (Press *et al.* 1990). Th gamma deviate generation algorithm available at the time was not very efficient especially for smaller values of the shape parameter, $k \leq 0.1$. For values of the shape parameter, $k \leq 0.1$, significant discrepancies were found in the critical values of the two gamma GOF test statistics obtained using the two gamma deviate generation algorithms: Whitaker (1974) and Marsaglia and Tsang (2000).

Therefore, for values of $k \leq 0.2$, critical values for the two gamma GOF tests have been re-generated and tables of critical values of the two gamma GOF tests have been updated in Appendix A. Specifically, for values of the shape parameter, $k \leq 0.1$, critical values of the two gamma GOF tests have been generated using the more efficient gamma deviate generation algorithm as described in Marsaglia and Tsang's (2000) and Best (1983). Detailed description about the implementation of Marsaglia and Tsang's algorithm to generate gamma deviates can be found in Kroese, Taimre, and Botev (2011). From a practical point of view, for values of k greater than 0.1, the simulated critical values obtained using Marsaglia and Tsang's algorithm (2000) are in general agreement with the critical values of the two GOF test statistics incorporated in ProUCL 4.1 for the various values of the sample size considered. Therefore, those critical values for values of $k > 0.1$ do not have to be updated.

Note: In March 2015 minor discrepancies were identified in critical values of the gamma GOF A-D tests, as summarized in Tables A1-A6 of ProUCL 5.0 Technical Guide. For example, for a specified sample size and level of significance, α, the critical values for GOF tests are expected to decrease as k increases. Due to inherent random variability in the simulated gamma data sets, critical values do not follow (deviations are minor occurring in 2nd or 3rd decimal places) this trend in a few cases. However, from a practical and decision making point of view those differences are minor (see below). These discrepancies can be eliminated by performing simulation experiments using more iterations. In ProUCL 5.1, these discrepancies in the critical values of gamma GOF tests have been fixed via interpolation.

For example, in Table A-3, for the A-D test, with significance level $\alpha = 0.05$ and $n=7$, critical values for $k=10$, 20, and 50 are 0.708, 0.707, and 0.708. Also, in Table A-4 for $n=200$ and $k=0.025$, the critical value is 0.070489, and for $n=200$, $k=0.05$, the critical value is 0.07466. Due to a lack of resources and time, the critical values have not been re-simulated; however, this value has been replaced by an interpolated value using simulated values for $k=0.025$ and $k=0.1$.

2.3 Estimation of Parameters of the Three Distributions Incorporated in ProUCL

Let μ_1 and σ_1^2 represent the mean and variance of the random variable, X, and μ and σ^2 represent the mean and variance of the random variable $Y = \log(X)$. Also, $\hat{\sigma}$ represents the standard deviation of the log-

transformed data. For both lognormal and gamma distributions, the associated random variable can take only positive values. It is typical of environmental data sets to consist of only positive concentrations.

2.3.1 Normal Distribution

Let *X* be a continuous random variable (e.g., lead concentrations in surface soils of a site), 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:

$$f(x;\mu_1,\sigma_1) = exp[-(x-\mu_1)^2/2\sigma_1^2]/(\sigma_1\sqrt{2\pi});\ -\infty < x < \infty \qquad (2\text{-}7)$$

For normally distributed data sets, it is well known (Hogg and Craig 1995) that the MVUEs of the mean, μ_1, and the variance, σ_1^2, are 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, based upon the Student's t-distribution is optimal. In practice, for normally distributed data sets, UCLs computed using Student's t-distribution, the modified t-distribution, and bootstrap-t method are in close agreement.

2.3.2 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:

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

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

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

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

$$\text{Skewness} = CV^3 + 3CV \qquad (2\text{-}12)$$

2.3.2.1 MLEs of the Parameters of a Lognormal Distribution

For lognormally distributed data sets, note that $\bar{y}$ and s_y ($=\hat{\sigma}$) are the MLEs of μ and σ, respectively. The MLE of any function of the parameters μ and σ^2 is obtained by substituting these MLEs in place of the parameters (Hogg and Craig 1995). 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 also printed on the Excel-type output spreadsheet generated by ProUCL.

2.3.2.2 Relationship between Skewness and Standard Deviation, σ

For a lognormal distribution, the CV (given by equation (2-11) above) and the skewness (given by equation (2-12)) depend only on σ. Therefore, in this Technical Guide and also in ProUCL software, the standard deviation, σ (s*d* of log-transformed variable, *Y*), or its MLE, s_y (=$\hat{\sigma}$), has been used as a measure of the skewness of lognormally distributed data sets and also of other data sets with positive values. The greater the *sd*, the greater are the CV and the skewness. For example, for a lognormal distribution with $\sigma = 0.5$, the skewness = 1.75; with σ =1.0, the skewness = 6.185; with $\sigma = 1.5$, the skewness = 33.468; and with σ = 2.0, the skewness = 414.36. The skewness of a lognormal distribution becomes unreasonably large as σ starts approaching and exceeding 1.5. 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; Singh, Singh, and Iaci 2002) that for smaller sample sizes (such as smaller than 50), and for values of σ or $\hat{\sigma}$ approaching and exceeding 1.5 to 1.75, the use of the H-statistic-based H-UCL results in impractical and unacceptably large values.

For positively skewed data sets, the various levels of skewness can be defined in terms σ or its MLE estimate, s_y. These levels are described as follows in Table 2-1. ProUCL software uses the sample sizes and skewness levels defined below to make suggestions/recommendations to select an appropriate UCL as an estimate of the EPC.

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

Standard Deviation of Logged Data	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
$2.0 \le \sigma < 3.0$	Very high skewness
$\sigma \ge 3.0$	Extremely high skewness

Note: *When data are mildly skewed with* $\sigma < 0.5$, the three distributions considered in ProUCL tend to yield comparable upper limits irrespective of the data distribution.

2.3.2.3 MLEs of the Quantiles of a Lognormal Distribution

For highly skewed ($\sigma > 1.5$) lognormally distributed populations, the population mean, μ_1, often exceeds the higher quantiles (80%, 90%, 95%) of the distribution. Therefore, the estimation of these quantiles is also of interest. This is especially true when one may want to use MLEs of the higher order quantiles such as 95%, 97.5%, etc. as estimates of the EPC. The formulae to compute these quantiles are 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, $P(X \leq x_p) = p$. If z_p is the p^{th} quantile of the standard normal random variable, Z, with $P(Z \leq z_p) = p$, 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:

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

It is expected that 95% of the observations coming from a lognormal LN(μ, σ^2) distribution would lie at or 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 (2-8).

Notes: 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 $\sigma > 3.29$ *(extremely highly skewed)*. ProUCL computes the *MLEs* of the 50% (median), 90%, 95%, and 99% percentiles of lognormally distributed data sets.

2.3.2.4 MVUEs of Parameters of a Lognormal Distribution

Even though the sample mean $\bar{x}$ is an unbiased estimator of the population mean, μ_1, it does not possess the minimum variance (MV). The MVUEs of μ_1 and σ_1^2 of a lognormal distribution are given as follows:

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

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

The series expansion of the function $g_n(x)$ is given in Bradu and Mundlak (1970), and Aitchison and Brown (1969). Tabulations of this function are also provided by Gilbert (1987). Bradu and Mundlak (1970) computed the MVUE of the variance of the estimate, $\hat{\mu}_1$,

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

The square root of the variance given by equation (2-16) is called the standard error (SE) of the estimate, $\hat{\mu}_1$, given by equation (2-14). The 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).

2.3.3 Estimation of the Parameters of a Gamma Distribution

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 MLE of k is quite complex and requires the computation of Digamma and Trigamma functions.

Several researchers (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 MLE method to estimate the shape and scale parameters of a gamma distribution is described below.

As before, let $x_1, x_2, ..., x_n$ be a random sample (e.g., representing constituent 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:

$$LogL(x_1, x_2, \ldots, x_n; k, \theta) = -nklog(\theta) - nlog\Gamma(k) + (k-1)\sum log(x_i) - \sum x_i/\theta \qquad (2\text{-}18)$$

To find the MLEs of k and θ, one differentiates the log-likelihood function as given in (2-18) with respect to k and θ, and sets the derivatives to zero. This results in the following two equations:

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

$$\hat{k}\hat{\theta} = \frac{1}{n}\sum x_i = \bar{x} \qquad (2\text{-}20)$$

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

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

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. An estimate of k can be computed iteratively by using the Newton-Raphson method (Press *et al.* 1990), leading to the following iterative equation:

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

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. Good approximate values for these two functions (Choi and Wette 1969) can be obtained using the following two approximations. For $k \geq 8$, these functions are approximated by:

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

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

For $k < 8$, one can use the following recurrence relations 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) through (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 initial 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) suggest using $\hat{k}$, given by (2-27) as a starting value of *k* for the iterative procedure, calculating $\hat{k}_l$ at the l^{th} iteration using 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 the Newton-Raphson method, as given by equation (2-22), is in close agreement with the one 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 MLEs of *k* and θ, and also bias-corrected estimates given by (2-29) and (2-30) of *k* and θ. The bias-corrected estimate (called *k star* and *theta star* in ProUCL graphs and output sheets) 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) below) of the mean of a gamma distribution.

Note on Bias Corrected Estimates, $\hat{k}^*$ and $\hat{\theta}^*$: As mentioned above, Choi and Wette (1969) concluded that the MLE, $\hat{k}$, of *k* is biased high. They suggested the use of the bias-corrected (Johnson, Kotz, and Balakrishnan 1994) estimate of *k* given by (2-29) above. However, recently the developers performed a simulation study to evaluate the bias in the MLE of the mean of a gamma distribution for various values of the shape parameter, *k* and sample size, *n*. For smaller values of *k* (e.g., <0.2), the bias in the mean estimate (in absolute value) and mean square error (MSE) based upon the biased corrected MLE, $\hat{k}^*$ are higher than those computed using the MLE estimate,$\hat{k}$; and for higher values of *k* (e.g., >0.2), the bias in the mean estimate and MSE computed using the biased corrected MLE, $\hat{k}^*$ are lower than those computed using the MLE, $\hat{k}$. For values of *k* around 0.2, the use of $\hat{k}^*$ and $\hat{k}$ yields comparable results for all values of the sample size. The bias in the mean estimate obtained using the MLE, $\hat{k}$, increases as *k* increases, and as

expected, bias and MSE decrease as the sample size increases. The results of this study will be published elsewhere.

At present for uncensored and left-censored data sets, ProUCL computes all gamma UCLs and other upper limits (Chapters 3, 4 and 5) using bias corrected estimates, $\hat{k}^*$ and $\hat{\theta}^*$ of k and θ. ProUCL generated output sheets display many intermediate results including $\hat{k}$ and $\hat{k}^*$; $\hat{\theta}$ and $\hat{\theta}^*$. Interested users may want to compute UCLs and other upper limits using MLE estimates, $\hat{k}$ and $\hat{\theta}$ of k and θ for values of k described in the above paragraph.

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 several parametric and nonparametric methods. ProUCL can compute a *(1 – α)*100* UCL (except for adjusted gamma UCL and Land's H-UCL) of the mean for any user selected confidence coefficient, *(1 – α),* lying in the interval [0.5, 1.0]. For the computation of the adjusted gamma UCL, three confidence levels, namely: 0.90, 0.95, and 0.99 are supported by the ProUCL software. An approximate gamma UCL can be computed for any level of significance in the interval [0.5, 1.0].

Parametric UCL Computation Methods in ProUCL include:

- Student's t-statistic (assumes normality or approximate normality) based UCL,
- Approximate gamma UCL (assumes approximate gamma distribution),
- Adjusted gamma UCL (assumes approximate gamma distribution),
- Land's H-Statistic UCL (assumes lognormality), and
- Chebyshev inequality based UCL: Chebyshev (MVUE) UCL obtained using MVUE of the parameters (assumes lognormality).

Nonparametric UCL Computation Methods in ProUCL include:

- Modified-t-statistic (modified for skewness) UCL,
- Central Limit Theorem (CLT) UCL to be used for large samples,
- Adjusted Central Limit Theorem UCL: adjusted-CLT UCL (adjusted for skewness),
- Chebyshev UCL: Chebyshev (Mean, *sd*) obtained using classical sample mean and standard deviation,
- Standard bootstrap UCL,
- Percentile bootstrap UCL,
- BCA bootstrap UCL,
- Bootstrap-t UCL, and
- Hall's bootstrap UCL.

For skewed data sets, Modified-t and adjusted CLT methods adjust for skewness. However, this adjustment is not adequate (Singh, Singh, and Iaci, 2002) for moderately skewed to highly skewed data sets (levels of skewness described in Table 2-1). Even though some UCL methods (e.g., CLT, standard bootstrap, and percentile bootstrap methods) do not perform well enough to provide the specified coverage to the population mean of skewed distributions. These methods have been included in ProUCL for comparison, academic, and research purposes. These comparisons are also necessary to demonstrate why the use of a Student's t-based UCL and Kaplan-Meier (KM) method based UCLs using t-critical values as suggested in some environmental books should be avoided. Additionally, the inclusion of these methods also helps the user to make better decisions. Based upon the sample size, *n*, data skewness, $\hat{\sigma}$, and data distribution, ProUCL makes suggestions regarding the use of one or more 95% UCL methods to estimate the EPC. For additional gudidance, the users may want to consult a statistician to select the most appropriate UCL95 to estimate an EPC.

It is noted that in the environmental literature, recommendations about the use of UCLs have been made without accounting for the skewness and sample size of the data set. Specifically, Helsel (2005, 2012) suggests the use t-statistic and percentile bootstrap method on robust regression on order statistics (ROS) and KM estimates to compute UCL95s without considering data skewness and sample size. For moderately skewed to highly skewed data sets, the use of such UCLs underestimates the population mean. These issues are illustrated by examples discussed in the following sections and also in Chapters 4 and 5.

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

The widely used 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. For normally distributed data sets, the test statistic given by equation (2-31) follows the Student's t-distribution with (*n* -1) *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:

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

For a normally (when the skewness is approximately 0) distributed data sets, 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 symmetric or mildly skewed (|skewness|<0.5) data sets, where the skewness is defined in Table 2-1. For moderately skewed data sets (e.g., when $\hat{\sigma}$, the *sd* of log-transformed data, starts approaching and exceeding 0.5), the UCL given by (2-32) fails to provide the desired coverage of the population mean. This is especially true when the sample size is smaller than 20-25 (graphs summarized in Appendix B). The situation gets worse (coverage much smaller) for higher values of the s*d*, $\hat{\sigma}$, or its *MLE*, s_y.

Notes: ProUCL 5.0 and later versions make a decision about the data distribution based upon both of the GOF test statistics: Lilliefors and Shapiro-Wilk GOF statistics for normal and lognormal distributions; and

A-D and K-S GOF test statistics for gamma distribution. Specifically, when only one of the two GOF statistic lead to the conclusion that data are normal (lognormal or gamma), ProUCL outputs the conclusion that the data set follows an approximate normal (lognormal, gamma) distribution; all decision statistics (parametric or nonparametric) are computed based upon this conclusion. Due to these changes, UCL(s) suggested by ProUCL 5.2 can differ from the UCL(s) suggested by ProUCL 4.1. Some examples illustrating these differences have been considered later in this chapter and also in Chapter 4.0.

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

It is well-known that the use of a lognormal distribution often yields unstable and unrealistic values of the decision statistics including UCLs and UTLs for moderately skewed to highly skewed lognormally distributed data sets; especially when the data set is of a small size (e.g., <30, 50, ...). Even though methods exist to compute 95% UCLs of the mean, UPLs and UTLs based upon gamma distributed data sets (Grice and Bain 1980; Wong 1993; Krishnamoorthy *et al*. 2008), those methods have not become popular due to their computational complexity and/or the lack of their availability in commercial software packages (e.g., Minitab 16). Despite the better performance (in terms of coverage and stability) of the decision making statistics based upon a gamma distribution, some practitioners tend to dismiss the use of gamma distribution based decision statistics by not acknowledging them (EPA 2009e; Helsel 2012b) and/or stating that the use of a lognormal distribution is easier to compute the various upper limits. Throughout this document, several examples have been used to illustrate these issues.

For gamma distributions, ProUCL software has both approximate (used for n>50) and adjusted (when $n \leq 50$) UCL computation methods. Critical values of the chi-square distribution and an estimate of the gamma shape parameter, k along with the sample mean are used to compute gamma UCLs. 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 the easy availability of statistical software to compute these estimates. It is noted that some of the gamma distribution based methods incorporated in ProUCL (e.g., prediction limits, tolerance limits) are also available in the R Script library.

Update in ProUCL 5.0 and Higher Versions: For gamma distributed data sets, all versions of ProUCL compute both adjusted and approximate gamma UCLs. However, in earlier versions of ProUCL, an adjusted gamma UCL was recommended for data sets of sizes ≤40 (instead of 50 as in ProUCL 5.1 and later), and an approximate gamma UCL was recommended for data sets of sizes>40, whereas ProUCL 5.1 and later suggests using approximate gamma UCL for sample sizes >50.

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 $\nu = 2nk$ degrees of freedom (*df*). When the shape parameter, k, is known, a uniformly most powerful test of size of α of the null hypothesis, H_0: $\mu_1 \geq C_s$, against the alternative hypothesis, H_A: $\mu_1 < C_s$, is to reject H_0 if $\bar{x}/C_x < \chi^2_{2nk}(\alpha)/2nk$. The corresponding *(1 – α) 100%* uniformly most accurate UCL for the mean, μ_1, is then given by the probability statement.

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

Where, $\chi^2_v(\alpha)$ denotes the cumulative percentage point of the chi-square distribution (e.g., α is the area in the left tail) with *v (=2nk) df*. That is, if *Y* follows χ^2_v, then $P(Y \leq \chi^2_v(\alpha)) = \alpha$. 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.

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

It should be pointed out that the UCL given by equation (2-34) is an approximate UCL without guarantee that the confidence level of *(1 – α)* will be achieved by this UCL. Simulation results summarized in Singh, Singh, and Iaci (2002) suggest that an approximate gamma UCL given by (2-34) does provide the specified coverage (95%) for values of $k > 0.5$. Therefore, for values of $k > 0.5$, one should use the approximate gamma UCL given by equation (2-34) to estimate the EPC.

For smaller sample sizes, 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 $(1 - \alpha)$. For $\alpha = 0.05$ (confidence coefficient of 0.95), $\alpha = 0.1$, and $\alpha = 0.01$, 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 Table 2-2. The adjusted *(1 – α) *100* UCL of the gamma mean, $\mu_1 = k\theta$, is given by the following equation:

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

Where β is given in Table 2-2 for $\alpha = 0.05$, 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 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, coverage probabilities for the mean associated with these UCLs do not depend upon the values of the scale parameter, θ.

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

*Note that for sample of size 5 (or less), when β becomes '0' for small α value of 0.01, it will not be possible to compute adjusted UCL as the denominator in equation (2-35) will become zero.

For gamma distributed data sets, Singh, Singh, and Iaci (2002) noted that the coverage probabilities provided by the 95% UCLs based upon bootstrap-t and Hall's bootstrap methods (discussed below) are in close agreement. For larger samples, these two bootstrap methods approximately provide the specified 95% coverage and for smaller data sets (from a gamma distribution), the coverage provided by these two methods is slightly lower than the specified level of 0.95.

Note 1: Gamma UCLs do not depend upon the standard deviation of the data set which gets distorted by the presence of outliers. Thus, unlike the lognormal distribution, outliers have reduced influence on the computation of the gamma distribution based upon decision statistics including the UCL of the mean—a fact generally not known to a typical user.

Note 2: For all gamma distributed data sets for all values of k and n, all modules and all versions of ProUCL compute the various upper limits based upon the mean and standard deviation obtained using the bias-corrected estimate, $\hat{k}^*$. As noted earlier, the estimate $\hat{k}^*$ does yield better estimates (reduced bias) for all values of k >0.2. For values of k <0.2, the differences between the various limits obtained using $\hat{k}$ and $\hat{k}^*$ are not that significant. However from a theoretical point of view, when k <0.2, it is desirable to compute the mean, standard deviation, and the various upper limits using the MLE estimate, $\hat{k}$. ProUCL generated output sheets display many intermediate results including $\hat{k}$ *and* $\hat{k}^*$*;* $\hat{\theta}$ *and* $\hat{\theta}^*$. Interested users may want to compute UCLs and other upper limits using MLE estimates, $\hat{k}$ and $\hat{\theta}$, of k and θ for values of k described in the above paragraph.

*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:

$$\text{UCL} = UCL = exp\left(\bar{y} + .05s_y^2 + s_y H_{1-\alpha}/\sqrt{n-1}\right) \quad (2\text{-}36)$$

Tables of H-statistic critical values can be found in Land (1975). When the population is lognormal, Land (1971) showed that theoretically the UCL given by equation (2-36) possesses optimal properties and is the uniformly most accurate unbiased confidence limit. However, in practice, the H-statistic based UCL can be quite disappointing and misleading, especially when the data set is not lognormal but skewed and/or consists of outliers, or represents a mixture data set coming from two or more populations (Singh, Singh, and Engelhardt 1997, 1999; Singh, Singh, and Iaci 2002). 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 s_y, which in turn inflates the H-UCL. Furthermore, it has been observed (Singh, Singh, Engelhardt 1997, 1999) that for samples of sizes smaller than 20-30 (sample size requirement also depends upon skewness), and for values of σ approaching and exceeding 1.0 (moderately skewed to highly skewed data), the use of the H-statistic results in impractical and unacceptably large UCL values.

Notes: In practice, many skewed data sets can be modeled by both gamma and lognormal distributions; however, there are differences in the properties and behavior of these two distributions. Decision statistics computed using the two distributions can differ significantly (see Example 2-2 below). It is noted that some recent documents (Helsel and Gilroy, 2012) incorrectly state that the two distributions are similar. Helsel (2012a, 2012b) suggests the use a lognormal distribution due its computational ease. However, one should not compromise the accuracy and defensibility of estimates and decision statistics by using easier methods which may underestimate (e.g., using a percentile bootstrap UCL based upon a skewed data set) or overestimate (e.g., H-UCL) the population mean. Computation of defensible estimates and decision statistics taking the sample size and data skewness into consideration is always recommended. For complicated and skewed data sets, several UCL computation methods (e.g., bootstrap-t, Chebyshev inequality, and Gamma UCL) are available in ProUCL to compute appropriate decision statistics (UCLs, UTLs) covering a wide-range of data skewness and sample sizes.

For lognormally distributed data sets, the coverage provided by the bootstrap-t 95% UCL is a little lower than the coverage provided by the 95% UCL based upon Hall's bootstrap method (Appendix B). However, it is noted that for lognormally distributed data sets, the coverage provided by these two bootstrap methods is significantly lower than the specified 0.95 coverage for samples of all sizes. This is especially true for moderately skewed to highly skewed (σ >1.0) lognormally distributed data sets. The H-statistic often results in unstable values of the UCL95, especially when the sample size is small, n<20, as shown in Examples 2-1 through 2-3.

Example 2-1. Consider the silver data set with n=56 (from NADA for R package [Helsel, 2013]). The normal GOF test graph is shown in Figure 2-1. It can be seen that the data set has an extreme outlier (an observation significantly different from the main body of the data set). The data set contains NDs, and therefore is considered in Chapter 4 and 5 again. Here this data set is considered assuming that all observations represent detected values. The data set does not follow a gamma distribution (Figure 2-3) but can be modeled by a lognormal distribution as shown in Figure 2-2, accommodating the outlier 560. The histogram shown in Figure 2-4 suggests that data are highly skewed. The *sd* of the logged data = 1.74. The various UCLs computed using ProUCL 5.0 are displayed in Table 2-3 (with outlier) and Table 2-4 (without outlier) following the Q-Q plots.

Figure 2-1. Normal Q-Q Plot of Raw Data in Original Scale

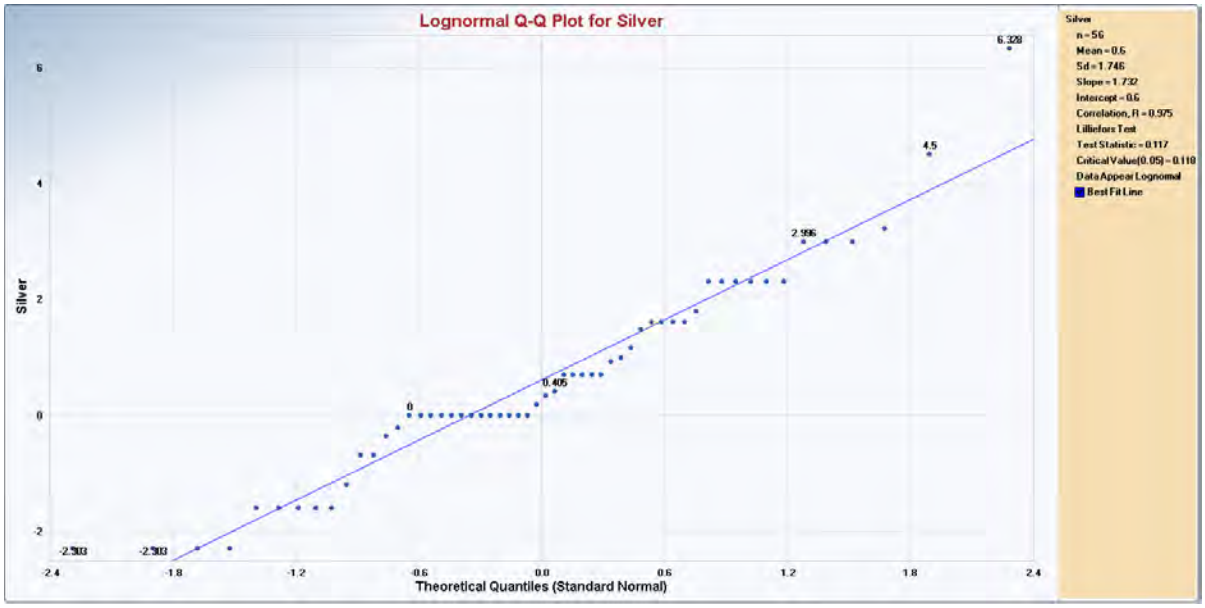

Figure 2-2. Lognormal Q-Q plot with GOF Test Statistics

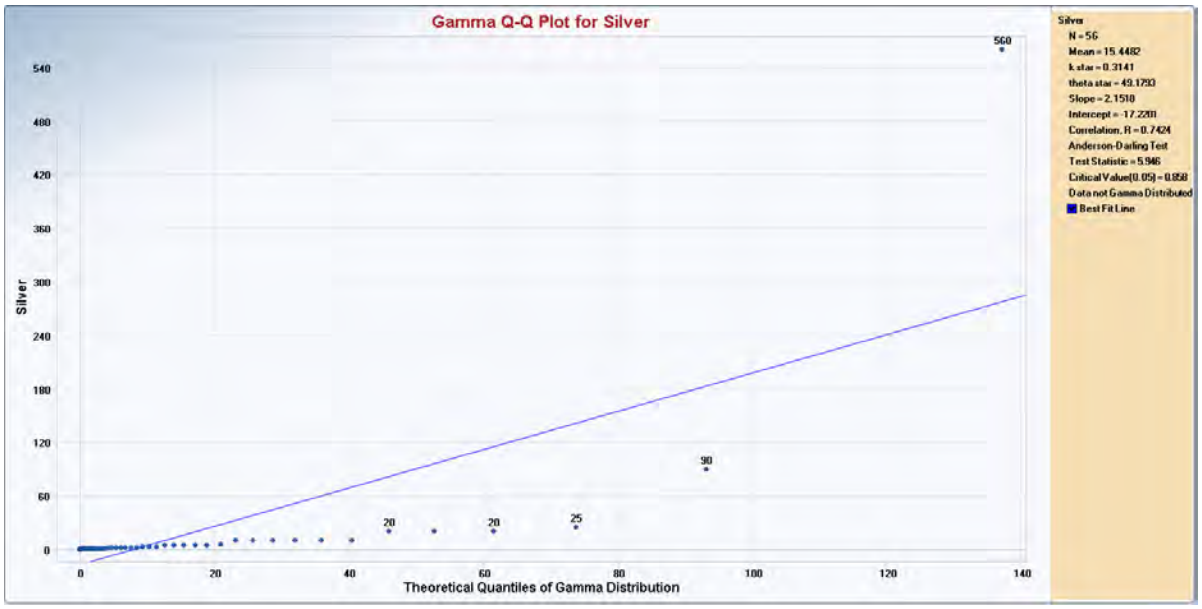

Figure 2-3. Gamma Q-Q plot with GOF Test Statistics

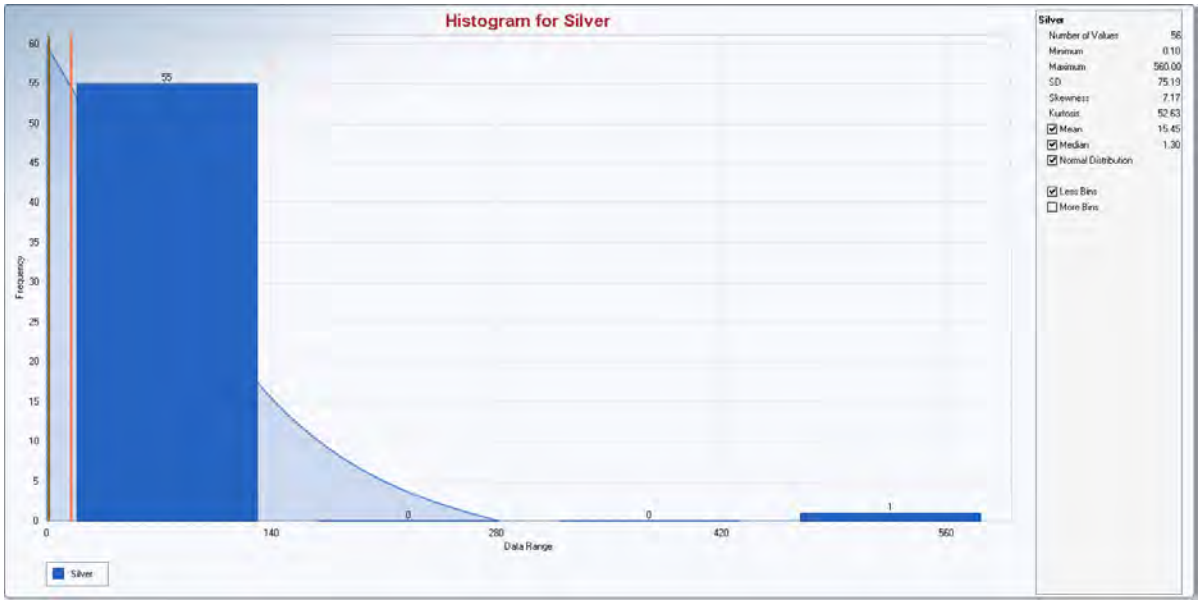

Figure 2-4. Histogram of Silver Data Set including Outlier 560.

In this case, the use of a lognormal UCL may underestimate the EPC. The BCA bootstrap UCL95 is 52.45 and other nonparametric UCLs (excluding the Bootstrap-t, Hall's Bootstrap, and Chebyshev UCLs) range from 31.98 to 35.5. If one insists that the outlier 560 represents a valid observation and comes from the same population, one may want to use a nonparametric BCA UCL95 or other non-parametric UCL to estimate the EPC. The recommendations from ProUCL version 5.1 are shown in Table 2-3. Note that

ProUCL version 5.2 no longer recommends the use of the H-UCL in such cases of small sample size (n < 75) where the appropriate distribution cannot be reliably determined. In such cases of small sample size and high skew, ProUCL version 5.2 does not provide a recommendation and instead encourages the user to contact a trained statistician for an appropriate UCL (Section 2.5.1).

Table 2-3. Lognormal and Nonparametric UCLs for Silver Data including the outlier 560

Silver			
General Statistics			
Total Number of Observations	56	Number of Distinct Observations	22
		Number of Missing Observations	0
Minimum	0.1	Mean	15.45
Maximum	560	Median	1.3
SD	75.19	Std. Error of Mean	10.05
Coefficient of Variation	4.868	Skewness	7.174
Lognormal GOF Test			
Shapiro Wilk Test Statistic	0.951	**Shapiro Wilk Lognormal GOF Test**	
5% Shapiro Wilk P Value	0.0464	Data Not Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.117	**Lilliefors Lognormal GOF Test**	
5% Lilliefors Critical Value	0.118	Data appear Lognormal at 5% Significance Level	
Data appear Approximate Lognormal at 5% Significance Level			
Lognormal Statistics			
Minimum of Logged Data	-2.303	Mean of logged Data	0.6
Maximum of Logged Data	6.328	SD of logged Data	1.746
Assuming Lognormal Distribution			
95% H-UCL	18.54	90% Chebyshev (MVUE) UCL	15.61
95% Chebyshev (MVUE) UCL	19.12	97.5% Chebyshev (MVUE) UCL	24
99% Chebyshev (MVUE) UCL	33.59		
Nonparametric Distribution Free UCLs			
95% CLT UCL	31.98	95% Jackknife UCL	32.26
95% Standard Bootstrap UCL	32.23	95% Bootstrap-t UCL	180.4
95% Hall's Bootstrap UCL	94.1	95% Percentile Bootstrap UCL	35.5
95% BCA Bootstrap UCL	52.45		
90% Chebyshev(Mean, Sd) UCL	45.59	95% Chebyshev(Mean, Sd) UCL	59.25
97.5% Chebyshev(Mean, Sd) UCL	78.2	99% Chebyshev(Mean, Sd) UCL	115.4
Suggested UCL to Use			
95% H-UCL	18.54		

The histogram without the outlier is shown in Figure 2-5. The data is positively skewed with skewness = 5.45. UCLs based upon the data set without the outlier are summarized in Table 2-4 as follows. A quick comparison of the results presented in Tables 2-3 and 2-4 reveals how the presence of an outlier affects the

various decision-making statistics. Refer to Sections 3.2, 7.1 and 7.2 for a discussion of how to appropriately handle outliers.

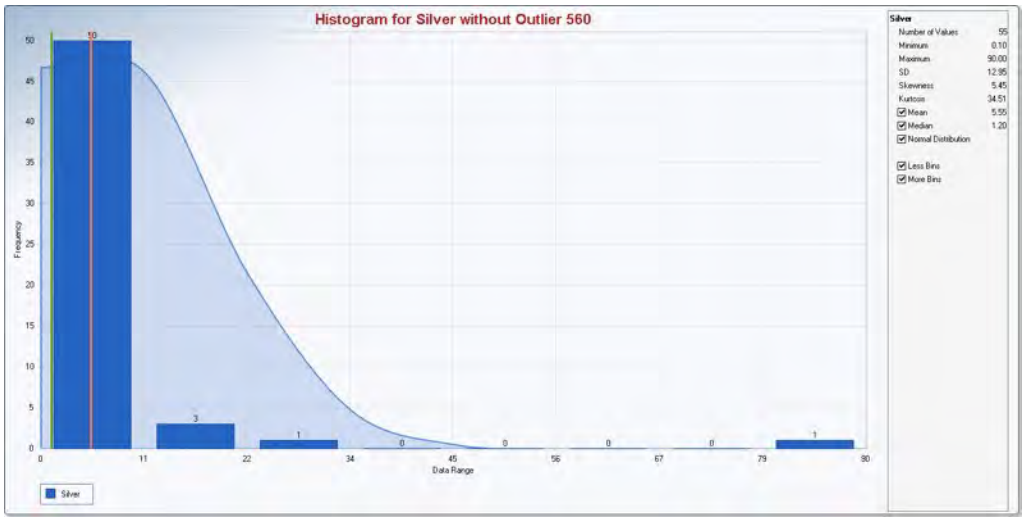

Figure 2-5. Histogram of Silver Data Set Excluding Outlier 560 for Example 2-1.

Table 2-4. Lognormal and Nonparametric UCLs Not Including the Outlier Observation 560

Silver			
General Statistics			
Total Number of Observations	55	Number of Distinct Observations	21
		Number of Missing Observations	0
Minimum	0.1	Mean	5.547
Maximum	90	Median	1.2
SD	12.95	Std. Error of Mean	1.746
Coefficient of Variation	2.334	Skewness	5.45
Lognormal GOF Test			
Shapiro Wilk Test Statistic	0.959	**Shapiro Wilk Lognormal GOF Test**	
5% Shapiro Wilk P Value	0.114	Data appear Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.122	**Lilliefors Lognormal GOF Test**	
5% Lilliefors Critical Value	0.119	Data Not Lognormal at 5% Significance Level	
Data appear Approximate Lognormal at 5% Significance Level			
Lognormal Statistics			
Minimum of Logged Data	-2.303	Mean of logged Data	0.496
Maximum of Logged Data	4.5	SD of logged Data	1.577
Assuming Lognormal Distribution			
95% H-UCL	11.11	90% Chebyshev (MVUE) UCL	10.13
95% Chebyshev (MVUE) UCL	12.26	97.5% Chebyshev (MVUE) UCL	15.22
99% Chebyshev (MVUE) UCL	21.04		
Nonparametric Distribution Free UCLs			
95% CLT UCL	8.419	95% Jackknife UCL	8.469
95% Standard Bootstrap UCL	8.371	95% Bootstrap-t UCL	12.12
95% Hall's Bootstrap UCL	19.2	95% Percentile Bootstrap UCL	8.642
95% BCA Bootstrap UCL	10.47		
90% Chebyshev(Mean, Sd) UCL	10.78	95% Chebyshev(Mean, Sd) UCL	13.16
97.5% Chebyshev(Mean, Sd) UCL	16.45	99% Chebyshev(Mean, Sd) UCL	22.92
Suggested UCL to Use			
95% H-UCL	11.11		

Example 2-2: The positively skewed data set consisting of 25 observations, with values ranging from 0.35 to 170, follows a lognormal or a gamma distribution. The data set is: 0.3489, 0.8526, 2.5445, 2.5602, 3.3706, 4.8911, 5.0930, 5.6408, 7.0407, 14.1715, 15.2608, 17.6214, 18.7690, 23.6804, 25.0461, 31.7720, 60.7066, 67.0926, 72.6243, 78.8357, 80.0867, 113.0230, 117.0360, 164.3302, and 169.8303.

The mean of the data set is 44.09. The data set is positively skewed with *sd* of log-transformed data = 1.68. The normal GOF results are shown in the Q-Q plot of Figure 2-6, it is noted that the data do not follow a normal distribution. The data set follows a lognormal or a gamma distribution as shown in Figures 2-7 and 2-8 and also in Tables 2-5 and 2-6. The various lognormal and nonparametric UCL95s (Table 2-5) and Gamma UCL95s (Table 2-6) are summarized.

The lognormal distribution based H UCL95 is 229.2 which is unacceptably higher than all other UCLs and an order of magnitude higher than the sample mean of 44.09. A more reasonable Gamma distribution based UCL95 of the mean is 74.27 (recommended by ProUCL).

The data set is highly skewed (Figure 2-6) with *sd* of the log-transformed data = 1.68; a Student's t-UCL of 61.66 and a nonparametric percentile bootstrap UCL95 of 60.32 may represent underestimates of the population mean.

The intent of the ProUCL software is to provide users with methods which can be used to compute reliable decision statistics required to make decisions which are cost-effective and protective of human health and the environment.

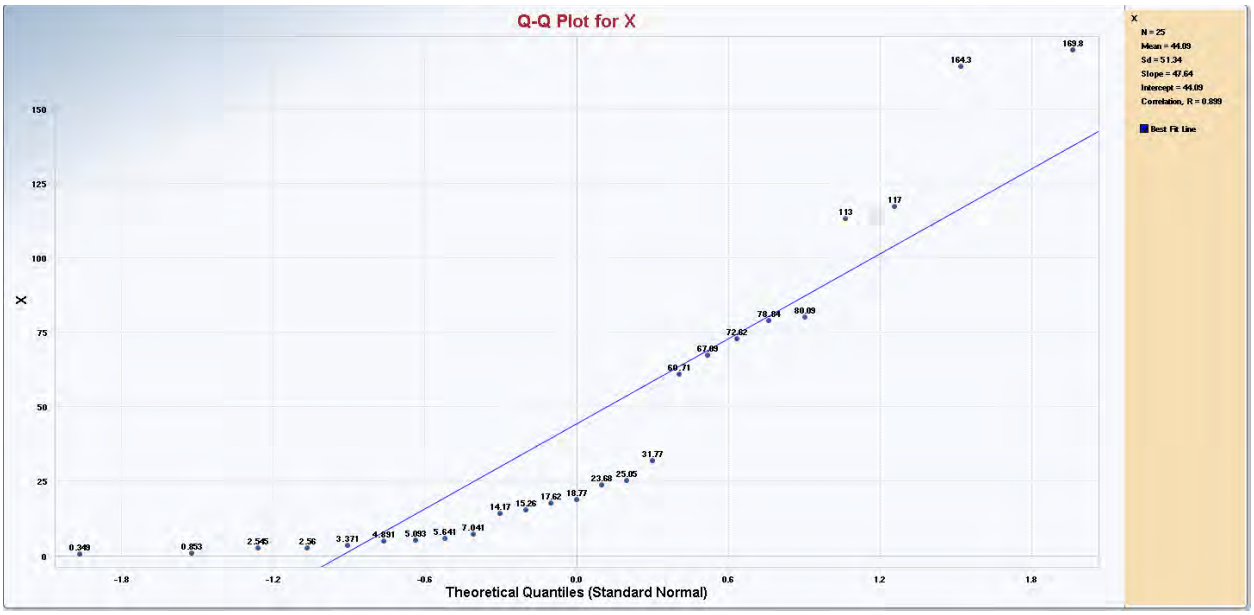

Figure 2-6. Normal Q-Q Plot of X

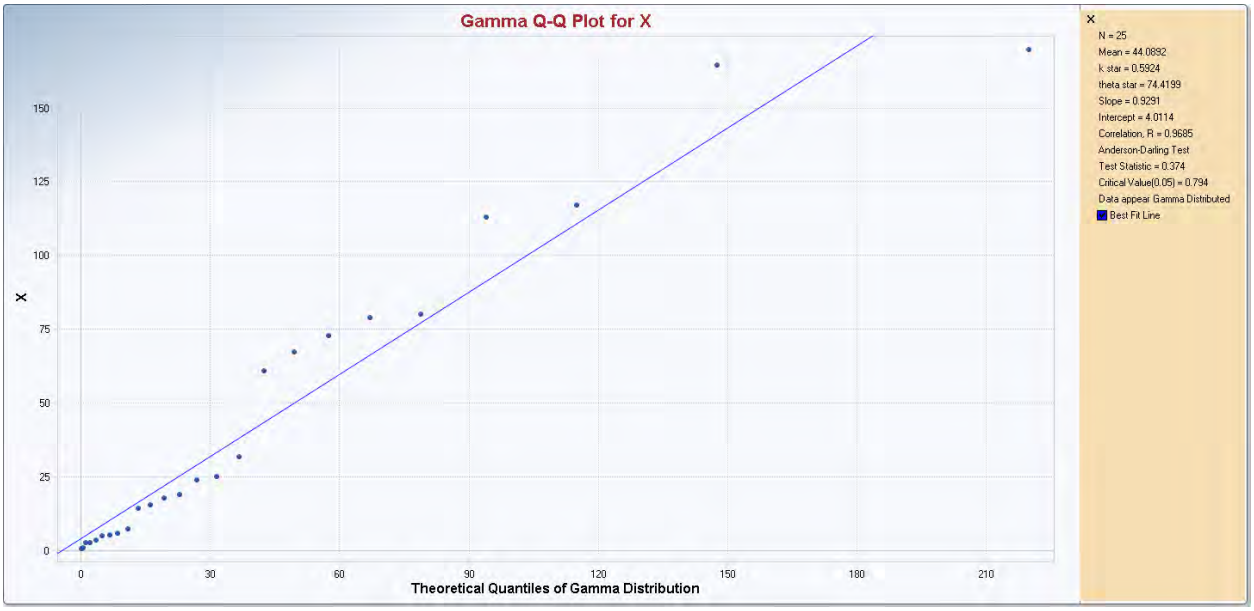

Figure 2-7. Gamma Q-Q Plot of X

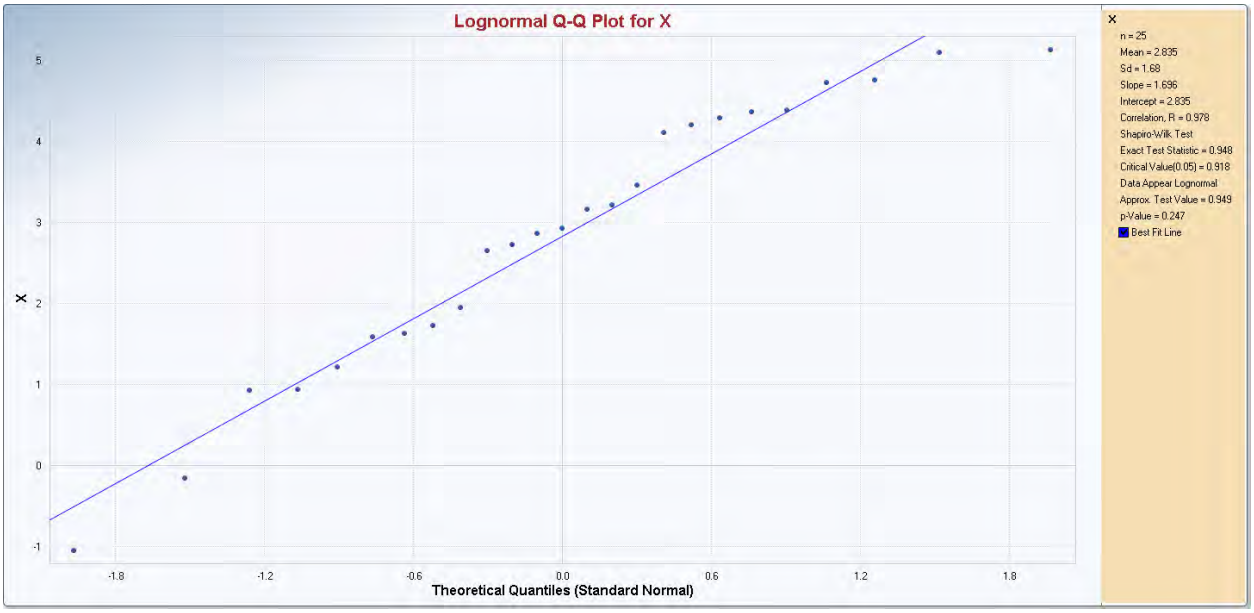

Figure 2-8. Lognormal Q-Q Plot of X

Table 2-5. Nonparametric and Lognormal UCL95

X			
General Statistics			
Total Number of Observations	25	Number of Distinct Observations	25
		Number of Missing Observations	0
Minimum	0.349	Mean	44.09
Maximum	169.8	Median	18.77
SD	51.34	Std. Error of Mean	10.27
Coefficient of Variation	1.164	Skewness	1.294
Lognormal GOF Test			
Shapiro Wilk Test Statistic	0.948	**Shapiro Wilk Lognormal GOF Test**	
5% Shapiro Wilk Critical Value	0.918	Data appear Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.135	**Lilliefors Lognormal GOF Test**	
5% Lilliefors Critical Value	0.177	Data appear Lognormal at 5% Significance Level	
Data appear Lognormal at 5% Significance Level			
Lognormal Statistics			
Minimum of Logged Data	-1.053	Mean of logged Data	2.835
Maximum of Logged Data	5.135	SD of logged Data	1.68
Assuming Lognormal Distribution			
95% H-UCL	229.2	90% Chebyshev (MVUE) UCL	140.6
95% Chebyshev (MVUE) UCL	176.3	97.5% Chebyshev (MVUE) UCL	225.8
99% Chebyshev (MVUE) UCL	323		
Nonparametric Distribution Free UCLs			
95% CLT UCL	60.98	95% Jackknife UCL	61.66
95% Standard Bootstrap UCL	60.57	95% Bootstrap-t UCL	65.58
95% Hall's Bootstrap UCL	62.55	95% Percentile Bootstrap UCL	60.32
95% BCA Bootstrap UCL	64.8		
90% Chebyshev(Mean, Sd) UCL	74.89	95% Chebyshev(Mean, Sd) UCL	88.85
97.5% Chebyshev(Mean, Sd) UCL	108.2	99% Chebyshev(Mean, Sd) UCL	146.3

Notes: The use of H-UCL is not recommended for moderately skewed to highly skewed data sets of smaller sizes (e.g., 30, 50, 70). ProUCL computes and outputs H-statistic based UCLs for historical and academic reasons. This example further illustrates that there are significant differences between a lognormal and a gamma model; 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.

Table 2-6. Gamma UCL95

X			
General Statistics			
Total Number of Observations	25	Number of Distinct Observations	25
		Number of Missing Observations	0
Minimum	0.349	Mean	44.09
Maximum	169.8	Median	18.77
SD	51.34	SD of logged Data	1.68
Coefficient of Variation	1.164	Skewness	1.294
Gamma GOF Test			
A-D Test Statistic	0.374	**Anderson-Darling Gamma GOF Test**	
5% A-D Critical Value	0.794	Data appear Gamma Distributed at 5% Significance Level	
K-S Test Statistic	0.113	**Kolmogrov-Smirnoff Gamma GOF Test**	
5% K-S Critical Value	0.183	Data appear Gamma Distributed at 5% Significance Level	
Data appear Gamma Distributed at 5% Significance Level			
Gamma Statistics			
k hat (MLE)	0.643	k star (bias corrected MLE)	0.592
Theta hat (MLE)	68.58	Theta star (bias corrected MLE)	74.42
nu hat (MLE)	32.15	nu star (bias corrected)	29.62
MLE Mean (bias corrected)	44.09	MLE Sd (bias corrected)	57.28
		Approximate Chi Square Value (0.05)	18.2
Adjusted Level of Significance	0.0395	Adjusted Chi Square Value	17.59
Assuming Gamma Distribution			
95% Approximate Gamma UCL	71.77	95% Adjusted Gamma UCL	74.27
Suggested UCL to Use			
95% Adjusted Gamma UCL	74.27		

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

It is well known that percentile bootstrap, standard bootstrap, and Student's t-statistic based UCL of the mean do not provide the desired coverage of a population mean (Johnson 1978, Sutton 1993, Chen 1995, Efron and Tibshirani 1993) of skewed data distributions. Several researchers including: Chen (1995), Johnson (1978), Kleijnen, Kloppenburg, and Meeuwsen (1986), and Sutton (1993) suggested the use of the modified-t-statistic and skewness adjusted CLT for testing the mean of a positively skewed distribution. The UCLs based upon the modified t-statistic and adjusted CLT methods were included in earlier versions of ProUCL (e.g., versions 1.0 and 2.0) for research and comparison purposes prior to the availability of Gamma distribution based UCLs in ProUCL 3.0 (2004). Singh, Singh, and Iaci (2002) noted that these two skewness adjusted UCL computation methods work only for mildly skewed distributions. These methods have been retained in later versions of ProUCL for academic reasons. The *(1 – α)*100* UCL of the mean based upon a modified t-statistic is given by:

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

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

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

This modification for a skewed distribution does not perform well even for mildly to moderately skewed data sets. Specifically, the UC*L* 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 2002). 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 to 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.

*2.4.5 (1 – α)*100 UCL of the Mean Based upon the Central Limit Theorem*

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

$$z_n = \frac{\bar{x}_n - \mu_1}{\sigma/\sqrt{n}} \quad (2\text{-}39)$$

has a standard normal limiting distribution. For large sample sizes, *n*, the sample mean, $\bar{x}$, has an approximate normal distribution irrespective of the underlying distribution function (Hogg and Craig 1995). The large sample requirement depends upon the skewness of the underlying distribution function of individual observations. The large sample requirement for the sample mean to follow a normal distribution increases with skewness. Specifically, for highly skewed data sets, even samples of size 100 may not be large enough for the sample mean to follow a normal distribution. This issue is illustrated in Appendix B. Since the CLT method requires no distributional assumptions, this is a nonparametric method. As noted by Hogg and Craig (1995), 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 *(1 – α)*100 UCL* of the mean:

$$UCL = \bar{x} + z_a s_x/\sqrt{n} \quad (2\text{-}40)$$

An often cited and used rule of thumb for a sample size associated with a CLT based method is n ≥ 30. However, this may not be adequate if the population is skewed, specifically when σ (*sd* of log-transformed variable) starts exceeding 0.5 to 0.75 (Singh, Singh, Iaci 2002). 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. Noting these observations, Chen (1995) proposed a refinement of the CLT approach, which makes a slight adjustment for skewness.

*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_a, in the upper limit of equation (2-40) is replaced by the following adjusted critical value (Chen 1995):

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

Thus, the adjusted- CLT $(1 - \alpha)$*100 UCL for the mean, μ_1, is given by

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

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

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

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.

Just like the modified-t-UCL, it is observed that the adjusted-CLT UCL also does not provide the specified coverage to the population mean when the population is moderately skewed, specifically when σ becomes larger than 0.75. This is especially true when the sample size is smaller than 20 to25. This large sample size requirement increases as the skewness (or 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 given by (2-42) does not provide adequate coverage to the mean of a gamma distribution, especially when the shape parameter (or its estimate) $k \leq 1.0$ and the sample size is small.

Notes: 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 even for mildly to moderately skewed (e.g., σ in [0.5, 1.0]) data sets. The coverage of the population mean provided by these UCLs becomes worse (much smaller than the specified coverage) for highly skewed data sets. These methods have been retained in ProUCL 5.1 and 5.2 for academic and research purposes.

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

Several commonly used UCL95 computation methods (e.g., Student's t-UCL, percentile and BCA bootstrap UCLs) fail to provide the specified coverage (e.g., 95%) to the population mean of skewed data sets. The use of a lognormal distribution based H-UCL (EPA 2006a, EPA 2009e) is still commonly used to estimate EPCs based upon lognormally distributed skewed data sets. However, the use of Land's H-statistic yields unrealistically large UCL95 values for moderately skewed to highly skewed data sets. On the other hand, when the mean of a logged data set is negative, the H-statistic tends to yield an impractically low value of H-UCL (See Example 2-1 above) especially when the sample size is large (e.g., > 30-50). To address some of these issues associated with lognormal H-UCLs, Singh, Singh, and Engelhardt (1997) proposed the use of the Chebyshev inequality to compute a UCL of the mean of skewed distributions. They noted that a Chebyshev UCL tends to yield stable, realistic, and conservative estimates of the EPCs. The use of the Chebyshev UCL has been adopted by the ITRC (2012 and 2020) to compute UCLs of the mean based upon data sets obtained using the incremental sampling methodology (ISM) approach. However, the use of the Chebyshev UCL has been found to yield unrealistically high estimates of the mean and fails to balance objectives of both coverage and accuracy in an appropriate way (Section 2.5.1.3). There are also

issues with the statistical theory behind this method (Section 2.5.1.3). ProUCL version 5.2 no longer recommends the Chebyshev UCL.

For moderately skewed data sets, the Chebyshev inequality yields conservative UCL95. But for highly skewed data sets, even a Chebyshev inequality fails to yield a UCL95 providing 95% coverage for the population mean (Singh, Singh, and Iaci 2002; Appendix B). To address these issues, ProUCL version 5.1 recommended a 97.5% or 99% Chebyshev UCL, which are typically even more egregious overestimates (Section 2.5.1.3). Since the use of the Chebyshev inequality tends to yield conservative UCL95s, especially for moderately skewed data sets of large sizes (e.g., >50), ProUCL 5.2 also outputs a UCL90 based upon the Chebyshev inequality. ProUCL version 5.2 displays but never recommends Chebyshev UCLs of any confidence level.

The two-sided Chebyshev theorem (Hogg and Craig 1995) 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 to the sample mean, $\bar{x}$ (with mean, μ_1 and variance, σ_1^2/n), to compute 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 = $\bar{x} + 4.47\sigma_1/\sqrt{n}$ represents 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 not guaranteed to be conservative. However, in practice, the use of the sample *sd* does yield conservative values of the UCL95 unless the data set is highly skewed with *sd* of the log-transformed data exceeding 2 to 2.5, and so forth. 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 as follows:

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

All versions of ProUCL compute 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, it is a nonparametric method. This UCL may be used to estimate the population mean, μ_1, when the data are not normal, lognormal, or gamma distributed, especially when *sd*, σ (or its estimate, s_y) becomes large such as > 1.5.

From simulation results summarized in Singh, Singh, and Iaci (2002) and graphical results presented in Appendix B, it is observed that for highly skewed gamma distributed data sets (with shape parameter $k <$ 0.5), the coverage provided by the Chebyshev 95% UCL (given by equation (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 tends to result in a higher (but stable) UCL of the mean of positively skewed gamma distributions.

Note about Chebyshev Inequality based UCLs: The developers of ProUCL have made significant efforts to make suggestions that allows the user to choose the most appropriate UCL95 to estimate the EPC. However, suggestions made in ProUCL may not cover all real world data sets, especially smaller data sets with higher variability. Based upon the results of the simulation studies and graphical displays presented in Appendix B, the developers noted that for smaller data sets with high variability (e.g., *sd* of logged data >1, 1.5, etc.) even a conservative Chebyshev UCL95 tends not to provide the desired 95% coverage to the population mean.

*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*

Earlier versions of ProUCL (when gamma UCLs were not available in ProUCL) used 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 a *(1 – α)*100* UCL for μ_1, which tends to be conservative; where $\hat{\mu}_1$ and $\hat{\sigma}_1(\hat{\mu}_1)$ are given by equations (2-14) and (2-16), respectively. This UCL is retained in ProUCL 5.1/5.2 for historical reasons and research purposes. ProUCL 5.2 does not make any recommendations based upon this version of Chebyshev UCL.

Notes: Many skewed data sets can be modeled both by a lognormal distribution as well as a gamma distribution. Since, the use of a lognormal distribution tends to yield 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 the mean, μ_1, and other upper limits such as UPLs and UTLs should be computed using a gamma distribution.

For a confidence coefficient of 0.95, ProUCL **UCLs/EPCs** module makes suggestions which are based upon the extensive experience of the developers of ProUCL with environmental statistical methods, published literature (Singh, Singh, and Engelhardt 1997, Singh and Nocerino 2002, Singh, Singh, and Iaci 2002, and Singh, Maichle, and Lee 2006) and procedures described in the various guidance documents. However, the project team is responsible for determining whether to use the suggestions made by ProUCL. This determination should be based upon the conceptual site model (CSM), expert site and regional knowledge. The project team may want to consult a statistician.

*2.4.9 (1 – α)*100 UCL of the Mean Using Bootstrap Methods*

Bootstrap methods (Efron 1981, 1982; Efron and Tibshirani 1993) are nonparametric statistical resampling techniques which can be used to reduce the bias in point estimates and construct approximate confidence intervals for parameters, such as the population mean, population percentiles. These methods do not require

any distributional assumptions and can be applied to a variety of situations. The bootstrap methods incorporated in ProUCL for computing upper limits include: the standard bootstrap method, percentile bootstrap method, BCA percentile bootstrap method, bootstrap-t method (Efron,1981, 1982; Hall 1988), and Hall's bootstrap method (Hall 1992; Manly 1997).

As before, let $x_1, x_2, \ldots, x_n$ represent a random sample of size n from a population with an unknown parameter, θ, and let $\hat{\theta}$ be an estimate of θ, which is a function of all n observations. Here, 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 lognormally distributed data sets.

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 each are drawn with replacement from a given data set of size n. 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 are used to compute an estimate of the SE of $\hat{\theta}$. A description of the bootstrap methods, illustrated by application to the population mean, μ_1, and the sample mean, $\bar{x}$, is given as follows.

Step 1. Let $(x_{i1}, x_{i2}, \ldots, x_{in})$ represent the i^{th} bootstrap sample of size n with replacement from the original data set, $(x_1, x_2, \ldots, x_n)$; denote the sample mean using this bootstrap sample 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, ROS 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 SE of the estimate, $\bar{x}$, is given by:

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\bar{x}_i - \bar{x}_B)^2} \qquad (2\text{-}54)$$

If some parameter, θ (e.g., the population median), other than the mean is of concern with an associated estimate (e.g., the sample median), then same steps described above are 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 those N estimates. The difference, $\bar{\theta}_B - \hat{\theta}$, provides an estimate of the bias in the estimate, $\hat{\theta}$, and an estimate of the *SE* of $\hat{\theta}$ is given by:

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

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

$$UCL = \hat{\theta} + z_\alpha \hat{\sigma}_B \qquad (2\text{-}56)$$

ProUCL computes the standard bootstrap UCL by using the population mean and sample mean, given by μ_1 and $\bar{x}$. The UCL obtained using the standard bootstrap method is quite similar to the UCL obtained using

the Student's t-statistic given by equation (2-32), and, as such, does not adequately adjust for skewness. For skewed data sets, the coverage provided by the standard bootstrap UCL is much lower than the specified coverage (e.g., 0.95).

Notes: Typically, bootstrap methods are not recommended for small data sets consisting of less than 10-15 distinct values. Also, it is not desirable to use bootstrap methods on larger (n > 500) data sets. For small data sets, several bootstrap re-samples could be identical and/or all values in a bootstrap re-sample could be identical; no statistical computations can be performed on data sets with all identical observations. For larger data sets, there is no need to perform and use bootstrap methods as a large data set is already representative of the population itself. Methods based upon normal approximations, applied to data sets of larger sizes (n > 500), yield good estimates and results. Also, for larger data, bootstrap methods can take a long time to compute statistics of interest.

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 of size n is used to generate the bootstrap distribution of the unknown population mean. In this method, the N bootstrapped means, $\bar{x}_i$, i:=1,2,...,N, are arranged in ascending order as $\bar{x}_{(1)} \le \bar{x}_{(2)} \le \ldots \le \bar{x}_{(N)}$. The *(1 – α)*100 UCL* of the population mean, μ_1, is given by the value that exceeds the $(1-\alpha)$*100 of the generated mean values. The 95% UCL of the mean is the 95th percentile of the generated means and is given by:

$$95\%\ Percentile\ UCL = 95^{th}\%\ \bar{x}_i;\ i: = 1, 2, \ldots, N \qquad (2\text{-}57)$$

For example $x_{(950)}$, when N = 1000, the bootstrap 95% percentile UCL is given by the 950th ordered mean value given by . It is well-known that for skewed data sets, the UCL95 of the mean based upon the percentile

bootstrap method does not provide the desired coverage (95%) for the population mean. The users of ProUCL and other software packages are cautioned about the suggested use of the percentile bootstrap method for computing UCL95s of the mean based upon skewed data sets. Noting the deficiencies associated with the upper limits (UCLs) computed using the percentile bootstrap method, researchers (Efron 1981; Hall 1988, 1992; Efron and Tibshirani 1993) have developed and proposed the use of skewness adjusted bootstrap methods. Simulations results and graphs presented in Appendix B verify that for skewed data sets, the coverage provided by the percentile bootstrap UCL95 and standard bootstrap UCL is much lower than the coverages provided by the UCL95s based upon the bootstrap-t and the 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 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 adjusts for bias in the estimate (Efron and Tibshirani 1993; and Manly 1997). Results and graphs summarized in Appendix B suggest that the BCA method does provide a slight improvement over the simple percentile and standard bootstrap methods. However, for skewed data sets (parametric or nonparametric), the improvement is not adequate enough and yields UCLs with a coverage probability much lower than the coverage provided by bootstrap-t and Hall's bootstrap methods. This is especially true when the sample size is small. For skewed data sets, the BCA method also performs better

than the modified-t-UCL. Based upon gamma distributed data sets, the coverage provided by the BCA 95%UCL approaches 0.95 as the sample size increases. For lognormally distributed data sets, the coverage provided by the BCA 95%UCL is 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:

$$BCA - UCL = \bar{x}^{(\alpha_2)} \tag{2-58}$$

Here $\bar{x}^{(\alpha_2)}$ is the α_2*100th percentile computed using *N* bootstrap means $\bar{x}_i$; *i*: *= 1, 2, …, N*. For example, when *N* = 2000, $\bar{x}^{(\alpha_2)}$= *(*α_2*N)th* ordered statistic of the *N* bootstrapped means, $\bar{x}_i$; *i*: *= 1, 2, …, N* denoted by $\bar{x}_{(\alpha_2 N)}$ represents a BCA-UCL; α_2 is given by the following probability statement:

$$\alpha_2 = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{\alpha}(\hat{z}_0 + z^{(1-\alpha)})}\right] \tag{2-59}$$

Φ(*z*) is the standard normal cumulative distribution function and $z^{(1-\alpha)}$ is the *100(1–α)th* percentile of a standard normal distribution. For example, $z^{(0.95)}$ = 1.645, and Φ(1.645) = 0.95. Also for equation (2-59), the $\hat{z}_0$ (bias correction factor) and $\hat{\alpha}$ (acceleration factor) are given as follows:

$$\hat{z}_0 = \Phi^{-1}\left[\frac{\#(\bar{x}_i < \bar{x})}{N}\right] \tag{2-60}$$

Here Φ^{-1} (*x*) is the inverse standard normal cumulative distribution function, e.g., Φ^{-1} (0.95) = 1.645; and # represents the number of bootstrap means, $\bar{x}_i$ (out of N means) less than the overall sample mean, $\bar{x}$.

$$\hat{\alpha} = \frac{\sum(\bar{x} - \bar{x}_{-i})^3}{6[\sum(\bar{x} - \bar{x}_{-i})^2]^{1.5}} \tag{2-61}$$

In (2-61), summation is being carried from *i* = 1 to *i* = *n*; $\bar{x}$ is the sample mean based upon all original *n* observation and $\bar{x}_{-i}$ is the mean of (*n*-1) observations without the *i*th observation, *i*: = 1, 2, …, *n*.

2.4.9.5 (1 – α)*100 UCL of the Mean Based upon the Bootstrap-t Method

The nonparametric bootstrap-t (Efron 1982) method uses the bootstrap approach to estimate quantiles of the pivotal t-statistic given by equation (2-31). Rather than using the quantiles/percentiles/critical values of the familiar Student's t-statistic, Hall (1988) proposed computing estimates of the quantiles of the statistic given by equation (2-31) directly from the data. Specifically, as in Steps 1 and 2 of Section 2.4.9.2 above, let $\bar{x}$ be the sample mean computed from the original data, and $\bar{x}_i$ and $s_{x,i}$ be the sample mean and sample standard deviation computed from the *i*th bootstrap sample. For N bootstrap sample, the *N* quantities $t_i = \sqrt{n}[(\bar{x}_i - \bar{x})/s_{x,i}$ are computed and sorted, yielding ordered quantities, *t*(1) ≤ *t*(2) ≤ … ≤ *t*($_N$). The estimate of the lower αth quantile of the pivotal quantity in equation (2-31) is $t_{(\alpha N)}$. For example, if *N* = 1000 bootstrap samples are generated, then the 50th ordered value, $t_{(50)}$, would be the bootstrap estimate of the lower 0.05th quantile of the pivotal quantity given in equation (2-31). Then a *(1–α)*100* UCL of the mean based upon the bootstrap-t-method is given as follows:

$$UCL = \bar{x} - t_{(\alpha N)}\frac{s_x}{\sqrt{n}} \tag{2-62}$$

Note the " – " sign in equation (2-62) is CORRECT.

From the simulation results summarized in Singh, Singh, and Iaci (2002) and in Appendix B, it is observed that for skewed data sets, the bootstrap-t method tends to yield more conservative (higher) UCL values than the other UCLs obtained using the Student's t, modified-t, adjusted-*CLT*, and other bootstrap methods described above. It is noted that for highly skewed ($k < 0.1$ or $\sigma > 2$) data sets of small sizes ($n < 10$ to 15), the bootstrap-t method performs better (in terms of coverage) than other (adjusted gamma *UCL*, or Chebyshev inequality *UCL*) UCL computation methods.

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 for CERCLA sites (EPA 2002a). In this method, $\bar{x}_i$, $s_{x,i}$, and $\hat{k}_{3i}$, the sample mean, the sample standard deviation, and the sample skewness, respectively, are computed from the i^{th} bootstrap re-sample (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 using the original data set of size n. The quantities, W_i and Q_i, given below are computed for the N bootstrap samples:

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

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

$$W(q_a) = 3\left(\left(1 + \hat{k}_3(q_a - \hat{k}_3/(6n))\right)^{1/3} - 1\right)/\hat{k}_3 \qquad (2\text{-}63)$$

In equation (2-63), $\hat{k}_3$ is computed using equation (2-43). Finally, the *(1 – α)*100 UCL* of the population mean based upon Hall's bootstrap method is given as follows:

$$\text{UCL} = \bar{x} - W(q_a)s_x \qquad (2\text{-}64)$$

For both lognormal and gamma distributions, bootstrap-t and Hall's bootstrap methods perform better than the other bootstrap methods, namely, the standard bootstrap method, simple percentile, and bootstrap BCA percentile methods. 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. This is true even for samples of larger sizes ($n \geq 100$). 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 (sd > 1.5, 2.0) data sets of small sizes ($n < 15$), Hall's bootstrap method and the bootstrap-t method perform better than the Chebyshev UCL, and for larger samples, the Chebyshev UCL performs better than Hall's and bootstrap-t methods.

Notes: The bootstrap-t and Hall's bootstrap methods sometimes yield 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 and the project team decides to use them in UCL computations, the use of alternative UCL computation methods (e.g., based upon the Chebyshev inequality) is suggested. These issues are examined in Example 2-3.

Also, when a data set follows a normal distribution without outliers, these nonparametric bootstrap methods, percentile bootstrap method, BCA bootstrap method and bootstrap-t method, will yield comparable results to the Student's t-UCL and modified-t UCL.

Moreover, when a data set is mildly skewed *sd* of logged data <0.5), parametric methods and bootstrap methods discussed in this chapter tend to yield comparable UCL values.

Example 2-3: Consider the pyrene data set with $n = 56$ selected from the literature (She 1997; Helsel 2005). The pyrene data set has been used in several chapters of this technical guide to illustrate the various statistical methods incorporated in ProUCL. The pyrene data set contains several NDs and will be considered again in Chapter 4. However, here, the data set is considered as an uncensored data set to discuss the issues associated with skewed data sets containing outliers; and how outliers can distort UCLs based upon bootstrap-t and Hall's bootstrap UCL computation methods. The Rosner outlier test (see Chapter 7) and normal Q-Q plot displayed in Figure 2-9 below confirm that the observation, 2982.45, is an extreme outlier. However, the lognormal distribution accommodated this outlier and the data set with this outlier follows a lognormal distribution (Figure 2-10). Note that the data set including the outlier does not follow a gamma distribution.

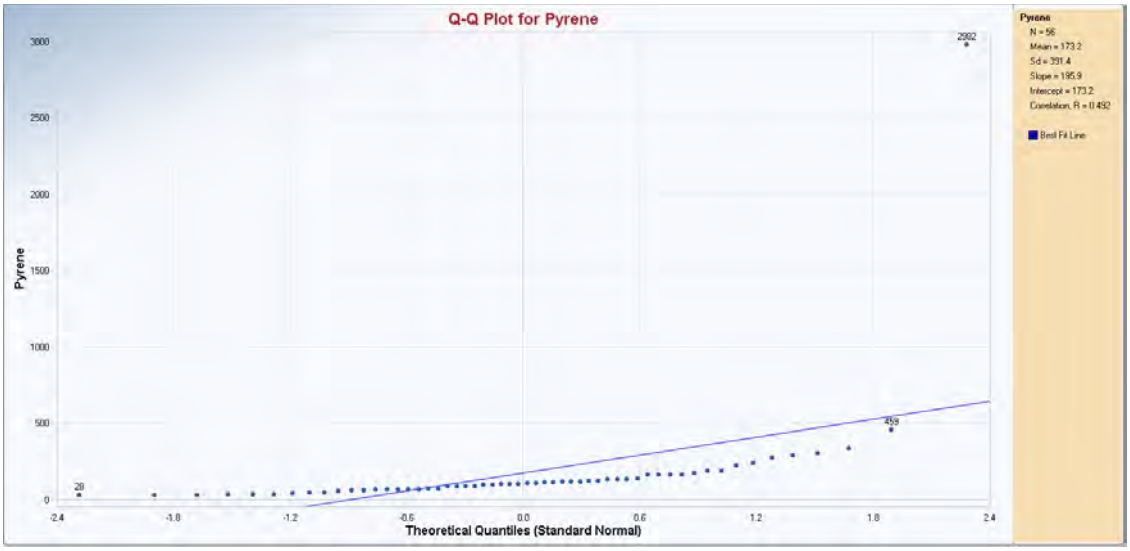

Figure 2-9. Normal Q-Q Plot of She's Pyrene Data Set

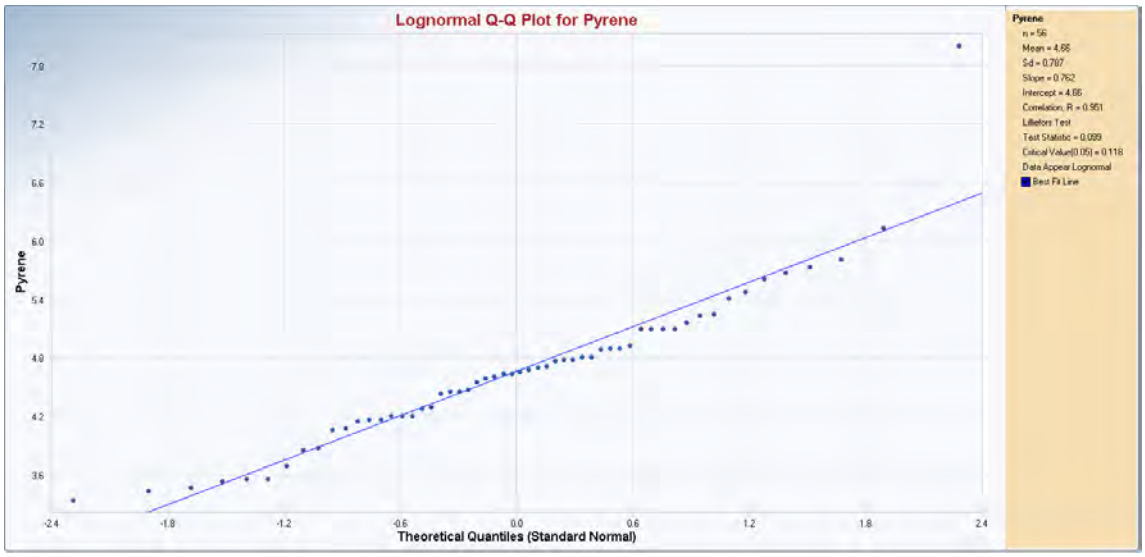

Figure 2-10. Lognormal Q-Q Plot of She's Pyrene Data Set

Several lognormal and nonparametric UCLs (with outlier) are summarized in Table 2-7 below.

Table 2-7. Nonparametric and Lognormal UCLs on Pyrene Data Set with Outlier 2982

Pyrene			
General Statistics			
Total Number of Observations	56	Number of Distinct Observations	44
		Number of Missing Observations	0
Minimum	28	Mean	173.2
Maximum	2982	Median	104
SD	391.4	Std. Error of Mean	52.3
Coefficient of Variation	2.26	Skewness	6.967
Lognormal GOF Test			
Shapiro Wilk Test Statistic	0.924	**Shapiro Wilk Lognormal GOF Test**	
5% Shapiro Wilk P Value	0.00174	Data Not Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.0992	**Lilliefors Lognormal GOF Test**	
5% Lilliefors Critical Value	0.118	Data appear Lognormal at 5% Significance Level	
Data appear Approximate Lognormal at 5% Significance Level			
Lognormal Statistics			
Minimum of Logged Data	3.332	Mean of logged Data	4.66
Maximum of Logged Data	8	SD of logged Data	0.787
Assuming Lognormal Distribution			
95% H-UCL	180.2	90% Chebyshev (MVUE) UCL	193.5
95% Chebyshev (MVUE) UCL	216.3	97.5% Chebyshev (MVUE) UCL	248.1
99% Chebyshev (MVUE) UCL	310.4		
Nonparametric Distribution Free UCLs			
95% CLT UCL	259.2	95% Jackknife UCL	260.7
95% Standard Bootstrap UCL	254.5	95% Bootstrap-t UCL	525.2
95% Hall's Bootstrap UCL	588.5	95% Percentile Bootstrap UCL	276.5
95% BCA Bootstrap UCL	336.7		
90% Chebyshev(Mean, Sd) UCL	330.1	95% Chebyshev(Mean, Sd) UCL	401.1
97.5% Chebyshev(Mean, Sd) UCL	499.8	99% Chebyshev(Mean, Sd) UCL	693.6

Looking at the mean (173.2), standard deviation (391.4), and SE (52.3) in the original scale, the H-UCL (180.2) may represent an underestimate of the population mean; a nonparametric UCL such as a BCA Bootstrap UCL Since there is an outlier present in the data set, both bootstrap-t (UCL=525.2) and Hall's bootstrap (UCL=588.5) methods yield elevated values for the UCL95. A similar pattern was noted in Example 2-1 where the data set included an extreme outlier.

Computations of UCLs without the Outlier 2982

The data set without the outlier follows both a gamma and lognormal distribution with *sd* of the log-transformed data = 0.649 suggesting that the data are moderately skewed. The gamma GOF test results are shown in Figure 2-11. The UCL output results for the pyrene data set without the outlier are summarized in Table 2-8. Since the data set is moderately skewed and the sample size of 55 is fairly large, all UCL methods (including bootstrap-t and Hall's bootstrap methods) yield comparable results. ProUCL suggested the use of a gamma UCL95. This example illustrates how the inclusion of even a single outlier affects all statistics of interest. For a discussion of how to properly handle outliers, refer to Sections 3.2, 7.1 and 7.2.

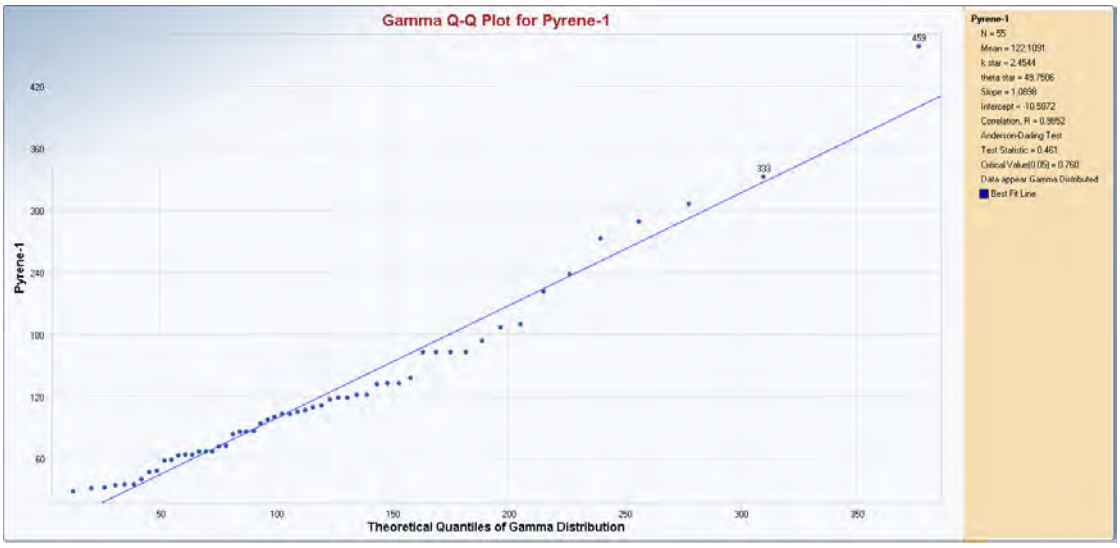

Figure 2-11. Gamma GOF Test on Pyrene Data Set without the Outlier

Table 2-8. Gamma, Nonparametric and Lognormal UCLs on Pyrene Data Set without Outlier=2982

Gamma GOF Test			
A-D Test Statistic	0.461	**Anderson-Darling Gamma GOF Test**	
5% A-D Critical Value	0.76	Detected data appear Gamma Distributed at 5% Significance Level	
K-S Test Statistic	0.0916	**Kolmogrov-Smirnoff Gamma GOF Test**	
5% K-S Critical Value	0.121	Detected data appear Gamma Distributed at 5% Significance Level	
Detected data appear Gamma Distributed at 5% Significance Level			
Gamma Statistics			
k hat (MLE)	2.583	k star (bias corrected MLE)	2.454
Theta hat (MLE)	47.27	Theta star (bias corrected MLE)	49.75
nu hat (MLE)	284.2	nu star (bias corrected)	270
MLE Mean (bias corrected)	122.1	MLE Sd (bias corrected)	77.94
		Approximate Chi Square Value (0.05)	232.9
Adjusted Level of Significance	0.0456	Adjusted Chi Square Value	232
Assuming Gamma Distribution			
95% Approximate Gamma UCL (use when n>=50)	141.5	95% Adjusted Gamma UCL (use when n<50)	142.1
Lognormal GOF Test			
Shapiro Wilk Test Statistic	0.976	**Shapiro Wilk Lognormal GOF Test**	
5% Shapiro Wilk P Value	0.552	Data appear Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.0553	**Lilliefors Lognormal GOF Test**	
5% Lilliefors Critical Value	0.119	Data appear Lognormal at 5% Significance Level	
Data appear Lognormal at 5% Significance Level			
Lognormal Statistics			
Minimum of Logged Data	3.332	Mean of logged Data	4.599
Maximum of Logged Data	6.129	SD of logged Data	0.649
Assuming Lognormal Distribution			
95% H-UCL	146.2	90% Chebyshev (MVUE) UCL	156.8
95% Chebyshev (MVUE) UCL	172.6	97.5% Chebyshev (MVUE) UCL	194.4
99% Chebyshev (MVUE) UCL	237.3		

Table 2-8 (continued). Gamma, Nonparametric and Lognormal UCLs on Pyrene Data Set without Outlier=2982

Nonparametric Distribution Free UCL Statistics			
Data appear to follow a Discernible Distribution at 5% Significance Level			
Nonparametric Distribution Free UCLs			
95% CLT UCL	141	95% Jackknife UCL	141.3
95% Standard Bootstrap UCL	141	95% Bootstrap-t UCL	146.2
95% Hall's Bootstrap UCL	145	95% Percentile Bootstrap UCL	141.5
95% BCA Bootstrap UCL	145.1		
90% Chebyshev(Mean, Sd) UCL	156.6	95% Chebyshev(Mean, Sd) UCL	172.2
97.5% Chebyshev(Mean, Sd) UCL	193.8	99% Chebyshev(Mean, Sd) UCL	236.4
Suggested UCL to Use			
95% Approximate Gamma UCL	141.5		

Example 2-4: Consider the chromium concentration data set of size 24 from a real polluted site to illustrate the differences in UCL95 suggested by ProUCL 4.1 and ProUCL 5.0/ProUCL 5.1. The data set is provided here in full as it has been also used in several examples in Chapter 3.

Table 2-9. Example Data Set for Chromium.

Aluminum	Arsenic	Chromium	Iron	Lead	Mn	Thallium	Vanadium
6280	1.3	8.7	4600	16	39	0.0835	12
3830	1.2	8.1	4330	6.4	30	0.068	8.4
3900	2	11	13000	4.9	10	0.155	11
5130	1.2	5.1	4300	8.3	92	0.0665	9
9310	3.2	12	11300	18	530	0.071	22
15300	5.9	20	18700	14	140	0.427	32
9730	2.3	12	10000	12	440	0.352	19
7840	1.9	11	8900	8.7	130	0.228	17
10400	2.9	13	12400	11	120	0.068	21
16200	3.7	20	18200	12	70	0.456	32
6350	1.8	9.8	7340	14	60	0.067	15
10700	2.3	14	10900	14	110	0.0695	21
15400	2.4	17	14400	19	340	0.07	28
12500	2.2	15	11800	21	85	0.214	25
2850	1.1	8.4	4090	16	41	0.0665	8
9040	3.7	14	15300	25	66	0.4355	24
2700	1.1	4.5	6030	20	21	0.0675	11
1710	1	3	3060	11	8.6	0.066	7.2
3430	1.5	4	4470	6.3	19	0.067	8.1
6790	2.6	11	9230	13	140	0.068	16
11600	2.4	16.4		98.5	72.5	0.13	
4110	1.1	7.6		53.3	27.2	0.068	
7230	2.1	35.5		109	118	0.095	
4610	0.66	6.1		8.3	22.5	0.07	

The chromium concentrations follow an approximate normal distribution (determined using the two normality tests) and also a gamma distribution. ProUCL 5.1 uses the conclusion based upon both (Shapiro-Wilk and Lilliefors) normality tests and ProUCL 4.1 uses the conclusion based only upon the Shapiro-Wilk test leading to the conclusion that the data set does not follow a normal distribution and suggested the use of gamma UCLs. UCL results computed and suggested by ProUCL 5.1 and ProUCL 4.1 are summarized as follows. Data are mildly skewed (with sd of logged data = 0.57), therefore, UCL95s obtained using normal and gamma distributions are comparable.

Table 2-10. UCLs Suggested by ProUCL 5.0 and later

General Statistics			
Total Number of Observations	24	Number of Distinct Observations	19
		Number of Missing Observations	0
Minimum	3	Mean	11.97
Maximum	35.5	Median	11
SD	6.892	Std. Error of Mean	1.407
Coefficient of Variation	0.576	Skewness	1.728

Normal GOF Test			
Shapiro Wilk Test Statistic	0.87	**Shapiro Wilk GOF Test**	
1% Shapiro Wilk Critical Value	0.884	Data Not Normal at 1% Significance Level	
Lilliefors Test Statistic	0.134	**Lilliefors GOF Test**	
1% Lilliefors Critical Value	0.205	Data appear Normal at 1% Significance Level	
Data appear Approximate Normal at 1% Significance Level			

Assuming Normal Distribution			
95% Normal UCL		**95% UCLs (Adjusted for Skewness)**	
95% Student's-t UCL	14.38	95% Adjusted-CLT UCL (Chen-1995)	14.81
		95% Modified-t UCL (Johnson-1978)	14.46

Suggested UCL to Use	
95% Student's-t UCL	14.38

UCLs Suggested by ProUCL 4.1

Gamma Distribution Test		Data Distribution	
k star (bias corrected)	3.128	Data appear Gamma Distributed at 5% Significance Level	
Theta Star	3.825		
MLE of Mean	11.97		
MLE of Standard Deviation	6.766		
nu star	150.2		
Approximate Chi Square Value (.05)	122.8	**Nonparametric Statistics**	
Adjusted Level of Significance	0.0392	95% CLT UCL	14.28
Adjusted Chi Square Value	121.1	95% Jackknife UCL	14.38
		95% Standard Bootstrap UCL	14.28
Anderson-Darling Test Statistic	0.208	95% Bootstrap-t UCL	15.19
Anderson-Darling 5% Critical Value	0.75	95% Hall's Bootstrap UCL	16.77
Kolmogorov-Smirnov Test Statistic	0.0925	95% Percentile Bootstrap UCL	14.37
Kolmogorov-Smirnov 5% Critical Value	0.179	95% BCA Bootstrap UCL	14.95
Data appear Gamma Distributed at 5% Significance Level		95% Chebyshev(Mean, Sd) UCL	18.1
		97.5% Chebyshev(Mean, Sd) UCL	20.75
Assuming Gamma Distribution		99% Chebyshev(Mean, Sd) UCL	25.96
95% Approximate Gamma UCL (Use when n >= 40)	14.63		
95% Adjusted Gamma UCL (Use when n < 40)	14.84		
Potential UCL to Use		Use 95% Adjusted Gamma UCL	14.84

Example 2-5: Consider another mildly skewed real-world data set consisting of lead (Pb) concentrations from a polluted site Questions were raised regarding ProUCL suggesting that the data are approximate normal and suggesting the use of the Student's t-UCL This example is included to illustrate that when data are mildly skewed (sd of logged data <0.5), the differences between UCLs computed using different distributions are not substantial from a practical point of view. The mildly skewed (with sd of logged data =0.47), zinc (Zn) data set of size 11 is given by: 38.9, 45.4, 40.1, 101.4, 166.7, 53.9, 57. 35.7, 43.2, 72.9, and 72.1. The Zn data set follows an approximate normal (using the Lilliefors test). As we know, the Lilliefors test works well for data sets of size >50; so it is valid to question why ProUCL suggests the use of a normal Student's t-UCL. This data set also follows a gamma (using both tests) and lognormal distribution (using both tests). Student's t-UCL95 suggested by ProUCL (using approximate normality) = 87.26, Gamma UCL95 (adjusted) = 93.23, Gamma UCL95 (approximate) = 88.75, and a lognormal UCL95 = 90.51. So all UCLs are comparable for this mildly skewed data set.

Note: When a data set follows all three distributions (when this happens, it is highly likely that data set is mildly skewed), one may want to use a UCL for the distribution with the highest *p*-value. Also when skewness in terms of *sd* of logged data is <0.5, all three distributions yield comparable UCLs.

Suggestions made by ProUCL are based upon simulations performed by the developers. A typical simulation study does not (cannot) cover all data sets of various sizes and skewness from the various distributions. The ProUCL Technical Guide provides sufficient guidance which can help a user select the most appropriate UCL as an estimate of the EPC. ProUCL makes these UCL suggestions to help a typical user select the appropriate UCL from the various available UCLs. Non-statisticians may want to seek help from a qualified statistician.

2.5 Suggestions and Summary

The suggestions provided by ProUCL for selecting an appropriate UCL of the mean are summarized in this section. These suggestions are made to help the users in selecting the most appropriate UCL to estimate the EPC which is routinely used in exposure assessment and risk management studies of the USEPA. A typical simulation study does not (cannot) cover all data sets of all sizes and skewness from all distributions. For an analyte (data set) with skewness (*sd* of logged data) near the end points of the skewness intervals described in decision tables, Table 2-12 and Table 2-13, the user may select the most appropriate UCL based upon expert site knowledge, toxicity of the analyte, and exposure risk associated with that analyte. ProUCL makes these UCL suggestions to help a typical user in selecting the appropriate UCL from the many available UCLs. Non-statisticians may want to seek help from a qualified statistician.

2.5.1 New in ProUCL 5.2

When calculating and determining appropriate UCLs, it is important to remember that the UCL is still an estimate of central tendency (specifically, of the mean). It is a common misconception that the UCL should be close to the upper extreme of the data set. All UCL calculations that can be computed from closed analytical formulas (i.e., not bootstrap-based methods) are based on some underlying distributional assumptions. For example, gamma UCLs work well if the underlying data follow a gamma distribution, and the H-UCL works well if the underlying data truly follow a lognormal distribution. However, if assumptions of a particular distributional form are violated, such formulas do not perform as intended. The exception is the t-UCL, which is very robust to deviations from assumptions of normality. For any realistic distribution of data under unbiased (representative) sampling, the distribution of the sample mean tends toward a normal distribution due the operation of the Central Limit Theorem (CLT). See Frost (2018) for discussion and examples. In cases of large skewness, the convergence to the normal distribution may be slow. However, for many non-normal distributions, a sample size of 30 gives quite a good normal approximation to the mean. Because the UCL is an estimated upper bound for the mean, the Central Limit Theorem (CLT) applies and the t-UCL can be applied to most data sets in cases where assumptions about normality of the underlying distributional form may be violated.

A number of improvements have been made to the decision logic for the recommendation of UCLs in ProUCL 5.2. The way in which goodness of fit (GOF) tests are used to select appropriate UCLs is modified. The Chebyshev UCL is no longer recommended, and the H UCL recommendation has changed. It is now recommended only when there is high confidence that the assumption of lognormality is met to a good approximation and for certain conditions of sample size and range of sample log scale standard deviations. This section provides information on why these updates were necessary. Note that in some cases, data may be too skewed or not numerous enough to determine an appropriate UCL. ProUCL 5.2 does not provide a recommendation in these cases, but encourages the user to: 1) Verify that the data were collected randomly (rather than through biased sampling, such as hot spot delineation sampling or best professional judgment sampling); 2) consider site knowledge that may explain why the data may be skewed (such as small areas of high concentrations); and 3) to contact a statistician if ProUCL cannot provide a recommendation.

2.5.1.1 Goodness of Fit Tests

ProUCL uses GOF tests to choose the most appropriate UCL. GOF tests place the burden of proof on rejecting a particular distribution, not confirming it. For example, if data "pass" a Shapiro-Wilk test for normality, this simply means there is not enough evidence to conclude the data are *not* normal, but that does not prove that the data are truly normal. Therefore, such tests should be used with caution, especially for data sets with small sample sizes. With small sample sizes, there is very little power in GOF tests to reject the null hypothesis of a particular distribution. Therefore, the smaller the sample size, the more likely the data will "pass" a distribution test for the incorrect distribution. On the other hand, very large sample sizes make it likely that the data will be rejected as coming from a specified distribution, even if that is truly the distribution from which the data were generated and no matter the false positive rate (Type 1 error) specified for the GOF test.

Initial simulations were conducted to explore the accuracy of goodness of fit tests. In this limited simulation study, data were simulated from a range of distributional forms and sample sizes. The following distributional forms were simulated (number of distinct distributions): normal (2), truncated normal (1), gamma (2), lognormal (5), Weibull (6), and non-parametric (mixture, 16). Most distributions were tested with sample sizes 30-100 in increments of 5, with a single iteration for each. A select few distributions were tested with sample sizes of 10, 50, 100, and 500, also with a single iteration for each.

As sample sizes increase, the rate of correct assignment of distribution is expected to increase up to a certain very large sample size, after which it is expected to increasingly reject the null hypothesis of a particular distribution even if the data are truly generated by that distribution. However, this sample size is not known as it depends on the particular test. Different data sets will give varying results; however, ProUCL was run on the initial limited set of simulations to explore the behavior of the goodness of fit tests. For sample sizes of less than 50, about 40% of data sets resulted in the correct conclusion of which distribution they follow across all distributional forms. For sample sizes of 50-70 and 71-100, ProUCL was able to assign the correct distribution in about 50% and 60% of cases, respectively, across all distributional forms.

ProUCL uses goodness of fit tests in sequential order: normal, gamma, and lognormal. The first test that the data pass identifies the distribution for ProUCL's decision logic. Table 2-11 shows the rates of correct conclusions by the true distributional form and the rates of correct and incorrect decision by identification of distributional class across all sample sizes tested. These simulations were run through ProUCL 5.1. The GOF tests were run with a Type 1 error rate of 5%.

Table 2-11. Ability of ProUCL 5.1 to choose the correct distributional form.

True Distribution Class (number of simulations tested)	Identified as Normal	Identified as Gamma	Identified as Lognormal	Identified as Other (Non-parametric)	Percent of True Distribution Class Correctly Identified
Normal (34)	34	0	0	0	100%
Gamma (34)	4	28	0	2	82%
Lognormal (95)	5	21	68	1	72%
Other (301) identified as Other (Non-parametric)	94	45	97	65	22%
Total (464)	137	94	165	68	42%
Percent Correct Identification of Class of Distribution at Each Step	25%	30%	41%	96%	
Percent Incorrect Identification of Class of Distribution at Each Step	75%	70%	59%	4%	

For data that were truly normal, the normality test was very effective for the data sets and sample sizes tested ($n \leq 500$). Thirty-four (34) out of 34 were classified as normal. However, 103 non-normal data sets were also classified as normal—four were gamma, five were lognormal, and 94 were other. So, only 25% of the distributions identified as normal were correctly classified. However, the data sets misclassified as normal had lower skewness and less severe departure from normality. In these cases, use of the t-UCL is appropriate based on the robustness of the 1-sample t-test.

Note however that for much larger sample sizes, goodness of fit tests may reject the null hypothesis of normality (or other distribution) even if the data fit the distribution well. Also note that true 100% accuracy is impossible, and that these results are merely an approximation.

For data that follow gamma or lognormal distributions, the accuracy in assigning the correct distribution declines due to the sequential order of tests. For example, for data sets simulated from a gamma distribution, 82% of data sets were correctly classified as gamma, while 12% of data sets were classified as normal and

the remainder were classified as non-parametric (that is, non-identified or other). On the other hand, of the 94 distributions identified as gamma, 28 were gamma, 21 were lognormal, and 45 were from other distributions. So, only 30% of the distributions identified as gamma were correctly classified. Since data sets with empirical distributions not too far from the normal distribution had already been identified as normal, those identified as gamma had moderate skewness and a shape and tail weight intermediate between the normal and lognormal.

The gamma distribution parameter estimation methods used by ProUCL to estimate Gamma distribution UCLs are functions of the arithmetic and geometric means of the data. The geometric mean is the arithmetic mean of the data in log-scale exponentiated to transform it back to the original scale. The H-UCL used for the lognormal distribution is a function of the mean and variance in log scale. While the mean is impacted by extreme observations, the variance is much more so. For moderately skewed data, the Gamma UCL procedures perform reasonably well. Very skewed data sets, with extreme values in the log scale, will very likely be rejected as Gamma and subsequently be tested for lognormality.

For data sets simulated from a lognormal distribution, 72% of data sets were correctly classified as lognormal, while 5% of data sets were classified as normal and 22% of data sets were classified as gamma. On the other hand, of distributions identified as lognormal, none were normal or gamma, and 97 were from other distributions. So, only 41% of the distributions identified as lognormal were correctly classified. Due to the limited power of goodness of fit tests to reject the null hypothesis of a particular distribution, especially with small sample sizes, the strategy used is most problematic for data identified as lognormal. For data sets that are smaller, highly skewed and not actually from a lognormal distribution, the performance of the H-UCL can be poor. It tends to produce unrealistically high UCL estimates, sometimes several orders of magnitude higher than the true population mean. Since this behavior by the H-UCL is driven by the sample having a large log scale variance, the problem can be caused either by the high end or the low end of the sample. In particular, the inappropriate handling of nondetects can cause the H-UCL procedure to give inappropriately large estimates.

For data sets following a distribution other than normal, gamma, or lognormal, only 22% of data sets were correctly identified as other or non-parametric.

Due to the inherent problems with goodness of fit tests, for version 5.2 of ProUCL the significance levels for each GOF test were modified in order to improve the recommendation of UCLs. Because the primary purpose of the GOF test is to inform this recommendation, significance levels are modified according to the performance of various UCL methods under deviations from the distributional forms they are based on. For example, the H-UCL performs poorly when the data are not truly lognormal. This results in data that are truly do not conform to a particular distributional form having unreasonable values for the H-UCL. In cases with relatively small sample sizes, it is likely the goodness of fit test will not show strong evidence against the null hypothesis of lognormality even for data whose distribution is far from lognormal. For example, in the initial simulations tested, of the data sets with sample sizes less than 75 that were simulated from mixture (non-parametric) distributions and did not pass the normal or gamma goodness of fit tests (n=88), only about 30% were correctly identified as not following to a particular distributional form, whereas about 70% were incorrectly identified as lognormal. To prevent this, lognormality is rejected with less evidence against the null hypothesis (α = 10%). By contrast, methods that assume normality (e.g., the t-UCL) are robust to deviations from this assumption, particularly since the mean of a data set is expected

to follow a normal distribution for a very wide range of the underlying distributions of the data. Therefore, ProUCL version 5.2 requires stronger evidence against normality to reject the null hypothesis in the Shapiro-Wilk and Lilliefors tests for normality (α = 1%). Because the gamma UCL methods perform reasonably well, the significance level for the gamma goodness of fit tests remains the same (α = 5%). With real data, the true distribution is never known. As such, any UCL that relies on a specific distributional form (e.g., the approximate gamma UCL, the adjusted gamma UCL, and the H-UCL) should be used with caution.

2.5.1.2 Modifications to Decision Logic

Historically, ProUCL has placed emphasis on achieving adequate coverage, but not on achieving an accurate estimate of the mean, in the sense of an upper bound for the mean that is as close as possible to the true mean while maintaining the desired coverage. Depending on the data, there are some UCLs in ProUCL (particularly Chebyshev and H) that can generate gross overestimates of the mean so that adequate coverage will almost certainly be achieved in these cases, but accuracy suffers. Although this philosophy ensures that the likelihood of one decision error will be small (i.e., Type I error, concluding a site is not contaminated when it is), such an overestimate is can result in a large opposite decision error (i.e., Type II error, concluding a site is contaminated when it is not). The objective should be to not only control for Type I error, but also to protect against large Type II errors. This requires balancing both objectives (coverage and accuracy) to select the most appropriate UCL method. See the flowcharts which summarize the decision logic, given in Appendix C.

2.5.1.3 Chebyshev UCL

The Chebyshev UCL can be highly conservative, resulting in gross overestimates of the mean. The Chebyshev UCL is based on the Chebyshev inequality, and became a popular choice of UCL because such an overestimate of the mean results in greater coverage than can be obtained from other UCL methods. In fact, simulated data from a variety of distributions and sample sizes have shown that the Chebyshev UCL can be many times the value of the true mean, especially for distributions that are highly skewed.

Initial simulations were conducted to investigate the accuracy of Chebyshev and H UCLs compared with other methods. The advantage of simulated data is that the true mean is known, so it is possible to determine coverage and accuracy of various UCLs. In this limited simulation study, data were simulated from a range of distributional forms and sample sizes, with an emphasis on distributions that might cause high UCL estimates using the Chebyshev and H UCL methods. The following distributional forms (number of distinct distributions): normal (2), truncated normal (1), gamma (2), lognormal (11), Weibull (6), and non-parametric (mixture, 13). Most distributions were tested with sample sizes 30-100 in increments of 5, with a single iteration for each. A select few distributions were tested with sample sizes of 10, 50, 100, and 500, also with a single iteration for each.

In one example of a simulated mixture of lognormal distributions with extreme skewness (50% $\mu = 3$, $\sigma = 2$, and 50% $\mu = 1$, $\sigma = 3$; n=35), ProUCL correctly determined the data were non-parametric, and recommended the 97.5% Chebyshev UCL, which was over 179 times the true mean. In some cases it may be even more of an overestimate, especially if the skewness is extreme or if the 99% Chebyshev UCL is recommended. In another example, 35 samples were taken from a lognormal distribution ($\mu = 1.5$, $\sigma = 3$),

and the recommended 99% Chebyshev UCL was over 395 times the true mean. In cases where ProUCL recommended the 95% Chebyshev UCL, the most extreme example is simulated from lognormal data with extreme skewness ($\mu = 3$, $\sigma = 3$, n=100), where the 95% Chebyshev UCL was over 33 times the value of the true mean. To prevent the recommendation of UCLs that are gross overestimates of the mean, ProUCL version 5.2 no longer recommends the Chebyshev UCL. A second simulation study was conducted to identify the optimal choice of UCL in cases of extremely skewed distributions, considering both accuracy and coverage (Section 2.5.1.5).

In probability theory, Chebyshev's inequality guarantees that in any data sample or probability distribution most values are close to the mean — the precise statement being that no more than $1/k^2$ of the distribution's values can be more than k standard deviations away from the mean, where the mean and standard deviation are the population values rather than the sample values. If a confidence level (α) is substituted for $1/k^2$, then the implied inequality can be made to look like a confidence interval statement, and this is the basis of the Chebyshev approach to UCL calculation.

While the Chebyshev inequality is based on the population values, the true values, of the mean and standard deviation, the Chebyshev UCL is necessarily based on sample estimates of the mean and standard deviation. For small sample sizes, the standard deviation may be substantially underestimated a significant fraction of the time, because the sampling distribution of the standard deviation is right-skewed even for the normal case and much more so as sample size decreases and as the parent distribution is also right-skewed. This causes the median of estimates of the standard deviation to be smaller than the true value. So, the sample standard deviation is too low most of the time resulting in UCLs which may not cover the true mean. This is balanced by a smaller proportion of relatively large standard deviation estimates which lead to quite large overestimates of the true mean. The result is that the guaranteed coverage of the Chebyshev inequality is not really guaranteed at all. Furthermore, on average it has poor accuracy because of the large overestimates.

The Chebyshev construction of a UCL is not based on a theoretically correct assumption, only a notion. There are many examples in the field of statistics for which methods are theoretically wrong but produce reasonable results. However, the Chebyshev UCL is an attempt at conservatism which, in most cases, is far too conservative, but simultaneously does not even guarantee the conservatism that is implied (i.e., in the most extreme cases, the Chebyshev falls short of the desired coverage).

Chebyshev's rule provides the most reasonable results, in the sense of getting nearly accurate coverage probability, for distributions that have highest probability in a very narrow band around the mean (i.e., in cases with a point mass at the mean), a small probability in a narrow band near the minimum value, and a similarly small probability in a narrow band near the maximum value. Such a distribution is rarely seen in environmental data (or with just about any non-artificial data). As probability is spread out across a wider range of possible values, Chebyshev's inequality becomes ever more conservative. This spread-out distribution is much more likely to be the case for environmental data, though the case that may come closest is in measuring a site that has moderate skew, with mostly low values (e.g., background) and a small subset of high values (e.g., contamination). Still, when applying Chebyshev's inequality to a sample average, the distribution of sample averages tends to be approximately normal (that is, a more spread-out distribution), and thus Chebyshev's inequality would lead to coverage probabilities much higher than intended, compared to applying the true standard deviation.

In practice, when utilizing an estimate of the standard deviation, it is impossible to know what the true coverage probability might be for an arbitrary data distribution. It is possible to concoct distributions for which a 95% Chebyshev UCL has actual confidence anywhere from 100% down to 0% for a fixed sample size. That is, the goal of coverage probability that is often associated with Chebyshev's UCL rule is not achieved – coverage probability for this rule can be made to be vanishingly small. With such a weak theoretical underpinning, the Chebyshev UCL is simply an arbitrary procedure that introduces some conservatism over the *t*-UCL (when sample size is larger than 2), and as noted above, conservatism may lead to better coverage probability, but trades off with poorer accuracy.

2.5.1.4 H UCL

The H-UCL, based on Land's H-statistic, also tends to grossly overestimate the mean in certain cases, depending on the distribution of the data. Typically, if the data are truly lognormal, the H-UCL provides a conservative but relatively accurate estimate of the mean. However, the H-UCL has been shown to result in unreasonable estimates of the UCL of the mean, especially in cases where the data are not truly lognormal, or in some cases when the skewness of the data is extreme. A previous simulation study (Singh, Singh, and Engelhardt, 1997) found that for lognormal data sets with high standard deviation (sd), σ, of the natural log-transformed data (e.g., σ exceeding 1.0 to 1.5), the H-UCL becomes unacceptably large, exceeding the 95% and 99% data quantiles, and even the maximum observed concentration, by orders of magnitude. The H-UCL was also very sensitive to a few low or a few high values. Although the specifics of this simulation study could not be verified, additional simulations were conducted on lognormal data sets with standard deviations of log-transformed data ranging from 0.5 to 8 and sample sizes ranging from 10 to 500. As an example, in one case of extreme skewness the H-UCL was over 17 orders of magnitude greater than the true mean (n=45, sdlog = 7). This initial set of simulations was used to inform a more in-depth simulation study.

2.5.1.5 Simulation Studies

Three simulation studies are used to inform the UCL recommendation for version 5.2 in cases where the data are classified as lognormal or non-parametric. The first study was conducted by Neptune in 2017 (Flagg et al 2017) and uses data simulated from lognormal distributions with standard deviations of log-transformed data between 0.5 and 1.7, along with gamma, truncated normal, and several mixture distributions. Sample sizes ranged from 5 to 30. Several different types of loss functions were tested in order to penalize UCLs for their distance from the true mean (accuracy), as well as their tendency to underestimate the true mean (coverage), with the choice of the best UCL being fairly consistent across various loss functions. Loss functions are used to balance the two objectives in order to minimize both Type I and Type II error. The optimal UCL in each case is the one that achieves the minimum value of the loss function. The results of the study showed that the t-UCL or the maximum of the t-UCL and the BCA UCL was the optimal choice in nearly all cases tested, with mild skewness and sample sizes less than 30. The *t*-UCL and BCA UCL were recommended in cases with small sample sizes (less than 75) classified as lognormal and mild skewness (sd log < 1.5)

A second smaller-scale simulation study was done recently in direct support of this update to focus on lognormal distributions with extreme skew, with standard deviations of log-transformed data ranging from 2.8 to 3.5 and sample sizes between 25 and 250. This study used the weighted sum of the mean squared

relative error in the difference between the true mean and UCL (as a penalty for accuracy) and the squared difference in logits between the true coverage and desired coverage (as a penalty for under- or over-coverage) for the loss function. Different penalty factors for coverage and accuracy were tested, with the results being fairly consistent across reasonable choices of penalty factors. In cases where the data were classified as non-parametric, the *t*-UCL was consistently shown to be the best choice across all sample sizes and skewness. Therefore, ProUCL recommends the *t*-UCL if the data are classified as non-parametric, particularly since the t-UCL is robust to deviations from normality.

The third simulation study was a large study that focused on generated lognormal data sets, which were filtered using the updated ProUCL 5.2 GoF test decision logic and were also classified as lognormal. The simulation used in this study generated 10,000 replicate data sets for each lognormal distribution used. These distributions have a common mean of 100 and a wide range of coefficients of variation (CVs) (25 values from 0.1 to 20) covering behavior from very slightly skewed to highly skewed. Since these are all lognormal distributions, the CV determines the standard deviation of logs of the values and vice versa. The CV is used as an index parameter for the populations simulated in order to easily fit simulations for other distributions (and mixtures) into the same framework.

The sample sizes of the simulated data sets range from 5 to 1,000 with 47 different values. The UCLs simulated include the Chebyshev 95% UCL, the Chebyshev 90% UCL, the H-UCL, the t-UCL, the skewed t-UCL, the adjusted Gamma UCL, the BCa bootstrap UCL, the bootstrap-t UCL, and Hall's bootstrap UCL. The UCLs had a target coverage level of 95%, except for the Chebyshev 90% UCL. The results, which took several days to compute using parallel computing with up to 50 CPU cores, give an accurate characterization of the behavior of the UCLs calculated. As a result of the modified GoF data filtering rules, the data sets were relatively close to the lognormal distribution. As a result, the H UCL performed very well in this study. The recommendation from this study was that, for data classified by ProUCL 5.2 as lognormal, use the H UCL when the sample size is greater than or equal to 28 or when the log-scale standard deviation is less than or equal to 1.5, and use the t-UCL otherwise. The technical report for this study is included as Appendix D to this report.

Additional studies are needed to determine how change in the GoF test decision logic would affect the choice of the most optimal UCL for a wider range of distributions, including mixture distributions. Such studies may be performed in the future to further improve ProUCL recommendations.

2.5.2 Recommendations by Distributional Form

UCL suggestions have been summarized for: 1) normally distributed data sets, 2) gamma distributed data sets, 3) lognormally distributed data sets, and 4) nonparametric data sets (data sets not following any of the three distributions available in ProUCL). For a given data set, an appropriate UCL can be computed by using more than one method. Therefore, depending upon the data size, distribution, and skewness, sometimes ProUCL may suggest more than one UCL. In such situations, the user may choose any of the suggested UCLs. If needed, the user may consult a statistician for additional insight. For an overview, see the flowcharts which summarize the decision logic, given in Appendix C.

2.5.2.1 Normally or Approximately Normally Distributed Data Sets

For normally distributed data sets, several methods such as: the Student's t-statistic, modified-t-statistic, and bootstrap-t computation methods yield comparable UCL95s providing coverage probabilities close to the nominal level, 0.95. 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.

2.5.2.2 Gamma or Approximately Gamma Distributed 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 (suggested by gamma Q-Q plots and gamma GOF tests), it is generally acceptable to use a 95% UCL based upon a gamma distribution to estimate the EPC.

- For gamma distributed data sets of sizes ≥ 50 with shape parameter, k>1, the use of the approximate gamma UCL95 is recommended to estimate the EPC.
- For gamma distributed data sets of sizes <50, with shape parameter, k >1, the use of the adjusted gamma UCL95 is recommended.
- For highly skewed gamma distributed data sets of small sizes (e.g., <15 or <20) and small values of the shape parameter, k (e.g., k < =1.0), a gamma UCL95 may fail to provide the specified 0.95 coverage for the population mean (Singh, Singh, and Iaci 2002); the use of a bootstrap-t UCL95 or Hall's bootstrap UCL95 is suggested for small highly skewed gamma distributed data sets to estimate the EPC. The small sample size requirement increases as skewness increases. That is as k decreases, the required sample size, n, increases. In the case Hall's bootstrap and bootstrap-t methods yield inflated and erratic UCL results (e.g., when outliers are present), the 95% UCL of the mean may be computed based upon the adjusted gamma 95% UCL.
- For highly skewed gamma distributed data sets of sizes ≥ 15 and small values of the shape parameter, k (k < 1.0), the adjusted gamma UCL95 (when available) may be used to estimate the EPC, otherwise one may want to use the approximate gamma UCL.
- For highly skewed gamma distributed data sets of sizes ≥ 50 and small values of the shape parameter, k (k < 1.0), the approximate gamma UCL95 may be used to estimate the EPC.

Notes: 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 (Efron and Tibshirani 1993). In the case Hall's bootstrap and bootstrap-t methods yield inflated and erratic UCL results, the 95% UCL of the mean may 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.

Table 2-12. Summary Table for the Computation of a 95% *UCL* of the Unknown Mean, μ_1, of a Gamma Distribution; Suggestions are made Based upon Biased Adjusted Estimates

$\hat{k}^*$ *(Skewness* **Bias Adjusted)**	**Sample Size, n**	**Suggestion**
$\hat{k}^* > 1.0$	n>=50	Approximate gamma 95% UCL (Gamma KM or GROS)
$\hat{k}^* > 1.0$	n<50	Adjusted gamma 95% UCL (Gamma KM or GROS)
$\hat{k}^* \leq 1.0$	n < 15	95% UCL based upon bootstrap-t, Hall's bootstrap, or Adjusted gamma 95% UCL (Gamma KM)*
$\hat{k}^* \leq 1.0$	n ≥ 15, n<50	Adjusted gamma 95% UCL (Gamma KM) if available, otherwise use approximate gamma 95% UCL (Gamma KM)
$\hat{k}^* \leq 1.0$	n ≥ 50	Approximate gamma 95% UCL (Gamma KM)

*In case the bootstrap-t method or Hall's bootstrap method yields an erratic, inflated, and unstable UCL value, the UCL of the mean should be computed using the adjusted gamma UCL method.

Note: Suggestions made in Table 2-12 are used for uncensored as well as left-censored data sets. This table is not repeated in Chapter 4. All suggestions have been made based upon bias adjusted estimates, $\hat{k}^*$ of k. When the data set is uncensored, use upper limits based upon the sample size and bias adjusted MLE estimates; and when the data set is left-censored, use upper limits based upon the sample size and biased adjusted estimates obtained using the KM method or GROS method provided $\hat{k}^*>1$. When $\hat{k}^*>1$, UCLs based upon the GROS method and gamma UCLs computed using KM estimates tend to yield comparable UCLs from a practical point of view. For an overview, see the flowcharts which summarize the decision logic, given in Appendix C.

2.5.2.3 Lognormally or Approximately Lognormally Distributed Skewed Data Sets

For lognormally, LN (μ, σ^2), distributed data sets, the H-statistic-based UCL provides the specified 0.95 coverage for the population mean for all values of σ. The recommendation is that, for data classified by ProUCL 5.2 as lognormal, use the H UCL when the sample size is greater than or equal to 28 or when the log-scale standard deviation is less than or equal to 1.5, and use the t-UCL otherwise. The technical report containing this recommendation is included as Appendix D to this report.

The Chebyshev (*MVUE*) UCL has been retained in ProUCL software for historical and information purposes. ProUCL 5.0 and higher versions do not suggest its use.

Table 2-13. Summary Table for the Computation of a UCL of the Unknown Mean, μ_1, of a Lognormal Population to Estimate the EPC

$\hat{\sigma}$ (Standard Deviation)	Sample Size, n	Suggestions
All *	n ≥ 28	H UCL
$\hat{\sigma} \geq 1.5$	$n < 28$	Student's *t*-UCL
$\hat{\sigma} < 1.5$	$n < 28$	H UCL

* Note that the H-UCL recommendation is based on simulations of lognormal distributions with up to 3.5. For extremely skewed distributions, it should be used with caution.

In the third simulation study, the Chebyshev 95% UCL performed poorly relative most other UCL estimators under all risk functions examined. The same was true for the Chebyshev 95% UCL except for the following case: sample size less than 28 and log-scale standard deviation less than 0.64. In this limited case, assuming that ProUCL had identified the data as lognormal, the Chebyshev 90% UCL minimized the average risk over a variety of loss functions. However, because of the many problems with the Chebyshev UCLs, ProUCL 5.2 does not recommend it. But ProUCL provides the Chebyshev UCLs, and while the use of the Chebyshev 90% UCL in this limited case is up to the user's discretion. There is a similar finding for Hall's bootstrap UCL, which can be used for sample size less than 28 and log-scale standard deviation greater than 2.0. See the flowcharts which summarize the recommendation decision logic and the discretionary choices in Appendix C.

For data sets with extreme skew, users are encouraged to examine the data and determine whether it may result from biased sampling or otherwise include small areas of high concentrations that may not be appropriate to include in a single EPC calculation. In these cases, users are encouraged to consult a statistician for proper comparisons.

2.5.2.4 Nonparametric Skewed Data Sets without a Discernible Distribution

For moderately and highly skewed data sets which are neither gamma nor lognormal, one can use a Student's *t*-UCL of the mean to estimate the EPC.

2.5.3 Summary of the Procedure to Compute a 95% UCL of the Unknown Population Mean, μ1, Based upon Full Uncensored Data Sets without Nondetect Observations

A summary of the process used to compute an appropriate UCL95 of the mean is summarized as follows. See also the flowcharts which summarize the decision logic, given in Appendix C.

Formal GOF tests are performed first so that based on the determined data distribution, an appropriate parametric or nonparametric UCL of the mean can be computed to estimate the EPC. ProUCL generates

formal GOF Q-Q plots to graphically evaluate the distribution (normal, lognormal, or gamma) of the data set.

For mildly skewed data sets with (sd of logged data) less than 0.5, all distributions available in ProUCL tend to yield comparable UCLs. Also, when a data set follows all three distributions in ProUCL, compute the upper limits based upon the distribution with highest *p*-value.

For a normally or approximately normally distributed data set, the user is advised to use a Student's t-distribution-based UCL of the mean.

For gamma or approximately gamma distributed data sets, the user is advised to: 1) use the approximate gamma UCL when biased adjusted MLE, $\hat{k}^*$ of $k > 1$ and $n \geq 50$; 2) use the adjusted gamma UCL when biased MLE, $\hat{k}^*$ of $k > 1$ and $n < 50$; 3) use the bootstrap-t method or Hall's bootstrap method when $\hat{k}^* \leq 1$ and the sample size, $n < 15$ (or <20, sample size requirement depends upon k); 4) use the adjusted gamma UCL (if available) for $\hat{k}^* \leq 1$ and sample size, $15 \leq n < 50$; and 5) use approximate gamma UCL when $\hat{k}^* \leq 1$ but $n \geq 50$. If the adjusted gamma UCL is not available, then use the approximate gamma UCL as an estimate of the EPC. When the bootstrap-t method or Hall's bootstrap method yields an erratic inflated UCL (when outliers are present) result, the UCL may be computed using the adjusted gamma UCL (if available) or the approximate gamma UCL.

For lognormally or approximately lognormally distributed data sets, ProUCL recommends a UCL computation method based upon the sample size, *n*, and standard deviation of the log-transformed data, $\hat{\sigma}$. These suggestions are summarized in Table 2-13.

For nonparametric data sets, which are not normally, lognormally, or gamma distributed, the Student's *t*-UCL is used to estimate the EPC.

Notes: It should be pointed out that when dealing with a small data set (e.g., <50), and the Lilliefors test suggests that data are normal and S-W test suggests that data are not normal, ProUCL will suggest that the data set follows an approximate normal distribution. However, for smaller data sets, Lilliefors test results may not be reliable, therefore the user is advised to review GOF tests for other distributions and proceed accordingly. It is emphasized, when a data set follows a distribution (distribution A) using all GOF tests, and also follows an approximate distribution (distribution B) using one of the available GOF tests, it is preferable to use distribution A over distribution B. However, when distribution A is a highly skewed (*sd* of logged data >1.0) lognormal distribution, use the guidance provided on the ProUCL generated output.

Finally, ProUCL makes suggestions about the use of one or more UCLs based upon the data distribution, sample size, and data skewness. Most of the suggestions made in ProUCL are based upon the simulation studies performed by the developers and their professional experience. However, simulations performed do not cover all real world scenarios and data sets. The users may use UCLs values other than those suggested by ProUCL based upon their own experiences and project needs.

CHAPTER 3

Computing Upper Limits to Estimate Background Threshold Values Based Upon Uncensored Data Sets without Nondetect Observations

3.1 Introduction

In background evaluation studies, site-specific (e.g., soils, groundwater) background level constituent concentrations are needed to compare site concentrations with background level concentrations also known as background threshold values (BTVs). The BTVs are estimated, based upon sampled data collected from reference areas and/or unimpacted site-specific background areas (e.g., upgradient wells) as determined by the project team. The first step in establishing site-specific background level constituent concentrations is to collect an appropriate number of samples from the designated background or reference areas. The **Stats/Sample Sizes** module of ProUCL software can be used to compute DQOs-based sample sizes. Once an adequate amount of data has been collected, the next step is to determine the data distribution. This is typically done using exploratory graphical tools (e.g., Q-Q plots) and formal GOF tests. Depending upon the data distribution, one will use a parametric or nonparametric methods to estimate BTVs.

In this chapter and also in Chapter 5 of this document, a BTV is a parameter of the background population representing an upper threshold (e.g., 95^{th} upper percentile) of the background population. When one is interested in comparing averages, a BTV may represent an average value of a background population which can be estimated by a UCL95 (e.g., Chapter 21 of EPA 2009 RCRA Guidance). However, in ProUCL guidance and in ProUCL software, a BTV represents an upper threshold of the background population. The **Upper Limits/BTVs** module of ProUCL software computes upper limits which are often used to estimate a BTV representing an upper threshold of the background population. With this definition of a BTV, an onsite observation in exceedance of a BTV estimate may be considered as not coming from the background population; such a site observation may be considered as exhibiting some evidence of contamination due to site-related activities. Sometimes, locations exhibiting concentrations higher than a BTV estimate are re-sampled to verify the possibility of contamination. Onsite values less than BTVs represent unimpacted locations and can be considered part of the background (or comparable to the background) population. This approach, comparing individual site or groundwater (GW) monitoring well (MW) observations with BTVs, is particularly helpful to: 1) identify and screen constituents/contaminants of concern (COCs); and 2) use after some remediation activities (e.g., installation of a GW treatment plant) have already taken place and the objective is to determine if the remediated areas have been remediated close enough to the background level constituent concentrations.

Background versus site comparisons can also be performed using two-sample hypothesis tests (see Chapter 6). However, BTV estimation methods described in this chapter are useful when not enough site data are available to perform hypotheses tests such as the two-sample t-test or the nonparametric Wilcoxon Rank Sum (WRS) test. When enough (more than 8 to10 observations) site data are available, hypotheses testing approaches can be used to compare onsite and background data or onsite data with some pre-established threshold or screening values. The single-sample hypothesis tests (e.g., t-test, WRS test, proportion test) are used when screening levels or BTVs are known or pre-established. The two-sample hypotheses tests

are used when enough data (at least 8-10 observations from each population) are available from background (e.g., upgradient wells) as well as site (e.g., monitoring wells) areas. This chapter describes statistical limits that may be used to estimate the BTVs for full uncensored data sets without any ND observations. Statistical limits for data sets consisting of NDs are discussed in Chapter 5.

Background data set needs to be evaluated for the presence of data caused by reporting and/or laboratory errors, and extreme values that are suspects of misrepresenting the observed population. Statistical outlier tests give probabilistic evidence for the "misfit" of extreme values. However, their drawback is that they assume normal distribution of the data without outliers. This is often not the case with environmental data, which tend to be naturally right-skewed. Therefore, statistical outlier tests available in ProUCL should only be used to identify potential suspect data points that require further investigation to gain an understanding of extreme values in the context of site processes, geology, and historical use. For example, extreme values may represent contamination from the site (hot spots). However, it is not unusual for a background to consist of different subpopulations due to the presence of varying soil types, textures, vegetation, historical use of the site, etc. It may have, therefore, have higher variability than expected in the planning process.

It is implicitly assumed that the background data set used to estimate BTVs represents a single statistical population. However, since outliers (well-separated from the dominant data) are inevitable in most environmental applications, some outliers such as the observations coming from populations other than the background population may also be present in a background data set. Outliers, when present, distort decision statistics of interest (e.g., upper prediction limits [UPLs], upper tolerance limits [UTLs]), which in turn may lead to incorrect remediation decisions that may not be cost-effective or protective of human health and the environment.

It is suggested that all relevant statistics be computed using the data sets with and without identified outliers. This extra step often helps the project team to see the potential influence of outlier(s) on the decision making statistics (UCLs, UPLs, UTLs) and 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 more accurate estimate(s) of the population parameters (e.g., mean, EPC, BTV) under consideration. In any case, suspect all observtions, including those identified as potential outliers, should be investigated from a scientific and quatily perspective. Since the treatment and handling of outliers in environmental applications is a subjective and controversial topic, the project team (including decision makers, site experts) may decide to treat outliers on a site-specific basis using all existing knowledge about the site and reference areas under investigation. A couple of classical outlier tests, incorporated in ProUCL, are discussed in Chapter 7.

A review of the environmental literature reveals that one or more of the following statistical upper limits are used to estimate BTVs:

- Upper percentiles
- Upper prediction limits (UPLs)
- Upper tolerance limits (UTLs)
- Upper Simultaneous Limits (USLs)

Note: The upper limits which are selected to estimate the BTV are dependent on the project objective (e.g., comparing a single future observation, or comparing an unknown number of observations with a BTV estimate). . ProUCL does not provide suggestions as to which estimate of a BTV is appropriate for a project; the appropriate upper limit is determined by the project team. Once the project team has decided on an upper limit (e.g., UTL95-95), a similar process used to select a UCL95 can be used for selecting a UTL95-95 from among the UTLs computed by ProUCL. The differences between the various limits used to estimate BTVs are not clear to many practitioners. Therefore, a detailed discussion about the use of the different limits with their interpretation is provided in the following sections. Since 0.95 is the most commonly used confidence coefficient (CC), these limits are described for a CC of 0.95 and coverage probability of 0.95 associated with a UTL. ProUCL can compute these limits for any valid combination of CC and coverage probabilities including some commonly used values of CC levels (0.80, 0.90, 0.95, 0.99) and coverage probabilities (0.80, 0.90, 0.95, 0.975).

Caution: To provide a proper balance between false positives and false negatives, the upper limits described above, especially a 95% USL (USL95), should be used only when the background data set represents a single environmental population without outliers (observations not belonging to background). Inclusion of multiple populations and/or outliers tends to yield elevated values of USLs (and also of UPLs and UTLs) which can result in a high number (and not necessarily high percentage) of undesirable false negatives, especially for data sets of larger sizes ($n > 30$).

Note on Computing Lower Limits: In many environmental applications (e.g., in GW monitoring), one needs to compute lower limits including: lower confidence limits (LCL) of the mean, lower prediction limits (LPLs), lower tolerance limits (LTLs), or lower simultaneous limit (LSLs). At present, ProUCL does not directly compute a LCL, LPL, LTL, or a LSL. For data sets with and without NDs, ProUCL outputs several intermediate results and critical values (e.g., khat, nuhat, tolerance factor K for UTLs, d2max for USLs) needed to compute the interval estimates and lower limits. For data sets with and without NDs, except for the bootstrap methods, the same critical value (e.g., normal z value, Chebyshev critical value, or t-critical value) can be used to compute a parametric LPL, LSL, or a LTL (for samples of size >30 to be able to use Natrella's approximation in LTL) as used in the computation of a UPL, USL, or a UTL (for samples of size >30). Specifically, to compute a LPL, LSL, and LTL (n>30) the '+' sign used in the computation of the corresponding UPL, USL, and UTL (n>30) needs to be replaced by the '-' sign in the equations used to compute UPL, USL, and UTL (n>30). For specific details, the user may want to consult a statistician. For data sets *without ND* observations, the Scout 2008 software package (EPA 2009d) can compute the various parametric and nonparametric LPLs, LTLs (all sample sizes), and LSLs.

3.1.1 Description and Interpretation of Upper Limits used to Estimate BTVs

Based upon a background data set, upper limits such as a 95% upper confidence limit of the 95th percentile (UTL95-95) are used to estimate upper threshold value(s) of the background population. It is expected that observations coming from the background population will lie below that BTV estimate with a specified CC. BTVs should be estimated based upon an "established" data set representing the background population under consideration.

Established Background Data Set: This data set represents background conditions free of outliers which potentially represent locations impacted by the site and/or other activities. An established background data

set should be representative of the environmental background population. This can be determined by using a normal Q-Q plot on a background data set. If there are no jumps and breaks in the normal Q-Q plot, the data set may be considered representative of a single environmental population. A single environmental background population here means that the background (and also the site) can be represented by a single geological formation, or by single soil type, or by a single GW aquifer etc. Outliers, when present in a data set, result in inflated values of many decision statistics including UPLs, UTLs, and USLs. The use of inflated statistics as BTV estimates tends to result in a higher number of false negatives.

However, when a site consists of various formations or soil types, separate background data sets may need to be established for each formation or soil type, therefore the project team may want to establish separate BTVs for different formations. When it is not feasible (e.g., due to implementation complexities) or desirable to establish separate background data sets for different geological formations present at a site (e.g., large mining sites), the project team may decide to use the same BTV for all formations.. In this case, a Q-Q plot of background data set collected from unimpacted areas may display discontinuities as concentrations in different formations may vary naturally. In these scenarios, use a Q-Q plot and outlier test only to identify outliers (well separated from the rest of the data) which may be excluded from the computation of BTV estimates.

Notes: The user specifies the allowable false positive error rate, α (=1-CC). The false negative error rate (declaring a location clean when in fact it is contaminated) is controlled by making sure that one is dealing with a defensible/established background data set representing a background population and the data set is free of outliers.

Let $x_1, x_2, \ldots, x_n$ represent sampled concentrations of an established background data set collected from some site-specific or general background reference area.

Upper Percentile, $x_{0.95}$: Based upon an established background data set, a 95th percentile represents that statistic such that 95% of the sampled data will be less than or equal to ($\leq$) $x_{0.95}$. It is expected that an observation coming from the background population (or comparable to the background population) will be $\leq x_{0.95}$ with probability 0.95. A parametric percentile takes data variability into account.

Upper Prediction Limit (UPL): Based upon an established background data set, a 95% UPL (UPL95) represents that statistic such that an independently collected observation (e.g., new/future) from the target population (e.g., background, comparable to background) will be less than or equal to the UPL95 with CC of 0.95. We are 95% sure that a *single future value* from the background population will be less than the UPL95 with CC= 0.95. A parametric UPL takes data variability into account.

In practice, many onsite observations are compared with a BTV estimate. The use of a UPL95 to compare many observations may result in a higher number of false positives; that is the use of a UPL95 to compare many observations just by chance tends to incorrectly classify observations coming from the background or comparable to background population as coming from the impacted site locations. For example, if many (e.g., 30) independent onsite comparisons (e.g., Ra-226 activity from 30 onsite locations) are made with the same UPL95, each onsite value may exceed that UPL95 with a probability of 0.05 just by chance. The overall probability, α_{actual} of at least one of those 30 comparisons being significant (exceeding BTV) just by chance is given by:

$$\alpha_{actual} = 1-(1-\alpha)^k = 1 - 0.95^{30} \sim 1-0.21 = 0.79 \text{ (false positive rate).}$$

This means that the probability (overall false positive rate) is 0.79 (and is not equal to 0.05) that at least one of the 30 onsite locations will be considered contaminated even when they are comparable to background. The use of a UPL95 is not recommended when multiple comparisons are to be made.

Upper Tolerance Limit (UTL): Based upon an established background data set, a UTL95-95 represents that statistic such that 95% of observations (current and future) from the target population (background, comparable to background) will be less than or equal to the UTL95-95 with CC of 0.95. A UTL95-95 represents a 95% UCL of the 95th percentile of the data distribution (population). A UTL95-95 is designed to simultaneously provide coverage for 95% of all potential observations (current and future) from the background population (or comparable to background) with a CC of 0.95. A UTL95-95 can be used when many (unknown) current or future onsite observations need to be compared with a BTV. A parametric UTL95-95 takes the data variability into account.

By definition a UTL95-95 computed based upon a background data set just by chance can classify 5% of background observations as not coming from the background population with CC 0.95. This percentage (false positive error rate) stays the same irrespective of the number of comparisons that will be made with a UTL95-95. However, when a large number of observations coming from the target population (background, comparable to background) are compared with a UTL95-95, the number of exceedances (not the percentage of exceedances) of UTL95-95 by background observations can be quite large. This implies that a larger number (but not greater than 5%) of onsite locations comparable to background may be falsely declared as requiring additional investigation which may not be cost-effective.

To avoid this situation, ProUCL provides a limit called USL which can be used to estimate the BTV provided the background data set represents a single population free of outliers. The use of a USL is not advised when the background data set may represent several geological formations/soil types.

Upper Simultaneous Limit (USL): Based upon an established background data set *free* of outiers and representing a single statistical population (representing a single formation, representing the same soil type, same aquifer), a USL95 represents that statistic such that *all* observations from the "established" background data set are less than or equal to the USL95 with a CC of 0.95. Outliers should be removed before computing a USL as outliers in a background data set tend to represent observations coming from a population other than the background population represented by the majority of observations in the data set. Since USL represents an upper limit on the largest value in the sample, that largest value should come from the same background population. A parametric USL takes the data variability into account. It is expected that all current or future observations coming from the background population (comparable to background population, unimpacted site locations) will be less than or equal to the USL95 with CC, 0.95 (Singh and Nocerino 2002). The use of a USL as a BTV estimate is suggested when a large number of onsite observations (current or future) need to be compared with a BTV.

The false positive error rate does not change with the number of comparisons, as the USL95 is designed to perform many comparisons simultaneously. Furthermore, the USL95 also has a built-in outlier test (Wilks 1963), and if an observation (current or future) exceeds the USL95, then that value definitely represents an outlier and does not come from the background population. The false negative error rate is controlled by

making sure that the background data set represents a single background population free of outliers. Typically, the use of a USL95 tends to result in a smaller number of false positives than a UTL95-95, especially when the size of the background data set is greater than 15.

3.1.2 Confidence Coefficient (CC) and Sample Size

This section briefly discusses sample sizes and the selection of CCs associated with the various upper limits used to estimate BTVs.

- Higher statistical limits are associated with higher levels of CCs. For example, a 95% UPL is higher than a 90% UPL.
- Higher values of a CC (e.g., 99%) tend to decrease the power of a test, resulting in a higher number of false negatives - dismissing contamination when present.

Therefore, the CC should not be set higher than necessary.

- Smaller values of the CC (e.g., 0.80) tend to result in a higher number of false positives (e.g., declaring contamination when it is not present).
- In most practical applications, choice of a 95% CC provides a good compromise between confidence and power.
- Higher level of uncertainty in a background data set (e.g., due to a smaller background data set) and higher values of critical values associated with smaller (n <15-20) samples tend to dismiss contamination as representing background conditions (results in higher number of false negatives; identifying a location that may be dirty as background). This is especially true when one uses UTLs and UPLs to estimate BTVs.
- Nonparametric upper limits based upon order statistics (e.g., the largest, the second largest, etc.) may not provide the desired coverage as they do not take data variability into account. Nonparametric methods are less powerful than the parametric methods; and they require larger data sets to achieve power comparable to parametric methods.

3.2 Treatment of Outliers

The inclusion of outliers in a background data set tends to yield distorted and inflated estimates of BTVs. A couple of classical outlier tests cited in environmental literature (Gilbert 1987; EPA 2006b, 2009; Navy 2002a, 2002b) are available in the ProUCL software. The drawback of these tests is that they assume the normal distribution for data set without outliers. This is mostly not the case for environmental data. Therefore, examination of data distribution needs to be performed before applying outlier test in ProUCL. If the data are not normally distributed, appropriate transformation needs to be applied to approximately normalize or symmetrize the data. An outlier test can then be applied to normalized data to identify potential outliers that need to be scientifically investigated. It is also recommended to supplement outlier tests with graphical displays such as box plots and/or Q-Q plots. Data should never be rejected based on outlier tests only, but when problems with the data are confirmed through scientific and quality investigation.

It is noted that nonparametric upper percentiles, UPLs and UTLs, are often represented by higher order statistics such as the largest value or the second largest value. When high outlying observations are present in a background data set, the higher order statistics may represent observations coming from the contaminated onsite/offsite areas. Decisions made based upon outlying observations or distorted parametric upper limits can be incorrect and misleading. Therefore, special attention should be given to outlying observations. The project team and the decision makers involved should decide about the proper disposition of outliers based on scientifical investigation, to include or not include them, in the computation of the decision making statistics such as the UCL95 and the UTL95-95. Sometimes, performing statistical analyses twice on the same data set, once using the data set with outliers and once using the data set without outliers, can help the project team in determining the proper disposition of high outliers. Examples elaborating on these issues are discussed in several chapters (Chapters 2, 4, 7) this document.

Notes: It should be pointed out that methods incorporated in ProUCL can be used on any data set with or without NDs and with or without outliers. Do not misinterpret that ProUCL is restricted and can only be used on data sets without outliers. It is not a requirement to exclude outliers before using any of the statistical methods incorporated in ProUCL. Statistics computed based upon a data set with outliers tend to be impacted by those outliers and may be less reflective of the population represented by the majority of the data set. The inflated decision statistics tend to represent the locations with those elevated observations rather than representing the dominant population. The outlying observations may be separately investigated to determine the reasons for their occurrences (e.g., errors or contaminated locations). It is suggested to compute the statistics with and without the outliers, and compare the potential impact of outliers on the decision making processes.

Let $x_1, x_2, ..., x_n$ represent concentrations of a contaminant/constituent of concern (COC) collected from some site-specific or general background reference area. The data are arranged in ascending order and the ordered sample (called order statistics) is denoted by $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$. The order statistics are used to define nonparametric estimates of upper percentiles, UPLs, UTLs and USLs. Also, let $y_i = \ln(x_i)$; $i = 1, 2, ..., n$, and $\bar{y}$ and s_y represent the mean and standard deviation (*sd*) of the log-transformed data. Statistical details of some parametric and nonparametric upper limits used to estimate BTVs are described in the following sections.

3.3 Upper p*100% Percentiles as Estimates of BTVs

In most statistical textbooks (e.g., Hogg and Craig 1995), the p^{th} (e.g., $p = 0.95$) sample percentile of the measured sample values is defined as that value, $\hat{x}_p$, such that p*100% of the sampled data set lies at or below it. The carat sign over x_p, indicates that it represents a statistic/estimate computed using the sampled data. The same use of the carat sign is found throughout this guidance document. The statistic $\hat{x}_p$ represents an estimate of the p^{th} population percentile. It is expected that about p*100% of the population values will lie below the p^{th} percentile. Specifically, $x_{0.95}$ represents an estimate of the 95th percentile of the background population.

*3.3.1 Nonparametric p*100% Percentile*

Nonparametric 95% percentiles are used when the background data (raw or transformed) do not follow a discernible distribution at some specified (e.g., $\alpha = 0.05, 0.1$) level of significance. Different software

packages (e.g., SAS, Minitab, and Microsoft Excel) use different formulae to compute nonparametric percentiles, and therefore yield slightly different estimates of population percentiles, especially when the sample size is small, such as less than 20-30. Specifically, some software packages estimate the pth percentile by using the p*nth order statistic, which may be a whole number between 1 and n or a fraction lying between 1 and n, while other software packages compute the pth percentile by the p*(n+1)th order statistic (e.g., used in ProUCL versions 4.00.02 and 4.00.04) or by the (pn+0.5) th order statistic. For example, if n = 20, and p = 0.95, then 20*0.95 = 19, thus the 19th ordered statistic represents the 95th percentile. If n = 17, and p = 0.95, then 17*0.95= 16.15, thus the 16.15th ordered value represents the 95th percentile. The 16.15th ordered value lies between the 16th and the 17th order statistics and can be computed by using a simple linear interpolation given by:

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

Earlier versions of ProUCL (e.g., ProUCL 4.00.02, 4.00.04) used the $p^*(n+1)^{th}$ order statistic to estimate the nonparametric p^{th} percentile. However, since most users are familiar with Excel and some consultants have developed statistical software packages using Excel, and at the request of some users, it was decided to use the same algorithm as incorporated in Excel to compute nonparametric percentiles. ProUCL 4.1 and higher versions compute nonparametric percentiles using the same algorithm as used in Excel 2007. This algorithm is used on data sets with and without ND observations.

Notes: From a practical point of view, nonparametric percentiles computed using the various percentile computation methods described in the literature are comparable unless the data set is small (e.g., n <20-30) and/or comes from a mixed population consisting of some extreme high values. No single percentile computation method should be considered superior to other percentile computation methods available in the statistical literature. In addition to nonparametric percentiles, ProUCL also computes several parametric percentiles described as follows.

*3.3.2 Normal p*100% Percentile*

The sample mean, $\bar{x}$. and *sd*, *s,* are computed first. For normally distributed data sets, the p^*100^{th} sample percentile is given by the following statement:

$$\hat{x}_p = \bar{x} + sz_p \qquad (3\text{-}2)$$

Here z_p is the p^*100^{th} 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 and background data are comparable, then it is expected that an observation coming from a population (e.g., site) comparable to the background population would lie at or below the $p^*100\%$ upper percentile, $\hat{x}_p$, with probability p.

*3.3.3 Lognormal p*100% Percentile*

To compute the p^{th} percentile, $\hat{x}_p$, of a lognormally distributed data set, the sample mean, $\bar{y}$, and *sd*, s_y, of log-transformed data, y are computed first. For lognormally distributed data sets, the p^*100^{th} percentile is given by the following statement:

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

z_p is the p*100th percentile of a standard normal, N(0,1), distribution.

3.3.4 Gamma p*100% Percentile

Since the introduction of a gamma distribution, G (k, θ), is relatively new in environmental applications, a brief description of the gamma distribution is given first; more details can be found in Section 2.3.3. The maximum likelihood estimator (MLE) equations to estimate gamma parameters, k (shape parameter) and θ (scale parameter), can be found in Singh, Singh, and Iaci (2002). A random variable (RV), X (arsenic 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:

$$f(x;k,\theta)=\begin{cases}\frac{1}{\theta^k\Gamma(k)}\cdot x^{k-1}e^{\frac{-x}{\theta}}, & x>0\\ 0, & \text{otherwise}\end{cases} \qquad (3\text{-}4)$$

The mean, variance, and skewness of a gamma distribution are: $\mu = k\theta$, 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., $k \geq 10$). In practice, k is not known and a normal approximation may be used even when the MLE estimate of k is as small as 6.

Let $\hat{k}$ and $\hat{\theta}$ represent the MLEs of k and θ respectively. 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 (df). 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. Once an α*100% percentile, $y_{(\alpha)\,2k}$, of a chi-square distribution with 2k df is obtained, the α*100% percentile for a gamma distribution is computed using the following equation:

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

3.4 Upper Tolerance Limits

A UTL (1-α)-p (e.g., UTL95-95) based upon an established background data set represents that limit such that p*100% of the observations (current and future) from the target population (background, comparable to background) will be less than or equal to UTL with a CC, (1-α). It is expected that p*100% of the observations belonging to the background population will be less than or equal to a UTL with a CC, (1-α). A UTL $(1-\alpha)$-p represents a $(1-\alpha)$ 100% UCL for the unknown p^{th} percentile of the underlying background population.

A UTL95-95 is designed to provide coverage for 95% of all observations potentially coming from the background or comparable to background population(s) with a CC of 0.95. A UTL95-95 will be exceeded by all (current and future) values coming from the background population less than 5% of the time with a CC of 0.95, that is 5 exceedances per 100 comparisons (of background values) can result just by chance for an overall CC of 0.95. Unlike a UPL95, a UTL95-95 can be used when many, or an unknown number of current or future onsite observations need to be compared with a BTV. A parametric UTL95-95 takes the data variability into account.

When a large number of comparisons are made with a UTL95-95, the number of exceedances (not the percentage of exceedances) of the UTL95-95 by those observations can also be large just by chance. This implies that just by chance, a larger number (but not larger than 5%) of onsite locations comparable to background can be greater than a UTL95-95 potentially requiring unnecessary investigation which may not be cost-effective. In order to avoid this situation, it is suggested to use a USL95 to estimate a BTV, provided the background data set represents a single statistical population, free of outliers.

3.4.1 Normal Upper Tolerance Limits

First, compute the sample mean, $\bar{x}$, and *sd*, *s*, using a defensible data set representing a single background population. For normally distributed data sets, an upper $(1 - \alpha)$*100% UTL with coverage coefficient, *p,* is given by the following statement.

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

Here, $K = K(n, \alpha, p)$ is the tolerance factor and depends upon the sample size, *n*, CC = $(1 - \alpha)$, and the coverage proportion = *p*. For selected values of *n, p,* and (1-α), 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 the non-central t-distribution. Also, some large sample approximations (Natrella 1963) are available to compute the *K* values for one-sided tolerance intervals (same for both UTLs and lower tolerance limits). The approximate value of *K* is also a function of the sample size, *n*, coverage coefficient, *p*, and the CC, $(1 - \alpha)$. For samples of small sizes, $n \leq 30$, ProUCL uses the tabulated (Hahn and Meeker 1991) *K* values. Tabulated K values are available only for some selected combinations of *p* (0.90, 0.95, 0.975) and $(1\text{-}\alpha)$ values (0.90, 0.95, 0.99). For sample sizes larger than 30, ProUCL computes the *K* values using the large sample approximations, as given in Natrella (1963). The Natrella's approximation seems to work well for samples of sizes larger than 30. ProUCL computes these *K* values for all valid values of *p* and $(1\text{-}\alpha)$ and samples of sizes as large as 5000.

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 *sd*, s_y, of the log-transformed data are computed. An upper $(1 - \alpha)$*100% tolerance limit with tolerance or coverage coefficient, *p* is given by the following statement:

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

The *K* factor in (3-7) is the same as the one used to compute the normal UTL.

Notes: Even though there in no back-transformation bias present in the computation of a lognormal UTL, a lognormal distribution based UTL is typically higher (sometimes unrealistically higher as shown in the following example) than other parametric and nonparametric UTLs; especially when the sample size is less than 20. Therefore, the use of lognormal UTLs to estimate BTVs should be avoided when skewness is high (*sd* of logged data > 1 or 1.5) and sample size is small (e.g., $n < 20$-30).

3.4.3 Gamma Distribution Upper Tolerance Limits

Positively skewed environmental data can often be modeled by a gamma distribution. ProUCL software has two goodness-of-fit tests: the Anderson-Darling (A-D) and Kolmogorov-Smirnov (K-S) tests for a gamma distribution. These GOF tests are described in Chapter 2. UTLs based upon normal approximation to the gamma distribution (Krishnamoorthy *et al*. 2008) have been incorporated in ProUCL 4.00.05 (EPA 2010d) and higher versions. Those approximations are based upon Wilson-Hilferty (WH)(Wilson and Hilferty 1931) and Hawkins-Wixley (HW) (Hawkins and Wixley 1986) approximations.

Note: It should be pointed out that the performance of gamma UTLs and gamma UPLs based upon these HW and WH approximations is not well-studied and documented. Interested researchers may want to evaluate the performance of these gamma upper limits based upon HW and WH approximations.

A description of method to compute gamma UTLs is given as follows.

Let $x_1, x_2, \ldots, x_n$ represent a data set of size n from a gamma distribution, $G(k, \theta)$ with shape parameter, k and scale parameter θ.

According to the WH approximation, the transformation, $Y = X^{1/3}$ follows an approximate normal distribution. The mean, μ and variance, σ^2 of the transformed normally distributed variable, Y are given as follows:

$$\mu = \left[\theta^{1/3}\Gamma(k + 1/3)\right]/\Gamma(k);\ and\ \sigma^2 = \left[\theta^{2/3}\Gamma(k + 2/3)\right]/\Gamma(k) - \mu^2$$

According to the HW approximation, the transformation, $Y = X^{1/4}$ follows an approximate normal distribution.

Let $\bar{y}$ and s_y represent the mean and *sd* of the observations in the transformed scale (Y).

Using the WH approximation, the gamma UTL (in original scale, X), is given by:

$$UTL = max\left(0, (\bar{y} + K * s_y)^3\right) \quad (3\text{-}8)$$

Similarly, using the HW approximation, the gamma UTL in original scale is given by:

$$UTL = (\bar{y} + K * s_y)^4 \quad (3\text{-}9)$$

The tolerance factor, K is defined earlier in (3-6) while computing a UTL based upon normal distribution.

Note: For mildly skewed to moderately skewed gamma distributed data sets, HW and WH approximations yield fairly comparable UTLs. However, for highly skewed data sets (k <0.*5-1.0*) with higher variability, the HW method tends to yield higher limits than the WH method. A couple of examples are discussed later in this chapter.

3.4.4 Nonparametric Upper Tolerance Limits

The computation of nonparametric UTLs and associated achieved confidence levels are described as follows. A nonparametric $UTL_{p,(1-\alpha)}$ =UTL *p-(1 - α)* providing coverage to *p**100% observations with CC, *(1 – α)* represents an *(1 – α)*100%* UCL for the p^{th} percentile of the target population under study. It is expected that about *p**100% of the observations (current and future) coming from the target population (e.g., background, comparable to background) will be ≤ $UTL_{p,(1-\alpha)}$ with CC, *(1 – α)*100*.

Let $x_{(1)} \leq x_{(2)} \leq \ldots x_{(r)} \leq \ldots \leq x_{(n)}$ represent n ordered statistics (arranged in ascending order) of a given data set, $x_1, x_2, \ldots, x_n$. A nonparametric UTL is computed by higher order statistics such as the largest, the second largest, the third largest, and so on. The order, *r* of the statistic, $x_{(r)}$ used to compute a nonparametric UTL depends upon the sample size, *n*, coverage probability, *p,* and the desired CC, *(1 - α)*. It is noted that in comparison with parametric UTLs, nonparametric UTLs require larger data sets to achieve the desired CC; a nonparametric *UTL p-(1 - α)* computed by order statistics often fails to achieve the specified CC, *(1 – α)*.

Note: Higher order statistics are used to compute nonparametric upper limits which do not account for data variability. Depending upon the data set size, those limits may not provide the specified coverage (e.g., 95% CC) to the parameter (BTV) of interest (e.g., 95% upper percentile of the population). Therefore, before using a nonparametric estimate of the BTV, one should make sure that the data set does not follow a known distribution. Specifically, when dealing with a data set with NDs, account for the NDs and determine the distribution of detected values instead of using a nonparametric UTL. If the detected data follow a parametric distribution, one may want to compute a UTL (and other upper limits) using that distribution and KM estimates. These issues are discussed in Chapter 5.

The formula to compute the order statistic, sample size, and CC achieved by nonparametric UTLs are described below. More details can be found in David and Nagaraja (2003), Conover (1999), Hahn and Meeker (1991), Wald (1963), Scheffe and Tukey (1944) and Wilks (1941).

Note: Just like UCLs, for mildly skewed nonparametric data sets with standard deviation of log-transformed data less than 0.5, one may use a normal distribution based UTLs and UPLs.

3.4.4.1 Determining the Order, r, of the Statistic, x(r), to Compute UTLp,(1-α)

Using the cumulative binomial probabilities, a number, *r*: $1 \leq r \leq n$, 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 – *α*). The binomial distribution (BD) based algorithm has been incorporated in ProUCL for data sets of sizes up to 2000. For data sets of size, *n* >2000, ProUCL computes the r^{th} (*r*: $1 \leq r \leq n$) order statistic by using the normal approximation (Conover, 1999) given by the equation (3-10).

$$r = np + z_{(1-\alpha)}\sqrt{np(1-p)} + 0.5 \qquad (3\text{-}10)$$

Depending upon the sample size, *p*, and *(1 - α)* the largest, the second largest, the third largest, and so forth order statistic is used to estimate the UTL. As mentioned earlier for a given data set of size *n*, the r^{th} order

statistic, $x_{(r)}$ may or may not achieve the specified CC, $(1 - \alpha)$. ProUCL uses the F-distribution based probability statement to compute the CC achieved by the UTL determined by the r^{th} order statistic.

3.4.4.2 Determining the Achieved Confidence Coefficient, CCachieve, Associated with x(r)

For a given data set of size, n, once the r^{th} order statistic, $x_{(r)}$, has been determined, ProUCL can be used to determine if a UTL computed using $x_{(r)}$ achieves the specified CC, $(1 - \alpha)$. ProUCL computes the achieved CC by using the following approximate probability statement based upon the F-distribution with ν_1 and ν_2 degrees of freedom.

$$CC_{Achieve} = (1 - \alpha_*) = Probability(F_{(\nu_1,\nu_2)} \leq f);\ \nu_1 = 2(n - r + 1), and\ \nu_2 = 2r$$

$$f = \frac{r(1-p)}{(n-r+1)p} \qquad (3\text{-}11)$$

For a given data set of size n, ProUCL first computes the order statistic that is used to compute a nonparametric UTL$p,(1-\alpha)$. Once the order statistic has been determined, ProUCL computes the CC actually achieved by that UTL.

3.4.4.3 Determining the Sample Size

For specified values of p and $(1 - \alpha)$, the minimum sample size can be computed using Scheffe and Tukey (1944) approximate sample size formula given by equation (3-12). The minimum sample size formula should be used before collecting any data/samples. Once the data set of size n has been collected, using the binomial distribution or approximate normal distribution, one can compute the order, r, of the statistic to compute a UTL. As mentioned earlier, the UTLs based upon order statistics often do not achieve the desired confidence level. One can use equation (3-11) to compute the CC achieved by a UTL.

$$n_{needed} = 0.25 * \chi^2_{2m,(1-\alpha)} * (1 + p)/(1 - p) + (m - 1)/2 \qquad (3\text{-}12)$$

In equation (3-12), $\chi^2_{2m,(1-\alpha)}$ represents the $(1 - \alpha)$ quantile of a chi-square distribution with $2m$ *df*. It should be noted that in addition to p and $(1 - \alpha)$, the Scheffe and Tukey (1944) approximate minimum sample size formula (3-12) also depends upon the order, r, of the statistic, $x_{(r)}$, used to compute the UTL$p, (1 - \alpha)$. Here m: $1 \leq m \leq n$; and $m=1$ when the largest value, $x_{(n)}$, is used to compute the UTL; and m=2, when the second largest value, $x_{(n-1)}$ is used to compute a UTL, and $m=n-r+1$ when the r^{th} order statistic, $x_{(r)}$, is used to compute a UTL. For example, if the largest sample value, $x_{(n)}$, is used to compute a UTL95-95, then a minimum sample size of 59 (see equation (3-12)) will be needed to achieve a confidence level of 0.95 providing coverage to 95% of the observations coming from the target population. A UTL95-95 estimated by the largest value and computed based upon a data set of size less than 59 may not achieve the desired confidence of 0.95 for the 95th percentile of the target population.

Note: The minimum sample size requirement of 59 cited in the literature is valid when the *largest* value, $x_{(n)}$ (with $m=1$) in the data set is used to compute a compute a UTL95-95. For example, when the largest order statistic, $x_{(n)}$ (with $m=1$) is used to compute a nonparametric UTL95-95, the approximate minimum sample size needed $0.25*5.99*1.95/0.05 \approx 58.4$ (using equation (3-12)) which is rounded upward to 59; and when the *second largest* value, $x_{(n-1)}$ (with $m=2$) is used to compute a UTL95-95, the approximate

minimum sample size needed = [(0.25*9.488*1.95)/0.05] + 0.5 ≈ 93. Similarly, to compute a UTL90-95 by the largest sample value, about 29 observations will be needed to provide coverage for 90% of the observations from the target population with CC = 0.95. Other sample sizes for various values of *p* and *(1-α)* can be computed using equation, (3-12). In environmental applications, the number of available observations from the target population is much smaller than 29, 59 or 93 and a UTL computed based upon those data sets may not provide specified coverage with the desired CC. For specified values of CC, *(1-α)* and coverage, *p*, ProUCL outputs the achieved CC by a computed UTL and the minimum sample size needed to achieve the pre-specified CC.

3.4.4.4 Nonparametric UTL Based upon the Percentile Bootstrap Method

A couple of bootstrap methods to compute nonparametric UTLs are also available in ProUCL. Like the percentile bootstrap UCL computation method, for data sets without a discernible distribution, one can use percentile bootstrap resampling method to compute $UTL_{p,(1-\alpha)}$ =UTL *p,(1 - α)*. The *N* bootstrapped nonparametric p^{th} percentiles, *p,(i:=1,2,...,N)*, are arranged in ascending order: $p_1 \leq p_2 \leq \ldots \leq p_N$. The $UTL_{p,(1-\alpha)}$ for the target population is given by the value that exceeds the *(1 – α)*100* of the *N* bootstrap percentile values. The UTL95-95 is the 95th percentile and is given by:

95% Percentile UTL = 95th percentile of p_i values; i: = 1, 2, ..., N

For example, when *N* = 1000, the ULT95-95 is given by the 950th order percentile value of the 1000 bootstrapped 95th percentiles. Typically, this method yields the largest value in the data set to compute a UTL which may not provide the desired coverage (e.g., 0.95) to the 95th population percentile.

3.4.4.5 Nonparametric UTL Based upon the Bias-Corrected Accelerated (BCA) Percentile Bootstrap Method

Like the percentile bootstrap method, one can use the BCA bootstrap method (Efron and Tibshirani 1993) to compute nonparametric UTLs. However, this method needs further investigation. This method is incorporated in ProUCL 4.00.04 and higher versions for interested users. In this method one replaces the sample mean, bootstrap means by the corresponding bootstrap percentiles. The details of the BCA bootstrap method are given in Section 2.4.9.4.

3.5 Upper Prediction Limits

Based upon a background data set, UPLs are computed for a single (UPL_1) and k (UPL_k) future observations. Additionally, in groundwater monitoring applications, an upper prediction limit of the mean of the k future observations, UPL_k (mean) is also used. A brief description of parametric and nonparametric upper prediction limits is provided in this section.

<u>UPL_1 for a Single Future Observation:</u> A UPL_1 computed based upon an established background data set represents that statistic such that a single future observation from the target population (e.g., background, comparable to background) will be less than or equal to the $UPL_1$95 with a CC of 0.95. A parametric UPL takes the data variability into account. A UPL_1 is designed for a *single future* observation comparison;

however in practice users tend to use $UPL_1 95$ to perform many future comparisons which results in a high number of false postives (observations declared contaminated when in fact they are clean).

When k>*1* future comparisons are made with a UPL_1, some of those future observations will exceed the UPL_1 just by chance, each with probability 0.05. For proper comparison, a UPL needs to be computed accounting for the number of comaprisons that will be performed. For example, if 30 independent onsite comparisons (e.g., Pu-238 activity from 30 onsite locations) are made with the same background $UPL_1 95$, each onsite value comparable to background may exceed that $UPL_1 95$ with probability 0.05. The overall probability of at least one of those 30 comparisons being significant (exceeding the BTV) just by chance is given by:

$$\alpha_{actual} = 1-(1-\alpha)^k = 1 - 0.95^{30} \sim 1-0.21 = 0.79 \text{ (false positive rate).}$$

This means that the probability (overall false positive rate) is 0.79 (and not 0.05) that at least one of the 30 onsite observations will be considered contaminated even when they are comparable to background. Similar arguments hold when multiple *(=j,* a positive integer*)* constituents are analyzed, and status (clean or impacted) of an onsite location is determined based upon *j* comparisons (one for each analyte). The use of a UPL_1 is not recommended when multiple comparisons are to be made.

3.5.1 Normal Upper Prediction Limit

The sample mean, $\bar{x}$, and the *sd*, *s*, are computed first based upon a defensible background data set. For normally distributed data sets, an upper *(1 – α)*100%* prediction limit is given by the following well known equation:

$$UPL = \bar{x} + t_{((1-\alpha),(n-1))} * s * \sqrt{(1 + 1/n)} \quad (3\text{-}13)$$

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

3.5.2 Lognormal Upper Prediction Limit

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

$$UPL = exp(\bar{y} + t_{((1-\alpha),(n-1))} * s_y * \sqrt{(1 + 1/n)}) \quad (3\text{-}14)$$

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

3.5.3 Gamma Upper Prediction Limit

Given a sample, $x_1, x_2, \ldots, x_n$ of size n from a gamma distribution $G(k,\theta)$, approximate (based upon WH and HW approximations described earlier in Section 3.4.3, Gamma Distribution Upper Tolerance Limits), *(1 – α)*100%* upper prediction limits for a future observation from the same gamma distributed population are given by:

$$\text{Wilson-Hilferty (WH) } UPL = max\left(0, \left(\bar{y} + t_{((1-\alpha),(n-1))} * s_y * \sqrt{1 + 1/n}\right)^3\right) \quad (3\text{-}15)$$

$$\textit{Hawkins-Wixley (HW) UPL} = \left(\bar{y} + t_{((1-\alpha),(n-1))} * s_y * \sqrt{1 + 1/n}\right)^4 \quad (3\text{-}16)$$

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

Note: As noted earlier, the performance of gamma UTLs and gamma UPLs based upon these WH and HW approximations is not well-studied. Interested researchers may want to evaluate their performances via simulation experiments. These approximations are also available in R script.

3.5.4 Nonparametric Upper Prediction Limit

A one-sided nonparametric UPL is simple to compute and is given by the following mth order statistic. One can use linear interpolation if the resulting number, m, given below does not represent a whole number (a positive integer).

$$UPL = X_{(m)}, \text{ where } m = (n + 1) * (1 - \alpha) \quad . \quad (3\text{-}17)$$

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

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

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 24th and 25th order statistics. However, if a 99% UPL needs to be computed, then m = 0.99 * 26 = 25.74, which exceeds 25, the sample size; for such cases, the highest order statistic is used to compute the 99% UPL of the background data set. The largest value(s) should be used with caution to estimate the BTVs.

Since nonparametric upper limits (e.g., UTLs, UPLs) are based upon higher order statistics, often the CC achieved by these nonparametric upper limits is much lower than the specified CC of 0.95, especially when the sample size is small.

3.5.4.1 Upper Prediction Limit Based upon the Chebyshev Inequality

Like a UCL of the mean, the Chebyshev inequality can be used to compute a conservative UPL and is given by the following equation:

$$UPL = \bar{x} + \left[\sqrt{((1/\alpha) - 1) * (1 + 1/n)}\right] s_x$$

This is a nonparametric method since the Chebyshev inequality does not require any distributional assumptions. It should be noted that just like the Chebyshev UCL, a UPL based upon the Chebyshev inequality tends to yield higher estimates of BTVs than the various other methods. This is especially true when skewness is mild (*sd* of log-transformed data is low < 0.75), and the sample size is large ($n > 30$). The user is advised to apply professional judgment before using this method to compute a UPL.

3.5.5 Normal, Lognormal, and Gamma Distribution based Upper Prediction Limits for k-Future Comparisons

A UPL_k95 computed based upon an established background data set represents that statistic such that k future (next, independent and not belonging to the current data set) observations from the target population (e.g., background, comparable to background) will be less than or equal to the UPL_k95 with a CC of 0.95. A UPL_k95 for k (≥1) future observations is designed to compare k future observations; we are 95% sure that "k" future values from the background population will be less than or equal to UPL_k95 with CC of 0.95. In addition to UPL_k, ProUCL also computes an upper prediction limit of the mean of k future observations, UPL_k (mean). A UPL_k (mean) is commonly used in groundwater monitoring applications. A UPL_k controls the false positive error rate by using the Bonferroni inequality based critical values to perform k future comparisons. These UPLs statisfy the relationship: $UPL_1 \leq UPL_2 \leq UPL_3 \leq \ldots \leq UPL_k$. ProUCL can compute an upper prediction limit for any number of , k, future observations.

A normal distribution based $UPL_k(1 - \alpha)$ for k future observations, $x_{n+1}, x_{n+2}, \ldots, x_{n+k}$ is given by the probability statement:

$$P\left(x_{n+1}, x_{n+2}, \ldots, x_{n+k} \leq \bar{x} + t_{((1-\alpha),(n-1))} s\sqrt{1+\frac{1}{n}}\right) = 1 - \alpha \qquad (3\text{-}18)$$

$$UPL_k = \bar{x} + s * t_{((1-\alpha/k),(n-1))}\sqrt{1+\frac{1}{n}}$$

$$UPL_k 95 = \left(\bar{x} + t_{((1-0.05/k),(n-1))} s\sqrt{1+\frac{1}{n}}\right)$$

For an example, a UPL_3 95 for 3 future observations: $x_{01}, x_{02}, x_{03}$is given by:

$$UPL_3 95 = \left(\bar{x} + t_{((1-0.05/3),(n-1))} s\sqrt{1+\frac{1}{n}}\right)$$

A lognormal distribution based UPL_k *(1 - α)* for k future observations, $x_{n+1}, x_{n+2}, \ldots, x_{n+k}$ is given by the following equation:

$$UPL_k = exp\left(\bar{y} + s_y * t_{((1-\alpha/k),(n-1))}\sqrt{1+\frac{1}{n}}\right)$$

A gamma distribution based UPL_k for the next k > *1* (k future observations) are computed similarly using the WH and HW approximations described in Section 3.4.3.

3.5.6 Proper Use of Upper Prediction Limits

It is noted that some users tend to use UPLs without taking their definition and intended use into consideration; this is an incorrect application of a UPL. Some important points to note about the proper use of UPL_1 and UPL_k for k>1 are described as follows.

- When a UPL_k is computed to compare k future observations collected from a site area or a group of MW within an operating unit (OU), it is assumed that the project team will make a decision about the status (clean or not clean) of the site (MWs in an OU) based upon those k future observations.
- The use of an UPL_k implies that a decision about the site-wide status will be made only after k comparisons have been made with the UPL_k. It does not matter if those k observations are collected (and compared) simultaneously or successively. The k observations are compared with the UPL_k as they become available and a decision (about site status) is made based upon those k observations.
- An incorrect use of a UPL_1 95 is to compare many (e.g., 5, 10, 20, etc.) future observations. This results in a higher than 0.05 false positive rate. Similarly, an inappropriate use of a UPL_{100} would be to compare less than 100 (i.e., 10, 20, or 50 observations) future observations. Using a UPL_{100} to compare 10 or 20 observations can potentially result in a high number of false negatives (a test with reduced power), declaring contaminated areas comparable to background.
- The use of other statistical limits such as 95%-95% UTLs (UTL95-95) is preferred to estimate BTVs and not-to-exceed values. The computation of a UTL does not depend upon the number of future comparisons which will be made with the UTL.

3.6 Upper Simultaneous Limits

An *(1 – α) * 100%* upper simultaneous limit (USL) based upon an established background data set is meant to provide coverage for <u>all</u> observations, x_i, $i = 1, 2, \ldots, n$ simultaneously in the background data set. It is implicitly assumed that the data set comes from a single background population and is free of outliers (established background data set). A USL95 represents that statistic such that all observations from the "established" background data set will be less than or equal to the USL95 with a CC of 0.95. A USL95 can be used to perform any number (unknown) of comparisons of future observations. The false positive error rate does not change with the number of comparisons as the purpose of the USL95 is to perform any number of comparisons simultaneously.

<u>Notes:</u> If a background population is established based upon a small data set; as one collects more observations from the background populations, some of the new background observations will exceed the largest value in the existing data set. In order to address these uncertainties, the use of a USL is suggested, provided the data set represents a single population without outliers.

3.6.1 Upper Simultaneous Limits for Normal, Lognormal and Gamma Distributions

The normal distribution based two-sided *(1 – α) 100%* simultaneous interval obtained using the first order Bonferroni inequality (Singh and Nocerino 1995, 1997) is given as follows:

$$P\left(\bar{x} - sd_\alpha^b \le x_i \le \bar{x} + sd_\alpha^b; i := 1,2,\ldots,n\right) = 1 - \alpha \qquad (3\text{-}19)$$

Here, $\left(d_\alpha^b\right)^2$ represents the critical value (obtained using the Bonferroni inequality) of the maximum Mahalanobis distance (Max (MDs)) for α level of significance (Singh 1993). The details about the Mahalanobis distances and computation of the critical values, $\left(d_\alpha^b\right)^2$, can be found in Singh (1993) and

Singh and Nocerino (1997). These values have been programmed in ProUCL version 4.1 and higher versions to compute USLs for any combination of the sample size, *n,* and CC, *(1 - α)*.

The normal distribution based, one-sided *(1 – α) 100% USL* providing coverage for all *n* sample observations is given as follows:

$$P\left(x_i \le \bar{x} + sd_{2\alpha}^{b}; i := 1,2,\ldots,n\right) = 1 - \alpha;$$

$$USL = \bar{x} + s * d_{2\alpha}^{b}; \qquad (3\text{-}20)$$

Here $\left(d_{2\alpha}^{b}\right)^2$ is the critical value of Max (MDs) for a 2*α level of significance.

The lognormal distribution based one-sided *(1 – α) 100%* USL providing coverage for all *n* sample observations is given by the following equation:

$$USL = exp\left(\bar{x} + s * d_{2\alpha}^{b}\right) \qquad (3\text{-}21)$$

A gamma distribution based (using WH approximation), one-sided *(1 – α) 100%* USL providing coverage to all sample observations is given by:

$$USL = max\left(0, \left(\bar{y} + d_{2\alpha}^{b} * s_y\right)^3\right)$$

A gamma distribution based (using the HW approximation), one-sided *(1 – α) 100%* USL providing coverage to all sample observations is given as follows:

$$USL = \left(\bar{y} + d_{2\alpha}^{b} * s_y\right)^4$$

Nonparametric USL: For nonparametric data sets, the largest value, $x_{(n)}$ is used to compute a nonparametric USL. Just like a nonparametric UTL, a nonparametric USL may fail to provide the specified coverage, especially when the sample size is small (e.g., <60). The confidence coefficient actually achieved by a USL can be computed using the same process as used for a nonparametric UTL described in Sections 3.4.4.2 and 3.4.4.3. Specifically, by substituting *r = n* in equation (3-11), the confidence coefficient achieved by a USL can be computed, and by substituting *m=1* in equation (3-12), one can compute the sample size needed to achieve the desired confidence.

Note: Nonparametric USLs, UTLs or UPLs should be used with caution; nonparametric upper limits are based upon order statistics and therefore do not take the variability of the data set into account. Often nonparametric BTVs estimated by order statistics do not achieve the specified CC unless the sample size is fairly large.

Dependence of UTLs and USLs on the Sample Size: For smaller samples (*n* <10), a UTL tends to yield impractically large values, especially when the data set is moderately skewed to highly skewed. For data sets of larger sizes, the critical values associated with UTLs tend to stabilize whereas critical values associated with a USL increase as the sample size increases. Specifically, a USL95 is less than a UTL95-95 for samples of sizes, *n* ≤16, they are equal/comparable for samples of size 17, and a USL95 becomes

greater than a UTL95-95 as the sample size becomes greater than 17. Some examples illustrating the computations of the various upper limits described in this chapter are discussed as follows.

Example 3-1. Consider the real data set used in Example 2-4 of Chapter 2 consisting of concentrations for several constituents of potential concern, including aluminum, arsenic, chromium (Cr), and lead. The computation of background statistics obtained using ProUCL for some of the metals are summarized as follows.

Upper Limits Based upon a Normally Distributed Data Set: The aluminum data set follows a normal distribution as shown in the following GOF Q-Q plot of Figure 3-1.

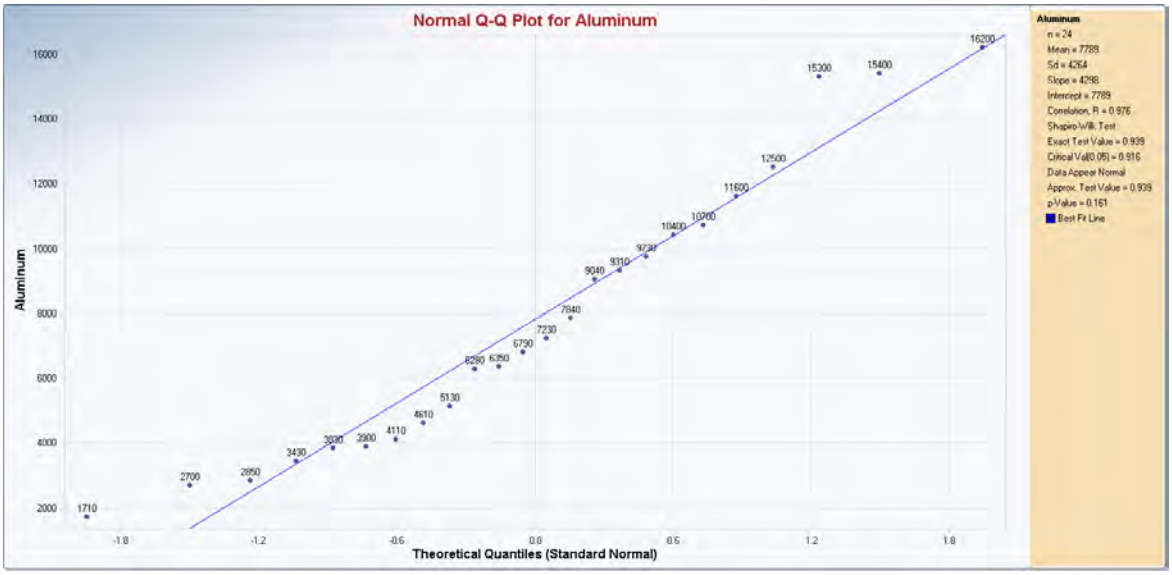

Figure 3-1. Normal Q-Q plot of Aluminum with GOF Statistics

From the normal Q-Q plot shown in Figure 3-1, it is noted that the 3 largest values are higher (but not extremely high) than the rest of the 21 observations. These observations may or may not come from the same population as the rest of the 21 observations.

Table 3-1. BTV Estimated Based upon All 24 Observations

Aluminum			
General Statistics			
Total Number of Observations	24	Number of Distinct Observations	24
Minimum	1710	First Quartile	4058
Second Largest	15400	Median	7010
Maximum	16200	Third Quartile	10475
Mean	7789	SD	4264
Coefficient of Variation	0.547	Skewness	0.542
Mean of logged Data	8.798	SD of logged Data	0.61
Critical Values for Background Threshold Values (BTVs)			
Tolerance Factor K (For UTL)	2.309	d2max (for USL)	2.644
Normal GOF Test			
Shapiro Wilk Test Statistic	0.939	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.916	Data appear Normal at 5% Significance Level	
Lilliefors Test Statistic	0.109	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.181	Data appear Normal at 5% Significance Level	
Data appear Normal at 5% Significance Level			
Background Statistics Assuming Normal Distribution			
95% UTL with 95% Coverage	17635	90% Percentile (z)	13254
95% UPL (t)	15248	95% Percentile (z)	14803
95% USL	19063	99% Percentile (z)	17708

The classical outlier tests (Dixon and Rosner tests) did not identify these 3 data points as outliers. The various upper limits have been computed with and without the 3 high observations and are summarized respectively, in Tables 3-1 and 3-2 as follows. The project team should make a determination based on scientific incvestigation of three extreme values of which statistics should be used to estimate BTVs.

Table 3-2. BTV Estimated Based upon 21 Observations without 3 Higher Values

Aluminum			
General Statistics			
Total Number of Observations	21	Number of Distinct Observations	21
		Number of Missing Observations	3
Minimum	1710	First Quartile	3900
Second Largest	11600	Median	6350
Maximum	12500	Third Quartile	9310
Mean	6669	SD	3215
Coefficient of Variation	0.482	Skewness	0.25
Mean of logged Data	8.676	SD of logged Data	0.549
Critical Values for Background Threshold Values (BTVs)			
Tolerance Factor K (For UTL)	2.371	d2max (for USL)	2.58
Normal GOF Test			
Shapiro Wilk Test Statistic	0.955	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.908	Data appear Normal at 5% Significance Level	
Lilliefors Test Statistic	0.12	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.193	Data appear Normal at 5% Significance Level	
Data appear Normal at 5% Significance Level			
Background Statistics Assuming Normal Distribution			
95% UTL with 95% Coverage	14291	90% Percentile (z)	10789
95% UPL (t)	12344	95% Percentile (z)	11957
95% USL	14964	99% Percentile (z)	14147

Example 3-2. As noted in Example 2-4, chromium concentrations follow a lognormal distribution. The lognormal GOF test is shown in Figure 3-2, and computation of background statistics using a lognormal model are shown in Table 3-3.

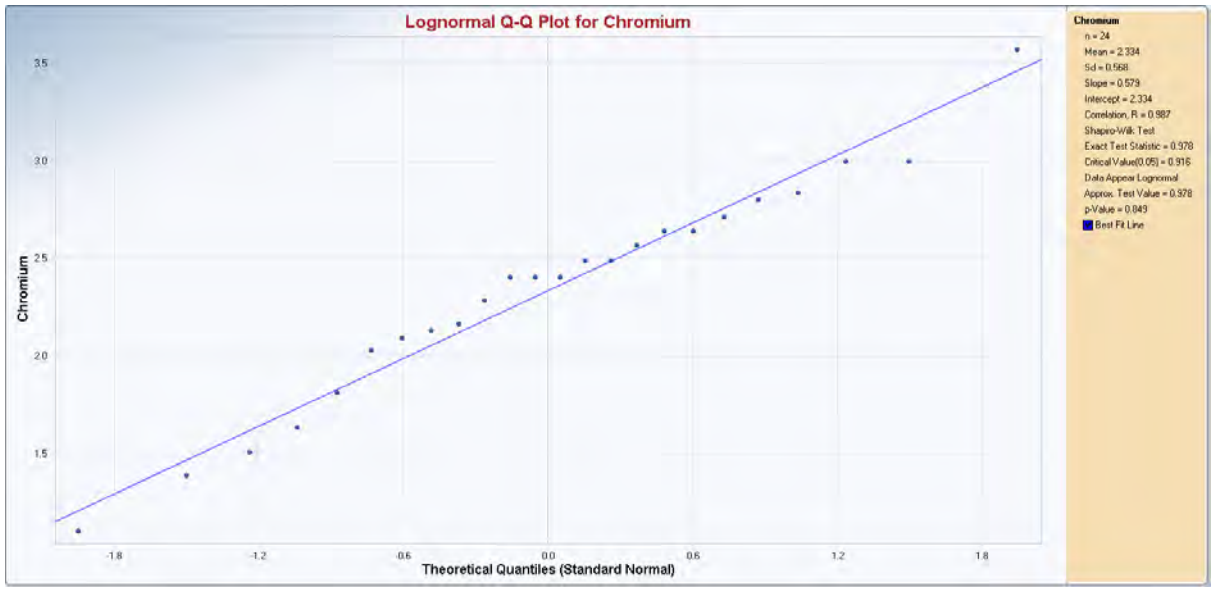

Figure 3-2. Lognormal Q-Q Plot of Chromium with GOF Statistics

Table 3-3. Lognormal Distribution Based UPLs, UTLs, and USLs

Chromium			
General Statistics			
Total Number of Observations	24	Number of Distinct Observations	19
Minimum	3	First Quartile	7.975
Second Largest	20	Median	11
Maximum	35.5	Third Quartile	14.25
Mean	11.97	SD	6.892
Coefficient of Variation	0.576	Skewness	1.728
Mean of logged Data	2.334	SD of logged Data	0.568
Critical Values for Background Threshold Values (BTVs)			
Tolerance Factor K (For UTL)	2.309	d2max (for USL)	2.644
Lognormal GOF Test			
Shapiro Wilk Test Statistic	0.978	**Shapiro Wilk Lognormal GOF Test**	
5% Shapiro Wilk Critical Value	0.916	Data appear Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.128	**Lilliefors Lognormal GOF Test**	
5% Lilliefors Critical Value	0.181	Data appear Lognormal at 5% Significance Level	
Data appear Lognormal at 5% Significance Level			
Background Statistics assuming Lognormal Distribution			
95% UTL with 95% Coverage	38.3	90% Percentile (z)	21.37
95% UPL (t)	27.87	95% Percentile (z)	26.27
95% UPL for Next 5 Observations	43.96	99% Percentile (z)	38.68
95% UPL for Mean of 5 Observations	16.66	95% USL	46.33

Example 3-3. Arsenic concentrations of the data set used in Example 2-4 follow a gamma distribution. The background statistics, obtained using a gamma model, are shown in Table 3-4. Figure 3-3 is the gamma Q-Q plot with GOF statistics.

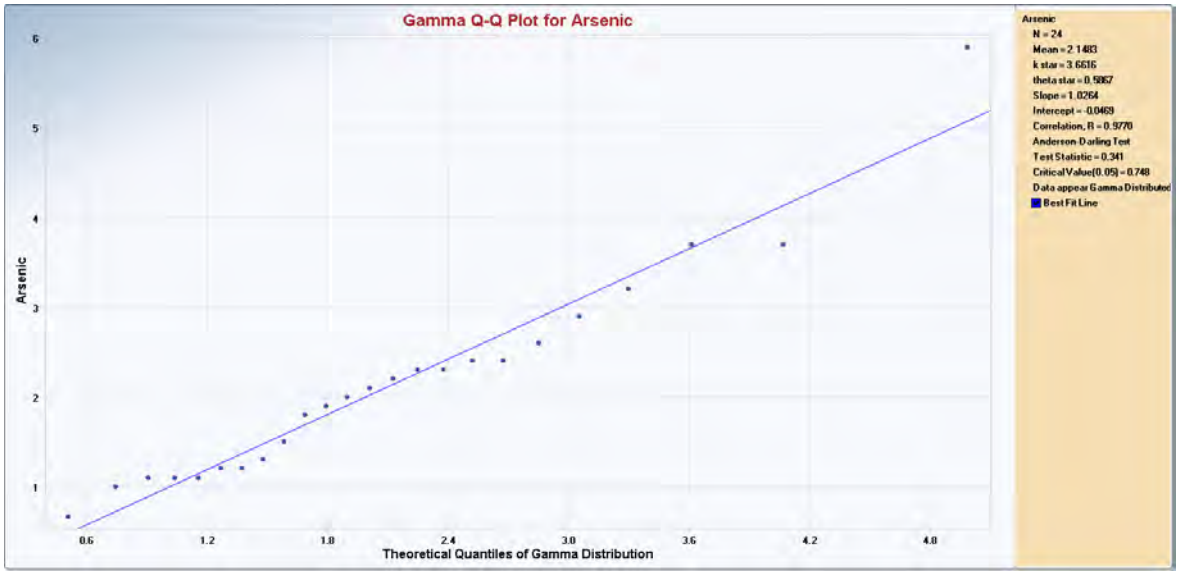

Figure 3-3. Gamma Q-Q plot of Arsenic with GOF Statistics

Table 3-4. Gamma Distribution Based UPLs, UTLs, and USLs

Arsenic			
General Statistics			
Total Number of Observations	24	Number of Distinct Observations	18
Minimum	0.66	First Quartile	1.2
Second Largest	3.7	Median	2.05
Maximum	5.9	Third Quartile	2.45
Mean	2.148	SD	1.159
Coefficient of Variation	0.54	Skewness	1.554
Mean of logged Data	0.639	SD of logged Data	0.51
Critical Values for Background Threshold Values (BTVs)			
Tolerance Factor K (For UTL)	2.309	d2max (for USL)	2.644
Gamma GOF Test			
A-D Test Statistic	0.341	**Anderson-Darling Gamma GOF Test**	
5% A-D Critical Value	0.748	Detected data appear Gamma Distributed at 5% Significance Level	
K-S Test Statistic	0.114	**Kolmogrov-Smirnoff Gamma GOF Test**	
5% K-S Critical Value	0.179	Detected data appear Gamma Distributed at 5% Significance Level	
Detected data appear Gamma Distributed at 5% Significance Level			
Gamma Statistics			
k hat (MLE)	4.153	k star (bias corrected MLE)	3.662
Theta hat (MLE)	0.517	Theta star (bias corrected MLE)	0.587
nu hat (MLE)	199.3	nu star (bias corrected)	175.8
MLE Mean (bias corrected)	2.148	MLE Sd (bias corrected)	1.123
Background Statistics Assuming Gamma Distribution			
95% Wilson Hilferty (WH) Approx. Gamma UPL	4.345	90% Percentile	3.654
95% Hawkins Wixley (HW) Approx. Gamma UPL	4.397	95% Percentile	4.264
95% WH Approx. Gamma UTL with 95% Coverage	5.382	99% Percentile	5.574
95% HW Approx. Gamma UTL with 95% Coverage	5.524		
95% WH USL	6.074	95% HW USL	6.294

Example 3-4. Lead concentrations of the data set used in Example 2-4 do not follow a discernible distribution. The various nonparametric background statistics for lead are shown in Table 3-5.

Table 3-5. Nonparametric UPLs, UTLs, and USLs for Lead in Soils

Lead			
General Statistics			
Total Number of Observations	24	Number of Distinct Observations	18
Minimum	4.9	First Quartile	10.43
Second Largest	98.5	Median	14
Maximum	109	Third Quartile	19.25
Mean	22.49	SD	26.83
Coefficient of Variation	1.193	Skewness	2.665
Mean of logged Data	2.743	SD of logged Data	0.771
Critical Values for Background Threshold Values (BTVs)			
Tolerance Factor K (For UTL)	2.309	d2max (for USL)	2.644
Nonparametric Distribution Free Background Statistics			
Data do not follow a Discernible Distribution (0.05)			
Nonparametric Upper Limits for Background Threshold Values			
Order of Statistic, r	24	95% UTL with 95% Coverage	109
Approximate f	1.263	Confidence Coefficient (CC) achieved by UTL	0.708
95% Percentile Bootstrap UTL with 95% Coverage	109	95% BCA Bootstrap UTL with 95% Coverage	109
95% UPL	106.4	90% Percentile	44.81
90% Chebyshev UPL	104.6	95% Percentile	91.72
95% Chebyshev UPL	141.8	99% Percentile	106.6
95% USL	109		

Note: As mentioned before, nonparametric upper limits are computed by higher order statistics, or by some value in between (based upon linear interpolation) the higher order statistics. In practice, nonparametric upper limits do not provide the desired coverage to the population parameter (upper threshold) unless the sample size is large. From Table 3-5, it is noted that a UTL95-95 is estimated by the maximum value in the data set of size 24. However, the CC actually achieved by UTL95-95 (and also by USL95) is only 0.708. *Therefore, one may want to use other upper limits such as 95% Chebyshev UPL = 141.8 to estimate a BTV.*

Note: As mentioned earlier, for symmetric and mildly skewed nonparametric data sets (when *sd* of logged data is <=0.5), one can use the normal distribution to compute percentiles, UPLs, UTLs and USLs.

Example 3-5: Why Use a Gamma Distribution to Model Positively Skewed Data Sets?

The data set considered in Example 2-2 of Chapter 2 is used to illustrate the deficiencies and problems associated with the use of a lognormal distribution to compute upper limits. The data set follows a lognormal as well as a gamma model; the various upper limits, based upon a lognormal and a gamma model, are summarized as follows. The data set is highly skewed with *sd* of logged data = 1.68. The largest value in the data set is 169.8, the UTL95-95 and UPL95 based upon a lognormal model are 799.7 and 319 both of which are significantly higher than the maximum value of 169.8. UTL95-95s based upon WH and HW approximations to gamma distributions are 245.3 and 285.6; UPLs based upon WH and HW

approximations are 163.5 and 178.2 which appear to represent more reasonable estimates of the BTV. These statistics are summarized in Table 3-6 (lognormal) and Table 3-7 (gamma) below.

Table 3-6. Background Statistics Based upon a Lognormal Model

X			
General Statistics			
Total Number of Observations	25	Number of Distinct Observations	25
Minimum	0.349	First Quartile	5.093
Second Largest	164.3	Median	18.77
Maximum	169.8	Third Quartile	72.62
Mean	44.09	SD	51.34
Coefficient of Variation	1.164	Skewness	1.294
Mean of logged Data	2.835	SD of logged Data	1.68
Critical Values for Background Threshold Values (BTVs)			
Tolerance Factor K (For UTL)	2.292	d2max (for USL)	2.663
Lognormal GOF Test			
Shapiro Wilk Test Statistic	0.948	**Shapiro Wilk Lognormal GOF Test**	
5% Shapiro Wilk Critical Value	0.918	Data appear Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.135	**Lilliefors Lognormal GOF Test**	
5% Lilliefors Critical Value	0.177	Data appear Lognormal at 5% Significance Level	
Data appear Lognormal at 5% Significance Level			
Background Statistics assuming Lognormal Distribution			
95% UTL with 95% Coverage	799.7	90% Percentile (z)	146.5
95% UPL (t)	319	95% Percentile (z)	269.7

Table 3-7. Background Statistics Based upon a Gamma Model

X			
General Statistics			
Total Number of Observations	25	Number of Distinct Observations	25
Minimum	0.349	First Quartile	5.093
Second Largest	164.3	Median	18.77
Maximum	169.8	Third Quartile	72.62
Mean	44.09	SD	51.34
Coefficient of Variation	1.164	Skewness	1.294
Mean of logged Data	2.835	SD of logged Data	1.68
Critical Values for Background Threshold Values (BTVs)			
Tolerance Factor K (For UTL)	2.292	d2max (for USL)	2.663
Gamma GOF Test			
A-D Test Statistic	0.374	**Anderson-Darling Gamma GOF Test**	
5% A-D Critical Value	0.794	Detected data appear Gamma Distributed at 5% Significance Level	
K-S Test Statistic	0.113	**Kolmogrov-Smirnoff Gamma GOF Test**	
5% K-S Critical Value	0.183	Detected data appear Gamma Distributed at 5% Significance Level	
Detected data appear Gamma Distributed at 5% Significance Level			
Gamma Statistics			
k hat (MLE)	0.643	k star (bias corrected MLE)	0.592
Theta hat (MLE)	68.58	Theta star (bias corrected MLE)	74.42
nu hat (MLE)	32.15	nu star (bias corrected)	29.62
MLE Mean (bias corrected)	44.09	MLE Sd (bias corrected)	57.28
Background Statistics Assuming Gamma Distribution			
95% Wilson Hilferty (WH) Approx. Gamma UPL	163.5	90% Percentile	115
95% Hawkins Wixley (HW) Approx. Gamma UPL	178.2	95% Percentile	159.4
95% WH Approx. Gamma UTL with 95% Coverage	245.3	99% Percentile	266.8
95% HW Approx. Gamma UTL with 95% Coverage	285.6		

CHAPTER 4

Computing Upper Confidence Limit of the Population Mean Based upon Left-Censored Data Sets Containing Nondetect Observations

4.1 Introduction

Nondetect (ND) observations are inevitable in most environmental data sets. It should be noted that the estimation of the mean and *sd*, and the computation of the upper limits (e.g., upper confidence limits [UCLs], upper tolerance intervals [UTLs]) are two different tasks. For left-censored data sets with NDs, in addition to the availability of good estimation methods, the availability of rigorous statistical methods which account for data skewness is needed to compute the decision making statistics such as UCLs, UTLs, and UPLs. For left-censored data sets consisting of multiple detection limits (DLs) or reporting limits (RLs), ProUCL 4.0 (2007) and its higher versions offer methods to: 1) impute NDs using regression on order statistics (ROS) methods; 2) perform GOF tests; 3) estimate the mean, standard deviation (*sd*), and standard error of the mean; and 4) compute skewness adjusted upper limits (e.g., UCLs, UTLs, UPLs). Based upon KM (Kaplan and Meier1958) estimates, and the distribution and skewness of detected observations, several upper limit computation methods which adjust for data skewness have also been incorporated in ProUCL.

For left-censored data sets with NDs, Singh and Nocerino (2002) compared the performances of the various estimation methods (in terms of bias and MSE) to estimate the population mean, μ_1, and *sd,* σ_1 including the MLE method (Cohen 1950, 1959), restricted MLE (RMLE) method (Perrson and Rootzen 1977); Expectation Maximization (EM) method (Gleit 1985), EPA Delta lognormal method (EPA 1991; Hinton 1993), Winsorization method (Gilbert 1987), and regression on order statistics (ROS) method (Helsel 1990). Singh, Maichle, and Lee (EPA 2006) performed additional simulation experiments to study and evaluate the performances (in terms of bias and MSE) of KM and ROS methods for estimating the population mean. They concluded that the KM method yields better estimates, in terms of bias, of population mean in comparison with other estimation methods including the LROS (ROS on logged data) method. Singh, Maichle, and Lee (EPA 2006) also studied the performances, in terms of coverage probabilities, of some parametric and nonparametric UCL computation methods based upon ROS, KM, and other estimation methods. They concluded that for skewed data sets, KM estimates based UCLs computed using bootstrap methods (e.g., BCA bootstrap, bootstrap-t) and Chebyshev inequality perform better than the Student's t statistic UCL and percentile bootstrap UCL computed using ROS and KM estimates as described in Helsel (2005, 2012) and incorporated in NADA packages (2013).

As mentioned above, computing good estimates of the mean and *sd* based upon left-censored data sets addresses only half of the problem. The main issue is computing decision statistics (UCL, UPL, UTL) which account for NDs as well as uncertainty and data skewness inherently present in environmental data sets. Until recently (ProUCL 4.0, 4.00.05, 4.1; Singh, Maichle, and Lee 2006), not much guidance was available on how to compute the various upper limits (UCLs, UPLs, UTLs) based upon skewed left-censored data sets with multiple DLs. For left-censored data sets, the existing literature (Helsel 2005, 2012) suggests computing upper limits using a Student's t-type statistic and percentile bootstrap methods on KM and LROS estimates without adjusting for data skewness. Environmental data sets tend to follow skewed

distributions, and UCL95s and other upper limits computed using methods described in Helsel (2005, 2012) will under estimate the population parameters of interest including EPCs and background threshold values.

In earlier versions of ProUCL (ProUCL versions 4 [2007, 2009, 2010]), all evaluated estimation methods including the poor performing methods (MLE and RMLE, and Winsorization methods) and better performing, in terms of bias in the mean estimate, estimation (KM method) and UCL computation methods (BCA bootstrap, bootstrap-t) were incorporated in ProUCL version 4 (2007, 2009, 2010). Currently, the KM estimation method is widely used in environmental applications to compute parametric (when detected data follow a known distribution) and nonparametric upper limits needed to estimate environmental parameters of interest such as the population mean and upper thresholds of a background population. Note that the KM method is now included in a recent EPA RCRA groundwater monitoring guidance document (2009).

Due to the poor performances and/or failure to correctly verify probability distributions for data sets with multiple DLs, the parametric MLE and RMLE methods, the normal ROS and the Winsorization estimation methods for computing upper limits are no longer available in ProUCL version 5.0/5.1/5.2. The normal ROS method is available only under the **Stats/Sample Sizes** module of ProUCL 5.0/5.1/5.2 to impute NDs based upon the normal distribution assumption for advanced users who may want to use the imputed data in other graphical and exploratory methods such as scatter plots, box plots, cluster analysis and principal component analysis (PCA). The estimation methods for computing upper limits retained in ProUCL 5.0/5.1/5.2 include the two ROS (lognormal, and gamma) methods and the KM method. The KM estimation method can be used on a wide-range of skewed data sets with multiple DLs and NDs exceeding detected observations. Also, the substitution methods such as replacing NDs by half of their respective DLs and the H-UCL method (EPA [2009e] recommends its use in Chapter 15) have been retained in ProUCL 5.0/5.1/5.2 for historical reasons, and academic and research purposes. Inclusion of the DL/2 method (substitution of ½ the DL for NDs) in ProUCL should not be inferred as a recommended method. The developers of ProUCL are not endorsing the use of the DL/2 estimation method or H-UCL computation method.

Note on the use of letter k (*k*): Not to get confused with the use of letter "k (*k*)" in this Chapter and in Chapters 2, 3, 4, and 5. Following the standard statistical terminology, "*k*" is used to denote the shape parameter of a gamma distribution, $G(k,\theta)$ as described in Chapter 2*;* "k" is used to represent future (next) observations (Chapter 3 and 5), and "k" is used to represent the number of ND observations present in a data set (Chapters 4 and 5).

Notes on Skewness of Left-Censored Data Sets: Skewness of a data set is measured as a function of *sd*, σ (or its estimate, $\hat{\sigma}$) of log-transformed data. Like uncensored full data sets, σ, or its estimate, $\hat{\sigma}$, of the log-transformed detected data is used to get an idea about the skewness of a data set consisting of ND observations. This information along with the distribution of detected observations is used to decide which UCL should be used to estimate the EPC and other upper limits for data sets consisting of both detects and NDs. For data sets with NDs, output sheets generated by ProUCL display the *sd*, $\hat{\sigma}$, of log-transformed data based upon detected observations. For a 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, *k* hat (and also the bias corrected MLE, *k* star) of the shape parameter, *k*, based upon detected observations.

4.2 Pre-processing a Data Set and Handling of Outliers

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” statistical population as a UCL is computed for the mean of a "single" statistical population. In addition to representing "wrong" values (e.g., typos, lab errors), outliers may also represent observations coming from population(s) significantly different from the dominant population whose parameters (mean, upper percentiles) we are trying to estimate based upon the available data set.

4.2.1 Assessing the Influence of Outliers and Disposition of Outliers

One can argue against “not using the outliers” while estimating the various environmental parameters such as the EPCs and BTVs. An argument can be made that outlying observations are inevitable and can be naturally-occurring (not impacted by site activities) in some environmental media (and therefore in data sets). For example, in groundwater applications, a few elevated values may be considered to be naturally occurring and as such may not represent the impacted MW data values.

To assess the influence of outliers on the various statistics (upper limits) of interest, it is suggested to compute all relevant statistics using data sets with outliers and without outliers, and then compare the results. This extra step often helps the project team/users to see the direct potential influence of outlier(s) on the various statistics of interest (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 (mean, EPC, BTV) under consideration.

4.2.2 Avoid Data Transformation

Data transformations are performed to achieve symmetry of the data set and be able to use parametric (normal distribution based) methods on transformed data. In most environmental applications, the cleanup decisions are made based on statistics and results computed in the original 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. Often, the back-transformed statistics (UCL of the mean) in the original scale suffer from an unknown amount of transformation bias; many times the transformation bias can be unacceptably large (for highly skewed data sets) leading to incorrect decisions.

The use of a gamma model does not require any data transformation therefore whenever applicable the use of a gamma distribution is suggested to model skewed data sets. In cases when a data set in the original scale cannot be modeled by a normal or a gamma distribution, it is better to use nonparametric methods rather than testing or estimating parameters in the transformed space. For data sets which do not follow a discernible parametric distribution, nonparametric and computer intensive bootstrap methods can be used to compute the upper limits needed to estimate environmental parameters. Several of those methods are available in ProUCL for data sets consisting of NDs with multiple DLs.

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

In addition to environmental scientists, ProUCL is also used by students and researchers. Therefore, for historical and comparison purposes, the substitution method of replacing NDs by half of the associated DLs (DL/2) is retained in ProUCL; that is the DL/2 GOF tests, UCL, UPL, and UTL computation methods have been retained in ProUCL 5.0 and newer for historical reasons, and comparison and academic purposes. For data sets with NDs, output sheets generated by ProUCL display a message suggesting that DL/2 is not a recommended method. It is suggested that the use of the DL/2 (t) UCL method (UCL computed using Student's t-statistic) be avoided when estimating a EPC or BTVs, unless the data set consists of only a small fraction of NDs (<5%) and the data are mildly skewed. The DL/2 UCL computation method does not provide adequate coverage (Singh, Maichle, and Lee 2006) for the population mean, even for censoring levels as low as 10% or 15%. This is contrary to statements (EPA 2006b) made that the DL/2 UCL method can be used for lower (≤ 20%) censoring levels. The coverage provided by the DL/2 (t) UCL method deteriorates fast as the censoring intensity, percentage of NDs, increases and/or data skewness increases.

4.2.4 Minimum Data Requirement

Whenever possible, it is suggested that a sufficient number of samples be collected to satisfy the requirements for the data quality objectives (DQOs) for the site. Often, in practice, it is not feasible to collect the number of samples as determined by DQOs-based sample size formulae. Therefore, some rule-of-thumb minimum sample size requirements are described in this section. At the minimum, collect a data set consisting of about 10 observations to compute reasonably reliable and accurate estimates of EPCs (UCLs) and BTVs (UPLs, UTLs). The availability of at least 15 to 20 observations is desirable to compute UCLs and other upper limits based upon re-sampling bootstrap methods. Some of these issues have also been discussed in Chapter 1 of this Technical Guide. However, from a theoretical point of view, ProUCL can compute various statistics (KM UCLs) based upon data sets consisting of at least 3 detected observations. The accuracy of the decisions based upon statistics computed using such small data sets remains questionable.

4.3 Goodness-of-Fit (GOF) Tests and Skewness for Left-Censored Data Sets

It is not easy to assess and verify the distribution of data sets with NDs, especially when multiple DLs are present and those DLs exceed the detected values. One can perform GOF tests on detected data and consider/expect that NDs (not the DLs) also follow the same distribution of detected data. For data sets with NDs, ProUCL has GOF tests for normal, lognormal, and gamma distributions which are also supplemented with graphical Q-Q plots. GOF tests in ProUCL include: 1) exclude all NDs; 2) replace NDs by their DL/2s; and 3) ROS methods. In the environmental literature (Helsel 2005, 2012), some other graphs such as censored probability plots have also been described. However, the usefulness of those graphs in the computation of decision making statistics is not clear. Some practitioners have criticized that ProUCL does not offer censored probability plots, therefore, even though those graphs do not provide additional useful information, ProUCL offers those graphs as well.

Formally, let x_1, x_2, ..., x_n (including k NDs and $(n-k)$ detected measurements) represent a random sample of n observations obtained from a population under investigation (e.g., background area, or an area of concern [AOC]). Out of the n observations, k: $1 \le k \le n$, values are reported as NDs lying below one or more

DLs, and the remaining *(n-k)* observations represent the detected values. Such data sets consisting of ND observations are called left-censored data sets. The *(n-k)* detected values are ordered and are denoted by $x_{(i)}$*; i: =k+1, k+2, ..., n.* The *k* ND observations are denoted by $x_{(ndi)}$ *; i:=1,2,...k.* The detected observations might come from a well-known parametric distribution such as a normal, a lognormal, or a gamma distribution, or from a population with a nondiscernible distribution. Using the Statistical Tests module of ProUCL, one can use GOF tests (described in Chapter 2) to assess the distribution of detected observations.

Like uncensored full data sets, for data sets with NDs, the skewness and data distribution of detected values plays an important role in selecting appropriate estimates of EPCs and BTVs. If the data set obtained by excluding the NDs is skewed, the data set consisting of all detects and NDs most likely will also be skewed. Therefore, for data sets with NDs, it is important to determine the distribution and skewness of the data set obtained by excluding the NDs. This information helps in selecting appropriate parametric or nonparametric methods to compute the various upper limits which account for NDs and adjust for data variability and skewness. For skewed data sets, a UCL (and other limits) of the mean computed using KM estimates in the t-statistic UCL equation or obtained using the percentile bootstrap method tend to fail to achieve the specified coverage for the population mean. One may also want to know the distribution of detects to determine which statistical methods should be used on the ROS or KM estimates when computing the various upper limits. There is no need to determine the plotting positions/percentiles when assessing the distribution of detected observations. Also, the use of the substitution DL/2 method yields a data set of size *n*, and GOF methods described in Chapter 2 can be used to determine the distribution of the data set thus obtained. Similarly, any of the GOF methods described in Chapter 2 can be used on the data set of size *n* obtained using a ROS method (normal, lognormal, and gamma). The ROS method is described in Section 4.5.

4.4 Nonparametric Kaplan-Meier (KM) Estimation Method

The KM estimation method (Kaplan and Meier 1958), also known as the product limit estimation (PLE) method, is a substitution method based upon a distribution function estimate, like the sample distribution function, except that the KM method adjusts for censoring. The KM method is commonly used in survival analysis (e.g., dealing with right-censored data associated with terminally ill patients) and various other biomedical applications. A brief description of the KM method to estimate the population mean and *sd*, and standard error (SE) of the mean for left-censored data sets is described in this section. For details, refer to Kaplan and Meier (1958) and the report prepared by Bechtel Jacobs Company for the DOE (2000). The properties of the KM method are well researched (Gillespie, Chen *et al.* 2010). Specifically, the KM estimator represents a consistent estimator and for large data sets the KM estimator is asymptotically efficient and normally distributed (Gu, Zhang 1993).

Formally, let $x_1, x_2, ..., x_n$ represent *n* data values of a left-censored data set. Let $\hat{\mu}_{KM}$ and $\hat{\sigma}^2_{KM}$ represent KM estimates of the mean and variance based upon such a data set with NDs. 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 detected 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 $x_i \leq x'_j$. Also, let $x_{(1)}$ denote the smallest x_i. Then

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

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

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

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

Note that in the last equality statement of $\tilde{F}(x)$ above, $\tilde{F}(x) = 0$ when $x_{(1)}$ is a detect, and is undefined when $x_{(1)}$ is a ND. An estimate of the population mean based upon the KM method is given as follows.

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

Using the PLE (or KM) method, an estimate of the SE of the mean is given by the following equation.

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

Where k = number of ND observations, and

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

The KM variance is computed as follows:

$$\hat{\sigma}^2_{KM} = \hat{\mu}_{(x^2)-KM} - (\hat{\mu}_{(x)-KM})^2 \qquad (4\text{-}3)$$

$$\hat{\mu}_{(x)-KM} = KM\ mean\ of\ the\ data, x$$

$$\hat{\mu}_{(x^2)-KM} = Km\ mean\ of\ the\ square\ of\ the\ data, x\ (second\ raw\ moment)$$

In addition to the KM mean, ProUCL computes both the SE of the mean given by (4-2) and the variance given by (4-3). The SE is used to estimate EPCs (e.g., UCLs) whereas the variance is used to compute BTV estimates (e.g., UTLs, USLs). The KM method in ProUCL can be used directly on left-censored environmental data sets without requiring any flipping of data and back flipping of the KM estimates and other statistics (e.g., flipping LCL to compute a UCL) which may be burdensome for most users and practitioners.

Note: Decision making statistics (e.g., UPLs and UTLs) used in background evaluations projects require good estimates of the population standard deviation, *sd*. The decision statistics (e.g., UTLs) obtained using the direct estimate of *sd* (Equation 4-3) and an indirect "back door" estimate of *sd* (Helsel 2012b) can differ significantly, especially for skewed data sets. An example illustrating this issue is described as follows.

Example 4-1 (Oahu Data Set): Consider the moderately skewed well-cited Oahu data set (Helsel 2012b). A direct KM estimate of the *sd* obtained using equation (4-3) is σ= 0.713; and an indirect KM estimate of *sd* = sqrt (24)*SE = 4.899 * 0.165 = 0.807 (Helsel 2012b, p 87). A UTL95-95 (direct) = 2.595 and a UTL95-95 (based upon indirect estimate of *sd*) = 2.812. The discrepancy between the two estimates of *sd* and upper limits (e.g., UTL95-95) computed using the two estimates increases with skewness.

Cautionary notes for NADA (2013) in R Users: It is well known that the KM method yields a good (in terms of bias) estimate of the population mean (Singh, Maichle, and Lee 2006). However, the use of KM estimates in the Student's t-statistic based UCL equation or percentile bootstrap method as included in NADA packages do not guarantee that those UCLs will provide the desired (e.g., 0.95) coverage for the population mean in all situations. Specifically, it is highly likely that for moderately skewed to highly skewed data sets (determined using detected values) the Student's t-statistic or percentile bootstrap method based UCLs computed using KM estimates will fail to provide the desired coverage to the population mean, as these methods do not account for skewness. Several UCL (and other limits) computation methods based upon KM estimates which adjust for data skewness are available in ProUCL 5.0 and newer; those methods were not available in ProUCL 4.1.

4.5 Regression on Order Statistics (ROS) Methods

In this guidance document and in ProUCL software, LROS represents the ROS (also known as robust ROS) method for a lognormal distribution and GROS represents the ROS method for a gamma distribution. The ROS methods impute NDs based upon a hypothesized distribution such as a gamma or a lognormal distribution. The "**Stats/Sample Sizes**" menu option of ProUCL can be used to impute and store imputed NDs along with the original detected values in additional columns generated by ProUCL. ProUCL assigns self-explanatory titles for those generated columns. It is a good idea to store the imputed values to determine the validity of the imputed NDs and assess the distribution of the complete data set consisting of detects and imputed NDs. As a researcher, one may want to have access to imputed NDs to be used by other methods such as regression analysis and PCA. Moreover, one cannot easily perform multivariate methods on data sets with NDs; and the availability of imputed NDs makes it possible for researchers to use multivariate methods on data sets with NDs. The developers believe that statistical methods to evaluate data sets with NDs require further investigation and research. Providing the imputed values along with the detected values may be helpful to practitioners conducting research in this area. For data sets with NDs, ProUCL also performs GOF tests on data sets obtained using the LROS and GROS methods. The ROS methods yield a data set of size n with $(n-k)$ original detected observations and k imputed NDs. The full data set of size n thus obtained can be used to compute the various summary statistics, and to estimate the EPCs and BTVs using methods described in Chapters 2 and 3 of this technical guidance document.

In a ROS method, the distribution (e.g., gamma, lognormal) of the $(n-k)$ detected observations is assessed first; and assuming that the k ND observations, $x_1, x_2, ..., x_k$ follow the same distribution (e.g., gamma or a lognormal distribution when used on logged data) of the $(n-k)$ detected observations, the NDs are imputed using an OLS regression line obtained using the $(n-k)$ pairs: (ordered detects, hypothesized quantiles). Earlier versions of ProUCL software also included the normal ROS (NROS) method for computing the various upper limits. The use of NROS on environmental data sets (with positive values) tends to yield unfeasible and negative imputed ND values; and the use of negative imputed NDs yields biased and incorrect results (e.g., UCL, UTLs). Therefore, the NROS method is no longer available in the **UCLs/EPCs** and **Upper Limits/BTVs** modules of ProUCL. Instead, when detected data follow a normal distribution, the use of KM estimates in normal equations is suggested for computing the upper limits as described in Chapters 2 and 3.

4.5.1 Computation of the Plotting Positions (Percentiles) and Quantiles

Before computing the n hypothesized (lognormal, gamma) quantiles, $q_{(i)}$; $i:=k+1, k+2,...,n$, and $q_{(ndi)}$; $i:= 1, 2, ..., k,$ the plotting positions (also known as percentiles) need to be computed for the n observations with k NDs and $(n-k)$ detected values. There are several methods available in the literature (Blom 1958; Barnett, 1976; Singh and Nocerino, 1995, Johnson and Wichern, 2002) to compute the plotting positions (percentiles). Note that plotting positions for the three ROS methods: LROS, GROS, and NROS are the same. For a full data set of size n, the most commonly used plotting position for the i^{th} observation (ordered) is given by $(i - 3/8) / (n + 1/4)$ or $(i - 1/2)/n; i:=1,2,...,n.$ These plotting positions are routinely used to generate Q-Q plots based upon full uncensored data sets (Singh 1993; Singh and Nocerino 1995; ProUCL 3.0 and higher versions). For the single DL case (with all observations below the DL reported as NDs), ProUCL uses Blom's percentiles, $(i - 3/8) / (n + 1/4)$ for normal and lognormal distributions, and uses empirical percentiles given by $(i - 1/2)/n$ for a gamma distribution. Specifically, for normal and lognormal distributions, once the plotting positions have been obtained, the n normal quantiles, $q_{(i)}$ are computed using the probability statement: $P(Z \leq q_{(i)}) = (i - 3/8) / (n + 1/4)$, i : = 1, 2, …, n , where Z represents a standard normal variate (SNV). The gamma quantiles are computed using the probability statement: $P(X \leq q_{(i)}) = (i - 1/2) /n$, i : = 1, 2, …, n , where X represents a gamma (~constant *chi-square) random variable.

In case multiple DLs are present with NDs potentially exceeding the detected observations, the plotting positions (percentiles) are computed using methods that adjust for multiple DLs. The details of the computation of such plotting positions (percentiles), $p_{i;}$ $i: =1, 2, ..., n$, for data sets with multiple DLs or with ND observations exceeding the DLs are given in Helsel (2005) and also in Singh, Maichle, and Lee (2006), a document that can be downloaded freely from the ProUCL website. The associated hypothesized quantiles, $q_{(i)}$ are obtained by using the following probability statements:

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

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

Once the n plotting positions have been computed, the n quantiles, $q_{(ndi)}$; $i:= 1, 2, ..., k,$ 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.

Example 4-2 (Pyrene Data Set): Using the well-cited She's (1997) pyrene data set (Helsel 2012b) of size n=56, the plotting positions (same for NROS, LROS, and GROS) and LROS and GROS quantiles (denoted by Q) generated by ProUCL are summarized in Table 4-1. The gamma quantiles are computed using the MLE estimates of shape and scale parameters.

4.5.2 Computing OLS Regression Line to Impute NDs

An ordinary least squares (OLS) regression model is obtained by fitting a linear straight line to the $(n-k)$ ordered (in ascending order) 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 ordered observations. The hypothesized quantiles are obtained for all of the n data values by using the hypothesized distribution for the $(n-k)$ detected observations. The quantiles associated with $(n-k)$ detected

values are denoted by $q_{(i)}$; $i:=k+1, k+2,...,n$, and the k quantiles associated with ND observations are denoted by $q_{(ndi)}$; $i:= 1, 2, ..., k.$.

An OLS regression line is obtained first by 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 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:

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

Table 4-1. Plotting Positions, Gamma and Lognormal (Normal) Quantiles (Q)

Pyrene	D_pyrene	Percentiles	Gamma-Q (Hat)	Normal-Q
28	0	0.01818162	4.339031664	-2.0928422
31	1	0.063635671	14.41837445	-1.5249509
32	1	0.090908101	20.46764835	-1.3351838
34	1	0.118180531	26.60142257	-1.1841314
35	0	0.048484321	11.07571144	-1.6597307
35	0	0.096968641	21.82204786	-1.2990194
40	1	0.163634582	37.09766155	-0.9796292
47	1	0.181816202	41.41222876	-0.9084654
48	1	0.199997822	45.80230241	-0.841629
58	0	0.109089721	24.54516765	-1.2313836
59	1	0.238013937	55.25114882	-0.7127057
63	1	0.257848432	60.33991965	-0.6499928
64	1	0.277682927	65.54790077	-0.5897387
64	1	0.297517422	70.88334133	-0.5315541
67	1	0.317351916	76.3549565	-0.4751166
67	1	0.337186411	81.97203137	-0.4201542
67	1	0.357020906	87.74452837	-0.3664333
72	1	0.376855401	93.68320235	-0.3137502
73	1	0.396689895	99.79972782	-0.2619243
84	1	0.41652439	106.1068416	-0.210793
86	0	0.218179443	50.27369978	-0.7783565
86	1	0.455406297	119.0785779	-0.1120136
87	1	0.474453708	125.7564975	-0.0640789
94	1	0.49350112	132.6689087	-0.016291
98	1	0.512548532	139.8343146	0.0314597
100	1	0.531595943	147.2733809	0.0792823
103	1	0.550643355	155.009301	0.1272869
103	1	0.569690767	163.0682381	0.1755869
105	1	0.588738178	171.4798683	0.2243003
107	1	0.60778559	180.2780505	0.2735521
110	1	0.626833001	189.501663	0.323477
111	1	0.645880413	199.1956546	0.374222
117	0	0.332463912	80.62092045	-0.4331197
119	1	0.67836071	216.9648821	0.4631197
119	1	0.691793595	224.8326032	0.5009408
122	0	0.35261324	86.44778239	-0.3782748
122	1	0.721551168	243.5316694	0.5874557
132	1	0.737875855	254.6417923	0.6368105
133	1	0.754200542	266.4543419	0.687768
133	1	0.770525229	279.0660434	0.7405777
138	1	0.786849916	292.5950126	0.7955388
163	0	0.200793651	45.99627612	-0.8387898
163	0	0.401587302	101.3388027	-0.2492407
163	0	0.602380952	177.7401595	0.2595148
163	1	0.812301587	315.8763629	0.8864096
174	0	0.410714286	104.2387681	-0.225708
187	1	0.837662338	342.4104955	0.9848957
190	1	0.853896104	361.6446136	1.0532907
222	1	0.87012987	383.1239177	1.1270053
238	1	0.886363636	407.448911	1.2074141
273	1	0.902597403	435.4987969	1.2964944
289	1	0.918831169	468.636129	1.3972525
306	1	0.935064935	509.1426864	1.5146142
333	1	0.951298701	561.2940357	1.6575784
459	1	0.967532468	634.6836503	1.8457049
2982	1	0.983766234	759.9003157	2.1386067

When ROS is used on transformed data (e.g., log-transformed), then ordered values, $x_{(i)}$; i: $= k + 1, k + 2, \ldots, n$ represent ordered detected data in that transformed scale (e.g., log-scale, Box-Cox (BC)-type transformation). Equation (4-4) is then used to impute or estimate the ND values. Specifically, for quantile, $q_{(ndi)}$ corresponding to the i^{th} ND, the imputed ND is given by $x_{(ndi)} = a + bq_{(ndi)}$; $i:=1,2,...k.$ When there is only a single DL and all values lying below the DL represent ND observations, then the quantiles corresponding to those ND values typically are lower than the quantiles associated with the detected observations. However, when there are multiple DLs, and when some of those DLs exceed detected values, then quantiles, $q_{(ndi)}$ corresponding to some of those ND values might become greater than the quantiles, $q_{(i)}$ associated with some of the detected values.

4.5.2.1 Influence of Outliers on Regression Estimates and Imputed NDs

Like all other statistics, it is well-known (Rousseeuw and Leroy 1987; Singh and Nocerino 1995; Singh and Nocerino 2002) that presence of outliers (detects) also distorts the regression estimates of slope and intercept which are used to impute NDs based upon a ROS method. It is noted that for skewed data sets with outliers, the imputed values computed using the ROS method on raw data in the original scale become negative (e.g., GROS method). Therefore, inclusion of outliers (e.g., impacted locations) can yield distorted statistics and upper limits computed using the ROS method. This issue is also discussed later in this chapter.

Note: It is noted that a linear regression line can be obtained even when only two detected observations are available. Therefore, methods (e.g., ROS) discussed here and incorporated in ProUCL can be used on data sets with 2 or more detected observations. However, to obtain a reliable OLS model (slope and intercept) and imputed NDs for computation of defensible upper limits, enough (> 4-6 as a rule of thumb, more are desirable) detected observations should be made available.

4.5.3 ROS Method for Lognormal Distribution

Let *Org* stand for the data in the original unit and *Ln* stand for the data in the natural logarithmic unit. The LROS method may be used when the log-transformed detected data follow a lognormal distribution. For the LROS method, the OLS model given by (4-4) is obtained using the log-transformed detected data and the corresponding normal quantiles. Using the OLS linear model on log-transformed, detected observations, the NDs in log-transformed scale are imputed corresponding to the *k* normal quantiles, $q_{(ndi)}$ associated with the ND observations which are back-transformed in original, *Org* scale by exponentiation.

4.5.3.1 Fully Parametric Log ROS Method

Once the *k* NDs have been imputed, the sample mean and *sd* can be computed using the back-transformation formula (El Shaarawi, 1989) given by equation (4-5) below. This method is called the 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* imputed ND (in log scale) values. Assuming lognormality, El-Shaarawi (1989) suggested estimating μ and σ by back-transformation using the following equations as one of the several ways of computing these estimates. The estimates given by equation (4-5) are neither unbiased nor have minimum variance (Gilbert 1987). Therefore, it is recommended to avoid the use of this version of ROS method on log-transformed data to compute UCL95s and other statistics. This method is not available in the ProUCL software.

$$\hat{\mu}_{Org} = exp(\hat{\mu}_{LN} + \hat{\sigma}^2_{LN}/2), and\ \hat{\sigma}^2_{Org} = \hat{\mu}^2_{Org}(exp(\hat{\sigma}^2_{LN}) - 1) \qquad (4\text{-}5)$$

4.5.3.2 Robust ROS Method on Log-Transformed Data

The robust ROS method is performed on log-transformed data as described above. In the robust ROS method, ND observations are first imputed in the log-scale, based upon a linear ROS model fitted to the log-transformed detects and normal quantiles. The imputed NDs are transformed back in the original scale by exponentiation. The process of using the ROS method based upon a lognormal distribution and imputing NDs by exponentiation does not yield negative estimates for ND values; perhaps that is why it got the name robust ROS (or LROS in ProUCL). This process yields a full data set of size *n*, and methods described in

Chapters 2 and 3 can be used to compute the decision statistics of interest including estimates of EPCs and BTVs. If the detected observations follow a lognormal, the data set consisting of detects and imputed NDs also follow a lognormal distribution. As expected, the process of imputing NDs using the LROS method does not reduce the skewness of the data set and therefore, appropriate methods need to be used to compute upper limits (Chapters 2 and 3) which provide specified (e.g., 0.95) coverage by adjusting for skewness.

Note: The use of the robust ROS method has become quite popular. Helsel (2012b) suggests the use of a classical t-statistic or a percentile bootstrap method to compute a UCL of the mean based upon the full data set obtained using the LROS method. These methods are also available in his NADA packages. However, these methods do not adjust for skewness and for moderately skewed to highly skewed data sets, and UCLs based upon these two methods fail to provide the specified coverage to the population mean. For skewed data sets, methods described in Chapter 2 can be used on LROS data sets to compute UCLs of the mean.

Example 4-3 (Oahu Data Set). Consider the Oahu arsenic data set of size 24 with 13 NDs. The detected data set of size 11 follows a lognormal distribution as shown in Figure 4-1; this graph simply represents a Q-Q plot of detects and does not account for NDs when computing quantiles. The censored probability plot is shown in Figure 4-2; its details can be found in the literature (Chapter 15 of Unified Guidance, EPA 2009e). A censored probability plot is also based upon detected observations and it computes quantiles by accounting for NDs. The LROS data set consisting of 11 detects and 13 imputed NDs also follows a lognormal distribution as shown in Figure 4-3. Summary statistics and LROS UCLs are summarized in Table 4-2.

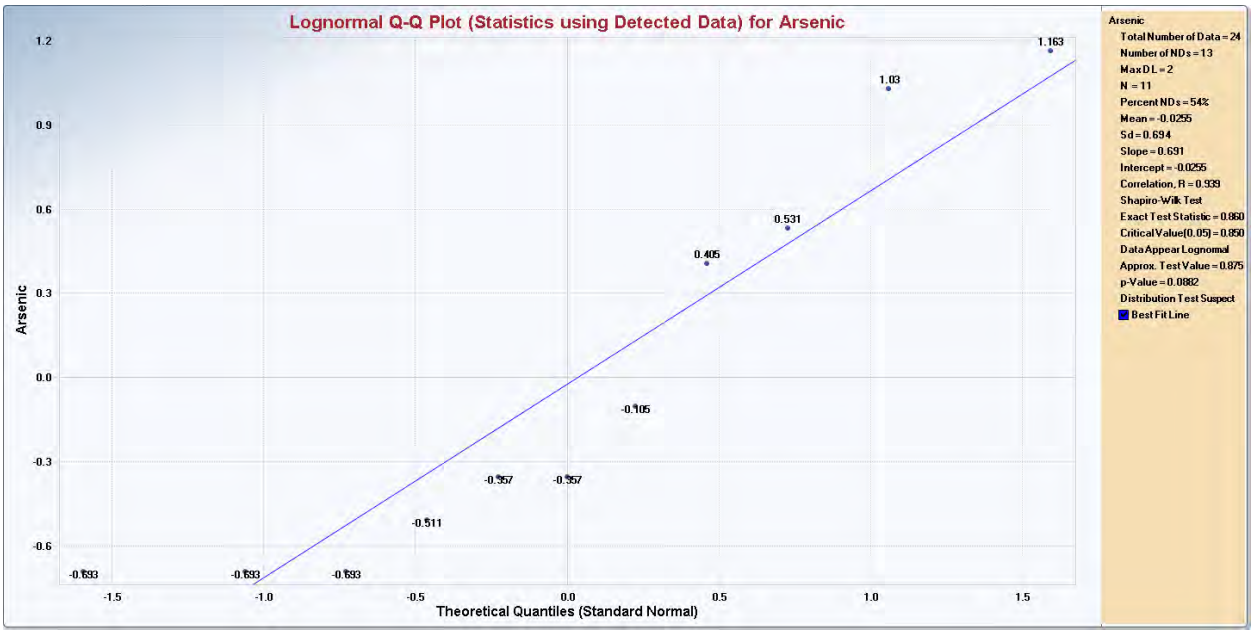

Figure 4-1. Lognormal GOF Test on Detected Oahu Data Set—Does not Account for NDs to Compute Quantiles

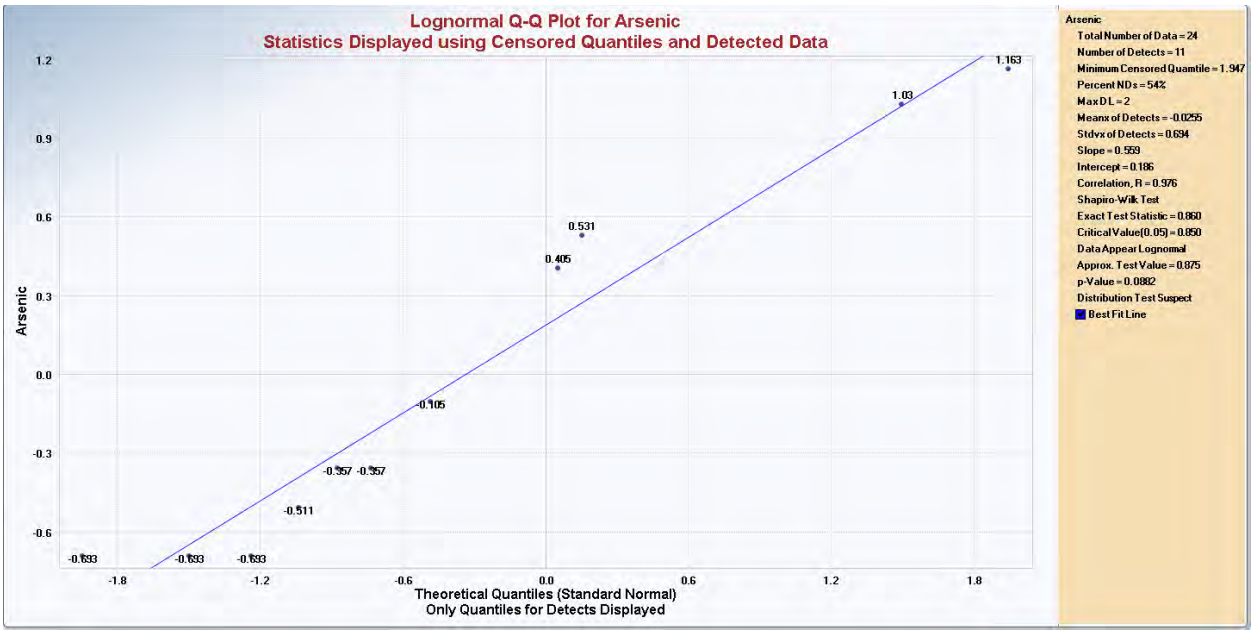

Figure 4-2. Lognormal Censored Probability Plot (Oahu Data)—Uses Only Detects but Accounts for NDs to Compute Quantiles

Note: The two graphs displayed in Figures 4-1 and 4-2 provide similar information about data distributions, as GOF tests simply use detected values (and not quantiles). Both graphs are okay without any preference.

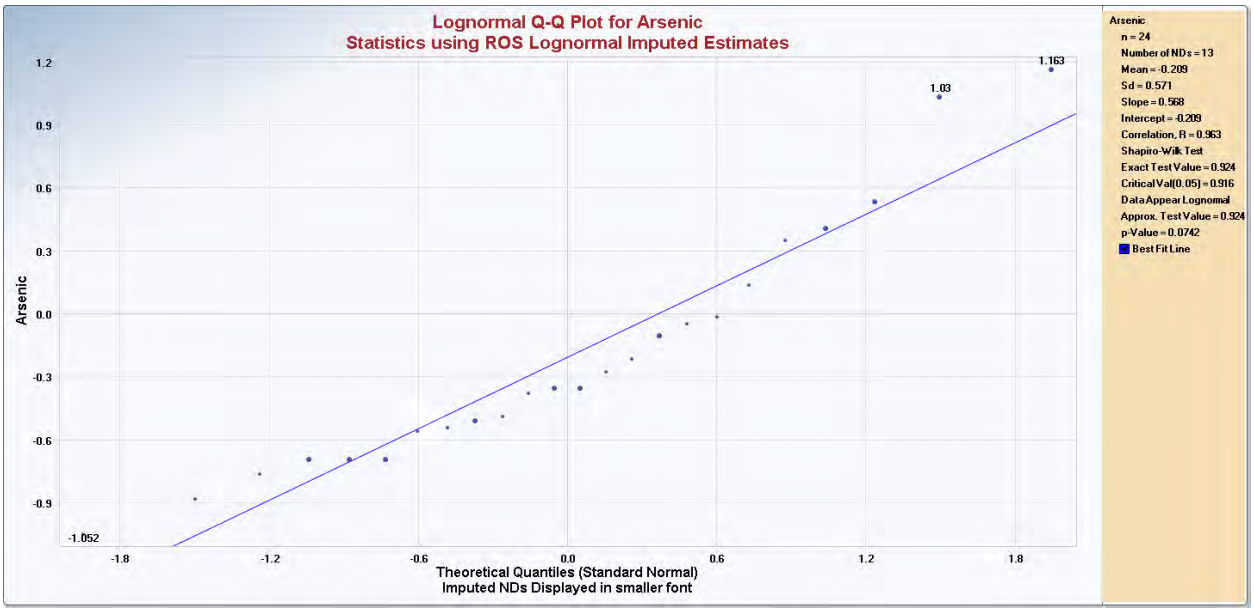

Figure 4-3. Lognormal GOF Test on LROS Data Obtained Using the Oahu Data Set

Table 4-2. Summary Statistics and UCL95 Based upon LROS data

Arsenic

General Statistics			
Total Number of Observations	24	Number of Distinct Observations	10
Number of Detects	11	Number of Non-Detects	13
Number of Distinct Detects	8	Number of Distinct Non-Detects	3
Minimum Detect	0.5	Minimum Non-Detect	0.9
Maximum Detect	3.2	Maximum Non-Detect	2
Variance Detects	0.931	Percent Non-Detects	54.17%
Mean Detects	1.236	SD Detects	0.965
Median Detects	0.7	CV Detects	0.78
Skewness Detects	1.322	Kurtosis Detects	0.517
Mean of Logged Detects	-0.0255	SD of Logged Detects	0.694

Lognormal GOF Test on Detected Observations Only		
Shapiro Wilk Test Statistic	0.86	**Shapiro Wilk GOF Test**
5% Shapiro Wilk Critical Value	0.85	Detected Data appear Lognormal at 5% Significance Level
Lilliefors Test Statistic	0.229	**Lilliefors GOF Test**
5% Lilliefors Critical Value	0.267	Detected Data appear Lognormal at 5% Significance Level
Detected Data appear Lognormal at 5% Significance Level		

Lognormal ROS Statistics Using Imputed Non-Detects			
Mean in Original Scale	0.972	Mean in Log Scale	-0.209
SD in Original Scale	0.718	SD in Log Scale	0.571
95% t UCL (assumes normality of ROS data)	1.224	95% Percentile Bootstrap UCL	1.22
95% BCA Bootstrap UCL	1.308	95% Bootstrap t UCL	1.375
95% H-UCL (Log ROS)	1.218		

The data set is moderately skewed with *sd* of logged detects equal to 0.694. All methods tend to yield comparable results. One may want to use a 95% BCA bootstrap UCL or a bootstrap-t UCL to estimate the EPC. However, the detected data follow a gamma distribution, therefore ProUCL recommends gamma UCLs as shown in the following section.

4.5.3.3 Gamma ROS Method

Many positively skewed data sets tend to follow a lognormal as well as a gamma distribution. Singh, Singh, and Iaci (2002) noted that the gamma distribution is better suited to model positively skewed environmental data sets. When a moderately skewed to highly skewed data set (uncensored data set or detected values in a left-censored data set) follows a gamma, as well as, a lognormal distribution, the use of a gamma distribution tends to result in more stable and realistic estimates of EPCs and BTVs (Examples 2-2 and 3-2, Chapters 2 and 3). Furthermore, when using a gamma distribution to compute decision statistics such as a UCL of the mean, one does not have to transform the data and back-transform the resulting UCL into the original scale.

Let $x_{(k+1)} \leq x_{(k+2)} \leq \ldots \leq x_{(n)}$ represent the *(n-k)* ordered detected values. If *(n-k)* detected observations follow a gamma distribution (can be verified using GOF tests in ProUCL) then the NDs can be imputed using the

OLS line (4-4) based upon (n - k) pairs given by: (n - k) gamma quantiles, ordered (n - k) detected observations). Let x_{nd1}, x_{nd2}, …, x_{ndk}, x_{k+1}, x_{k+2}, …, x_n be a random sample (with k NDs and $(n\text{-}k)$ detects) of size n where the detected $(n\text{-}k)$ observations follow a gamma distribution, $G(k,\theta)$.

Note: Not to get confused with k, the shape parameter of a gamma distribution, G(k,θ), which is different from k, the number of ND observations. Due to these notations used in the statistical literature and also in ProUCL software and output sheet, the same letter k is used for the shape parameter of a gamma distribution and number of NDs.

The n plotting positions, p_i; $i:=1,2,\ldots,n$ used to compute the gamma quantiles are computed for each observation (detected and nondetected) using the methods described earlier in Section 4.5.1. To compute n gamma quantiles associated with the n plotting positions (percentiles, empirical probabilities), one needs to estimate the gamma parameters, k and θ based upon the $(n\text{-}k)$ detected values. This process may have some effect on the accuracy of the estimated gamma quantiles (which use an estimated value of the shape parameter, k), and consequently on the accuracy of the imputed NDs. The availability of enough (at least 8-10) detected gamma distributed observations is suggested to compute the estimates of k and θ.

Let $\hat{k}$ and $\hat{\theta}$ represent the MLEs of k and θ, respectively, based upon detected data.

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

$$\int_0^{z_{0i}} f(\chi^2_{2\hat{k}} = (i - 1/2)/n; \quad i:= \ 1, 2, \ldots, n \quad \text{(Single DL Case)} \tag{4-6}$$

$$\int_0^{z_{0i}} f(\chi^2_{2\hat{k}}) d\chi^2_{2\hat{k}} \ = \ p_i; \ \ i:= 1, 2, \ldots, n \quad \text{(Multiple DL Case)} \tag{4-7}$$

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 above. The process of computing plotting positions, p_i, $i:=1,2,\ldots,n$, for left-censored data sets with multiple DLs has been incorporated in ProUCL. The inverse chi-square algorithm function (AS91) from Best and Roberts (1975) has been used to compute the inverse chi-square percentage points, z_{0i}, as given by the above equations. Using the OLS line (4-4) fitted to the (n - k) detected pairs, one can impute the k NDs resulting in a full data set of size $n = k + (n - k)$.

Notes about GROS for smaller values of k (e.g., ≤): In the ProUCL 5.0 Technical Guide (and its earlier versions) and ProUCL software, a suggestion was made that GROS may not be used when the shape parameter, k is less than 0.1 or less than 0.5. However, during late 2014, some users pointed out that k should be higher. Therefore, starting with version of ProUCL 5.1 now suggests that GROS may not be used for values of $k \leq 1.0$. It should be pointed out that the GROS algorithm incorporated in ProUCL works well for values of k > 2.

The GROS method incorporated in ProUCL does not appear to work well for smaller values of k or its MLE estimate, $\hat{k}$ (e.g., ≤1). The algorithm used to compute gamma quantiles is not efficient enough and does not perform well for smaller values of k. The developers thus far have not found time to look into this issue. In

January 2015, the developers of ProUCL requested the statistical community (via the American Statistical Association's section on environmental statistics and/or personal communication) to provide code/algorithms which may be used to improve the computation of gamma quantiles for smaller values of *k*.

For now, GROS may not be used when the data set with detected observations (used to compute OLS regression line) consists of outliers and/or is highly skewed (e.g., estimated values of *k* are small such as <=1.0). When the estimated value (MLE) of the shape parameter, *k*, based upon detected data is small (<= 1.0), or when the data set consists of many tied NDs at multiple DLs with a high percentage of NDs (>50%), the GROS tends to not perform well and often yields negative imputed NDs, due to outliers distorting the OLS regression. Since environmental concentration data are non-negative, one needs to replace the imputed negative values by a small positive value such as 0.1, 0.001. In ProUCL, negative imputed values are replaced by 0.01. *The use of such imputed values tends to yield inflated values of sd, UCLs, and BTV estimates (e.g., UPLs, UTLs)*.

Preferred Method: Alternatively, when detected data follow a gamma distribution, one can use KM estimates (described above) in gamma distribution based equations to compute UCLs (and other limits) which account for data skewness, unlike KM estimates when used in normal UCL equations. This hybrid gamma-KM method for computing upper limits is available in ProUCL. The details are provided in Section 4.6. The hybrid KM-gamma method yields reasonable UCLs and accounts for NDs as well as data skewness as demonstrated in Example 4-4.

Note: It is noted that when $\hat{k}^*>1$, UCLs based upon the GROS method and gamma UCLs computed using KM estimates tend to yield comparable UCLs from practical a point of view. This can also be seen in Example 4-4 below.

Example 4-4 (Oahu Data Set Continued): The detected data set of size 11 follows a gamma distribution as shown in Figure 4-4. The GROS data consisting of 11 detects and 13 imputed NDs also follows a gamma distribution as shown in Figure 4-5. Summary statistics and GROS UCLs are summarized in Table 4-3 following Figure 4-5. Since the data set is only mildly skewed all methods (GROS and Hybrid KM-Gamma) yield comparable results.

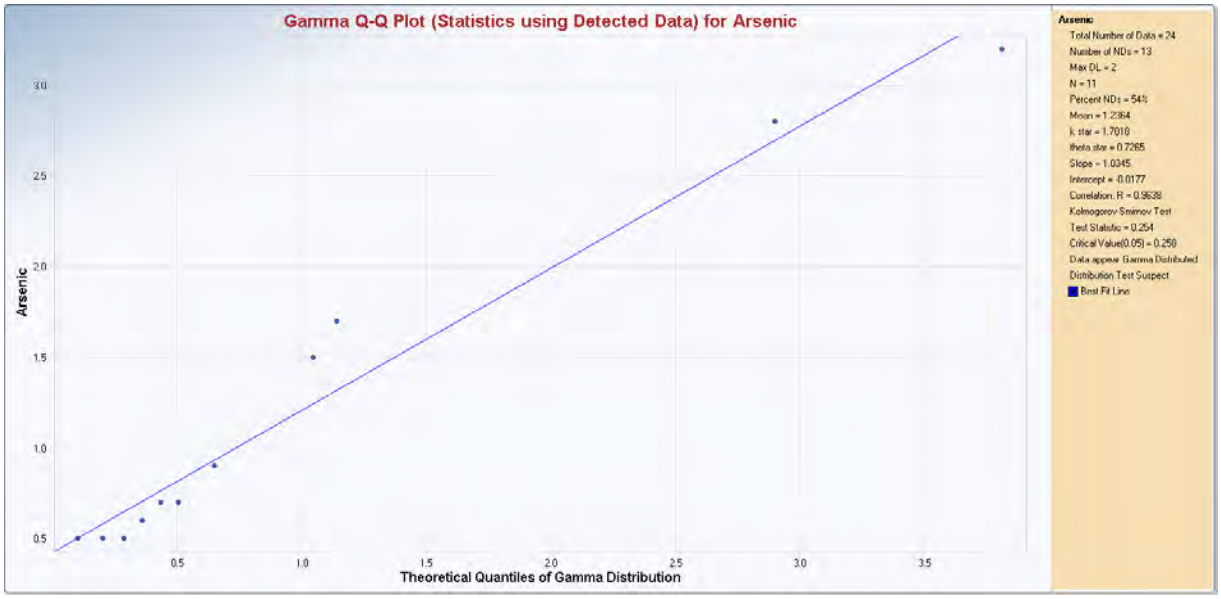

Figure 4-4. Gamma GOF Test on Detected Concentrations of the Oahu Data Set

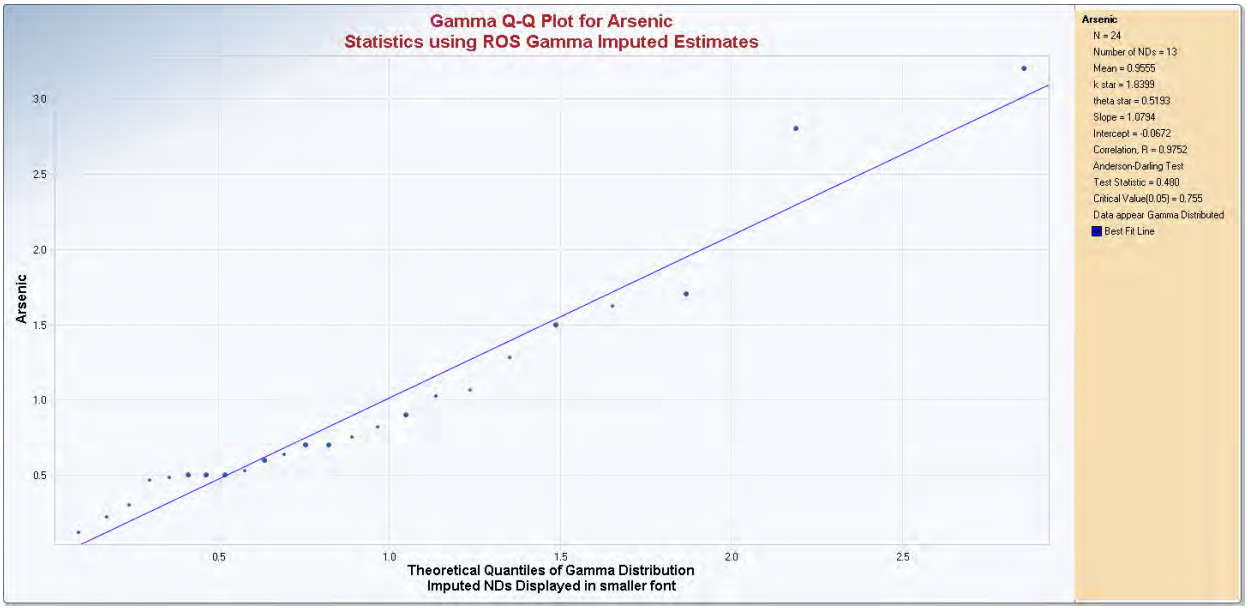

Figure 4-5. Gamma GOF Test on GROS Data Obtained Using the Oahu Data Set

Table 4-3. Summary Statistics and UCL95 Based upon Gamma ROS data

Minimum	0.119	Mean	0.956
Maximum	3.2	Median	0.7
SD	0.758	CV	0.793
k hat (MLE)	2.071	k star (bias corrected MLE)	1.84
Theta hat (MLE)	0.461	Theta star (bias corrected MLE)	0.519
nu hat (MLE)	99.41	nu star (bias corrected)	88.32
MLE Mean (bias corrected)	0.956		
MLE Sd (bias corrected)	0.704	Adjusted Level of Significance (β)	0.0392
Approximate Chi Square Value (88.32, α)	67.65	Adjusted Chi Square Value (88.32, β)	66.38
95% Gamma Approximate UCL (use when n>=50)	1.247	95% Gamma Adjusted UCL (use when n<50)	1.271
Kaplan-Meier (KM) Statistics Using Normal Critical Values			
Mean	0.949	Standard Error of Mean	0.165
SD	0.713	95% KM (BCA) UCL	1.192
95% KM (t) UCL	1.231	95% KM (Percentile Bootstrap) UCL	1.219
95% KM (z) UCL	1.22	95% KM Bootstrap t UCL	1.374
90% KM Chebyshev UCL	1.443	95% KM Chebyshev UCL	1.667
97.5% KM Chebyshev UCL	1.977	99% KM Chebyshev UCL	2.588
Gamma Kaplan-Meier (KM) Statistics			
k hat (KM)	1.771	nu hat (KM)	85.02
Approximate Chi Square Value (85.02, α)	64.77	Adjusted Chi Square Value (85.02, β)	63.53
95% Gamma Approximate KM-UCL (use when n>=50)	1.246	95% Gamma Adjusted KM-UCL (use when n<50)	1.27
Suggested UCL to Use			
95% KM (t) UCL	1.231	95% GROS Adjusted Gamma UCL	1.271
95% Adjusted Gamma KM-UCL	1.27		

ProUCL suggests using GROS UCL of 1.27.

4.6 A Hybrid KM Estimates and Distribution of Detected Observations Based Approach to Compute Upper Limits for Skewed Data

The KM method yields good estimates of the population mean and *sd*. Since it is hard to verify and justify the distribution of an entire left-censored data set consisting of detects and NDs with multiple DLs, it is suggested that the KM method be used to compute estimates of the mean, *sd*, and standard error of the mean. Depending upon the distribution and skewness of detected observations, one can use KM estimates in parametric upper limit computation formulae to compute upper limits including UCLs, UPLs, UTLs, and USLs. The use of this hybrid approach will yield more appropriate skewness adjusted upper limits than those obtained using KM estimates in normal distribution based UCL and UTL equations. Depending upon the distribution of detected data, ProUCL computes upper limits using KM estimates in parametric (normal, lognormal, and gamma) equations to compute the various upper limits. The use of this hybrid approach has also been suggested in Chapter 15 of EPA (2009e) to compute upper limits using KM estimates in the lognormal distribution based equations to compute the various upper limits.

ProUCL computes a 95% UCL of the mean based upon the KM method using: 1) the standard normal critical value, z_α and Student's t-critical value, $t_{\alpha,(n-1)}$; 2) bootstrap methods including the percentile bootstrap method, the bias-corrected accelerated (BCA) bootstrap method, and bootstrap-t method, and 3) the Chebyshev inequality. Additionally, when detected observations of a left- censored data set follow a gamma

or a lognormal distribution, ProUCL also computes KM UCLs and other upper limits using a lognormal or a gamma distribution. The use of these methods yields skewness adjusted upper limits. For a gamma distributed detected data, UCLs based upon the GROS and gamma distribution on KM estimates are generally in good agreement unless the data set is highly skewed (with estimated values of shape parameter, $k \leq 1$), or contains of outliers, or consists of many NDs (e.g., >50%) with NDs tied at multiple DLs. The various UCL computation formulae based upon KM estimates and incorporated in ProUCL are described as follows.

4.6.1 Detected Data Set Follows a Normal Distribution

Based upon Student's t-statistic, a 95% UCL of the mean based upon the KM estimates is as follows:

$$KM\ UCL95\ (t) = \hat{\mu} + t_{.95,(n-1)}\sqrt{\hat{\sigma}^2_{SE}} \qquad (4\text{-}8)$$

The above KM UCL (t) represents a good estimate of the EPC when detected data are normally distributed or mildly skewed. However, KM UCLs, computed using a normal or t-critical value, do not account for data skewness. The various bootstrap methods for left-censored data described in Section 4.7 can also be used on KM estimates to compute UCLs of the mean.

4.6.2 Detected Data Set Follows a Gamma Distribution

For highly skewed gamma distributed left-censored data with a large percentage of NDs and several NDs tied at multiple RLs, the GROS method tends to yield impractical, negative imputed values for NDs. It is also well known that the OLS estimates get distorted by outliers, therefore, GROS estimates and upper limits also get distorted when outliers are present in a data set.

In order to avoid these situations, one can use the gamma distribution on KM estimates to compute the various upper limits provided the detected data follow a gamma distribution. Using the properties of the gamma distribution, an estimate of the shape parameter, *k*, is computed based upon a KM mean and a KM variance. The mean and variance of a gamma distribution are given as follows:

Mean=$k^*\theta$, and

Variance = $k^*\theta^2$

Substituting a KM mean, $\hat{\mu}_{KM}$, and a KM variance, $\hat{\sigma}^2_{KM}$, in the above equations, an estimate of the shape parameter, *k*, is computed by using the following equation:

$$\hat{k} = \hat{\mu}_{KM}/\hat{\sigma}^2_{KM}$$

Using $\hat{\mu}_{KM}$, $\hat{\sigma}^2_{KM}$, *n*, and $\hat{k}$ in equations (2-34) and (2-35), gamma distribution based approximate and adjusted UCLs of the mean can be computed. Similarly, for gamma distributed left-censored data sets with detected observations following a gamma distribution, KM mean and KM variance estimates can be used to compute gamma distribution based upper limits described in Chapter 3. ProUCL computes gamma

distribution and KM estimates based UCLs and upper limits to estimate BTVs when detected data follow a gamma distribution.

Notes: It should be noted that the KM method does not require concentration data to be positive. In radio chemistry, the DLs (or minimum detectable concentration [MDC]) for the various radionuclides are often reported as negative values. Statistical models such as a gamma distribution cannot be used on data sets consisting of negative values. However, the hybrid gamma-KM method described above can be used on radionuclides data provided detected activities are all positive and follow a gamma distribution. One can compute KM estimates using the entire data sets consisting of negative NDs and detected positive values. Those KM estimates can be used to compute gamma UCLs described above provided $\hat{\mu}_{KM}$>0.

4.6.3 Detected Data Set Follows a Lognormal Distribution

The EPA RCRA (2009) guidance document suggests computing KM estimates on logged data and computing a lognormal H-UCL based upon the H-statistic. ProUCL computes lognormal and KM estimates based UCLs and upper limits to estimate BTVs when detected data follow a lognormal distribution. Like uncensored lognormally distributed data sets, for moderately skewed to highly skewed left-censored data sets, the use of a lognormal distribution on KM estimates tends to yield unrealistically high values of the various decision statistics; especially when the data sets are of sizes less than 30 to 50.

Example 4-5 (Oahu Data Set Continued): It was noted earlier that the detected Oahu data set follows a gamma as well as a lognormal distribution. The hybrid normal, lognormal and gamma UCLs obtained using the KM estimates are summarized in Table 4-4 as follows.

The hybrid Gamma UCL is 1.27, close to the UCL obtained using the GROS method of 1.271 (Example 4-4). The H-UCL as suggested in EPA (2009e) is 1.155 which appears to be a little lower than the other LROS BCA bootstrap UCL of 1.308 (Table 4-2).

Table 4-4. UCL95 Based on Hybrid KM Method and Normal, Lognormal and Gamma Distribution

Kaplan-Meier (KM) Statistics using Normal Critical Values and other Nonparametric UCLs			
Mean	0.949	Standard Error of Mean	0.165
SD	0.713	95% KM (BCA) UCL	1.228
95% KM (t) UCL	1.231	95% KM (Percentile Bootstrap) UCL	1.21
95% KM (z) UCL	1.22	95% KM Bootstrap t UCL	1.368
90% KM Chebyshev UCL	1.443	95% KM Chebyshev UCL	1.667
97.5% KM Chebyshev UCL	1.977	99% KM Chebyshev UCL	2.588
Gamma GOF Tests on Detected Observations Only			
A-D Test Statistic	0.787	**Anderson-Darling GOF Test**	
5% A-D Critical Value	0.738	Detected Data Not Gamma Distributed at 5% Significance Level	
K-S Test Statistic	0.254	**Kolmogrov-Smirnoff GOF**	
5% K-S Critical Value	0.258	Detected data appear Gamma Distributed at 5% Significance Level	
Detected data follow Appr. Gamma Distribution at 5% Significance Level			
Gamma Statistics on Detected Data Only			
k hat (MLE)	2.257	k star (bias corrected MLE)	1.702
Theta hat (MLE)	0.548	Theta star (bias corrected MLE)	0.727
nu hat (MLE)	49.65	nu star (bias corrected)	37.44
MLE Mean (bias corrected)	1.236	MLE Sd (bias corrected)	0.948
Gamma Kaplan-Meier (KM) Statistics			
k hat (KM)	1.771	nu hat (KM)	85.02
Approximate Chi Square Value (85.02, α)	64.77	Adjusted Chi Square Value (85.02, β)	63.53
95% Gamma Approximate KM-UCL (use when n>=50)	1.246	95% Gamma Adjusted KM-UCL (use when n<50)	1.27
UCLs using Lognormal Distribution and KM Estimates when Detected data are Lognormally Distributed			
KM Mean (logged)	-0.236	95% H-UCL (KM -Log)	1.155
KM SD (logged)	0.547	95% Critical H Value (KM-Log)	2.023
KM Standard Error of Mean (logged)	0.137		

Example 4-6. A real data set of size 55 with 18.8% NDs is considered next. The data set can be downloaded from the ProUCL website. The minimum detected value is 5.2 and the largest detected value is 79000, *sd* of detected logged data is 2.79 suggesting that the data set is highly skewed. The detected data follow a gamma as well as a lognormal distribution as shown in Figures 4-6 and 4-7. It is noted that GROS data set with imputed values follows a gamma distribution and LROS data set with imputed values follows a lognormal distribution (results not included).

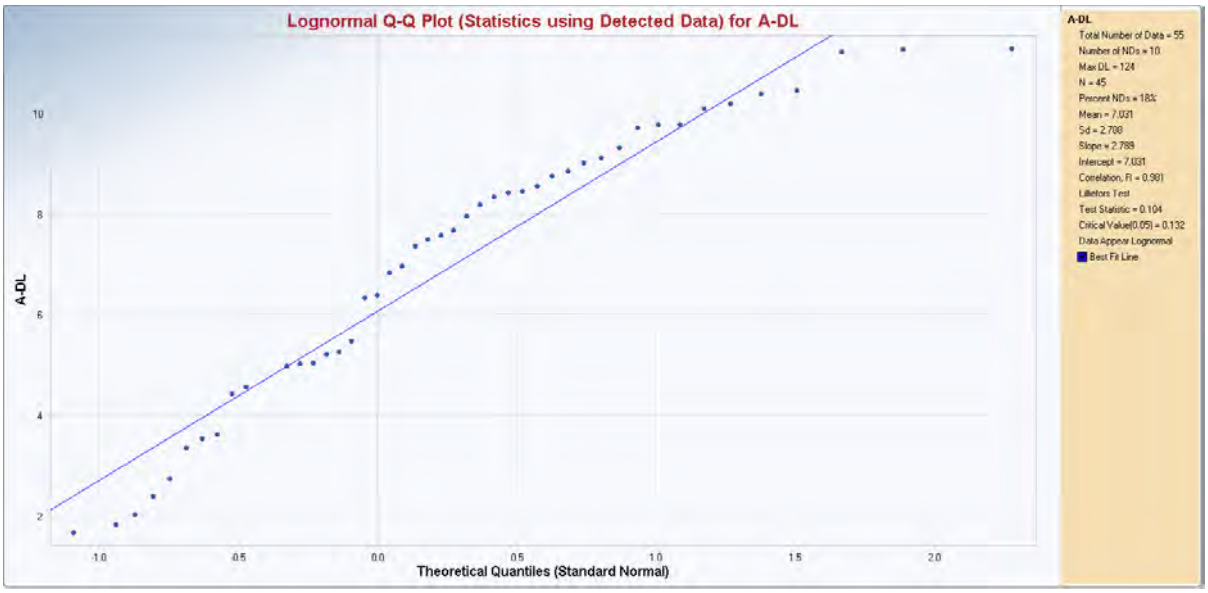

Figure 4-6. Lognormal GOF Test on Detected TRS Data Set

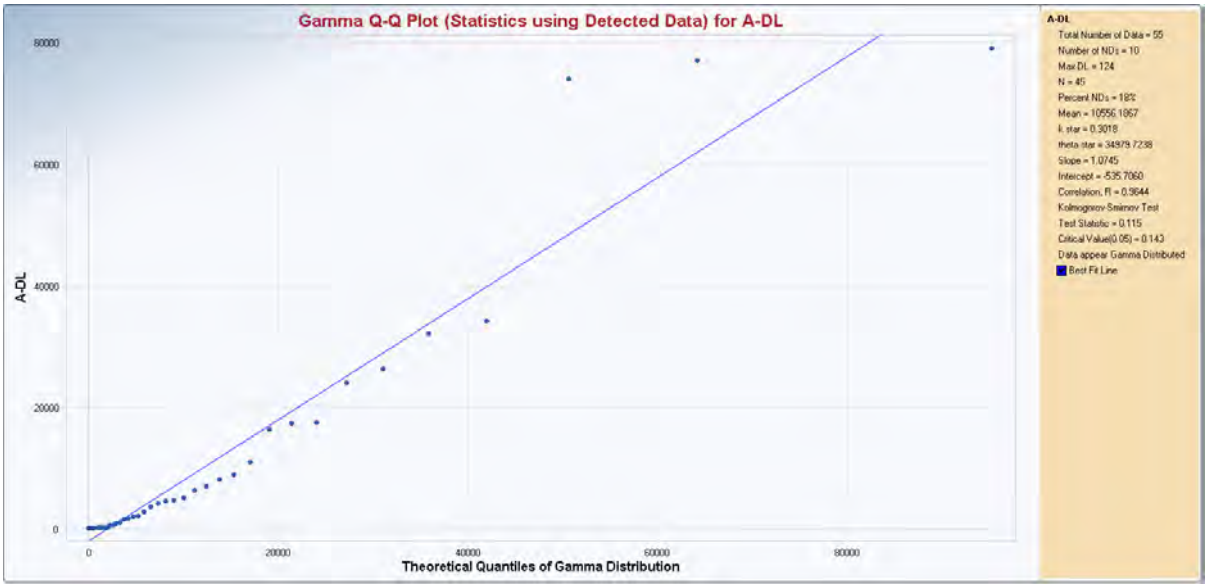

Figure 4-7. Gamma GOF Test on Detected TRS Data Set

Table 4-5. Statistics and UCL95s Obtained Using Gamma and Lognormal Distributions

A-DL

General Statistics			
Total Number of Observations	55	Number of Distinct Observations	53
Number of Detects	45	Number of Non-Detects	10
Number of Distinct Detects	45	Number of Distinct Non-Detects	8
Minimum Detect	5.2	Minimum Non-Detect	3.8
Maximum Detect	79000	Maximum Non-Detect	124
Variance Detects	3.954E+8	Percent Non-Detects	18.18%
Mean Detects	10556	SD Detects	19886
Median Detects	1940	CV Detects	1.884
Skewness Detects	2.632	Kurtosis Detects	6.496
Mean of Logged Detects	7.031	SD of Logged Detects	2.788

Kaplan-Meier (KM) Statistics using Normal Critical Values and other Nonparametric UCLs			
Mean	8638	Standard Error of Mean	2488
SD	18246	95% KM (BCA) UCL	13396
95% KM (t) UCL	12802	95% KM (Percentile Bootstrap) UCL	12792
95% KM (z) UCL	12731	95% KM Bootstrap t UCL	14509
90% KM Chebyshev UCL	16102	95% KM Chebyshev UCL	19483
97.5% KM Chebyshev UCL	24176	99% KM Chebyshev UCL	33394

Gamma GOF Tests on Detected Observations Only			
A-D Test Statistic	0.591	**Anderson-Darling GOF Test**	
5% A-D Critical Value	0.86	Detected data appear Gamma Distributed at 5% Significance Level	
K-S Test Statistic	0.115	**Kolmogrov-Smirnoff GOF**	
5% K-S Critical Value	0.143	Detected data appear Gamma Distributed at 5% Significance Level	
Detected data appear Gamma Distributed at 5% Significance Level			

Gamma Statistics on Detected Data Only			
k hat (MLE)	0.307	k star (bias corrected MLE)	0.302
Theta hat (MLE)	34333	Theta star (bias corrected MLE)	34980
nu hat (MLE)	27.67	nu star (bias corrected)	27.16
MLE Mean (bias corrected)	10556	MLE Sd (bias corrected)	19216

Gamma Kaplan-Meier (KM) Statistics			
k hat (KM)	0.224	nu hat (KM)	24.66
Approximate Chi Square Value (24.66, α)	14.35	Adjusted Chi Square Value (24.66, β)	14.14
95% Gamma Approximate KM-UCL (use when n>=50)	14844	95% Gamma Adjusted KM-UCL (use when n<50)	15066

Table 4-5 (continued). Statistics and UCL95s Obtained Using Gamma and Lognormal Distributions

Gamma ROS Statistics using Imputed Non-Detects			
Minimum	0.1	Mean	8637
Maximum	79000	Median	588
SD	18415	CV	2.132
k hat (MLE)	0.198	k star (bias corrected MLE)	0.199
Theta hat (MLE)	43697	Theta star (bias corrected MLE)	43402
nu hat (MLE)	21.74	nu star (bias corrected)	21.89
MLE Mean (bias corrected)	8637	MLE Sd (bias corrected)	19361
		Adjusted Level of Significance (β)	0.0456
Approximate Chi Square Value (21.89, α)	12.26	Adjusted Chi Square Value (21.89, β)	12.06
95% Gamma Approximate UCL (use when n>=50)	15426	95% Gamma Adjusted UCL (use when n<50)	15675
Lognormal GOF Test on Detected Observations Only			
Shapiro Wilk Test Statistic	0.939	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.945	Detected Data Not Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.104	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.132	Detected Data appear Lognormal at 5% Significance Level	
Detected Data appear Approximate Lognormal at 5% Significance Level			
Lognormal ROS Statistics Using Imputed Non-Detects			
Mean in Original Scale	8638	Mean in Log Scale	5.983
SD in Original Scale	18414	SD in Log Scale	3.391
95% t UCL (assumes normality of ROS data)	12793	95% Percentile Bootstrap UCL	12853
95% BCA Bootstrap UCL	13904	95% Bootstrap t UCL	15032
95% H-UCL (Log ROS)	1855231		
UCLs using Lognormal Distribution and KM Estimates when Detected data are Lognormally Distributed			
KM Mean (logged)	6.03	95% H-UCL (KM -Log)	1173988
KM SD (logged)	3.286	95% Critical H Value (KM-Log)	5.7
KM Standard Error of Mean (logged)	0.449		
Nonparametric Distribution Free UCL Statistics			
Detected Data appear Gamma Distributed at 5% Significance Level			
Suggested UCL to Use			
95% KM (Chebyshev) UCL	19483	95% GROS Approximate Gamma UCL	15426
95% Approximate Gamma KM-UCL	14844		

From Table 4-5, it is noted that the percentile bootstrap method on LROS method as described in Helsel (2012b) yields a lower value of the UCL95 = 12797, which is comparable to a KM (t)-UCL =12802. The student's t statistic based upper limits (e.g., KM (t)-UCL) do not adjust for data skewness; the two UCLs, bootstrap LROS UCL and KM(t)-UCL, appear to represent underestimates of the population mean. As expected, H-UCL on the other hand, resulted in impractically large UCL values (using both the LROS and KM methods). Based upon the data skewness, ProUCL suggested three UCLs (e.g., Gamma UCL = 15426) out of several UCL methods available in the literature and incorporated in ProUCL software.

4.6.3.1 Issues Associated with the Use of Lognormal distribution to Compute a UCL of Mean for *Data Sets with Nondetects*

Some drawbacks associated with the use of the lognormal distribution based UCLs on data sets with NDs are discussed next.

Example 4-7. Consider the benzene data set (Benzene-H-UCL-RCRA.xls) of size 8 used in Chapter 21 of the RCRA Unified Guidance document (EPA 2009e). The data set consists of one ND value with DL of 0.5 ppb. In the RCRA guidance, the ND value was replaced by 0.5/2=0.25 to compute a lognormal H-UCL. In this example, lognormal 95% UCLs (H-UCLs) are computed replacing the ND by the DL (0.5) and also replacing the ND by DL/2=0.25. Normal and lognormal GOF tests using DL/2 for the ND value are shown in Figures 4-8 and 4-9 as follows.

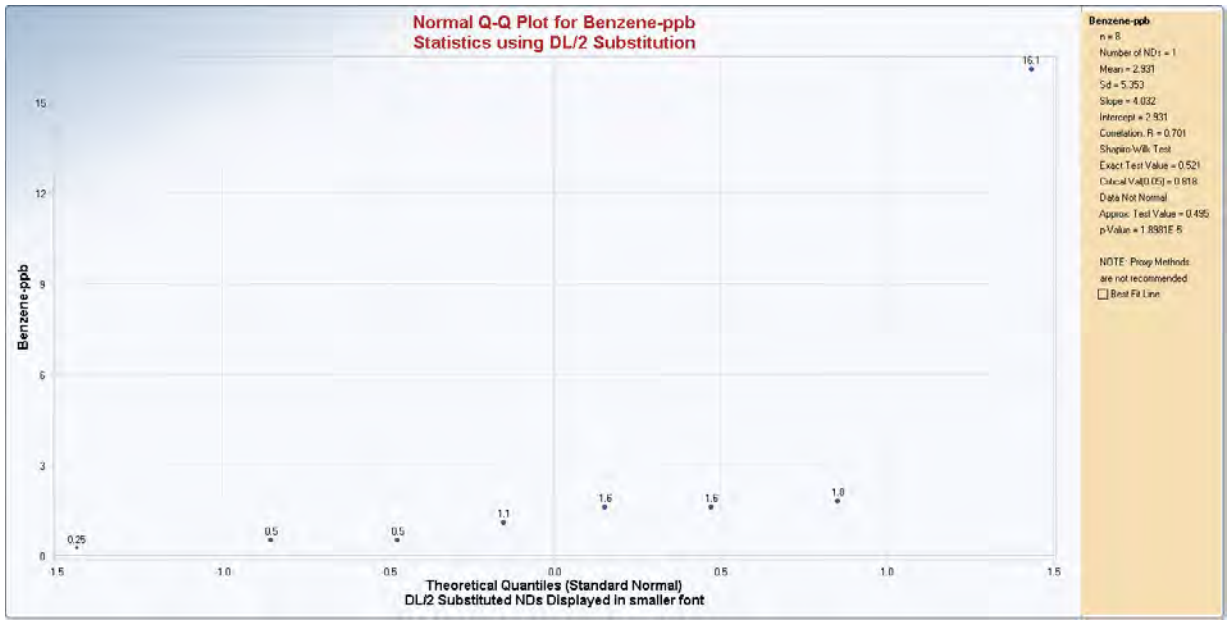

Figure 4-8. Normal Q-Q Plot on Benzene Data with ND Replaced by DL/2

From the above Q-Q plot, it is easy to see that observation 16.1 ppb represents an outlier. The Dixon test on logged data suggests that 2.779 (=ln(16.1)) is an outlier and observation 16.1 is an outlier in the original scale. The outlier, 2.779 was accommodated by the lognormal distribution resulting in the conclusion that the data set follows a lognormal distribution (Figure 4-9).

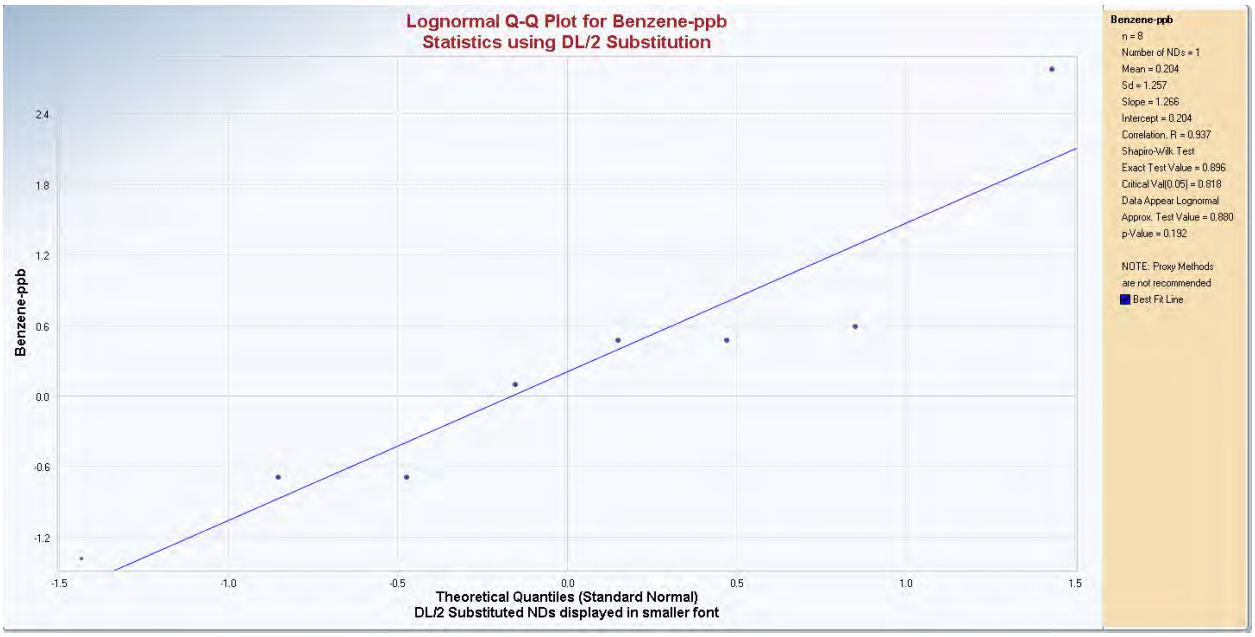

Figure 4-9. Lognormal Q-Q Plot on Benzene Data with ND Replaced by DL/2

4.6.3.1.1 Impact of Using DL and DL/2 for Nondetects on UCL95 Computations

Lognormal distribution based H-UCLs computed by replacing ND by DL and by DL/2 are respectively given in Tables 4-6 and 4-7 below.

Table 4-6. Lognormal 95% UCL (H-UCL) - Replacing ND by DL (=0.5)

Lognormal GOF Test			
Shapiro Wilk Test Statistic	0.803	**Shapiro Wilk Lognormal GOF Test**	
5% Shapiro Wilk Critical Value	0.818	Data Not Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.273	**Lilliefors Lognormal GOF Test**	
5% Lilliefors Critical Value	0.313	Data appear Lognormal at 5% Significance Level	
Data appear Approximate Lognormal at 5% Significance Level			
Lognormal Statistics			
Minimum of Logged Data	-0.693	Mean of logged Data	0.29
Maximum of Logged Data	2.779	SD of logged Data	1.152
Assuming Lognormal Distribution			
95% H-UCL	13.62	90% Chebyshev (MVUE) UCL	5.191
95% Chebyshev (MVUE) UCL	6.496	97.5% Chebyshev (MVUE) UCL	8.306
99% Chebyshev (MVUE) UCL	11.86		

Table 4-7. Lognormal 95% UCL (H-UCL) - Replacing ND by DL/2 (=0.25)

Lognormal GOF Test			
Shapiro Wilk Test Statistic	0.896	**Shapiro Wilk Lognormal GOF Test**	
5% Shapiro Wilk Critical Value	0.818	Data appear Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.255	**Lilliefors Lognormal GOF Test**	
5% Lilliefors Critical Value	0.313	Data appear Lognormal at 5% Significance Level	
Data appear Lognormal at 5% Significance Level			
Lognormal Statistics			
Minimum of Logged Data	-1.386	Mean of logged Data	0.204
Maximum of Logged Data	2.779	SD of logged Data	1.257
Assuming Lognormal Distribution			
95% H-UCL	18.86	90% Chebyshev (MVUE) UCL	5.514
95% Chebyshev (MVUE) UCL	6.952	97.5% Chebyshev (MVUE) UCL	8.948
99% Chebyshev (MVUE) UCL	12.87		

Note: 95% H-UCL (with ND replaced by DL/2) computed by ProUCL is in agreement with results summarized in Chapter 21 of the RCRA Guidance (EPA 2009e). However, it should be noted that the UCL computed using the DL for ND is 13.62, and the UCL computed using DL/2 for ND is 18.86. Substitution by DL/2 resulted in a data set with higher variability and a UCL higher than the one obtained using the DL method. These two UCLs differ considerably confirming that the use of substitution methods should be avoided.

From results summarized above, it is noted that replacing NDs reported as <DL (=0.5) by DL/2 = 0.25 resulted in an increase in the *sd* of the logged data from 1.152 to 1.257 which resulted in an increase in the H-critical value. The minor increase in the *sd* of logged data coupled with an increase in the H-critical value resulted in an unacceptable increase in the H-UCL, from 13.62 to 18.86. This gives another reason to avoid the use of the lognormal distribution to compute decision statistics. UCLs represent estimates of population means; inclusion of one outlier 16.1 resulted in a UCL95 of 18.86 (or 13.36) which appears to more closely represent the largest value of the data set rather than the average. This issue is illustrated as follows in Section *4.6.3.1.2.*

4.6.3.1.2 Impact of Outlier, 16.1 ppb on UCL95 Computations

The benzene data set without the outlier follows a normal distribution, and normal distribution based UCL95s are summarized below in Tables 4-8 (KM estimates), 4-9 (ND by DL), and 4-10 (ND by DL/2).

Table 4-8. Normal 95% UCL Computed using KM Estimates

Normal GOF Test on Detects Only			
Shapiro Wilk Test Statistic	0.847	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.788	Detected Data appear Normal at 5% Significance Level	
Lilliefors Test Statistic	0.265	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.362	Detected Data appear Normal at 5% Significance Level	
Detected Data appear Normal at 5% Significance Level			
Kaplan-Meier (KM) Statistics using Normal Critical Values and other Nonparametric UCLs			
Mean	1.086	Standard Error of Mean	0.225
SD	0.544	95% KM (BCA) UCL	N/A
95% KM (t) UCL	1.523	95% KM (Percentile Bootstrap) UCL	N/A

Table 4-9. Normal 95% UCL Computed by Replacing ND by DL = 0.5

Normal GOF Test			
Shapiro Wilk Test Statistic	0.814	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.803	Data appear Normal at 5% Significance Level	
Lilliefors Test Statistic	0.269	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.335	Data appear Normal at 5% Significance Level	
Data appear Normal at 5% Significance Level			
Assuming Normal Distribution			
95% Normal UCL		**95% UCLs (Adjusted for Skewness)**	
95% Student's-t UCL	1.517	95% Adjusted-CLT UCL (Chen-1995)	1.454
		95% Modified-t UCL (Johnson-1978)	1.518

Table 4-10. Normal 95% UCL Computed by Replacing ND by DL/2 = 0.25

Normal GOF Test			
Shapiro Wilk Test Statistic	0.875	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.803	Data appear Normal at 5% Significance Level	
Lilliefors Test Statistic	0.236	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.335	Data appear Normal at 5% Significance Level	
Data appear Normal at 5% Significance Level			
Assuming Normal Distribution			
95% Normal UCL		**95% UCLs (Adjusted for Skewness)**	
95% Student's-t UCL	1.516	95% Adjusted-CLT UCL (Chen-1995)	1.436
		95% Modified-t UCL (Johnson-1978)	1.515

Note: The recommended UCL is the KM UCL= 1.523. It is noted that normal UCLs are not influenced by changing a single ND from 0.5 (UCL95=1.517) to 0.25 (UCL95=1.516). Normal UCL95s without the outlier appear to represent more realistic estimates of the EPC (population mean). The Lognormal UCL based upon the data set with the outlier represents the outlying value(s) rather than representing the population mean.

4.7 Bootstrap UCL Computation Methods for Left-Censored Data Sets

The use of bootstrap methods has become popular with the easy access to fast personal computers. As described in Chapter 2, 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 data set of n observations. The process is repeated a large number of times, N (e.g., 1000-2000), and each time an estimate, $\hat{\theta}$ of θ (e.g., mean) is computed. These estimates are used to compute an estimate of the SE of the estimate, $\hat{\theta}$. Just as for the full uncensored data sets without any NDs, for left-censored data sets, the bootstrap resamples are obtained with replacement. An indicator variable, I (1 = detected value, and 0 = nondetected value), is tagged to each observation in a bootstrap sample (Efron 1981).

Singh, Maichle, and Lee (EPA 2006) studied the performances, in terms of coverage probabilities, of four bootstrap methods for computing UCL95s for data sets with ND observations. The four bootstrap methods included the standard bootstrap method, the bootstrap-t method, the percentile bootstrap method, and the bias-corrected accelerated (BCA) bootstrap method (Efron and Tibshirani 1993; Manly 1997). Some bootstrap methods, as incorporated in ProUCL, for computing upper limits on left-censored data sets are briefly discussed in this section.

4.7.1 Bootstrapping Data Sets with Nondetect Observations

As before, 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 (used to compute bootstrap UTLs), x_p, that needs to be estimated from the sampled data set with ND observations. Let $\hat{\theta}$ be an estimate of θ, which is a function of k ND 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 robust ROS, gamma ROS, or KM estimate 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 methods described above. The bootstrap method can then be used to compute a UCL of the percentile, also known as upper tolerance limit. The computations of upper tolerance limits are discussed in Chapter 5.

An indicator variable, I (taking only two values: 1 and 0), is assigned to each observation (detected or nondetected) when dealing with left-censored data sets (Efron 1981; Barber and Jennison 1999). The indicator variables, $I_{j:}\ j:=1,2,\ldots,n$, represent the detection status of the sampled observations, x_j ; $j: = 1, 2,\ldots, n$. A large number, N (1000, 2000) of two-dimensional bootstrap resamples, $(x_{iJ}, I_{iJ}), j:= j: = 1, 2,\ldots, N,$ and $i: = 1, 2,\ldots, n,$ 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. The two-dimensional bootstrap process keeps track of the detection status of each observation in a bootstrap re-sample. In this setting, the DLs are fixed as entered in the data set, and the number of NDs vary from bootstrap sample to bootstrap sample. There may be k_1 NDs in the first bootstrap sample, k_2 NDs in the second sample, ..., and k_N NDs in the N^{th} bootstrap sample. Since the sampling is conducted with replacement, the number of NDs, k_i, i: = 1, 2, ..., N, in a bootstrap re-sample 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 ND estimation methods (e.g., KM, ROS) to compute the statistics of interest (e.g., mean, *sd*, upper limits). It is possible that all (or most) observations in a bootstrap re-sample are the same. This is specifically true, when one is

dealing with small data sets. To avoid such situations (with all equal values) it is suggested that there be at least 15 to 20 (preferably more) observations in the data set. As noted in Chapter 2, it is not advisable to compute statistics based upon a bootstrap resample consisting of only a few detected values such as < 4-5.

Let $\hat{\theta}$ be an estimate of θ based upon the original left-censored data set of size n; if the parameter, θ, represents the population mean, then a reasonable choice for the estimate, $\hat{\theta}$, can be the sample ROS mean, or sample KM mean. Similarly, calculate the *sd* using one of these methods for left-censored data sets. The following two steps are common to all bootstrap methods incorporated in the ProUCL software.

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 ND is chosen in the resample). Compute an estimate of the mean (e.g., 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 the population mean. Denote these estimates (e.g., 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 ROS means 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 $\bar{\theta}_B$ (instead of $\bar{x}_B$). The estimate, $\bar{\theta}_B$ is the arithmetic mean of the N bootstrap estimates (e.g., KM mean, or ROS mean) given by $\hat{\theta}_i$, $i:=1,2,\ldots N$. If the estimate, $\hat{\theta}$, represents the KM estimate of, θ, then $\hat{\theta}_i$ (denoted by $\bar{x}_i$ in the above paragraph) also represents the KM mean based upon the i^{th} bootstrap resample. The difference, $\bar{\theta}_B - \hat{\theta}$, provides an estimate of the bias of the estimate, $\hat{\theta}$. After these two steps, a bootstrap procedure (percentile, BCA, or bootstrap-t) is used similarly to the conventional bootstrap procedure on a full uncensored data set as described in Chapter 2.

Notes: Just like for small uncensored data sets, for small left-censored data sets (<8-10) with only a few distinct values (2 or 3), it is not advisable to use bootstrap methods. In these scenarios, ProUCL does not compute bootstrap limits. However, due to the complexity of decision tables and lack of enough funding, there could be some rare cases where ProUCL may recommend a bootstrap method based UCL which is not computed by ProUCL (due to lack of enough data).

4.7.1.1 UCL of Mean Based upon Standard Bootstrap Method

Once the desired number of bootstrap samples and estimates has been obtained following the two steps described above, a UCL of the 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:

$$UCL = \hat{\theta} + z_{\alpha}\hat{\sigma}_B \quad (4\text{-}11)$$

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 of the mean of skewed (e.g., lognormal and gamma) data distributions (populations).

4.7.1.2 UCL of Mean Based upon Bootstrap-t Method

*A (1 – α)*100% UCL* of the mean based upon the bootstrap-t method is given as follows.

$$UCL = \bar{x} - t_{(\alpha N)}\frac{s_x}{\sqrt{n}} \quad (4\text{-}12)$$

It should be noted that the mean and *sd* used in equation (4-12) represent estimates (e.g., KM estimates, ROS estimates) obtained using original left-censored data set. Similarly, the *t*-cutoff value used in equation (4-12) is computed using the pivotal *t*-values based upon KM estimates or some other estimates obtained using bootstrap re-samples. 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, the bootstrap-t method sometimes results in unstable and erratic UCL values, especially in the presence of outliers (Efron and Tibshirani 1993). Therefore, the bootstrap-t method should be used with caution. In case this method results in erratic unstable UCL values. Additional suggestions on this topic are offered in Chapter 2.

4.7.1.3 Percentile Bootstrap Method

A detailed description of the percentile bootstrap method is given in Chapter 2. For left-censored data sets, sample means are computed for each bootstrap sample using a selected method (e.g., KM, ROS), which are arranged in ascending order. The 95% UCL of the mean is the 95th percentile and is given by:

$$95\%\ Percentile - UCL = 95^{th}\%\ \bar{x}_i;\ i: = 1, 2, ..., N \quad (4\text{-}13)$$

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

Singh, Maichle and Lee (2006) noted 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 and yields UCLs with a coverage probability lower than the specified coverage of 0.95. The BCA UCL for a selected estimation method (e.g., KM, ROS) is given by the following equation:

$$(1- \alpha)*100\%\ UCL_{PROC} = BCA - UCL = \bar{x}_{PROC}^{\alpha_2} \quad (4\text{-}14)$$

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., KM, DL/2, ROS) mean estimation methods. Here α_2 is given by the following probability statement:

$$\alpha_2 = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1-\hat{\alpha}(\hat{z}_0 + z^{(1-\alpha)})}\right] \quad (4\text{-}15)$$

$\Phi(Z)$ 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:

$$\hat{z}_0 = \Phi^{-1}\left[\frac{\#(\bar{x}_{PROC,i} < \bar{x}_{PROC})}{N}\right],\ i\text{: } = 1, 2, ..., N \quad (4\text{-}16)$$

$\Phi^{-1}(x)$ is the inverse 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:

$$\hat{\alpha} = \frac{\sum(\bar{x}_{PROC} - \bar{x}_{-i,PROC})^3}{6\left[\sum(\bar{x}_{PROC} - \bar{x}_{-i,PROC})^2\right]^{1.5}} \quad (4\text{-}17)$$

Summation in the above equation is being carried from $i = 1$ to 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 the PROC mean of $(n-1)$ observations without the i^{th} observation, *i: = 1, 2, ..., n*.

4.8 (1-α)*100% UCL Based upon Chebyshev Inequality

The use of the Chebyshev-type inequality (as used in Chapter 2) based UCLs has been suggested to provide better coverage to the population mean of skewed data distributions. The two-sided Chebyshev theorem (Dudewicz and Misra 1988) states that given a random variable, X, with finite mean and *sd*, μ_1 and σ_1, we have:

$$P(-k\sigma_1 \le X - \mu_1 \le k\sigma_1) \ge 1 - 1/k^2.$$

A $(1-\alpha)*100$ *UCL* of population mean, μ_1, can be obtained by:

$$UCL = \bar{x} + \sqrt{((1/\alpha) - 1)}s_x/\sqrt{n}. \quad (4\text{-}20)$$

In the above UCL equation, the sample mean and *sd* are computed using one of the many estimation methods for left-censored data sets with ND observations as described in earlier sections of this chapter. The UCL95 based upon Chebyshev inequality (with KM estimates) yields a conservative UCL of the mean.

Example 4-8. Pyrene Data Set(continued): A great deal of discussion has been provided in the literature (Helsel 2005, 2012; Helsel 2013 [NADA Package for R]) about estimation of mean and standard deviation based upon this data set; however, not much guidance is provided on how to compute upper limits such as a UCL of the mean for this data set. This data set is used here to illustrate the various bootstrap UCL computation methods incorporated in ProUCL, and how one can compute a UCL95 based upon this left-

censored data set. This data set also illustrates the impact of a few outliers on the various estimates and statistics. Table 4-11a has statistics computed using the outlier, 2982, and Table 4-11b has statistics computed without the outlier. It is noted that the detected data with the outlier does not follow a gamma or a lognormal distribution however, the detected data set without the outlier follows a lognormal distribution.

Table 4-11a. Statistics Computed Using Outlier=2982

Pyrene			
General Statistics			
Total Number of Observations	56	Number of Distinct Observations	44
Number of Detects	45	Number of Non-Detects	11
Number of Distinct Detects	39	Number of Distinct Non-Detects	8
Minimum Detect	31	Minimum Non-Detect	28
Maximum Detect	2982	Maximum Non-Detect	174
Variance Detects	189219	Percent Non-Detects	19.64%
Mean Detects	190.1	SD Detects	435
Median Detects	103	CV Detects	2.288
Skewness Detects	6.282	Kurtosis Detects	41
Mean of Logged Detects	4.711	SD of Logged Detects	0.805
Kaplan-Meier (KM) Statistics using Normal Critical Values and other Nonparametric UCLs			
Mean	164.1	Standard Error of Mean	52.65
SD	389.4	95% KM (BCA) UCL	271.8
95% KM (t) UCL	252.2	95% KM (Percentile Bootstrap) UCL	261
95% KM (z) UCL	250.7	95% KM Bootstrap t UCL	507.5
90% KM Chebyshev UCL	322	95% KM Chebyshev UCL	393.6
97.5% KM Chebyshev UCL	492.9	99% KM Chebyshev UCL	687.9
Lognormal ROS Statistics Using Imputed Non-Detects			
Mean in Original Scale	163.2	Mean in Log Scale	4.537
SD in Original Scale	393.1	SD in Log Scale	0.843
95% t UCL (assumes normality of ROS data)	251.1	95% Percentile Bootstrap UCL	262.6
95% BCA Bootstrap UCL	322.1	95% Bootstrap t UCL	507.8
95% H-UCL (Log ROS)	170.4		

UCLs computed using the KM method and percentile bootstrap and t-statistic are 261 and 252.2. The corresponding UCLs obtained using the LROS method are 262.6 and 251.2, which appear to underestimate the population mean. The H-UCL based upon the LROS method is unrealistically lower (170.4) than the other UCLs. Depending upon the data skewness (sd of detected logged data =0.81), one can use the Chebyshev UCL95 (or Chebyshev UCL90) to estimate the EPC. Note that as expected, the presence of one outlier resulted in a bootstrap-t UCL95 significantly higher than the various other UCLs. Table 4-11b has UCLs computed without the outlier. Exclusion of the outlier resulted in all comparable UCL values. Any of those UCLs can be used to estimate the EPC.

Table 4-11b. Statistics Computed without Outlier=2982

Pyrene			
General Statistics			
Total Number of Observations	55	Number of Distinct Observations	43
Number of Detects	44	Number of Non-Detects	11
Number of Distinct Detects	38	Number of Distinct Non-Detects	8
Minimum Detect	31	Minimum Non-Detect	28
Maximum Detect	459	Maximum Non-Detect	174
Variance Detects	8226	Percent Non-Detects	20%
Mean Detects	126.6	SD Detects	90.7
Median Detects	103	CV Detects	0.716
Skewness Detects	1.795	Kurtosis Detects	3.489
Mean of Logged Detects	4.636	SD of Logged Detects	0.637
Kaplan-Meier (KM) Statistics using Normal Critical Values and other Nonparametric UCLs			
Mean	112.9	Standard Error of Mean	11.84
SD	86.03	95% KM (BCA) UCL	134
95% KM (t) UCL	132.7	95% KM (Percentile Bootstrap) UCL	132.4
95% KM (z) UCL	132.3	95% KM Bootstrap t UCL	135.3
90% KM Chebyshev UCL	148.4	95% KM Chebyshev UCL	164.5
97.5% KM Chebyshev UCL	186.8	99% KM Chebyshev UCL	230.7
Lognormal GOF Test on Detected Observations Only			
Shapiro Wilk Test Statistic	0.973	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.944	Detected Data appear Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.0965	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.134	Detected Data appear Lognormal at 5% Significance Level	
Detected Data appear Lognormal at 5% Significance Level			
Lognormal ROS Statistics Using Imputed Non-Detects			
Mean in Original Scale	112.4	Mean in Log Scale	4.49
SD in Original Scale	86.61	SD in Log Scale	0.677
95% t UCL (assumes normality of ROS data)	132	95% Percentile Bootstrap UCL	133
95% BCA Bootstrap UCL	135.7	95% Bootstrap t UCL	137
95% H-UCL (Log ROS)	134.9		
UCLs using Lognormal Distribution and KM Estimates when Detected data are Lognormally Distributed			
KM Mean (logged)	4.491	95% H-UCL (KM -Log)	135
KM SD (logged)	0.676	95% Critical H Value (KM-Log)	2.013
KM Standard Error of Mean (logged)	0.0956		

The data set is not highly skewed with *sd* = 0.64 of logged detected data. Most methods (including H-UCL) yield comparable results. Based upon data skewness, ProUCL recommends the use of a UCL95 based upon the KM BCA method (highlighted in blue in Table 4-11b).

4.9 Saving Imputed NDs Using Stats/Sample Sizes Module of ProUCL

Using this option, NDs are imputed based upon the selected distribution (normal, lognormal, or gamma) of the detected observations. Using the menu option, "**Imputed NDs using ROS Methods**" ProUCL can be used to impute and save imputed NDs along with the original data in additional columns automatically generated by ProUCL. ProUCL assigns self-explanatory titles for those generated columns. This option is available in ProUCL for researchers and advanced users who want to experiment with the full data sets consisting of detected and imputed ND observations for other applications (e.g., ANOVA, PCA).

4.10 Parametric Methods to Compute UCLs Based upon Left-Censored Data Sets

Some researchers have suggested that parametric methods such as the expectation maximization (EM) method and maximum likelihood method (MLE) cited earlier in this chapter would perform better than the GROS method for data sets with NDs. As reported in ProUCL guidance and on ProUCL generated output sheets, the developers do realize that the GROS method does not perform well when the shape parameter, k, or its MLE estimate is small (≤ 1). The GROS method appears to work fine when k is large (> 2). However, for data sets with NDs and with many DLs, the developers are not sure if parametric methods such as the MLE method and the EM method perform better than the GROS method and other methods available in ProUCL. More research needs to be conducted to verify these statements. As noted earlier, it is not easy (perhaps not possible in most cases) to correctly assess the distribution of a data set containing NDs with multiple censoring points, a common occurrence in environmental data sets. If distributional assumptions are incorrect, the decision statistics computed using this incorrect distribution may also be incorrect. To the best of our knowledge, the EM method can be used on data sets with a single DL. Earlier versions of ProUCL (e.g., ProUCL 4.0, 2007) had some parametric methods including the MLE and RMLE methods; those methods were excluded from later versions of ProUCL due to their poor performances.

The research in this area is limited; to the best of our knowledge, parametric methods (MLE and EM) for data sets with multiple censoring points are not well-researched. The enhancement of these parametric methods to accommodate left-censored data sets with multiple DLs will be a big achievement in environmental statistical literature. The developers will be happy to include contributed better performing methods in ProUCL.

4.11 Summary and Suggestions

Most of the parametric methods including the MLE, the RMLE, and the EM method assume that there is only one DL. Like parametric estimates computed using uncensored data sets, MLE and EM estimates obtained using a left-censored data set are influenced by outliers, especially when a lognormal model is used. These issues are illustrated by an example as follows.

Example 4-9: Consider a left-censored data set of size 25 with multiple censoring points: <0.24, <0.24, <1, <0.24, <15, <10, <0.24, <22, <0 .24, < 5.56, <6.61, 1.33, 168.6, 0.28, 0.47, 18.4, 0.48, 0.26, 3.29, 2.35, 2.46, 1.1, 51.97, 3.06, and 200.54. The data set appears to have 2 extreme outliers and 1 intermediate outlier as can be seen from Figure 4-10. From Figure 4-10 and the results of the Rosner outlier test performed on the data set, it can be concluded that the 3 high detected values represent outliers. The Shapiro-Wilk test results performed on detected data shown in Figure 4-11 (censored probability plot) suggest that the detected data set (with outliers) follows a lognormal distribution accommodating the outliers.

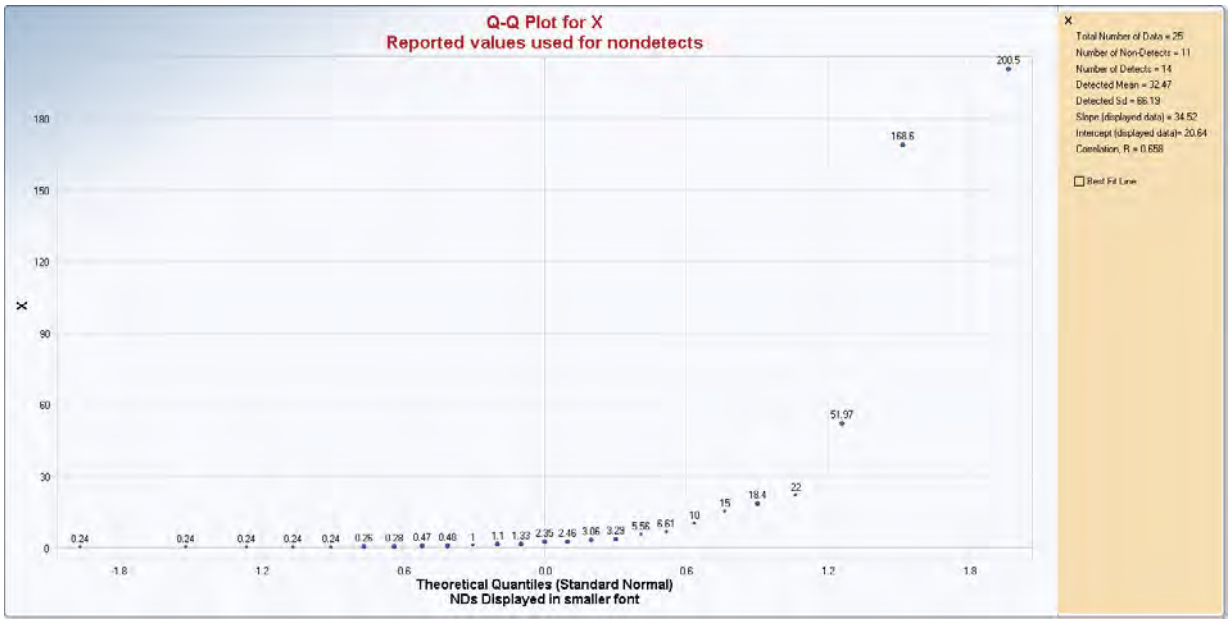

Figure 4-10. Exploratory Q-Q Plot to Identify Outliers Showing All Detects and Nondetects

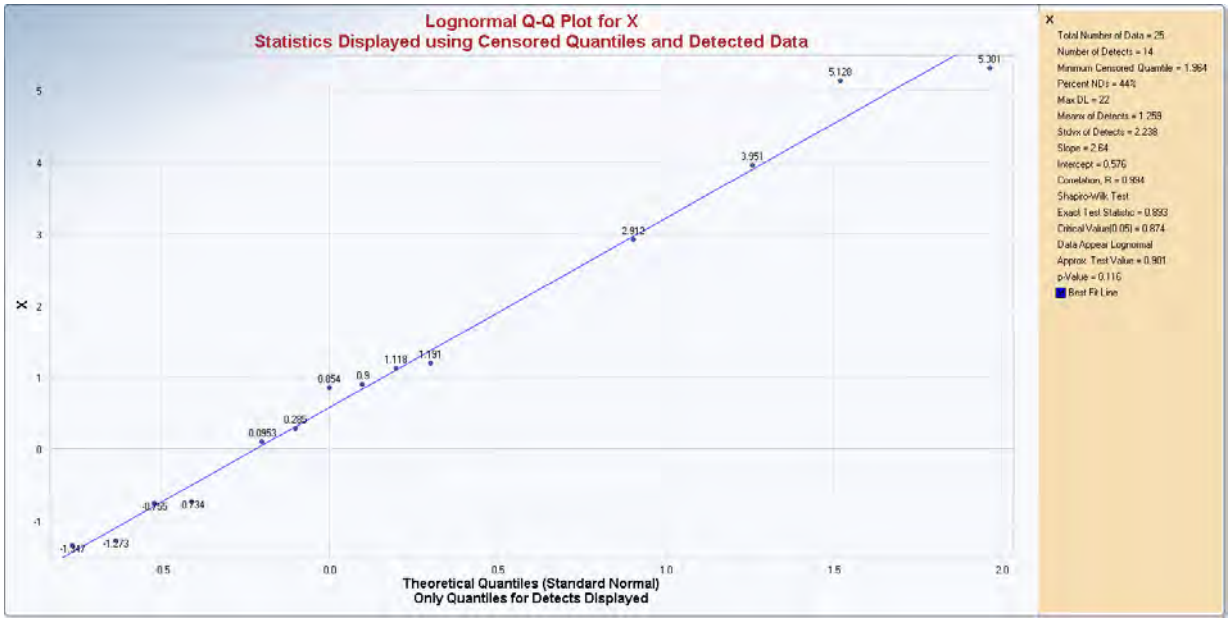

Figure 4-11. Censored Q-Q Plot Showing GOF Test Results on Detected Log-transformed Data

Table 4- 12. Statistics Computed with Outliers

From File: ND-Data-for MLE-1.xls

General Statistics for Uncensored Dataset

Variable	NumObs	# Missing	Minimum	Maximum	Mean	SD	SEM	MAD/0.675	Skewness	Kurtosis	CV
X	25	0	0.24	200.5	20.64	50.81	10.16	3.128	3.095	8.876	2.462

Percentiles for Uncensored Dataset

Variable	NumObs	# Missing	10%ile	20%ile	25%ile(Q1)	50%ile(Q2)	75%ile(Q3)	80%ile	90%ile	95%ile	99%ile
X	25	0	0.24	0.256	0.28	2.35	10	15.68	39.98	145.3	192.9

Table 4-13. Nonparametric estimates of the mean and *sd* using the KM method.

From File: ND-Data-for MLE-1.xls

General Statistics for Censored Datasets (with NDs) using Kaplan Meier Method

Variable	NumObs	# Missing	Num Ds	NumNDs	% NDs	Min ND	Max ND	KM Mean	KM Var	KM SD	KM CV
X	25	0	14	11	44.00%	0.24	22	18.48	2528	50.28	2.72

MLE estimates of the mean and *sd* obtained using Minitab 16, UCL95, and a 95%-95% upper tolerance limit based upon a lognormal distribution are summarized as follows. ML estimates in log scale are given in Table 4-14.

Table 4-14. ML Estimates in Log-Scale (with outliers).

Parameter	Estimate	Standard Error	Upper Bound
Location	-0.247900	0.641686	0.807580
Scale	2.71896	0.530176	3.74710

Log Likelihood = -58. 151; MLE estimates in original raw scale are (back transformation):

Mean = 31. 45,

SE of mean = 43.1279

UCL95 = 300.041

The use of the log-transformation has resulted in inflated estimates, mean = 31.45, UCL95 = 300.41, and a UTL95-95 = 346.54. The estimate of the mean based upon a data set with NDs should be smaller (e.g., KM mean = 18.48) than the mean estimate obtained using all NDs at their reported DLs, 20.64. For this left-censored data set, the MLE of the mean based upon a lognormal distribution is 31.45 which appears to be incorrect. This is because the back-transformed mean of the log-transformed data (i.e., the geometric mean) is not equivalent to the arithmetic mean. This is the challenge of correctly interpreting statistics resulting from a transformation.

Statistics Computed without Outliers

Detected data without the 2 extreme outliers also follow a lognormal distribution. MLE estimates, UCL95, UTL95-95 computed without the outliers and lognormal distribution (using Minitab) are:

Estimates in log scale are provided as follows:

Table 4-15. ML Estimates in Log-Scale (without outliers).

Parameter	Estimate	Standard Error	95% Upper Bound
Location	-0.561639	-0.561639	0.28616
Scale	2.02381	0.421546	2.85079

Log Likelihood = -38.56; MLE estimates in original raw scale are:

Mean = 4.42,

SE of mean = 3.688

UCL95 = 17.433

UTL95-95 = 63.42

Substantial differences are noted in the UCL95s ranging from 300.04 to 17.43, and in the UTL95-95s ranging from 346.54 to 63.42.

It is not easy to verify the data distribution of a left-censored data set consisting of detects and NDs with multiple DLs, therefore some poor performing estimation methods including the parametric MLE methods and the Winsorization method are not retained in ProUCL 4.1 and higher versions. Emphasis is given on the use of nonparametric UCL computation methods and hybrid parametric methods based upon KM estimates which account for data skewness in the computation of UCL95s. It is recommended that one avoid the use of transformations to achieve symmetry while computing the upper limits based upon left-censored data sets. It is not easy to correctly interpret statistics computed in the transformed scale. Moreover, the results and statistics computed in the original scale do not suffer from transformation bias.

When the *sd* of the log-transformed data, σ, becomes >1.0, avoid the use of a lognormal model even when the data appear to be lognormally distributed. Its use often results in unrealistic statistics of no practical merit (Singh, Singh, and Engelhard 1997; Singh, Singh, and Iaci 2002). It is also recommended the user identifies potential outliers representing observations coming from population(s) different from the dominant population and investigate them separately. Decisions about the disposition of outliers should be made by all interested members of the project team.

It is recommended that the use of the DL/2 (t) UCL method be avoided, as the DL/2 UCL does not provide the desired coverage (for any distribution and sample size) for the population mean, even for censoring levels as low as 10% and 15%. This is contrary to the conjecture and assertion (EPA 2006a) made that the DL/2 method can be used for lower (≤ 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 document and ProUCL software.

The use of the KM estimation method is a preferred method as it can handle multiple DLs. Therefore, the use of KM estimates is suggested for computing decision statistics based upon methods which adjust for data skewness. However, the KM estimation method may cause underestimation of the mean and should be used with caution in cases where the KM UCL is less than the mean using substitution of half the DL. Depending upon the data set size, distribution of the detected data, and data skewness, the various nonparametric and hybrid KM UCL95 methods including the KM *t*-UCL, KM BCA UCL, KM H UCL, KM bootstrap-t UCL, and KM Gamma UCLs based upon the KM estimates provide good coverages for the population mean. Suggestions regarding the selection of a 95% UCL of the mean are provided to help the user select the most appropriate 95% UCL, and are similar to the suggestions given for data without NDs (refer to Appendeix A). It is advised that the project team collectively determine which UCL will be most appropriate for their site project. For additional insight, the user may want to consult a statistician.

CHAPTER 5

Computing Upper Limits to Estimate Background Threshold Values Based upon Data Sets Consisting of Nondetect (ND) Observations

5.1 Introduction

As described in Chapter 3, a BTV considered in this chapter represents an upper threshold parameter (e.g., 95th) of the background population; which is used to perform point-by-point comparisons of onsite observations. Estimation of BTVs and comparison studies require the computation of UPLs and UTLs based upon left-censored data sets containing ND observations. Not much guidance is available in the statistical literature on how to compute UPLs and UTLs based upon left-censored data sets of varying sizes and skewness levels. Like UCLs, the use of Student's t-statistic and percentile bootstrap methods based UPLs and UTLs are difficult to defend for moderately skewed to highly skewed data sets with standard deviation (*sd*) of the log-transformed data exceeding 0.75-1.0. Since it is not easy to reliably perform GOF tests on left-censored data sets; emphasis is given on the use of distribution-free nonparametric methods including the KM, Chebyshev inequality, and other computer intensive bootstrap methods to compute upper limits needed to estimate BTVs.

All BTV estimation methods for full uncensored data sets as described in Chapter 3 can be used on data sets consisting of detects and imputed NDs obtained using ROS methods (e.g., GROS and LROS). 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 ND observations.

5.2 Treatment of Outliers in Background Data Sets with NDs

Just like full uncensored data sets, a few outlying observations present in a left-censored data set tend to yield distorted estimates of the population parameters (means, upper percentiles, OLS estimates) of interest. OLS regression estimates (slope and intercept) become distorted (Rousseeuw and Leroy 1987; Singh and Nocerino 1995) by the presence of outliers. Specifically, in the presence of outliers, the ROS method performed on raw data (e.g., GROS) tends to yield unfeasible imputed negative values for ND observations. Singh and Nocerino (2002) suggested the use of robust regression methods to compute regression estimates needed to impute NDs based upon ROS methods. Robust regression methods are beyond the scope of ProUCL. It is therefore suggested that potential outliers be identified after evaluating the distribution of the data sets. Identified potential outliers then need to be scientifically evaluated before proceeding with the computation of the various BTV estimates as described in this chapter. As mentioned in earlier chapters, upper limits may be affected by extreme values in the data set. It is therefore recommended that relevant statistics be computed using data sets with outliers and without outliers for comparison. This extra step helps the project team to see the potential influence of outlier(s) on the various decision-making statistics (e.g., UCLs, UPLs, UTLs); and helps the project team in making 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 more accurate estimate(s) of the population parameters (e.g., mean, EPC, BTV) under consideration.

A couple of classical outlier tests (Dixon and Rosner tests) are available in the ProUCL software. These tests can be used on data sets with or without ND observations. However, these tests assume normal distribution of data set without outliers. This is often not the case with environmental data, which tend to be naturally right-skewed. Therefore, a distribution of the data needs to be verified before outlier tests are applied. If the data are not normally distributed, they should be symmetrized or approximately normalized by using an appropriate transformation before ProUCL outlier tests are applied. Additionally, one can use box plots and/or Q-Q plots to visually identify extreme values and distribution of the data set. It should be pointed out, that for environmental applications, it is the identification of high outliers (perhaps representing contaminated locations and hot spots) that is important. The occurrence of ND (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 Estimating BTVs Based upon Left-Censored Data Sets

This section describes methods for computing upper limits (UPLs, UTLs, USLs, upper percentiles) that may be used to estimate BTVs and other not-to-exceed levels from data sets with ND observations. Several Student's t-type statistic and normal z-scores based methods have been described in the literature (Helsel 2005; Millard and Neerchal 2002; USEPA 2007, 2010d, 2011) to compute UPLs and UTLs based upon statistics (e.g., mean, *sd*) obtained using MLE, KM, or ROS methods. The methods used to compute upper limits (e.g., UPL, UTL, and percentiles) based upon a Student's t-type statistic are also described in this chapter; however, the use of such methods is not recommended for moderately skewed to highly skewed data sets. These methods may yield reasonable upper limits (e.g., with proper coverage) for normally distributed and mildly skewed to moderately skewed data sets with the *sd* of the detected log-transformed data less than 1.0.

Singh, Maichle, and Lee (EPA 2006) demonstrated that the use of the t-statistic and the percentile bootstrap method on moderately to highly skewed left-censored data sets yields UCL95s with coverage probabilities much lower than the specified CC, 0.95. A similar pattern is expected in the behavior and properties of the various other upper limits (e.g., UTLs, UPLs) used in the decision making processes of the USEPA. It is anticipated that the performance (in terms of coverages) of the percentile bootstrap and Student's t-type upper limits (e.g., UPLs, UTLs) computed using the KM and ROS estimates for moderately skewed to highly skewed left-censored data sets (sd of detected logged data >1) would also be less than acceptable. For skewed data sets, the use of the gamma distribution on KM estimates (when applicable) or nonparametric methods, which account for data skewness, is suggested for computing BTV estimates. A brief description of those methods is provided in the following sections.

5.3.1 Computing Upper Prediction Limits (UPLs) for Left-Censored Data Sets

This section describes some parametric and nonparametric methods for computing UPLs for left-censored data sets.

5.3.1.1 UPLs Based upon Normal Distribution of Detected Observations and KM Estimates

When detected observations in a data set containing NDs follow a normal distribution (which can be verified by using the **GOF** module of ProUCL), one may use the normal distribution on KM estimates to compute the various upper limits needed to estimate BTVs (also available in ProUCL 4.1). A *(1 – α)*100* UPL for a future (or next) observation (observation not belonging to the current data set) can be computed using the following KM estimates based equation:

$$\text{UPL}_1 = \hat{\mu}_{KM} + t_{((1-\alpha),(n-1))}\sqrt{\hat{\sigma}_{KM}(1+1/n)} \qquad (5\text{-}1)$$

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

(*df*). 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 $\text{UPL}_1$95 with probability 0.95. A similar equation can be developed for upper prediction limits for future k observations (described in Chapter 3) and the mean of k future observations.

5.3.1.2 UPL Based upon the Chebyshev Inequality

The Chebyshev inequality can be used to compute a reasonably conservative UPL and is given as follows:

$$UPL = \bar{x} + \left[\sqrt{((1/\alpha)-1)*(1+1/n)}\right]s_x \qquad (5\text{-}2)$$

The mean, $\bar{x}$, and *sd*, s_x, used in the above equation are computed using one of the estimation methods (e.g., KM) for left-censored data sets. Just like the Chebyshev UCL, a UPL based upon the Chebyshev inequality tends to yield higher estimate of BTVs than the other methods. This is especially true when skewness is moderately mild (*sd* of log-transformed data is low < 0.75), and the sample size is large *n* > 30). It is advised to apply professional/expert judgment before using this method to compute a UPL.

5.3.1.3 UPLs Based upon ROS Methods

As described earlier, ROS methods first impute *k* ND values using an OLS linear regression model (Chapter 4). This results in a full data set of size *n*. For ROS methods (gamma, lognormal), ProUCL generates additional columns consisting of *(n - k)* detected values and *k* imputed values of the *k* ND observations present in a data set. Once, the ND observations have been imputed, the user may use any of the available parametric and nonparametric BTVs estimation methods for full data sets (without NDs), as described in Chapter 3. Those BTV estimation methods are not repeated here. The users may want to review the behavior of the various ROS methods as described in Chapter 4.

5.3.1.4 UPLs when Detected Data are Gamma Distributed

When detected data follow a gamma distribution, methods described in Chapter 3 can be used on KM estimates to compute gamma distribution based upper prediction limits for future k≥1 observations. These limits are described below when k=1.

$$\textit{Wilson-Hilferty (WH) UPL} = max\left(0, \left(\bar{y}_{KM} + t_{((1-\alpha),(n-1))} * s_{yKM} * \sqrt{1 + 1/n}\right)^3\right)$$

$$\textit{Hawkins-Wixley (HW) UPL} = \left(\bar{y}_{KM} + t_{((1-\alpha),(n-1))} * s_{yKM} * \sqrt{1 + 1/n}\right)^4$$

Here $t_{((1-\alpha),(n-1))}$ is a critical value from the Student's t-distribution with (n–1) degrees of freedom (*df*), and KM estimates are computed based upon the transformed *y* data as described in Chapter 3. All detects and NDs are transformed to y-space to compute the KM estimates.

One of the advantages of using this method is that one does not have to impute NDs based upon the data distribution using LROS or GROS method.

5.3.1.5 UPLs when Detected Data are Lognormally Distributed

When detected data follow a lognormal distribution, methods described in Chapter 3 can be used on KM estimates to compute lognormal distribution based upper prediction limits for future k≥1 observations. These limits are described below when k=1. An upper *(1 – α)*100%* lognormal UPL is given by the following equation:

$$\textit{UPL} = exp\left(\bar{y} + t_{((1-\alpha),(n-1))} * s_y * \sqrt{1 + 1/n}\right)$$

Here $t_{((1-\alpha),(n-1))}$ is a critical value from Student's t-distribution with (n–1) *df*, and $\bar{y}$ and s_y represent the KM mean and *sd* based upon the log-transformed data (detects and NDs), y. All detects and NDs are transformed to y-space to compute the KM estimates.

*5.3.2 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.

5.3.2.1 Upper Percentiles Based upon Standard Normal Z-Scores

In a left-censored data set, when detected data are normally distributed, one can use normal percentiles and KM estimates (or some other estimates such as ROS estimates) of the mean and *sd* to compute the p^{th} percentile given as 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, 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.

5.3.2.2 Upper Percentiles when Detected Data are Lognormally Distributed

When detected data follow a lognormal distribution, methods described in Chapter 3 can be used on the KM estimates to compute lognormal distribution based upper percentiles. The lognormal distribution based p^{th} percentile based upon KM estimates is given as follows:

$$\hat{x}_p = e^{(\bar{y}+s_y z_p)}$$

In the above equation, $\bar{y}$ and s_y represent the KM mean and *sd* based upon the log-transformed data (detects and NDs), y. All detects and NDs are transformed to y-space to compute the KM estimates.

5.3.2.3 Upper Percentiles when Detected Data are Gamma Distributed

When detected data are gamma distributed, gamma percentiles can be computed similarly using the HW and WH approximations to compute KM estimates. According to the WH approximation, the transformed detected data $Y = X^{1/3}$ follow an approximate normal distribution; and according to the HW approximation, the transformed detected data $Y = X^{1/4}$ follow an approximate normal distribution. Let $\bar{y}$ and s_y represent the KM mean and *sd* of the transformed data (detects and NDs), y. The percentiles based upon the WH and HW transformations respectively are given as follows:

$$\hat{x}_p = max\begin{cases} 0, \\ \left(\bar{y} + z_p \cdot s_y\right)^3 \end{cases}$$

$$\hat{x}_p = \left(\bar{y} + z_p * s_y\right)^4$$

Alternatively, following the process described in Section 4.6.2, one can use KM estimates to compute KM estimates, $\hat{k}$ and $\hat{\theta}$ of the shape, *k* and scale, θ parameters of the gamma distribution, and use the chi-square distribution to compute gamma percentiles using the equation: X = Y * θ /2, where Y follows a chi-square distribution with $2\hat{k}$ degrees of freedom (*df*). This method does not require HW or WH approximations to compute gamma percentiles. Once an α*100% percentile, $y_\alpha = y_{(\alpha)\, 2k}$, of a chi-square distribution with $2\hat{k}$ *df* is obtained, the α*100% percentile for a gamma distribution is computed using the equation: $x_\alpha = y_\alpha * \hat{\theta}/2$. ProUCL computes gamma percentiles using this equation based upon KM estimates.

5.3.2.4 Upper Percentiles Based upon ROS Methods

As noted in Chapter 4, all ROS methods first impute k ND values using an OLS linear regression (Chapter 4) assuming a specified distribution of detected observations. This process results in a full data set of size *n* consisting of *k* imputed NDs and *(n*-k*)* detected original values. For ROS methods (normal, gamma, lognormal), ProUCL generates additional columns consisting of the *(n-k)* detected values, and k imputed ND values. Once, the ND observations have been imputed, an experienced user may use any of the parametric or nonparametric percentile computation methods for full uncensored data sets as described in Chapter 3. Those methods are not repeated in this chapter.

5.3.3 Computing Upper Tolerance Limits (UTLs) for Left-Censored Data Sets

UTL computation methods for data sets consisting of NDs are described in this section.

5.3.3.1 UTLs Based on KM Estimates when Detected Data are Normally Distributed

Normal distribution based UTLs computed using KM estimates may be used when the detected data are normally distributed (can be verified using **GOF** module of ProUCL) or moderately to mildly skewed, with the *sd* of log-transformed detected data, σ, less than 0.5-0.75. An upper $(1-\alpha)$*100% tolerance limit with tolerance or coverage coefficient, p, is given by the following statement:

$$UTL = \hat{\mu}_{KM} + K_{n,\alpha,p} * \sqrt{\hat{\sigma}^2_{KM}} \quad (5\text{-}4)$$

Here $K = K(n, \alpha, p)$ is the tolerance factor used to compute upper tolerance limits and depends upon the sample size, n, CC = $(1-\alpha)$, and the coverage proportion = p. The K critical values are based upon the non-central t-distribution, and have been tabulated extensively in the statistical literature (Hahn and Meeker 1991). For samples of sizes larger than 30, one can use Natrella's approximation (1963) to compute the tolerance factor, $K = K(n, \alpha, p)$.

5.3.3.2 UTLs Based on KM Estimates when Detected Data are Lognormally Distributed

When detected data follow a lognormal distribution, methods described in Chapter 3 can be used on KM estimates to compute lognormal distribution based upper tolerance limits. An upper $(1-\alpha)$*100% tolerance limit with tolerance or coverage coefficient, p, is given by the following statement:

$$UTL = exp(\bar{y} + K * s_y)$$

The K factor in the above equation is the same as the one used to compute the normal UTL; $\bar{y}$ and s_y represent the KM mean and *sd* based upon the log-transformed data. All detects and NDs are transformed to y-space to compute KM estimates.

5.3.3.3 UTLs Based on KM Estimates when Detected Data are Gamma Distributed

According to the WH approximation, the transformed detected data $Y = X^{1/3}$ follow an approximate normal distribution; and according to the HW approximation, the transformed detected data $Y = X^{1/4}$ follow an approximate normal distribution when detected X data are gamma distributed. Let $\bar{y}$ and s_y represent the KM mean and *sd* based upon transformed data (detects and NDs), Y.

Using the WH approximation, the gamma UTL (in original scale, X), is given by:

$$UTL = max\left(0, (\bar{y} + K * s_y)^3\right)$$

Similarly, using the HW approximation, the gamma UTL in original scale is given by:

$$UTL = (\bar{y} + K * s_y)^4$$

5.3.3.4 UTLs Based upon ROS Methods

As noted in Chapter 4, all ROS methods first impute k ND values using an OLS linear regression line assuming a specified distribution of detected and nondetected observations. This process results in a full data set of size *n* consisting of k imputed NDs and *(n*-k*)* detected original values. For ROS methods (normal, gamma, lognormal), ProUCL generates additional columns consisting of the *(n*-k*)* detected values, and k imputed ND values. Once, the ND observations have been imputed, an experienced user may use any of the parametric or nonparametric UTL computation methods for full data sets as described in Chapter 3. Those methods are not repeated in this chapter.

Note: In the Stats/Sample Sizes module, using the **General Statistics** option for data sets with NDs, for information and summary purposes, percentiles are computed using detects and nondetects, where reported DLs are used for NDs. Those percentiles do not account for NDs. However, KM method based upper limits such as the UTL95-95 account for NDs; therefore, sometimes, a UTL95-95 computed based upon a ND method (e.g., KM method) may be lower than the 95% percentile computed using the **General Statistics** option of **Stats/Sample Sizes** module.

5.3.4 Computing Upper Simultaneous Limits (USLs) for Left-Censored Data Sets

Parametric and nonparametric USL computation methods for are described as follows.

5.3.4.1 USLs Based upon Normal Distribution of Detected Observations and KM Estimates

When detected observations follow a normal distribution (can be verified by using the **GOF** module of ProUCL), one can use the normal distribution on KM estimates to compute a USL95.

A one-sided *(1 – α)* 100% USL providing *(1 – α) 100%* coverage for all sample observations is given by:

$$USL = \hat{\mu}_{KM} + d^{b}_{2\alpha} * \sqrt{\hat{\sigma}^{2}_{KM}}$$

Here $\left(d^{b}_{2\alpha}\right)^{2}$ is the critical value of Max (Mahalanobis Distances) for *2*α* level of significance.

5.3.4.2 USLs Based upon Lognormal Distribution of Detected Observations and KM Estimates

When detected data follow a lognormal distribution, methods described in Chapter 3 can be used on the KM estimates to compute lognormal distribution based USLs. Let $\bar{y}$ and s_y represent the KM mean and *sd* of the log-transformed data (detects and NDs), *y*; a *(1 – α) 100%* USL is given by as follows:

$$USL = exp(\bar{y} + s_y * d^{b}_{2\alpha})$$

5.3.4.3 USLs Based upon Gamma Distribution of Detected Observations and KM Estimates

According to the WH approximation, the transformed detected data $Y = X^{1/3}$ follow an approximate normal distribution; and according to the HW approximation, the transformed detected data $Y = X^{1/4}$ follow an approximate normal distribution. Let $\bar{y}$ and s_y represent the KM mean and *sd* of the transformed data

(detects and NDs), *y*. A gamma distribution based (using WH approximation), one-sided *(1 – α) 100%* USL is given by:

$$USL = max\left(0, \left(\bar{y} + d_{2\alpha}^{b} * s_y\right)^3\right)$$

A gamma distribution based (HW approximation) one-sided (1 – α) 100% USL is given as follows:

$$USL = \left(\bar{y} + d_{2\alpha}^{b} * s_y\right)^4$$

5.3.4.4 USLs Based upon ROS Methods

Once, the ND observations have been imputed, one can use parametric or nonparametric USL computation methods for full data sets as described in Chapter 3.

Example 5-1 (Oahu Data Set). The detected data are only moderately skewed (*sd* of logged detects = 0.694) and follow a lognormal as well as a gamma distribution. The various upper limits computed using ProUCL 5.1 are listed in Tables 5-1 through 5-3 as follows.

Table 5-1. Nonparametric and Normal Upper Limits Using KM Estimates

Arsenic

General Statistics			
Total Number of Observations	24	Number of Missing Observations	0
Number of Distinct Observations	10		
Number of Detects	11	Number of Non-Detects	13
Number of Distinct Detects	8	Number of Distinct Non-Detects	3
Minimum Detect	0.5	Minimum Non-Detect	0.9
Maximum Detect	3.2	Maximum Non-Detect	2
Variance Detected	0.931	Percent Non-Detects	54.17%
Mean Detected	1.236	SD Detected	0.965
Mean of Detected Logged Data	-0.0255	SD of Detected Logged Data	0.694

Critical Values for Background Threshold Values (BTVs)			
Tolerance Factor K (For UTL)	2.309	d2max (for USL)	2.644

Normal GOF Test on Detects Only			
Shapiro Wilk Test Statistic	0.777	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.85	Data Not Normal at 5% Significance Level	
Lilliefors Test Statistic	0.273	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.267	Data Not Normal at 5% Significance Level	
Data Not Normal at 5% Significance Level			
Kaplan Meier (KM) Background Statistics Assuming Normal Distribution			
Mean	0.949	SD	0.713
95% UTL95% Coverage	2.595	95% KM UPL (t)	2.196
95% KM Chebyshev UPL	4.121	90% KM Percentile (z)	1.863
95% KM Percentile (z)	2.122	99% KM Percentile (z)	2.608
95% KM USL	2.834		

Note that the upper limits, based upon the gamma and lognormal distribution, are comparable. The upper limits computed using KM estimates based upon normal equations are slightly lower than other upper limits which adjust for data skewness. Table 5-1 mostly contains normal distribution based upper limits computed using KM estimates as described in Helsel (2012b) irrespective of the distribution of the detected data. The detected data follow a gamma distribution as shown in Table 5-2 below. A gamma UTL95-95 using KM estimates = 2.66 (WH); and a UTL95-95 based upon the GROS method is 3.15 (WH). From Table 5-3, a lognormal UTL95-95 using KM estimates = 2.79, and a UTL95-95 using the LROS method =3.03.

Table 5-2. Upper Limits Using GROS, KM Estimates and Gamma Distribution of Detected Data

Gamma GOF Tests on Detected Observations Only		
A-D Test Statistic	0.787	**Anderson-Darling GOF Test**
5% A-D Critical Value	0.738	Data Not Gamma Distributed at 5% Significance Level
K-S Test Statistic	0.254	**Kolmogrov-Smirnoff GOF**
5% K-S Critical Value	0.258	Detected data appear Gamma Distributed at 5% Significance Level
Detected data follow Appr. Gamma Distribution at 5% Significance Level		

Gamma Statistics on Detected Data Only			
k hat (MLE)	2.257	k star (bias corrected MLE)	1.702
Theta hat (MLE)	0.548	Theta star (bias corrected MLE)	0.727
nu hat (MLE)	49.65	nu star (bias corrected)	37.44
MLE Mean (bias corrected)	1.236		
MLE Sd (bias corrected)	0.948	95% Percentile of Chisquare (2k)	8.503

Minimum	0.119	Mean	0.956
Maximum	3.2	Median	0.7
SD	0.758	CV	0.793
k hat (MLE)	2.071	k star (bias corrected MLE)	1.84
Theta hat (MLE)	0.461	Theta star (bias corrected MLE)	0.519
nu hat (MLE)	99.41	nu star (bias corrected)	88.32
MLE Mean (bias corrected)	0.956	MLE Sd (bias corrected)	0.704
95% Percentile of Chisquare (2k)	8.964	90% Percentile	1.895
95% Percentile	2.328	99% Percentile	3.291

The following statistics are computed using Gamma ROS Statistics on Imputed Data					
Upper Limits using Wilson Hilferty (WH) and Hawkins Wixley (HW) Methods					
	WH	HW		WH	HW
95% Approx. Gamma UTL with 95% Coverage	3.149	3.299	95% Approx. Gamma UPL	2.384	2.436
95% Gamma USL	3.676	3.915			

The following statistics are computed using gamma distribution and KM estimates					
Upper Limits using Wilson Hilferty (WH) and Hawkins Wixley (HW) Methods					
k hat (KM)	1.771		nu hat (KM)	85.02	
	WH	HW		WH	HW
95% Approx. Gamma UTL with 95% Coverage	2.661	2.685	95% Approx. Gamma UPL	2.087	2.077
95% Gamma USL	3.051	3.107			

Table 5-3. Upper Limits Using LROS method and KM Estimates and Lognormal Distribution of Detected Data

Lognormal GOF Test on Detected Observations Only			
Shapiro Wilk Test Statistic	0.86	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.85	Detected Data appear Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.229	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.267	Detected Data appear Lognormal at 5% Significance Level	
Detected Data appear Lognormal at 5% Significance Level			
Background Lognormal ROS Statistics Assuming Lognormal Distribution Using Imputed Non-Detects			
Mean in Original Scale	0.972	Mean in Log Scale	-0.209
SD in Original Scale	0.718	SD in Log Scale	0.571
95% UTL95% Coverage	3.032	95% BCA UTL95% Coverage	3.2
95% Bootstrap (%) UTL95% Coverage	3.2	95% UPL (t)	2.202
90% Percentile (z)	1.686	95% Percentile (z)	2.075
99% Percentile (z)	3.062	95% USL	3.671
Statistics using KM estimates on Logged Data and Assuming Lognormal Distribution			
KM Mean of Logged Data	-0.236	95% KM UTL (Lognormal)95% Coverage	2.792
KM SD of Logged Data	0.547	95% KM UPL (Lognormal)	2.056
95% KM Percentile Lognormal (z)	1.942	95% KM USL (Lognormal)	3.354

Example 5-2. A real data set of size 55 with 18.8% NDs is considered next. This data was used in Chapter 4 to illustrate the differences in UCLs computed using a lognormal and a gamma distribution. This data set is considered here to illustrate the merits of the gamma distribution based upper limits. It can be seen that the detected data follow a gamma as well as a lognormal distribution. The minimum detected value is 5.2 and the largest detected value is 79000. The *sd* of the detected logged data is 2.79 suggesting that the detected data set is highly skewed. Relevant statistics and upper limits including a UPL95, UTL95-95, and UCL95 have been computed using both the gamma and lognormal distributions. The gamma GOF Q-Q plot is shown as follows.

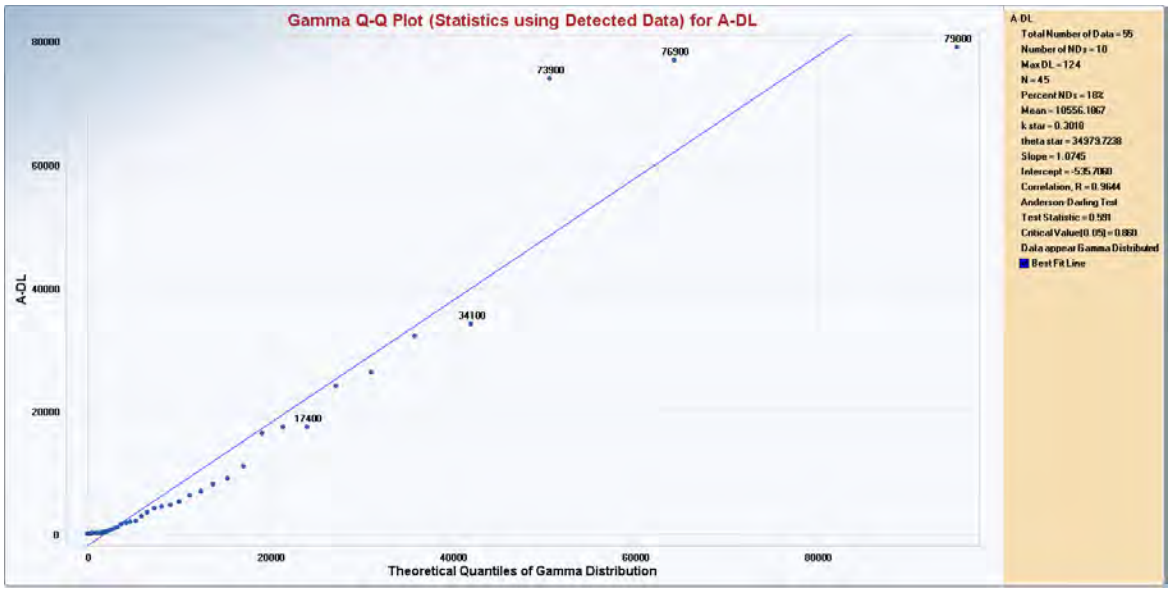

Figure 5-1. Gamma Q-Q Plot.

Table 5-4. Summary Statistics for Data Set of Example 5-2

A-DL			
General Statistics			
Total Number of Observations	55	Number of Missing Observations	0
Number of Distinct Observations	53		
Number of Detects	45	Number of Non-Detects	10
Number of Distinct Detects	45	Number of Distinct Non-Detects	8
Minimum Detect	5.2	Minimum Non-Detect	3.8
Maximum Detect	79000	Maximum Non-Detect	124
Variance Detected	3.954E+8	Percent Non-Detects	18.18%
Mean Detected	10556	SD Detected	19886
Mean of Detected Logged Data	7.031	SD of Detected Logged Data	2.788
Critical Values for Background Threshold Values (BTVs)			
Tolerance Factor K (For UTL)	2.036	d2max (for USL)	2.994

Mean of detects (=10556) reported above ignores all 18.18% NDs.

Table 5-5. KM Method Based Estimates of the Mean, SE of the Mean, and *sd*

Mean	8638
SD	18246
Standard Error of Mean	2488

KM mean (= 8638) reported above accounts for 18.18% NDs reported in the data set.

Notes: Direct estimate of KM *sd* = 18246

Indirect Estimate of KM *sd* (Helsel 2012b) = 18451.5

The gamma GOF test results on detected data and various upper limits including UCLs obtained using the GROS method and gamma distribution on KM estimates are provided in Tables 5-6 through 5-9; and the lognormal GOF test results on detected data and the various upper limits obtained using the LROS method and lognormal distribution on KM estimates are provided in Tables 5-10 and 5-11. Table 5-12 is a summary of the main upper limits computed using the lognormal and gamma distribution of the detected data.

Table 5-6. Upper Limits Using GROS, KM Estimates and Gamma Distribution of Detected Data

Gamma GOF Tests on Detected Observations Only			
A-D Test Statistic	0.591	**Anderson-Darling GOF Test**	
5% A-D Critical Value	0.86	Detected data appear Gamma Distributed at 5% Significance Level	
K-S Test Statistic	0.115	**Kolmogrov-Smirnoff GOF**	
5% K-S Critical Value	0.143	Detected data appear Gamma Distributed at 5% Significance Level	
Detected data appear Gamma Distributed at 5% Significance Level			
Gamma Statistics on Detected Data Only			
k hat (MLE)	0.307	k star (bias corrected MLE)	0.302
Theta hat (MLE)	34333	Theta star (bias corrected MLE)	34980
nu hat (MLE)	27.67	nu star (bias corrected)	27.16
MLE Mean (bias corrected)	10556		
MLE Sd (bias corrected)	19216	95% Percentile of Chisquare (2k)	2.756

Table 5-7. Upper Limits Computed Using Gamma ROS Method

Gamma ROS Statistics using Imputed Non-Detects			
Minimum	1.121	Mean	8642
Maximum	79000	Median	588
SD	18412	CV	2.13
k hat (MLE)	0.247	k star (bias corrected MLE)	0.246
Theta hat (MLE)	35001	Theta star (bias corrected MLE)	35193
nu hat (MLE)	27.16	nu star (bias corrected)	27.01
MLE Mean (bias corrected)	8642	MLE Sd (bias corrected)	17440
95% Percentile of Chisquare (2k)	2.39	90% Percentile	25972
95% Percentile	42055	99% Percentile	84976
The following statistics are computed using Gamma ROS Statistics on Imputed Data			
Upper Limits using Wilson Hilferty (WH) and Hawkins Wixley (HW) Methods			
	WH	HW	WH	HW
95% Approx. Gamma UTL with 95% Coverage	47429	54346	95% Approx. Gamma UPL	33332	35476

Table 5-8. Upper Limits Computed Using Gamma Distribution and KM Estimates

The following statistics are computed using gamma distribution and KM estimates				
Upper Limits using Wilson Hilferty (WH) and Hawkins Wixley (HW) Methods				
k hat (KM)	0.224		nu hat (KM)	24.66
	WH	HW	WH	HW
95% Approx. Gamma UTL with 95% Coverage	46978	54120	95% Approx. Gamma UPL	32961	35195

Table 5-8. 95% UCL of the Mean Based upon GROS Method

		Adjusted Level of Significance (β)	0.0456
Approximate Chi Square Value (27.01, α)	16.16	Adjusted Chi Square Value (27.01, β)	15.93
95% Gamma Approximate UCL (use when n>=50)	14445	95% Gamma Adjusted UCL (use when n<50)	14651

Table 5-9. 95% UCL of the Mean Using Gamma Distribution on KM Estimates

Gamma Kaplan-Meier (KM) Statistics			
k hat (KM)	0.224	nu hat (KM)	24.66
Approximate Chi Square Value (24.66, α)	14.35	Adjusted Chi Square Value (24.66, β)	14.14
95% Gamma Approximate KM-UCL (use when n>=50)	14844	95% Gamma Adjusted KM-UCL (use when n<50)	15066

Table 5-10. Upper Limits Using LROS and KM Estimates and Lognormal Distribution of Detected Data

Lognormal GOF Test on Detected Observations Only			
Shapiro Wilk Test Statistic	0.939	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.945	Data Not Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.104	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.132	Detected Data appear Lognormal at 5% Significance Level	
Detected Data appear Approximate Lognormal at 5% Significance Level			
Background Lognormal ROS Statistics Assuming Lognormal Distribution Using Imputed Non-Detects			
Mean in Original Scale	8638	Mean in Log Scale	5.983
SD in Original Scale	18414	SD in Log Scale	3.391
95% UTL95% Coverage	394791	95% BCA UTL95% Coverage	77530
95% Bootstrap (%) UTL95% Coverage	77530	95% UPL (t)	121584
90% Percentile (z)	30572	95% Percentile (z)	104784
99% Percentile (z)	1056400	95% USL	10156719
Statistics using KM estimates on Logged Data and Assuming Lognormal Distribution			
KM Mean of Logged Data	6.03	95% KM UTL (Lognormal)95% Coverage	334181
KM SD of Logged Data	3.286	95% KM UPL (Lognormal)	106741
		95% KM Percentile Lognormal (z)	92417

Table 5-11. 95% UCL of the mean Using LROS and Lognormal Distribution on KM Estimates Methods

Lognormal ROS Statistics Using Imputed Non-Detects			
Mean in Original Scale	8638	Mean in Log Scale	5.983
SD in Original Scale	18414	SD in Log Scale	3.391
95% t UCL (assumes normality of ROS data)	12793	95% Percentile Bootstrap UCL	12676
95% BCA Bootstrap UCL	13762	95% Bootstrap t UCL	14659
95% H-UCL (Log ROS)	1855231		
UCLs using Lognormal Distribution and KM Estimates when Detected data are Lognormally Distributed			
KM Mean (logged)	6.03	95% H-UCL (KM -Log)	1173988
KM SD (logged)	3.286	95% Critical H Value (KM-Log)	5.7
KM Standard Error of Mean (logged)	0.449		

Nonparametric upper percentiles are: 9340 (80%), 25320 (90%), 46040 (95%), and 77866 (99%). Other upper limits, based upon the gamma and lognormal distribution, are described in Table 5-12. All computations have been performed using the ProUCL software. In the following Table 5-6, method proposed/described in the literature have been cited in the Reference Method of Calculation column.

Table 5-12. Summary of Upper Limits Computed using Gamma and Lognormal Distribution of Detected Data: Sample Size = 55, No. of NDs=10, % NDs = 18.18%

Upper Limits	Gamma Distribution		Lognormal Distribution	
	Result	**Reference/ Method of Calculation**	**Result**	**Reference/ Method of Calculation**
Min (detects)	5.2	--	1.65	Logged
Max (detects)	79,000	--	11.277	Logged
Mean (KM)	8,638	--	6.3	Logged
Mean (ROS)	8,642	--	8,638	--
95% Percentile (ROS)	42,055	--	104,784	--
UPL95 (ROS)	33,332	WH- ProUCL	121,584	Helsel (2012b), EPA (2009e)
UTL95-95 (ROS)	47,429	WH- ProUCL	394,791	Helsel (2012b), EPA (2009e)
UPL95 (KM)	32,961	WH-ProUCL	106,741	EPA (2009e)
UTL95-95 (KM)	46,978	WH-ProUCL	334,181	EPA(2009e)
UCL95 (ROS)	14,445	ProUCL	14,659	bootstrap-t, ProUCL 5.0
			12,676	percentile bootstrap, Helsel (2012b)
UCL (KM)	14,844	ProUCL	1,173,988	H-UCL, KM mean and *sd* on logged data - EPA (2009e)

The statistics listed in Tables 5-4 through 5-11, and summarized in Table 5-13 demonstrate the need and merits of using the gamma distribution for computing practical and meaningful estimates (upper limits) of the decision parameters (e.g., mean, upper percentile) of interest.

Example 5.3. The benzene data set (Benzene-H-UCL-RCRA.xls) of size 8 used in Chapter 21 of the RCRA Unified Guidance document (EPA 2009e) was used in Section 4.6.3.1 to address some issues associated

with the use of lognormal distribution to compute a UCL of mean for data sets with nondetects. The benzene data set is used in this example to illustrate similar issues associated with the computation of UTLs and UPLs based upon lognormal distribution using substitution methods. Lognormal distribution based upper limits using ROS and KM methods are summarized in Table 5-13.

Table 5-13. Lognormal 95%-95% Upper Limits based upon LROS and KM Estimates

Lognormal GOF Test on Detected Observations Only		
Shapiro Wilk Test Statistic	0.829	**Shapiro Wilk GOF Test**
5% Shapiro Wilk Critical Value	0.803	Detected Data appear Lognormal at 5% Significance Level
Lilliefors Test Statistic	0.304	**Lilliefors GOF Test**
5% Lilliefors Critical Value	0.335	Detected Data appear Lognormal at 5% Significance Level
Detected Data appear Lognormal at 5% Significance Level		

Background Lognormal ROS Statistics Assuming Lognormal Distribution Using Imputed Non-Detects			
Mean in Original Scale	2.913	Mean in Log Scale	0.0926
SD in Original Scale	5.364	SD in Log Scale	1.443
95% UTL95% Coverage	109.2	95% BCA UTL95% Coverage	16.1
95% Bootstrap (%) UTL95% Coverage	16.1	95% UPL (t)	19.95
90% Percentile (z)	6.976	95% Percentile (z)	11.79
99% Percentile (z)	31.52	95% USL	20.6
Statistics using KM estimates on Logged Data and Assuming Lognormal Distribution			
KM Mean of Logged Data	0.29	95% KM UTL (Lognormal)95% Coverage	41.42
KM SD of Logged Data	1.077	95% KM UPL (Lognormal)	11.65
95% KM Percentile Lognormal (z)	7.865	95% KM USL (Lognormal)	11.93

The data set has only one ND with a DL of 0.5. Lognormal upper limits computed by replacing the ND by DL and DL/2, respectively are given in Tables 5-14 and 5-15.

Table 5-14. Lognormal Distribution Based Upper Limits using DL (=0.5) for ND

Mean of logged Data	0.29	SD of logged Data	1.152

Lognormal GOF Test		
Shapiro Wilk Test Statistic	0.803	**Shapiro Wilk Lognormal GOF Test**
5% Shapiro Wilk Critical Value	0.818	Data Not Lognormal at 5% Significance Level
Lilliefors Test Statistic	0.273	**Lilliefors Lognormal GOF Test**
5% Lilliefors Critical Value	0.313	Data appear Lognormal at 5% Significance Level
Data appear Approximate Lognormal at 5% Significance Level		
Background Statistics assuming Lognormal Distribution		
95% UTL with 95% Coverage	52.5	90% Percentile (z)	5.849
95% UPL (t)	13.53	95% Percentile (z)	8.888
95% USL	13.88	99% Percentile (z)	19.48

Table 5-15. Lognormal Distribution Based Upper Limits using DL/2 (=0.25) for ND

Mean of logged Data	0.204	SD of logged Data	1.257

Lognormal GOF Test			
Shapiro Wilk Test Statistic	0.896	**Shapiro Wilk Lognormal GOF Test**	
5% Shapiro Wilk Critical Value	0.818	Data appear Lognormal at 5% Significance Level	
Lilliefors Test Statistic	0.255	**Lilliefors Lognormal GOF Test**	
5% Lilliefors Critical Value	0.313	Data appear Lognormal at 5% Significance Level	
Data appear Lognormal at 5% Significance Level			
Background Statistics assuming Lognormal Distribution			
95% UTL with 95% Coverage	67.44	90% Percentile (z)	6.142
95% UPL (t)	15.34	95% Percentile (z)	9.699
95% USL	15.78	99% Percentile (z)	22.85

Note: Even though UPLs and UTLs computed using the lognormal distribution do not suffer from transformation bias, a minor increase in the *sd* of logged data (from 1.152 to 1.257 above) causes a significant increase in upper limits, especially in UTLs (from 52.5 to 67.44) computed using a small data set (<15-20). This is particularly true when the data set contains outliers.

Impact of Outlier, 16.1 ppb on the Computations of Upper Limits

Benzene data set without the outlier, 16.1 ppb, follows a normal distribution, and normal distribution based upper limits without the outlier 16.1 are summarized as follows in Tables 5-16 (KM estimates), 5-17 (ND by DL), and 5-18 (ND by DL/2).

Table 5-16. Normal Distribution Based Upper Limits Computed Using KM estimates

Normal GOF Test on Detects Only			
Shapiro Wilk Test Statistic	0.847	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.788	Detected Data appear Normal at 5% Significance Level	
Lilliefors Test Statistic	0.265	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.362	Detected Data appear Normal at 5% Significance Level	
Detected Data appear Normal at 5% Significance Level			
Kaplan Meier (KM) Background Statistics Assuming Normal Distribution			
Mean	1.086	SD	0.544
95% UTL95% Coverage	2.933	95% KM UPL (t)	2.215
95% KM Chebyshev UPL	3.619	90% KM Percentile (z)	1.782
95% KM Percentile (z)	1.98	99% KM Percentile (z)	2.35
95% KM USL	2.139		

Table 5-17. Normal Distribution Based Upper Limits Computed using DL for ND

Normal GOF Test			
Shapiro Wilk Test Statistic	0.814	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.803	Data appear Normal at 5% Significance Level	
Lilliefors Test Statistic	0.269	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.335	Data appear Normal at 5% Significance Level	
Data appear Normal at 5% Significance Level			
Background Statistics Assuming Normal Distribution			
95% UTL with 95% Coverage	3.081	90% Percentile (z)	1.838
95% UPL (t)	2.305	95% Percentile (z)	2.052
95% USL	2.224	99% Percentile (z)	2.452

Note: DL (=0.5) has been used for the ND value (does not accurately account for its ND status). Therefore, upper limits are slightly higher than those computed using KM estimates.

Table 5-18. Normal Distribution Based Upper Limits Computed using DL/2 for ND

Normal GOF Test			
Shapiro Wilk Test Statistic	0.875	**Shapiro Wilk GOF Test**	
5% Shapiro Wilk Critical Value	0.803	Data appear Normal at 5% Significance Level	
Lilliefors Test Statistic	0.236	**Lilliefors GOF Test**	
5% Lilliefors Critical Value	0.335	Data appear Normal at 5% Significance Level	
Data appear Normal at 5% Significance Level			
Background Statistics Assuming Normal Distribution			
95% UTL with 95% Coverage	3.206	90% Percentile (z)	1.863
95% UPL (t)	2.368	95% Percentile (z)	2.094
95% USL	2.28	99% Percentile (z)	2.526

Note: DL/2 (=0.25) has been used for the ND value (does not accurately account for its ND status). The use of DL/2 has increased the variance slightly which causes a slight increase in the various upper limits. Therefore, upper limits are slightly higher than those computed using KM estimates and using DL for the ND value. Based upon the benzene data set, normal UTL95-95 (= 2.93) computed using KM estimates appears to represent a more realistic estimate of background threshold value.

Example 5-4. The manganese (Mn) data set used in Chapter 15 of the Unified RCRA Guidance (2009) has been used here to demonstrate how LROS method generates elevated BTVs.

Table 5-19. Summary statistics for Example 5-4.

General Statistics for Censored Datasets (with NDs) using Kaplan Meier Method											
Variable	NumObs	# Missing	Num Ds	NumNDs	% NDs	Min ND	Max ND	KM Mean	KM Var	KM SD	KM CV
Mn	25	0	19	6	24.00%	2	5	19.87	641	25.32	1.274

General Statistics for Raw Dataset using Detected Data Only											
Variable	NumObs	# Missing	Minimum	Maximum	Mean	Median	Var	SD	MAD/0.675	Skewness	CV
Mn	19	0	3.3	106.3	25.46	12.6	752.7	27.44	9.34	1.942	1.078

Percentiles using all Detects (Ds) and Non-Detects (NDs)										
Variable	NumObs	# Missing	10%ile	20%ile	25%ile(Q1)	50%ile(Q2)	75%ile(Q3)	80%ile	90%ile	95%ile	99%ile
Mn	25	0	2.52	5	5	10	21.6	25.06	50.52	72.48	99.32

The detected data follow a lognormal distribution, the maximum value in the data set is 106, and using the LROS method (robust ROS method), one gets a 99% percentile = 183.4, and a UTL of 175. These statistics are summarized in Table 5-20.

The detected data also follows a gamma distribution. Gamma-KM method based upper limits are summarized as follows. The Gamma UTL95-95s (KM) are 92.5 (WH) and 99.32 (HW) and the 99% percentiles are: 94.42 (WH) and 101.8 (HW). The Gamma UTL (KM) appears to represent a reasonable estimate of BTV. These BTV estimates are summarized in Table 5-21.

Table 5-20. LROS and Lognormal KM Method Based Upper Limits

Background Lognormal ROS Statistics Assuming Lognormal Distribution Using Imputed Non-Detects			
Mean in Original Scale	19.83	Mean in Log Scale	2.277
SD in Original Scale	25.87	SD in Log Scale	1.261
95% UTL95% Coverage	175.6	95% BCA UTL95% Coverage	106.3
95% Bootstrap (%) UTL95% Coverage	106.3	95% UPL (t)	88.06
90% Percentile (z)	49.1	95% Percentile (z)	77.64
99% Percentile (z)	183.4	95% USL	280.4
Statistics using KM estimates on Logged Data and Assuming Lognormal Distribution			
KM Mean of Logged Data	2.309	95% KM UTL (Lognormal)95% Coverage	151
KM SD of Logged Data	1.182	95% KM UPL (Lognormal)	79.12
95% KM Percentile Lognormal (z)	70.31	95% KM USL (Lognormal)	234.1

Table 5-21. Gamma KM Method Based Upper Limits

The following statistics are computed using gamma distribution and KM estimates					
Upper Limits using Wilson Hilferty (WH) and Hawkins Wixley (HW) Methods					
	k hat (KM)	0.616		nu hat (KM)	30.79
	WH	HW		WH	HW
95% Approx. Gamma UTL with 95% Coverage	92.4	99.32	95% Approx. Gamma UPL	63.96	65.76
95% KM Gamma Percentile	59.5	60.7	95% Gamma USL	115.8	128.4

Notes: Even though one can argue that there is no transformation bias when computing lognormal distribution based UTLs and UPLs, the use of a lognormal distribution on data with or without NDs often yields inflated values which are not supported by the data set used to compute them. Therefore, its use including LROS method should be avoided.

Before using a nonparametric BTV estimate, one should make sure that the detected data do not follow a known distribution. When dealing with a data set with NDs, it is suggested to account for NDs and determine the distribution of detected values instead of using a nonparametric UTL as used in Example 17-4 on page 17-21 of Chapter 17 of the EPA Unified Guidance, 2009. If detected data follow a parametric distribution, one may want to compute a UTL using that distribution and KM estimates; this approach will account for data variability instead defaulting to higher order statistics.

5.3.5 Summary and Recommendation

It is recommended that outliers confirmed as suspect by investigation not be used in the computation of decision-making statistics. The decision-making statistics (e.g., UCLs, UTLs, UPLs) should be computed using observations representing the population. The use of a lognormal distribution should be avoided in computing upper limits (UCLs, UTLs, UPLs) based upon data sets with *sd* of detected logged data for moderately skewed to highly skewed data sets of sizes smaller than 20-30. It is reasonable to state that, like uncensored data sets without NDs, the minimum sample size requirement increases as the skewness increases.

The project team should collectively make a decision about the disposition of outliers. It is often helpful to compute decision statistics (upper limits) and hypothesis test statistics twice: once including outliers, and once without outliers. By comparing the upper limits computed with and without outliers, the project team can determine which limits are more representative of the site conditions under investigation.

5.4 Computing Nonparametric Upper Limits Based upon Higher Order Statistics

For full data sets without any discernible 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 DLs, ProUCL computes USLs, UTLs, UPLs, and upper percentiles by using nonparametric methods as described in Chapter 3. Since, nonparametric UTLs, UPLs, USLs, and upper percentiles are represented by higher order statistics (or by some value in between higher order statistic obtained using linear interpolation) every effort should be made to make sure that those higher order statistics do not represent observations coming from population(s) other than the dominant (e.g., background) population under study.

CHAPTER 6

Single and Two-sample Hypotheses Testing Approaches

Both single-sample and two-sample hypotheses testing approaches are used to make cleanup decisions at polluted sites, and compare constituent concentrations of two (e.g., site versus background) or more (GW in MWs) populations. This chapter provides 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. For interested users, this chapter presents a brief description of the mathematical formulations of the various parametric and nonparametric hypotheses testing approaches as incorporated in ProUCL. ProUCL software provides hypotheses testing approaches for data sets with and without ND observations. For data sets containing multiple nondetects, a new two-sample hypothesis test, the Tarone-Ware (T-W; 1978) test has been incorporated in the current ProUCL, versions 5.0 and 5.1. The developers of ProUCL recommend supplementing statistical test results with graphical displays. It is assumed that the users have collected an appropriate amount of good quality (representative) data, perhaps based upon data quality objectives (DQOs). The **Stats/Sample Sizes** module can be used to compute DQOs based sample sizes needed to perform the hypothesis tests described in this chapter.

6.1 When to Use Single Sample Hypotheses Approaches

When pre-established background threshold values and not-to-exceed values (e.g., USGS background values, Shacklette and Boerngen 1984) exist, there is no need to establish, or collect a background or reference data set. Specifically, when not-to-exceed action levels or average cleanup standards are known, one-sample hypotheses tests can be used to compare onsite data with known and pre-established threshold values, provided enough onsite data needed to perform the hypothesis tests are available. When the number of available site observations is less than 4-6, one might perform point-by-point site observation comparisons with a BTV; and when enough onsite observations (> 8 to 10, more are preferable) are available, it is suggested to use single-sample hypothesis testing approaches. Some recent EPA guidance documents (EPA 2009e) also recommend the availability of at least 8-10 observations to perform statistical inference. Some minimum sample size requirements related to hypothesis tests are also discussed in Chapter 1 of this Technical Guide.

Depending upon the parameter (e.g., the average value, μ_0, or a not-to-exceed action level, A_0), representing a known threshold value, one can use single-sample hypothesis tests for the population mean (t-test, sign test) or single-sample tests for proportions and percentiles. Several single-sample tests listed below are available in ProUCL.

One-Sample t-Test: This test is used to compare the site mean,μ, with some specified cleanup standard, C_s (μ_0), where C_s represents a specified value of the true population mean, μ. The Student's t- test or UCL of the mean is used (assuming normality of site data, or when the sample size is larger than 30, 50, or 100) to verify the attainment of cleanup levels at a polluted sites (EPA 1989a, 1994). Note that the large sample size requirement (n= 30, 50, or 100) depends upon the data skewness. Specifically, as skewness increases measured in terms of the *sd, σ,* of the log-transformed data, the large sample size requirement also increases to be able to apply the normal distribution and Student's t-statistic, due to the central limit theorem (CLT).

One-Sample Sign Test or Wilcoxon Signed Rank (WSR) Test: These tests are nonparametric tests which can also handle ND observations, provided all NDs and therefore their associated DLs are less than 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 the similar location measure.

One-Sample Proportion Test or Percentile Test: When a specified cleanup standard, A_0, such as a preliminary remediation goal (PRG), or a compliance limit (CL) represents an upper threshold value of a constituent concentration distribution rather than the mean threshold value, μ, a test for proportion or a test for percentile (e.g., UTL95-95, UTL95-90) can be used to compare exceedances to the actionable level. The proportion, p, of exceedances of A_0 by site observations are compared to some pre-specified allowable proportion, P_0, of exceedances. One scenario where this test may be applied is following remediation activities at an AOC. The proportion test can also handle NDs provided all NDs are below the action level, A_0.

It is beneficial to use DQO-based sampling plans to collect an appropriate amount of 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 observations) should be made available to perform the single-sample hypotheses tests listed above. As mentioned before, if only a small number of 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. Individual point-by-point observations are not to be compared with the average cleanup or threshold level, C_s. The estimated sample mean, such as a UCL95, is compared with a threshold representing an average cleanup standard.

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 is 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 MWs. The **Stats/Sample Sizes** module of ProUCL can be used to compute DQO-based sample sizes for two-sample parametric and nonparametric hypothesis testing approaches. 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) in sample size determinations. That is, a larger number (>10 to 15) of representative background (or site) samples may need to be collected from larger background (or site) areas to capture the greater inherent heterogeneity/variability typically present in larger areas.

The two-sample hypotheses approaches are used when the site parameters (e.g., mean, shape, distribution) are compared with the background parameters (e.g., mean, shape, distribution). Specifically, two-sample hypotheses testing approaches can be used to compare the average (also medians or upper tails) constituent concentrations of two or more populations such as the background population and the potentially contaminated site areas. 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 T-W test are included in ProUCL. Some details of those methods are described in this chapter for interested users. It is recommended that statistical results and test statistics be supplemented with graphical displays, such as the

multiple Q-Q plots and side-by-side box plots as graphical displays do not require any distributional assumptions and are not influenced by outlying observations and NDs.

Data Types: Analytical data sets collected from the two (or more) populations should be of the same type obtained using similar analytical methods and sampling equipment. Additionally, site and background data should be all discrete or all composite (obtained using the same design, pattern, and number of increments), and should be collected from the same medium (soil) at comparable depth levels (e.g., all surface samples or all subsurface samples) and time (e.g., during the same quarter in groundwater applications). Good sample collection methods and sampling strategies are described in Gerlach, R. W., and J. M. Nocerino (2003) and the ITRC ISM guidance documents (2012 and 2020).

6.3 Statistical Terminology Used in Hypotheses Testing Approaches

The first step in developing a hypothesis test is to state the problem in 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 based on the computed hypothesis test statistic (e.g., t-statistic, WMW test statistic). The statistical terminologies including error rates, hypotheses statements, Form 1, Form 2, and two-sided tests, are explained in terms of two-sample hypotheses testing approaches. Similar terms apply to all parametric and nonparametric single-sample and two-sample hypotheses testing approaches. Additional details may be found in EPA guidance documents (2002b, 2006b), and MARSSIM (2000) or in statistical text books including Bain and Engelhardt (1992), Hollander and Wolfe (1999), and Hogg and Craig (1995).

Two forms, Form 1 and Form 2, of the statistical hypothesis test are useful for environmental applications. The null hypothesis in the first form (Form 1) states that the mean/median concentration of the potentially impacted site area *does not exceed the mean/median of the background concentration.* The null hypothesis in the second form (Form 2) of the test is that the concentrations of the impacted site area *exceed the background concentrations by a substantial difference, S, with* $S \geq 0$.

Formally, let $X_1, X_2, \ldots, X_n$ represent a random sample of size n collected from Population 1 (e.g., downgradient MWs or a site AOC) with mean (or median) μ_X, and $Y_1, Y_2, \ldots, Y_m$ represent a random sample of size *m* from Population 2 (upgradient MWs or a background area) with mean (or median) μ_Y. Let $\Delta = \mu_X - \mu_Y$ represent the difference between the two means (or medians).

6.3.1 Test Form 1

The null hypothesis (H_0): The mean/median of Population 1 (constituent concentration in samples collected from potentially impacted areas (or monitoring wells)) is less than or equal to the mean/median of Population 2 (concentration in samples collected from background (or upgradient wells) areas) with H_0: $\Delta \leq 0$.

The alternative hypothesis (H_A*)*. The mean/median of Population 1 (constituent concentration in samples collected from potentially impacted areas) is greater than the mean of Population 2(background areas) with H_A: $\Delta > 0$.

When performing 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/median does exceed background mean/median concentration. Otherwise, the null hypothesis cannot be rejected based on the available data, and the mean/median concentration found in the potentially impacted site areas is considered equivalent and comparable to that of the background areas.

6.3.2 Test Form 2

The null hypothesis (H_0): The mean/median of Population 1 (constituent concentration in potentially impacted areas) exceeds the mean/median or Population 2 (background concentrations) by more than S units. Symbolically, the null hypothesis is written as H_0: $\Delta \geq S$, where $S \geq 0$.

The alternative hypothesis (H_A): The mean/median of Population 1 (constituent concentration in potentially impacted areas) does not exceed the mean/median of Population 2 (background constituent concentration) by more than S (H_A: $\Delta < S$).

Here, S is the background investigation level. When S>0, Test Form 2 is called Test Form 2 with substantial difference, S. Some details about this hypothesis form can be found in the background guidance document for CERCLA sites (EPA 2002b).

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, 2002b). 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 constituent 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 Rates and Confidence Levels

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, Δ, and the degree of decision errors that is acceptable. There are two ways to err when analyzing sampled data (Table 6-1) to derive conclusions about population parameters.

Type I Error: Based on the observed collected 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, based upon the collected data, 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 6-1. 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 two error probabilities, α and β.

Confidence level 100(1– α)%: As the confidence level is lowered (or alternatively, as α is increased), the likelihood of committing a Type I error increases.

Power 100(1 – β)%: 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 in the interval (0, 1) can be selected for these two parameters, as the demand for precision increases, the number of samples and the associated cost (sampling and analytical 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, or increased risk to human health and the environment.

There is an inherent tradeoff between the probabilities of committing a Type I or a 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). The choice of α determines the probability of the Type I error. The smaller the α-value, the less likely to incorrectly reject the null hypothesis (H_0). However, a smaller value for α also means lower power with decreased probability of detecting a difference when one exists. The most commonly used α value is 0.05. With $\alpha = 0.05$, the chance of finding a significance difference that does not really exist is only 5%. In most situations, this probability of error is considered acceptable.

Suggested values for the Two Types of Error Rates: Typically, the following values for error probabilities are selected as the minimum recommended performance measures:

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% (α = 0.10) and the power should be at least 80% (β = 0.20).

Seriousness of the Two Types of Error Rates:

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 a Type II error, 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 sets follow a known statistical distribution (mostly normal); and that the data sets are statistically independent with no expected spatial and temporal trends in the data sets. Many statistical tests (e.g., two-sample t-test) 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 of data are known as *parametric* tests. The most commonly used distribution for tests involving environmental data is the normal distribution. It is noted that GOF tests which are used to determine data set's distribution (e.g., S-W test for normality) often fail if there are not enough observations, if the data contain multiple populations, or if there is a high proportion of NDs 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 may be considered a small sample. However, in practice, many times it may not be possible, due to resource constraints, to collect data sets of sizes greater than 10. This is especially true for background data sets, as the decision makers often do not want to collect many background samples. Sometimes they want to make cleanup decisions based upon data sets of sizes even smaller than 10. Statistics computed based upon small data sets of sizes < 5 cannot be considered reliable to derive important decisions affecting human health and the environment.

6.5 Nonparametric Hypotheses Tests

Statistical tests that do not assume a specific statistical form for the data 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 NDs and outlying observations. In two-sample comparisons (e.g., t-test), 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 with different shapes and variability, then nonparametric tests may be used to perform site versus background comparisons. Typically, nonparametric tests and statistics require larger size data sets to derive correct conclusions. Several two-sample nonparametric hypotheses tests, the WMW test, Gehan test, and Tarone-Ware (T-W) test, are available in ProUCL. Like the Gehan test, the T-W test is used for data sets containing NDs with multiple RLs. The T-W test was new in ProUCL 5.0.

The relative performances of different testing procedures can 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 based upon the given data sets of sampled observations. The *p*-value of a test is sometimes called the critical level or the significance level of the test. Whenever possible, critical values and *p*-values have been computed using the exact or approximate distribution of the test statistics (e.g., GOF tests, t-test, Sign test, WMW test, Gehan test, M-K trend 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. Not all nonparametric methods are robust and resistant to outliers. Specifically, nonparametric upper limits used to estimate BTVs can get affected and misrepresented by outliers. This issue has been discussed earlier in Chapter 3 of this Technical Guide.

- If a parametric test for comparing means is applied to data from a non-normal population and the sample size is large, then a 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 the CLT can become quite large, sometimes larger than 100. A brief simulation study elaborating on the sample size requirements to apply the CLT on skewed data sets is given in Appendix B. For moderately skewed (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. Unless the population distribution is 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 on a data set 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, nonparametric tests for comparing means are only slightly less powerful than parametric tests with large samples.
- If a parametric test is applied on a data set from a non-normal population and the sample size is small (< 20 data points), then the *p*-value may be inaccurate because the CLT does not apply in this case.
- If a nonparametric test is applied to a data set 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.

Notes: It is suggested that the users supplement their test statistics and conclusions by using graphical displays for visual comparisons of two or more data sets. ProUCL software has side-by-side box plots and multiple Q-Q plots that can be used to graphically compare two or more data sets with and without ND observations.

6.6 Single Sample Hypotheses Testing Approaches

This section describes the mathematical formulations of parametric and nonparametric single-sample hypotheses testing approaches incorporated in ProUCL software. For the sake of interested users, some

directions to perform these hypotheses tests are described as follows. The directions are useful when the user wants to manually perform these tests.

6.6.1 The One-Sample t-Test for Mean

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 mean level (cleanup standard representing a mean concentration level). The **Stats/Sample Sizes** module of ProUCL can be used to determine the minimum number of observations needed to achieve the desired DQOs. The collected sample should be a random sample representing the AOC under investigation.

6.6.1.1 Limitations and Robustness of One-Sample t-Test

The one-sample t-test is not robust in the presence of outliers and may not yield reliable results in the presence of ND observations. Do not use this test when dealing with data sets containing NDs. Some nonparametric tests described below may be used in cases where NDs are present in a data set. This test may yield reliable results when performed on mildly or moderately skewed data sets. Note that levels of skewness are discussed in Chapters 3 and 4. The use of a t-test should be avoided when data are highly skewed (*sd* of log-transformed data exceeding 1, 1.5), even when the data set is of a large size such as n=100.

6.6.1.2 Directions for the One-Sample t-Test

Let $x_1, x_2, \ldots, x_n$ represent a random sample (analytical results) of size, n, collected from a population (AOC). The use of the One-Sample t-Test requires that the data set follows a normal distribution; that is when using a typical software package (e.g., Minitab), the user needs to test for the normality of the data set. For the sake of users and to make sure that users do not skip this step, ProUCL verifies normality of the data set automatically.

STEP 1: Specify an average cleanup goal or action level, μ_0 (C_s), and choose one of the following combination of null and alternative hypotheses:

Form 1: H_0: site $\mu \leq \mu_0$ vs. H_A: site $\mu > \mu_0$

Form 2: H_0: site $\mu \geq \mu_0$ vs. H_A: site $\mu < \mu_0$

Two-Sided: H_0: site $\mu = \mu_0$ vs. H_A: site $\mu \neq \mu_0$.

Form 2 with substantial difference, S: H_0: site $\mu \geq \mu_0 + S$ vs. H_A: site $\mu < \mu_0 + S$, here $S > 0$.

STEP 2: Calculate the test statistic:

$$t_0 = \frac{\bar{x} - \mu_0 - S}{\frac{sd}{\sqrt{n}}} \qquad (6\text{-}1)$$

In the above equation, S is assumed to be equal to "0", except for Form 2 with substantial difference.

STEP 3: Use Student's t-table (ProUCL computes them) to find the critical value $t_{n-1,\ 1-\alpha}$

Conclusion:

Form 1:

If $t_0 > t_{n-1,\alpha}$, then reject the null hypothesis that the site population mean is less than the cleanup level, μ_0

Form 2:

If $t_0 < -t_{n-1,\alpha}$, then reject the null hypothesis that the site population mean exceeds the cleanup level, μ_0

Two-Sided:

If $|t_0| > t_{n-1,\ \alpha/2}$, then reject the null hypothesis that the site population mean is same as the cleanup level, μ_0

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. Here, $t_{n-1,\alpha}$ represents the critical value from t-distribution with $(n-1)$ degrees of freedom (df) such that the area to the right of $t_{n-1,\alpha}$ under the t-distribution probability density function is α.

6.6.1.3 P-values

In addition to computing critical values (some users still like to use critical values for a specified α), ProUCL computes exact or approximate *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 set, 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 t-test can be found in any statistical text book such as Daniel (1995). ProUCL computes *p*-values for t-tests associated with each form of the null hypothesis. Specifically, 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.

6.6.1.4 Relation between One-Sample Tests and Confidence Limits of the Mean or Median

There has been some confusion among the users whether to use a LCL or a UCL of the mean to determine if the remediated site areas have met the cleanup standards. There is a direct relation between one sample hypothesis tests and confidence limits of the mean or median. For example, depending upon the hypothesis test form, a t-test is related to the upper or lower confidence limit of the mean, and a Sign test is related to the confidence limits of the median. In confirmation sampling, either a one sample hypothesis test (e.g., t-test, WSR test) or a confidence interval of the mean (e.g., LCL, UCL) can be used. Both approaches result in the same conclusion.

These relationships have been illustrated for the t-test and the LCLs and upper UCLs for normally distributed data sets. The use of a UCL95 to determine if a polluted site has attained the cleanup standard, μ_0, after remediation is very common. If a UCL95 < μ_0, then it is concluded that the site meets the standard. The conclusion based upon the UCL or LCL, or the interval (LCL, UCL) is derived from hypothesis test

statistics. For an example, while using a 95% lower confidence limit (LCL95), one is testing hypothesis test Form 1, and when using UCL95, one is testing hypothesis Form 2.

For a normally distributed data set: $x_1, x_2, \ldots, x_n$ (e.g., collected after excavation), the UCL95 and LCL95 are given as follows:

$$UCL95 = \bar{x} + t_{n-1,0.05} * sd/\sqrt{n}, \text{ and}$$

$$LCL95 = \bar{x} - t_{n-1,0.05} * sd/\sqrt{n}$$

Objective: Does the site average, μ, meet the cleanup level, μ_0?

Form 1: H_0: site $\mu \leq \mu_0$ vs. H_A: site $\mu > \mu_0$

Form 2: H_0: site $\mu \geq \mu_0$ vs. H_A: site $\mu < \mu_0$

Two-Sided: H_0: site $\mu = \mu_0$ vs. a H_A: site $\mu \neq \mu_0$.

Based upon the t-test, conclusions are:

Form 1:

If $t > t_{n-1, 0.05}$, then reject the null hypothesis in favor of the alternative hypothesis

Form 2:

If $t_0 < -t_{n-1, 0.05}$, then reject the null hypothesis in favor of the alternative hypothesis

Two-Sided:

If $|t_0| > t_{n-1, 0.025}$, then reject the null hypothesis that the site population mean is same as the cleanup level

Here $t_{n-1, 0.05}$ represents a critical value from the right tail of the t-distribution with (n-1) degrees of freedom such that area to right of $t_{n-1, 0.05}$ is 0.05.

For Form 1, we have:

Reject H_0 if $t > t_{n-1,0.05}$, that is reject the null hypothesis when

$$\bar{x} > \mu_0 + t_{n-1,0.05} * sd/\sqrt{n}$$

Equivalently reject the null hypothesis and conclude that site has not met the cleanup standard when

$$\bar{x} - t_{n-1,0.05} * sd/\sqrt{n} > \mu_0;$$ or when LCL95>cleanup goal, μ_0.

The site is concluded dirty when LCL95> μ_0.

For Form 2, we have:

Reject H_0 if $t < -t_{n-1,0.05}$, that is reject the null hypothesis when

$$\bar{x} < \mu_0 - t_{n-1,0.05} \cdot \frac{sd}{\sqrt{n}}$$

Equivalently reject the null hypothesis and conclude that site meets the cleanup standard when

$$\bar{x} < \mu_0 - t_{n-1,0.05} \cdot \frac{sd}{\sqrt{n}} < \mu_0 \qquad \text{or}$$

$$\text{UCL95} < \mu_0$$

The site is concluded clean when $UCL95 < \mu_0$.

6.6.2 The One-Sample Test for Proportions

The one-sample test for proportions represents a test for evaluating the difference between the population proportion, *P*, and a specified threshold proportion, P_0. Based upon the sampled data set and sample proportion, *p*, of exceedances of a pre-specified action level, A_0, by the *n* sample observations (e.g., onsite observations); the objective is to determine if the population proportion (of exceedances of the threshold value, A_0) exceeds the pre-specified proportion level, P_0. This proportion test is equivalent to a sign test (described next), when $P_0 = 0.5$. The Stats/Sample Sizes module of ProUCL can be used to determine the minimum sample size needed to achieve pre-specified DQOs.

6.6.2.1 Limitations and Robustness

Normal approximation to the distribution of the test statistic is applicable when both (nP_0) and $n(1-P_0)$ are at least 5. For smaller data sets, ProUCL uses the 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, provided all ND values (DLs, reporting limits) 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 (data set) of size, *n*, from a population (e.g., the site (e.g., exposure area) under investigation. Let A_0 represent a compliance limit or an action level to be met by site data. It is expected (e.g., after remediation) that the proportion of site observations exceeding the action level, A_0, is smaller than the specified proportion, P_0.

Let B = number of site values in the data set exceeding the action level, A_0. A typical observed sample value of B (based upon a data set) is denoted by b. It is noted that the random variable, B follows a binomial distribution (BD) ~ B(*n, P*) with n equal to the number of trials and *P* being the unknown population proportion (probability of success). Under the null hypothesis, the variable B follows a binomial distribution (BD) ~ B(n, P_0).

The sample proportion, p=b/n = (number 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_A: $P > P_0$

Form 2: H_0: $P \geq P_0$ vs. H_A: $P < P_0$

Two-Sided: H_0: $P = P_0$ vs. H_A: $P \neq P_0$

STEP 2: Calculate the test statistic:

$$z_0 = \frac{p+c-P_0}{\sqrt{P_0(1-P_0)/n}} \quad (6\text{-}2)$$

Where $c = \begin{cases} \frac{-0.5}{n} \; if, p > P_0 \\ \frac{0.5}{n} if, p < P_0 \end{cases}$ and $p = \frac{x(\#\ of\ site\ values > A_0)}{n}$

Here *c* is the continuity correction factor for use of the normal approximation.

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 (automatically computed by ProUCL) approximation may be used to compute the critical values (*z*-values) and *p*-values.

STEP 4: Conclusion described for the approximate test based upon the normal approximation:

Form 1: If $z_0 > z_\alpha$, 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 $z_0 < -z_\alpha$, then reject the null hypothesis that the population proportion, *P*, is more than the specified proportion, P_0.

Two-Sided: If $|z_0| > z_{\alpha/2}$, then reject the null hypothesis that the population proportion, *P*, is the same as the specified proportion, P_0.

Here, z_α represents the critical value of a standard normal variable, Z, such that area to the right of z_α under the standard normal curve is α.

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 the proportion test based upon large sample normal approximation can be found in any statistical text book

such as Daniel (1995). ProUCL computes large sample *p*-values for the proportion test associated with each form of null hypothesis.

6.6.2.3 Use of the Exact Binomial Distribution for Smaller Samples

ProUCL 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 checks for the availability of appropriate amount of data, and performs the tests using a normal approximation or the exact binomial distribution accordingly.

STEP 1: When the sample size is small (e.g., < 30), and either (nP_0), or $n\ (1 - P_0)$ is less than 5, one should use the exact BD to perform this test. ProUCL performs this test based upon the BD, when the above conditions are not satisfied. In such cases, ProUCL 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 the 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: $P(B \geq b) = \alpha$, or equivalently, $P(B < b) = (1 - \alpha)$. 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 a BD(n, P_0). The critical value, b 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: 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, both are chosen to be equal to $\alpha /2$; b_1 and b_2 are given by the probability statements: $P(B \leq b_1) \sim \alpha/2$, and $P(B > b_2) \sim \alpha/2$. B being 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: The probability statements for computing a p-value for a proportion test based upon BD can be found in Conover (1999). Using the BD, ProUCL computes *p*-values for the 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 computations. There are some variations in the literature regarding the computation of *p*-values for a proportion test based upon the exact BD. Therefore, the *p*-value computation procedure as incorporated in ProUCL is described below.

Let b be the calculated value of the binomial random variable, B under the null hypothesis. ProUCL computes the *p*-values using the following probability (Prob) statements:

Form 1: p-value = Prob(B ≥ b)

Form 2: p-value = Prob(B ≤ b)

Two-sided Alternative:

For b > (n - b): P-value = 2* Prob(B ≤ b)

For b ≤ (n - b): P-value = 2*Prob(B ≥ b)

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 *(e.g., representing the desired median value)*. Like the WSR test, the Sign test can also be used on paired data to compare the location parameters of two dependent populations. This test makes no distributional assumptions. The Sign test is used when the data are not symmetric and the sample size is small (EPA, 2006). The Stats/Sample Sizes module of ProUCL can be used to determine minimum number of observations needed to achieve pre-specified DQOs associated with the Sign test.

6.6.3.1 Limitations and Robustness

Like the Proportion test, the Sign test can also be used on data sets with NDs, provided all values reported as NDs are smaller than the cleanup level/action level, C. For data sets with NDs, the process to perform a Sign test is the same as that for data sets without NDs, provided DLs associated with all NDs are less than the cleanup level. Per EPA guidance document (2006), all NDs exceeding the action level are discarded from the computation of Sign test statistic; also all observations, detects and NDs equal to the action level are discarded from the computation of the Sign test statistic. Discarding of observations (detects and NDs) will have an impact on the power of the test (reduced power). ProUCL has the Sign test for data sets with NDs as described in USEPA (2006). However, the performance of the Sign test on data sets with NDs requires some evaluation.

6.6.3.2 Sign Test in the Presence of Nondetects

A principal requirement when applying the sign test is that the cleanup level, C, should be greater than the largest ND value; in addition all observations (detects and NDs) equal to the action level and all NDs greater than or equal to the action level are discarded from the computation of the Sign test statistic.

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, S ≥ 0 represents the substantial difference used in Form 2 hypothesis tests.

STEP 1: Let $\tilde{\mu}_X$ be the site population median.

State the following null and the alternative hypotheses:

Form 1: H_0: $\tilde{\mu}_X \leq C$ vs. H_A: $\tilde{\mu}_X > C$

Form 2: H_0: $\tilde{\mu}_X \geq C$ vs. H_A: $\tilde{\mu}_X < C$

Two-Sided: H_0: $\tilde{\mu}_X = C$ vs. H_A: $\tilde{\mu}_X \neq C$

Form 2 with substantial difference, *S*: H_o: $\tilde{\mu}_X \geq C + S$ vs. H_A: $\tilde{\mu}_X < C + S$

STEP 2: Calculate the n differences, $d_i = x_i - C$. If some of the $d_i = 0$, then reduce the sample size until all the remaining $|d_i| > 0$. This means that all observations (detects and NDs) tied at C are ignored from the computation. Compute the binomial random variable, *B* representing the number of $d_i > 0$, $i = 1,2,\ldots,n$. Note that under the null hypothesis, the binomial random variable, B follows a binomial distribution (BD) ~ BD (n, ½) where n represents the reduced sample size after discarding observations as described above. Thus, one can use the exact BD to compute the critical values and *p*-values associated with this test.

STEP 3: For n ≤ 40, ProUCL computes the exact BD based test statistic, *B*; and

For n > 40, one may use the approximate normal test statistic given by,

$$z_0 = \frac{B - \frac{n}{2} - S}{\sqrt{\frac{n}{4}}}. \qquad (6\text{-}3)$$

The substantial difference, S =0, except for Form 2 hypotheses with substantial difference.

STEP 4: For n ≤ 40, one can use the BD table as given in EPA (2006). These critical values are automatically computed by ProUCL) to calculate the critical values. For n > 40, use the normal approximation and the associated normal z critical values.

STEP 5: Conclusion when n ≤ 40 (following EPA 2006):

Form 1: If $B \geq B_{UPPER}(n, 2\alpha)$, then reject the null hypothesis that the population median is less than the cleanup level, C.

Form 2: If $B \leq B_{UPPER}(n, 2\alpha)$, then reject the null hypothesis that the population median is more than the cleanup level.

Two-Sided: If $B \geq B_{UPPER}(n, \alpha)$ or $B \leq B_{UPPER}(n, \alpha) - 1$, then reject the null hypothesis that the population median is comparable to the cleanup level, C.

Form 2 with substantial difference, *S*: If $B \leq B_{UPPER}(n, 2\alpha)$, then reject the null hypothesis that the population median is more than the cleanup level, C + substantial difference, S.

ProUCL calculates the critical values and p-values based upon the BD (n, ½) for both small samples and large samples.

Conclusion: Large Sample Approximation when n>40

Form 1: If $z_0 > z_\alpha$, then reject the null hypothesis that population median is less than the cleanup level, C.

Form 2: If $z_0 <- z_\alpha$, then reject the null hypothesis that the population median is greater than the cleanup level, C.

Two-Sided: If $|z_0| > z_{\alpha/2}$, then reject the null hypothesis that the population median is comparable to the cleanup level, C.

Form 2 with substantial difference, S: If $z_0 <- z_\alpha$, then reject the null hypothesis that the population median is more than the cleanup level, C + substantial difference, S.

Here, z_α represents the critical value of a standard normal distribution (SND) such that area to the right of z_α under the standard normal curve is α.

P-Values for One-Sample Sign Test

ProUCL 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 evaluating the difference between the location parameter (mean or median) of a population and a fixed cleanup standard such as C, with C_s representing a location value. It can also be used to compare the medians of paired populations (e.g., placebo versus treatment). Hypotheses about parameters of paired populations require that data sets of equal sizes are collected from the two populations.

6.6.4.1 Limitations and Robustness

For symmetric distributions, the WSR test appears to be more powerful than the Sign test. However, WSR test tends to yield incorrect results in the presence of many tied values. On data sets with NDs, the process to perform a WSR test is the same as that for data sets without NDs once all NDs are assigned some surrogate value. However, like the Sign test, not much guidance is available in the literature for performing WSR test on data sets consisting of ND observations. The WSR test for data sets with NDs as described in USEPA (2006) and incorporated in ProUCL requires further investigation especially when multiple DLs with NDs exceeding the detects are present in the data set.

For data sets with NDs with a single DL, DL, a surrogate value of DL/2 is used for all ND values (EPA, 2006). The presence of multiple DLs makes this test less powerful. It is suggested not to use this test when multiple DLs are present with NDs exceeding the detected values. Per EPA (2006) guidance, when multiple DLs are present, then all detects and NDs less than the largest DL may be censored which tends to reduce the power of the test. In ProUCL, all NDs including the largest ND value are replaced by half of their respective reporting limit values. All detected values are used as reported.

6.6.4.2 Wilcoxon Signed Rank (WSR) Test in the Presence of Nondetects

Following the suggestions made in the EPA guidance document (2006), ProUCL uses the following process to perform WSR test in the presence of NDs.

For left-censored data sets with a single DL (it is preferred to have all detects greater than the NDs), it is suggested (EPA, 2006) to replace all NDs by DL/2. This suggestion (EPA, 2006) has been used in the WSR test as incorporated in ProUCL software. Specifically, if there are *k* ND values with the same DL, then they are considered as "ties" and are assigned the average rank for this group.

The presence of multiple DLs makes this test less powerful. When multiple DLs are present, then all NDs are replaced by half of their respective DLs. All detects are used as reported.

6.6.4.3 Directions for the Wilcoxon Signed Rank Test

Let $x_1, x_2, \ldots, x_n$ represent a random sample of size, n collected from a site area under investigation, and C represent the cleanup level.

STEP 1: State/select one of the following null hypotheses:

Form 1: H_0: Site location ≤ C vs. H_A: Site location > C

Form 2: H_0: Site location ≥ C vs. H_A: Site location < C

Two-Sided: H_0: Site location = C vs. H_A: Site location ≠ C

Form 2 with substantial difference, S: H_0: Site location ≥ C + S vs. H_a: Site location < C + S, here S ≥ 0.

STEP 2: Calculate the deviations, $d_i = x_i - C$. If some $d_i = 0$, then reduce the sample size until all $|d_i| > 0$. That is, ignore all observations with $d_i = 0$.

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 $n > 20$, use a normal approximation and compute the test statistic given by

$$z_0 = \frac{T^+ - n(n+1)/4}{\sqrt{var(T^+)}} \qquad (6\text{-}4)$$

Here $var(T^+)$ is the variance of T^+ and is given by

$$var(T^+) = \frac{n(n+1)(2n+1)}{24} - \frac{1}{48}\sum_{j=1}^{g} t_j(t_j^2 - 1)$$; g = number of tied groups.

STEP 6: Conclusion when $n \leq 20$:

Form 1: Larger values of the test statistic, T^+, will cause the rejection of the Form 1 null hypothesis. That is if $T^+ \geq \frac{n(n+1)}{2}$- $w_\alpha = w_{(1-\alpha)}$, then reject the null hypothesis that the location parameter is less than the cleanup level, C.

Form 2: Smaller values of the test statistic will cause the rejection of the Form 2 null hypothesis. If $T^+ \leq w_\alpha$, then reject the null hypothesis that the location parameter is greater than the cleanup level, C.

Two-Sided Alternative: If $T^+ \geq \frac{n(n+1)}{2} - w_{\alpha/2}$ or $T^+ \leq w_{\alpha/2}$, then reject the null hypothesis that the location parameter is comparable to the action level, C.

Form 2 with substantial difference, S: If $T^+ \leq w_\alpha$, then reject the null hypothesis that the location parameter is more than the cleanup level, C + the substantial difference, S.

Notes: In the above, w_α represents the α^{th} quantile (lower α^{th} critical value) of the distribution of the test statistic T^+. The upper α^{th} critical value, $w_{(1-\alpha)}$ ($=(1-\alpha)^{th}$ quantile of the test statistic, T^+, as needed for the Form 1 hypothesis is given as follows:

$$P(T^+ \leq w_{1-\alpha}) = 1 - \alpha, \text{ with}$$

$$w_{1-\alpha} = n(n+1)/2 - w_\alpha$$

The lower critical values (quantiles of the test statistic, T^+) for $\alpha \leq 0.5$ are tabulated in the various statistics books (e.g., Conover, 1999; Hollander and Wolfe, 1999) and Technical Guidance document (EPA 2006b). The upper quantiles used in the Form 1 hypothesis or two-sided hypothesis are obtained using the equation described above.

Conclusion when n > 20:

Form 1: If $z_0 > z_\alpha$, then reject the null hypothesis that location parameter is less than the cleanup level, C.

Form 2: If $z_0 < -z_\alpha$, then reject the null hypothesis that the location parameter is greater than the cleanup level, C.

Two-Sided: If $|z_0| > z_{\alpha/2}$, then reject the null hypothesis that the location parameter is comparable to the cleanup level, C.

Form 2 with substantial difference, S: If $z_0 < -z_\alpha$, then reject the null hypothesis that the location parameter is more than the cleanup level, *C* + the substantial difference, S.

It should be noted that WSR can be used to compare medians (means when data are symmetric) of two correlated (paired) data sets.

Notes: The critical values, w_α as tabulated in EPA (2006b) have been programmed in ProUCL. For smaller data sets with $n \leq 20$ the p-values are computed using the BD; and for larger data sets with $n > 20$ the normal approximation is used to compute the critical values and p-values.

Example 6-1: Consider the aluminum and thallium concentrations of the real data set used in Example 2-4 of Chapter 2. Please note that the aluminum data set follows a normal distribution and the thallium data set does not follow a discernible distribution. One-sample t-test (Form 2), Proportion test (2-sided) and WRS test (Form 1) results are shown below.

Table 6-2. Single-sample t-Test, H_0: Aluminum Mean Concentration ≥10000

Date/Time of Computation	3/9/2013 8:46:40 AM
From File	SuperFund.xls
Full Precision	OFF
Confidence Coefficient	95%
Substantial Difference	0.000
Action Level	10000.000
Selected Null Hypothesis	Mean >= Action Level (Form 2)
Alternative Hypothesis	Mean < the Action Level

Aluminum

One Sample t-Test

Raw Statistics

Number of Valid Observations	24
Number of Distinct Observations	24
Minimum	1710
Maximum	16200
Mean	7789
Median	7010
SD	4264
SE of Mean	870.4

H0: Sample Mean >= 10000 (Form 2)

Test Value	-2.54
Degrees of Freedom	23
Critical Value (0.05)	-1.714
P-Value	0.00915

Conclusion with Alpha = 0.05

Reject H0, Conclude Mean < 10000

P-Value < Alpha (0.05)

Conclusion: Reject the null hypothesis and conclude that mean aluminum concentration <10000.

Table 6-3. Single-Sample Proportion Test
(*H_0*: Proportion, P, of exceedances by thallium values exceeding the action level of 0.2 is equal to 0.1, *vs.* *H_A*: Proportion of exceedances is not equal to 0.1).

Confidence Coefficient	95%
User Specified Proportion	0.100 (P0 of Exceedances of Action Level)
Action/compliance Limit	0.200
Select Null Hypothesis	Sample Proportion, P of Exceedances of Action Level = User Specified Proportion (2 Sided Alternative)
Alternative Hypothesis	Sample Proportion, P of Exceedances of Action Level <> User Specified Proportion

Thallium

One Sample Proportion Test

Raw Statistics

Number of Valid Observations	24
Number of Distinct Observations	18
Minimum	0.066
Maximum	0.456
Mean	0.147
Median	0.07
SD	0.133
SE of Mean	0.0271
Number of Exceedances	6
Sample Proportion of Exceedances	0.25

H0: Sample Proportion = 0.1

Approximate P-Value	0.0349

Conclusion with Alpha = 0.05

Reject H0. Conclude Sample Proportion <> 0.1

Conclusion: Proportion of thallium concentrations exceeding 0.2 is not equal to 0.1.

Table 6-4. Single-sample WRS Test (H_0: Median of thallium concentrations ≤0.2)

Confidence Coefficient	95%
Substantial Difference	0.000
Action Level	0.200
Selected Null Hypothesis	Mean/Median <= Action Level (Form 1)
Alternative Hypothesis	Mean/Median > the Action Level

Thallium

One Sample Wilcoxon Signed Rank Test

Raw Statistics

Number of Valid Observations	24
Number of Distinct Observations	18
Minimum	0.066
Maximum	0.456
Mean	0.147
Median	0.07
SD	0.133
SE of Mean	0.0271
Number Above Action Level	6
Number Equal Action Level	0
Number Below Action Level	18
T-plus	93
T-minus	207

H0: Sample Mean/Median <= 0.2 (Form 1)

Large Sample z-Test Statistic	-1.644
Critical Value (0.05)	1.645
P-Value	0.95

Conclusion with Alpha = 0.05

Do Not Reject H0, Conclude Mean/Median <= 0.2

P-Value > Alpha (0.05)

Conclusion: Do not reject the null hypothesis and conclude that median of thallium concentrations < 0.2.

Example 6-2: Consider the blood lead-levels data set discussed in the environmental literature (Helsel, 2013). The data set consists of several NDs. The box plot shown in Figure 6-1 suggests that median of lead concentrations is less than the action level. The WSR tests the null hypothesis: Median lead concentrations in blood ≥ action level of 0.1

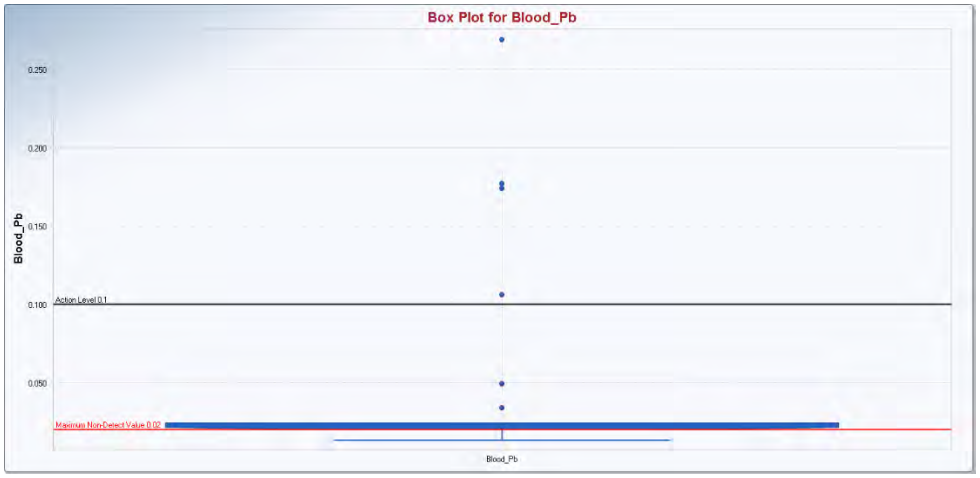

Figure 6-1. Box Plot of Lead in Blood Data Comparing Pb Concentrations with the Action Level of 0.1

Table 6-5. One-Sample Wilcoxon Signed Rank Test for Example 6-2

Blood_Pb

One Sample Wilcoxon Signed Rank Test

Raw Statistics

Number of Valid Data	27
Number of Distinct Data	13
Number of Non-Detects	19
Number of Detects	8
Percent Non-Detects	70.37%
Minimum Non-detect	0.0137
Maximum Non-detect	0.02
Minimum Detect	0.0235
Maximum Detect	0.269
Mean of Detects	0.107
Median of Detects	0.0776
SD of Detects	0.0911
Median of Processed Data used in WSR	0.01
Number Above Action Level	4
Number Equal Action Level	0
Number Below Action Level	23
T-plus	39
T-minus	339

H0: Sample Median >= 0.1 (Form 2)

Large Sample z-Test Statistic	-3.667
Critical Value (0.05)	-1.645
P-Value	1.2291E-4

Conclusion with Alpha = 0.05

Reject H0, Conclude Mean/Median < 0.1

P-Value < Alpha (0.05)

Conclusion: Both the graphical display and the WSR test suggest that median of lead concentrations in blood is less than 0.1.

6.7 Two-sample Hypotheses Testing Approaches

The use of parametric and nonparametric two-sample hypotheses testing approaches is quite common in environmental applications including site versus background comparison studies. Several of those approaches for data sets with and without ND observations have been incorporated in the ProUCL software. Additionally some graphical methods (box plots and Q-Q plots) for data sets with and without NDs are also available in ProUCL to visually compare two or more populations.

Student's two-sample t-test is used to compare the means of the two independently distributed normal populations such as the potentially impacted site area and a background reference area. Two cases arise: 1) the variances (dispersion) of the two populations are comparable, and 2) the variances of the two populations are not comparable. Generally, a t-test is robust and not sensitive to minor 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 site samples and $Y_1, Y_2, \ldots, Y_m$ represent background samples that are collected at random from the two independent populations. The validity of random samples and independence assumptions may be confirmed by reviewing the procedures described in EPA (2006b). Let $\overline{X}$ and $\overline{Y}$ represent the sample means of the two data sets. Using the GOF tests (available in ProUCL 5.2 under Statistical Tests Module), one needs to verify that the two data sets are normally distributed. If both *m* and *n* are large (and the data are *mildly* to moderately skewed), one may make this assumption without further verification (due to the CLT). If the data sets are highly skewed (skewness discussed in Chapters 3 and 4), the use of nonparametric tests such as the WMW test supplemented with graphical displays is preferable.

6.7.1.2 Limitations and Robustness

The two-sample t-test with equal variances is fairly robust to violations of the assumption of normality. However, if the investigator has tested and rejected normality or equality of variances and sample sizes are small, then nonparametric procedures such as the WMW may be applied. It is suggested that a t-test not be used on log-transformed data sets as a t-test on log-transformed data tests the equality of medians and not the equality of means. For skewed distributions there are significant differences between mean and median. The Student's t- test assumes the equality of variances of the two populations under comparison; if the two variances are not equal and the normality assumption of the means is valid, then the Satterthwaite's t-test (described below) can be used.

In the presence of NDs, it is suggested to use a Gehan test or T-W test. Sometimes, users tend to use a t-test on data sets obtained by replacing all NDs by surrogate values, such as respective DL/2 values, or DL values. The use of such methods can yield incorrect results and conclusions. *The use of substitution methods (e.g., DL/2) should be avoided.*

6.7.1.3 Guidance on Implementing the Student's Two-sample t-Test

The number of site (Population 1), *n* and background (Population 2), *m* measurements required to conduct the two-sample t-test should be calculated based upon appropriate DQO procedures (EPA [2006a, 2006b]). In case, it is not possible to use DQOs, or to collect as many samples as determined using DQOs, one may want to follow the minimum sample size requirements as described in Chapter 1. The Stats/Sample Sizes module of ProUCL can be used to determine DQOs based sample sizes. ProUCL also has an F-test to verify the equality of two variances. ProUCL automatically performs this test to verify the equality of two dispersions. The user should review the output for the equality of variances test conclusions before using one of the two tests: Student's t-test or Satterthwaite's t-test. If some measurements appear to be unusually large compared to the majority of the measurements in the data set, then a test for outliers (Chapter 7) should be conducted. Once any identified outliers have been investigated to determine if they are mistakes or errors and, if necessary, discarded, the site and background data sets should be re-tested for normality using formal GOF tests and normal Q-Q plots.

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.

6.7.1.4 Directions for the Student's Two-sample t-Test

Let $X_1, X_2, \ldots, X_n$ represent a random sample collected from a site area (Population 1) and $Y_1, Y_2, \ldots, Y_m$ represent a random data set collected from another independent population such as a background population. The two data sets are assumed to be normally distributed or mildly skewed.

STEP 1: State the following null and the alternative hypotheses:

Form 1: $H_0: \mu_x - \mu_y \leq 0$ *vs.* $H_A: \mu_x - \mu_y > 0$

Form 2: $H_0: \mu_x - \mu_y \geq 0$ *vs.* $H_A: \mu_x - \mu_y < 0$

Two-Sided: $H_0: \mu_x - \mu_y = 0$ *vs.* $H_A: \mu_x - \mu_y \neq 0$

Form 2 with substantial difference, *S*: $H_0: \mu_x - \mu_y \geq S$ *vs.* $H_A: \mu_x - \mu_y < S$

STEP 2: Calculate the sample mean $\bar{x}$ and the sample variance S_x^2 for the site (e.g., Population 1, Sample 1) data and compute the sample mean $\bar{y}$ and the sample variance S_y^2 for the background data (e.g., Population 2, Sample 2).

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 test. Calculate the pooled *sd*, S_p and the t-test statistic, t_0:

$$s_p = \sqrt{\frac{(n-1)s_x^2+(m-1)s_y^2}{(m-1)+(n-1)}} \quad (6\text{-}5)$$

$$t_0 = \frac{(\bar{x}-\bar{y})-S}{s_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \quad (6\text{-}6)$$

Here S = 0, except when used in Form 2 hypothesis with substantial difference, S ≥ 0.

STEP 4: Compute the critical value $t_{m+n-2,1-\alpha}$ such that *100(1 – α) %* of the t-distribution with *(m + n - 2) df* is below $t_{m+n-2,1-\alpha}$.

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 (comparable) 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 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 is 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 variances.

6.7.2.1 Limitations and Robustness

In the presence of NDs, replacement by a surrogate value such as the DL or DL/2gives biased results. As mentioned above, the use of these substitution methods should be avoided. Instead the use of nonparametric tests such as the Gehan test or Tarone-Ware test is suggested when the data sets consist of NDs. In cases where the assumptions of normality of means are violated, the use of nonparametric tests such as the WMW test is preferred.

6.7.2.2 Directions for the Satterthwaite Two-sample t-Test

Let $X_1, X_2, \ldots, X_n$ represent random site (Population 1) samples and $Y_1, Y_2, \ldots, Y_m$ represent random background (Population 2) samples collected from two independent populations.

STEP 1: State the following null and the alternative hypotheses:

Form 1: H_0: $\mu_x - \mu_y \leq 0$ vs. H_A: $\mu_x - \mu_y > 0$

Form 2: H_0: $\mu_x - \mu_y \geq 0$ *vs.* H_A: $\mu_x - \mu_y < 0$

Two-Sided: H_0: $\mu_x - \mu_y = 0$ *vs.* H_A: $\mu_x - \mu_y \neq 0$

Form 2 with substantial difference, *S*: H_0: $\mu_x - \mu_y \geq S$ *vs.* H_A: $\mu_x - \mu_y < S$

STEP 2: Calculate the sample mean $\bar{x}$ and the sample variance S_x^2 for the site data and compute the sample mean $\bar{y}$ and the sample variance S_y^2 for the background data.

STEP 3: Use the F-test described below (in ProUCL) to verify if the variances of the two populations are comparable. Compute the t-statistic:

$$t_0 = \frac{(\bar{x}-\bar{y})-S}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}} \qquad (6\text{-}7)$$

Here S = 0, except when used in Form 2 hypothesis with substantial difference, S ≥ 0.

STEP 4: Use a t-table (ProUCL computes them) to find the critical value $t_{1-\alpha}$ such that $100(1-\alpha)\%$ of the t-distribution with *df* degrees of freedom is below $t_{1-\alpha}$, where the Satterthwaite's Approximation for *df* is given by:

$$df = \frac{\left[\frac{s_x^2}{n}+\frac{s_y^2}{m}\right]^2}{\frac{s_x^4}{n^2(n-1)}+\frac{s_y^2}{m^2(m-1)}} \qquad (6\text{-}8)$$

STEP 5: Conclusion:

Form 1: If $t_0 > t_{df,\ 1-\alpha}$, then reject the null hypothesis that the site (Population 1) mean is less than or equal (comparable) to the background (Population 2) mean.

Form 2: If $t_0 < -t_{df,\ 1-\alpha}$, then reject the null hypothesis that the site (Population 1) mean is greater than or equal to the background (Population 2) mean.

Two-Sided: If $|t_0| > t_{df,\ 1-\alpha/2}$, then reject the null hypothesis that the site (Population 1) mean is comparable to the background (Population 2) mean.

Form 2 with substantial difference, S: If $t_0 < -t_{df,\ 1-\alpha}$, then reject the null hypothesis that the site (Population 1) mean is greater than or equal to the background (Population 2) 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. ProUCL computes (based upon an appropriate t-distribution) *p*-values for two-sample t-tests associated with each form of the 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 that verifies the assumption of the equality of two variances. This assumption is needed to perform a simple two-sample Student's t-test described above.

6.8.1 The F-Test for the Equality of Two-Variances

An F-test is used to verify whether the 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 represent independent random samples from two normal populations. The F-test for equality of variances is sensitive to departures from normality. There are other statistical tests such as the Levene's test (1960) which also tests the equality of the variances of two normally distributed populations. However, the inclusion of the Levene test will not add any new capability to the software. Therefore, taking the budget constraints into consideration, the Levene's test has not been incorporated in the ProUCL software.

Moreover, it should be noted that, although it makes sense to first determine if the two variances are equal or unequal, this is not a requirement to perform a t-test. The t-distribution based confidence interval or test for μ_1 - μ_2 based on the pooled sample variance does not perform better than the approximate confidence intervals based upon Satterthwaite's test. Hence testing for the equality of variances is not required to perform a two-sample t-test. The use of Welch-Satterthwaite's or Cochran's method is recommended in all situations (see, for example, F. Hayes [2005]).

6.8.1.1 Directions for the F-Test

Let $X_1, X_2, \ldots, X_n$ represent the *n* data points from site (Population 1) and $Y_1, Y_2, \ldots, Y_m$ represent the *m* data points from background (Population 2). To manually perform an F-test, one can 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 the F- distribution (ProUCL computes them), find a cutoff, $U = f_{1-\alpha/2}(k, q)$ associated with the F distribution with *k* and *q* degrees of freedom for some significance level, α. If the calculated F value > *U*, conclude that the variances of the two populations are not equal.

P-Values for Two-sample Dispersion Test for Equality of Variances

ProUCL 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 sets.

Example 6-3: Consider a real manganese data set collected from an upgradient well (Well 1) and two downgradient MWs (Wells 2 and 3). The side-by-side box plots comparing concentrations of the three wells are shown in Figure 6-2. The two-sample t-test comparing the manganese concentrations of the two downgradient MWs are summarized in Table 6-6.

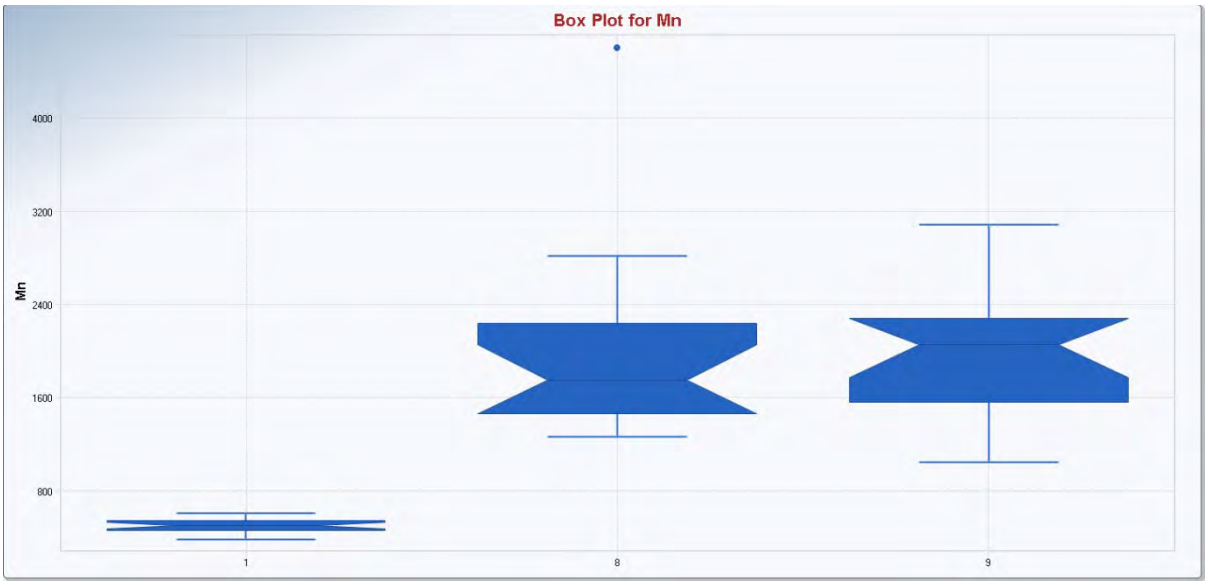

Figure 6-2. Box Plots Comparing Concentrations of Three Wells: One Upgradient and Two Downgradient

Table 6-6. T-Test Comparing Mn in MW8 vs. MW9
H_0: Mean Mn concentrations of MW 8 and MW9 are comparable

Selected Null Hypothesis	Sample 1 Mean = Sample 2 Mean (Two Sided Alternative)
Alternative Hypothesis	Sample 1 Mean <> Sample 2 Mean

Sample 1 Data: Mn-89(8)
Sample 2 Data: Mn-89(9)

Raw Statistics

	Sample 1	Sample 2
Number of Valid Observations	16	16
Number of Distinct Observations	16	15
Minimum	1270	1050
Maximum	4600	3080
Mean	1998	1968
Median	1750	2055
SD	838.8	500.2
SE of Mean	209.7	125

Sample 1 vs Sample 2 Two-Sample t-Test

H0: Mean of Sample 1 = Mean of Sample 2

		t-Test	Lower C.Val	Upper C.Val	
Method	DF	Value	t (0.025)	t (0.975)	P-Value
Pooled (Equal Variance)	30	0.123	-2.042	2.042	0.903
Welch-Satterthwaite (Unequal Varian	24.5	0.123	-2.064	2.064	0.903

Pooled SD: 690.548

Conclusion with Alpha = 0.050

Student t (Pooled): Do Not Reject H0, Conclude Sample 1 = Sample 2

Welch-Satterthwaite: Do Not Reject H0, Conclude Sample 1 = Sample 2

Test of Equality of Variances

Variance of Sample 1	703523
Variance of Sample 2	250190

Numerator DF	Denominator DF	F-Test Value	P-Value
15	15	2.812	0.054

Conclusion with Alpha = 0.05

Two variances appear to be equal

Conclusion: The variances of the two populations are comparable, both the t-test and Satterthwaite test lead to the conclusion that there are no significant differences in the mean manganese concentrations of the two downgradient monitoring wells.

6.9 Nonparametric Tests

When the data do not follow a discernible distribution, the use of parametric statistical tests may lead to inaccurate conclusions. Additionally, if the data sets contain outliers or ND values, an additional level of uncertainty is faced when conducting parametric tests. Since most environmental data sets tend to consist of observations from two or more populations including some outliers and ND values, it is unlikely that the current wide-spread use of parametric tests is justified, given that these tests may be adversely affected by outliers and by the assumptions made for handling ND values. Several nonparametric tests have been incorporated in ProUCL that can be used on data sets consisting of ND observations with single and multiple DLs.

6.9.1 The Wilcoxon-Mann-Whitney (WMW) Test

The Mann-Whitney (M-W) (or WMW) test (Bain and Engelhardt, 1992) is a nonparametric test used for determining whether a difference exists between the site and the background population distributions. This test is also known as the WRS test. The WMW test statistic 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/shapes of the two distributions are roughly the same (comparable).

6.9.1.1 Advantages and Disadvantages

The main advantage of the WMW test is that 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 WMW test statistic for large sample sizes. The WMW test may be used on data sets with NDs provided the DL or the reporting limit (RL) is the same for all NDs. If NDs with multiple DLs are present, then the largest DL is used for all ND observations. Specifically, the WMW test handles ND values by treating them as ties. Due to these constraints, other tests such as the Gehan test and theTarone-Ware test are better suited to perform two-sample tests on data sets consisting of NDs. The WMW test is more resistant to outliers than two-sample t-tests discussed earlier. It should be noted that 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 the other population (e.g., background). Like all other tests, it is suggested that the WMW test results be supplemented with graphical displays.

6.9.1.2 WMW Test in the Presence of Nondetects

If there are *t* ND values with a single DL, then they are considered as “ties” and are assigned the average rank for this group. If more than one DL is present in the data set, then WMW test censors all of the observations below the largest DL, and are treated as NDs at the largest DL. This of course results in loss of power associated with WMW test.

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 statistically and spatially 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 probability distribution of the measurements from a site area (Population 1) is similar to (e.g., including variability, shape) the probability distribution of measurements collected from a background or reference area (Population 2). The assumption of equal variances of the two regions: site vs. background should also be evaluated using descriptive statistics and graphical displays such as side-by-side box plots. The WMW test may result in an incorrect conclusion if the assumption of equality of variability is not met.

6.9.1.4 Directions for the WMW Test when the Number of Site and Background Measurements is small (n ≤ 20 or m ≤20)

Let $X_1, X_2, \ldots, X_n$ represent systematic and random site samples (Group 1, Sample 1) and $Y_1, Y_2, \ldots, Y_m$ represent systematic and random background samples (Group 2, Sample 2) collected from two independent populations. It should be noted that instead of 20, some texts suggest to use 10 as a small sample size for the two populations.

STEP 1: Let $\tilde{\mu}_x$ represent site (Population 1) median and $\tilde{\mu}_y$ represent the background (Population 2) median. State the following null and the alternative hypotheses:

Form 1: H_0: $\tilde{\mu}_x - \tilde{\mu}_y \leq 0$ *vs.* H_A: $\tilde{\mu}_x - \tilde{\mu}_y > 0$

Form 2: H_0: $\tilde{\mu}_x - \tilde{\mu}_y \geq 0$ *vs.* H_A: $\tilde{\mu}_x - \tilde{\mu}_y < 0$

Two-Sided: H_0: $\tilde{\mu}_x - \tilde{\mu}_y = 0$ *vs.* H_A: $\tilde{\mu}_x - \tilde{\mu}_y \neq 0$

Form 2 with substantial difference, S: H_0: $\tilde{\mu}_x - \tilde{\mu}_y \geq S \; vs. \; H_A$: $\tilde{\mu}_x - \tilde{\mu}_y < S$

It should be noted that when the Form 2 hypothesis is used with substantial difference, S, the value S is added to all observations in the background data set before ranking the combined data set of size (n+m) as described in the following.

STEP 2: List and rank the pooled data set of size, $N = n + m$ 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 a rank of 1 to the smallest value among the pooled data, a 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 those 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 (NDs) occur (say, < 10%), and if all such values are less than the smallest detected measurement in the pooled data set, then treat all NDs as tied values at the reported DL or at an arbitrary (when no DL is reported) 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). Today with the availability of advanced technologies and instruments, instead of reporting NDs as less-than values, NDs are typically reported at DL levels below which the instrument cannot accurately measure the concentrations present in a sample. The use of DLs is particularly helpful when NDs are reported with multiple DLs (RLs).
- If between 10% and 40% of the pooled data set are reported as NDs, and all are less than the smallest detected measurement, then one may use the approximate WMW test procedure described below provided enough (e.g., $n > 10$ and $m > 10$) data are available. However, the use of the WMW test is not recommended in the presence of multiple DLs or RLs with NDs larger than the detected values.

STEP 3: Calculate the sum of the ranks of the n site measurements. Denote this sum by W_S and then calculate the Mann-Whitney (M-W), U-statistic as follows:

$$U = W_s - n(n+1)/2 \qquad (6\text{-}9)$$

The test proposed by Wilcoxon based upon the rank sum, W_s is called the WRS test. The test based upon the U-statistic given by (6-9) was proposed by Mann and Whitney and is called the WMW test. These two tests are equivalent tests and yield the same results and conclusions. ProUCL outputs both statistics; however the conclusions are derived based upon the U-statistic and its critical and p-values. Mean and variance of the U-statistic are given as follows:

$$E(U) = nm/2$$

$$Var(U) = nm(n+m+1)/12$$

Notes: Note the difference between the definitions of U and W_s. Obviously the critical values for W_s and U are different. However, critical values for one test can be derived from the critical values of the other test by using the relationship given by the above equation (6-9). These two tests (WRS test and WMW test) are equivalent tests, and the conclusions derived by using these test statistics are equivalent. For data sets of small sizes (with m or n <20), ProUCL computes exact as well as normal distribution based approximate critical values. For large samples with n and m both greater than 20, ProUCL computes normal distribution based approximate critical values and p-values.

STEP 4: For specific values of n, m, and α, find an appropriate WMW critical value, w_α, from the table as given in EPA (2006) and also in Daniel (1995). These critical values have been incorporated in the ProUCL software.

STEP 5: Conclusion:

Form 1: If $U \geq nm - w_{\alpha}$, then reject the null hypothesis that the site population median is less than or equal to the background population median.

Form 2: If $U \leq w_{\alpha}$, then reject the null hypothesis that the site population median is greater than or equal to the background population median.

Two-Sided: If $U \geq nm - w_{\alpha/2}$ or $U \leq w_{\alpha/2,}$ then reject the null hypothesis that the site population median (location) is comparable to that of the background population median (location).

Form 2 with substantial difference, S: If $U \leq w_{\alpha}$, then reject the null hypothesis that the site population median is greater than or equal to the background population median + the substantial difference, S. 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 Samples

For small samples, ProUCL computes only approximate (as computed for large samples) *p*-values for the WMW test. Details of computing approximate *p*-values are given in the next section for larger data sets. 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.

6.9.1.5 Directions for the WMW Test when the Number of Site and Background Measurements is Large (n > 20 and m > 20)

It should be noted that some texts suggest that both *n* and *m* needs to be ≥10 to be able to use the large sample approximation. ProUCL uses large sample approximations when *n*>20 and *m*>20.

STEP 1: As before, 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*: $H_0: \tilde{\mu}_x - \tilde{\mu}_y \geq S \; vs. \; H_A: \tilde{\mu}_x - \tilde{\mu}_y < S$

Note that when the Form 2 hypothesis is used with substantial difference, S, the value S is added to all observations in the background data set before ranking the combined data set of size (n+m). For data sets with NDs, the Form 2 hypothesis test with substantial difference, S is not incorporated in ProUCL.

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. All observations tied at a give value, x_0, are assigned the average rank of the observations tied at x_0. The same process is used for all tied values.

The WMW test is not recommended when many NDs observations with multiple DLs and /or NDs exceeding the detected values are present in the data sets. Other tests such as the T-W and Gehan tests also available in ProUCL are better suited for data sets consisting of many NDs with multiple DLs and/or NDs exceeding detected values.

It should however be noted these nonparametric tests (WMW test, Gehan test, and T-W test) assume that the shape (variability) of the two data distributions (e.g., background and site) are comparable. If this assumption is not met, these tests may lead to incorrect test statistics and conclusions.

STEP 3: Calculate the sum of the ranks of the site (Population 1) measurements. Denote this sum by W_s. ProUCL computes the WMW test statistics by adjusting for tied observations using equation (6-11); that is the large sample variance of the WMW test statistic is computed using equation (6-11) which adjusts for ties.

STEP 4: When no ties are present, calculate the approximate WMW test statistic, Z_0 as follows:

$$Z_0 = \frac{W_s - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \qquad (6\text{-}10)$$

The above test statistic, Z_0 is equivalent to the following approximate Z_0 statistic based upon the Mann-Whitney U-statistic:

$$Z_0 = \frac{U - nm/2}{\sqrt{\frac{nm}{12}(n+m+1)}}$$

When ties are present in the combined data set of size $(n+m)$, the adjusted large sample approximate test value, Z_0 is computed by using the following equation:

$$Z_0 = \frac{W_s - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm}{12}\left\{(n+m+1) - \frac{\sum_{j=1}^{g} t_j(t_j^2-1)}{(n+m)(n+m-1)}\right\}}} \qquad (6\text{-}11)$$

Here g represents the number of tied groups and t_j is the number of tied values in the j^{th} group.

STEP 5: For large data sets with both n and $m \geq 20$, ProUCL computes an approximate test statistic given by equations (6-10) and (6-11) and computes a normal distribution-based p-value and critical value, z_α, where z_α is the upper $\alpha*100$ critical value of the standard normal distribution and is given by the probability statement: $P(Z > z_\alpha) = \alpha$.

STEP 6: Conclusion for Large Sample Approximations:

Form 1: If $Z_0 > z_\alpha$, 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 $Z_0 < -z_\alpha$, then reject the null hypothesis that the site population mean is greater than or equal to the background population mean.

Two-Sided: If $|Z_0| > z_{\alpha/2}$, then reject the null hypothesis that the site population mean is same as the background population mean.

Form 2 with substantial difference, S: If $Z_0 < -z_\alpha$, 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 Samples

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. Based upon the normal approximation, ProUCL computes *p*-values for each form of the null hypothesis of the 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.

Example 6-4. The data set used here can be downloaded from the ProUCL website. The data set consists of several tied observations. The test results are summarized in Table 6-7.

Table 6-7. WMW Test Comparing Location Parameters of X3 versus Y3
Null hypothesis: Location Parameter of X3 > Location Parameter of Y3

Selected Null Hypothesis	Sample 1 Mean/Median >= Sample 2 Mean/Median (Form 2)
Alternative Hypothesis	Sample 1 Mean/Median < Sample 2 Mean/Median

Sample 1 Data: X3
Sample 2 Data: Y3

Raw Statistics

	Sample 1	Sample 2
Number of Valid Observations	24	25
Number of Distinct Observations	18	19
Minimum	5.687	1.85
Maximum	31.2	79.06
Mean	17.38	39.8
Median	17.56	44.63
SD	7.421	19.39
SE of Mean	1.515	3.878

Wilcoxon-Mann-Whitney (WMW) Test

H0: Mean/Median of Sample 1 >= Mean/Median of Sample 2

Sample 1 Rank Sum W-Stat	396
Standardized WMW U-Stat	-4.093
Mean (U)	300
SD(U) - Adj ties	49.97
Approximate U-Stat Critical Value (0.05)	-1.645
P-Value (Adjusted for Ties)	2.1298E-5

Conclusion with Alpha = 0.05

Reject H0, Conclude Sample 1 < Sample 2

P-Value < alpha (0.05)

Conclusion: Based upon the WMW test results, the null hypothesis is rejected, and it is concluded that the median of X3 is significantly less than the median of Y3. This conclusion is also supported by the box plots shown in following figure.

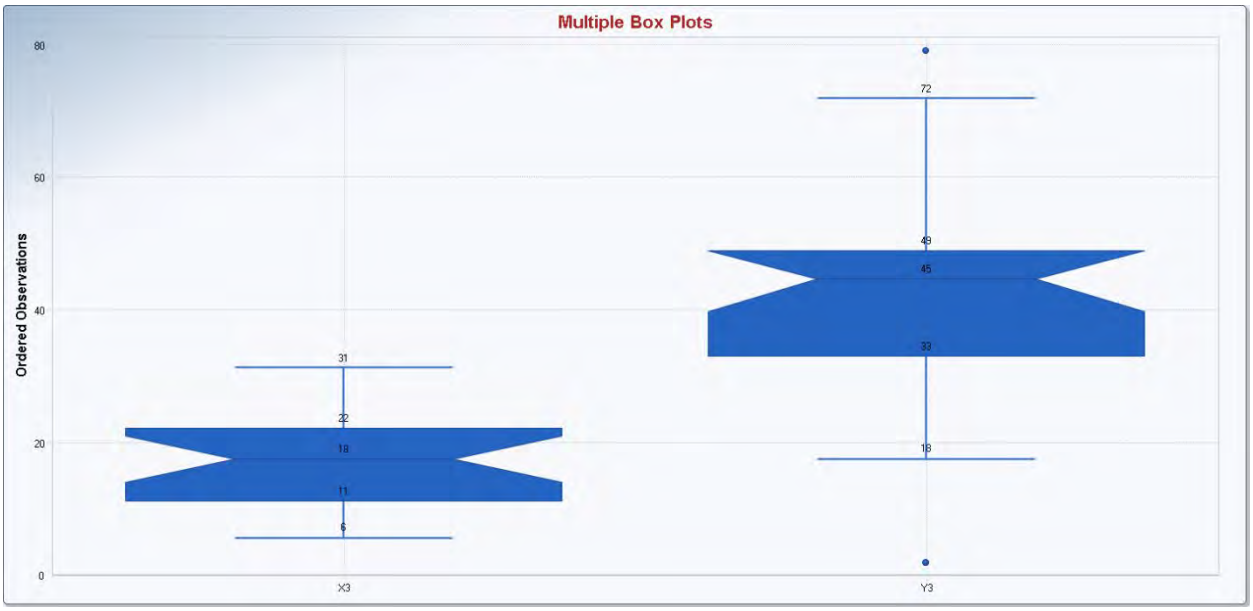

Figure 6-3. Box Plots Comparing Values of Two Groups used in Example 6-4.

Note about Quantile Test: For smaller data sets, the Quantile test as described in EPA documents ((1994, 2006 a) and Hollander and Wolfe (1999) is available in ProUCL 4.1 (see ProUCL 4.1 Technical Guide). In the past, some of the users incorrectly have used this test for larger data sets. Due to lack of resources, this test has not been expanded for data sets of all sizes. Therefore, to avoid confusion and its misuse for large data sets, the Quantile test was not included in ProUCL 5.0 and newer. Interested users may use R script to perform the Quantile test.

6.9.2 Gehan Test

The Gehan test (Gehan 1965) is one of several nonparametric tests that have been proposed to test for the differences between two populations when the data sets have multiple censoring points and DLs. Among these tests, Palachek *et al.* (1993) indicate that they selected the Gehan test primarily because: 1) it was the easiest to explain, 2) other methods (e.g., Tarone-Ware test) generally behave comparably, and 3) it reduces to the WRS test, a relatively well-known test to environmental professionals. The Gehan test as described here is available in the ProUCL software.

6.9.2.1 Limitations and Robustness

The Gehan test can be used when the background or site data sets contain many NDs with varying DLs. This test also assumes that the variabilities of the two data distributions (e.g., background vs. site, monitoring wells) are comparable.

The Gehan test is somewhat tedious to perform by hand. The use of a computer 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.

The Gehan test is used when many ND observations or multiple DLs are 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 human health and the environment. For data sets of sizes ≥ 10, the normal distribution based approximate Gehan's test statistic is described as follows.

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 and $Y_1, Y_2, \ldots, Y_m$ represent background data from the background population. Like the WMW test, this test also assumes that the variabilities of the two distributions (e.g., background vs. Site, MW1 vs. MW2) are comparable. Since we are dealing with data sets consisting of many NDs, the use of graphical methods such as the side-by-side box plots and multiple Q-Q plots is also desirable to compare the spread/variability of the two data distributions. For data sets of sizes larger than 10 (recommended), 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: $\tilde{\mu}_x - \tilde{\mu}_y \leq 0$ *vs.* H_A: $\tilde{\mu}_x - \tilde{\mu}_y > 0$

Form 2: H_0: $\tilde{\mu}_x - \tilde{\mu}_y \geq 0$ *vs.* H_A: $\tilde{\mu}_x - \tilde{\mu}_y < 0$

Two-Sided: H_0: $\tilde{\mu}_x - \tilde{\mu}_y = 0$ *vs.* H_A: $\tilde{\mu}_x - \tilde{\mu}_y \neq 0$

For data sets with NDs, the Form 2 hypothesis test with substantial difference, S is not incorporated in ProUCL. The user may want to adjust their background data sets accordingly to perform this hypothesis test form.

STEP 2: List the combined *m* background and *n* site measurements, including the ND values, from smallest to largest, where the total number of combined samples is $N = m + n$. The DLs associated with the ND (or less-than values) observations are used when listing the *N* data values 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:

$$G = \frac{\sum_{i=1}^{N} h_i \alpha(R_i)}{\left[mn \sum_{i=1}^{N} \frac{[\alpha(R_i)]^2}{N(N-1)} \right]^{1/2}} \quad (6\text{-}12)$$

Where $\begin{cases} h_i = 1 \\ h_i = 0 \end{cases}$ or

$h_i = 1$ 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 the approximate normal distribution of the *G*-statistic:

Form 1: If $G \geq z_{1-\alpha}$, then reject the null hypothesis that the site population median is less than or equal to the background population median.

Form 2: If $G \leq -z_{1-\alpha}$, then reject the null hypothesis that the site population median is greater than or equal to the background population median.

Two-Sided: If $|G| \geq z_{1-\alpha/2}$, then reject the null hypothesis that the site population median is same as the background population median.

P-Values for Two-sample Gehan Test

For the Gehan's test, *p*-values are computed using a normal approximation for the Gehan's *G*-statistic. The *p*-values can be computed using the simple procedure as used for computing large sample *p*-values for the two-sample nonparametric WMW test. ProUCL computes *p*-values for the Gehan test for each form of the 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 Tarone-Ware (T-W) Test

Like the Gehan test, the T-W test (1978) is a nonparametric test which can be used to test for the differences between the distributions of two populations (e.g., two sites, site versus background, two monitoring wells) when the data sets have multiple censoring points and DLs. The T-W test as described below has been incorporated in ProUCL 5.0 and newer. It is noted that the Gehan and T-W tests yield comparable test results.

6.9.3.1 Limitations and Robustness

The T-W test can be used when the background and/or site data sets contain multiple NDs with different DLs and NDs exceeding detected values.

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 (e.g., high DLs due to dilution effects) rather than an indication that the null hypothesis is rejected.

Like the Gehan test, the T-W test can be used when many ND observations or multiple DLs may be present in the two data sets; conclusions derived using this test may not be reliable when dealing with samples of small sizes (<10). Like the Gehan test, the T-W test described below is based upon the normal approximation of the T-W statistic and should be used when enough (e.g., $m \geq 10$ and $n \geq 10$) site and background (or monitoring well) data are available.

6.9.3.2 Directions for the Tarone-Ware Test when m ≥ 10 and n ≥ 10

Let $X_1, X_2, \ldots, X_n$ represent n data points from the site population and $Y_1, Y_2, \ldots, Y_m$ represent sample data from the background population. Like the Gehan test, this test also assumes that the variabilities of the two data distributions (e.g., background vs. site, monitoring wells) are comparable. One may use exploratory graphical methods to informally verify this assumption. Graphical displays are not affected by NDs and outlying observations.

STEP 1: Let $\tilde{\mu}_x$ represent the site and $\tilde{\mu}_y$ represent the background population medians. The following null and alternative hypotheses can be tested:

Form 1: $H_0: \tilde{\mu}_x - \tilde{\mu}_y \leq 0$ vs. $H_A: \tilde{\mu}_x - \tilde{\mu}_y > 0$

Form 2: $H_0: \tilde{\mu}_x - \tilde{\mu}_y \geq 0$ *vs.* $H_A: \tilde{\mu}_x - \tilde{\mu}_y < 0$

Two-Sided: $H_0: \tilde{\mu}_x - \tilde{\mu}_y = 0$ *vs.* $H_A: \tilde{\mu}_x - \tilde{\mu}_y \neq 0$

STEP 2: Let N denote the number of distinct detected values in the combined background and site data set of size ($n+m$) including the ND values. Arrange the N distinct detected measurements in the combined data set in ascending order from smallest to largest. Note that N will be less than $n+m$. Let $z_1 < z_2 < z_3 < \ldots < z_N$ represent N distinct ordered detected values in the data set of size, $(n+m)$.

STEP 3: Determine the N ranks, $R_1, R_2, \ldots, R_N$, for the N ordered distinct detected data values: $z_1 < z_2 < z_3 < \ldots < z_N$ in the combined data set of size ($n+m$).

STEP 4: Count the number, $n_i, i=1,2, \ldots, N$ of detects and NDs (reported as DLs or reporting limits) less than or equal to z_i in the combined data set of size ($n+m$). For each distinct detected value, z_i compute c_i = number of detects exactly equal to z_i ; $i=1,2,\ldots.N$

STEP 5: Repeat Step 4 on the site data set. That is count the number, $m_i, i=1,2,\ldots.N$ of detects and NDs (reported as DLs or reporting limits) less than or equal to z_i in site data set of size, (n). Also, for each distinct

detected value, z_i, compute d_i = number of detects in the site data set exactly equal to z_i; $i=1,2,....N$. Finally, compute, l_i ,$i=1,2,....N$, the number of detects and NDs (reported as DLs or reporting limits) less than or equal to z_i in background data set of size (m).

STEP 6: Compute the expected value and variance of detected values in the site data set of size, n, using the following equations:

$$E_{site}(\text{Detection}) = c_i \cdot \frac{m_i}{n_i} \qquad (6\text{-}13)$$

$$V_{site}(\text{Detection}) = \frac{c_i \cdot (n_i - c_i) \cdot m_i \cdot l_i}{n_i^2 \cdot (n_i - 1)} \qquad (6\text{-}14)$$

STEP 7: Compute the normal approximation of the TW test statistic using the following equation:

$$T - W = \frac{\sum_{i=1}^{N} \sqrt{n_i}(d_i - E_{Site}(Detection))}{\sqrt{\sum_{i=1}^{N}(V_{Site}(Detection))}} \qquad (6\text{-}15)$$

STEP 8: Conclusion based upon the approximate normal distribution of the T-W statistic:

Form 1: If $T\text{-}W \geq z_{1-\alpha}$, then reject the null hypothesis that the site population median is less than or equal to the background population median.

Form 2: If $T\text{-}W \leq -z_{1-\alpha}$, then reject the null hypothesis that the site population median is greater than or equal to the background population median.

Two-Sided: If $|T\text{-}W| \geq z_{1-\alpha/2}$, then reject the null hypothesis that the site population median is same as the background population median.

P-Values for Two-sample T-W Test

Critical values and p-values for the T-W test are computed following the same procedure as used for the Gehan test. ProUCL computes normal distribution based approximate critical values and p-values for the T-W test for each form of the 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 data set used in the computations.

Example 6-5. The copper (Cu) and zinc (Zn) concentrations data with NDs (from Millard and Deverel 1988) collected from groundwater of the two zones, Alluvial Fan and Basin Trough, is used to perform the Gehan and T-W tests using ProUCL 5.0. Box plots comparing Cu in the two zones are shown in Figure 6-4 and box plots comparing Zn concentrations in the two zones are shown in Figure 6-5.

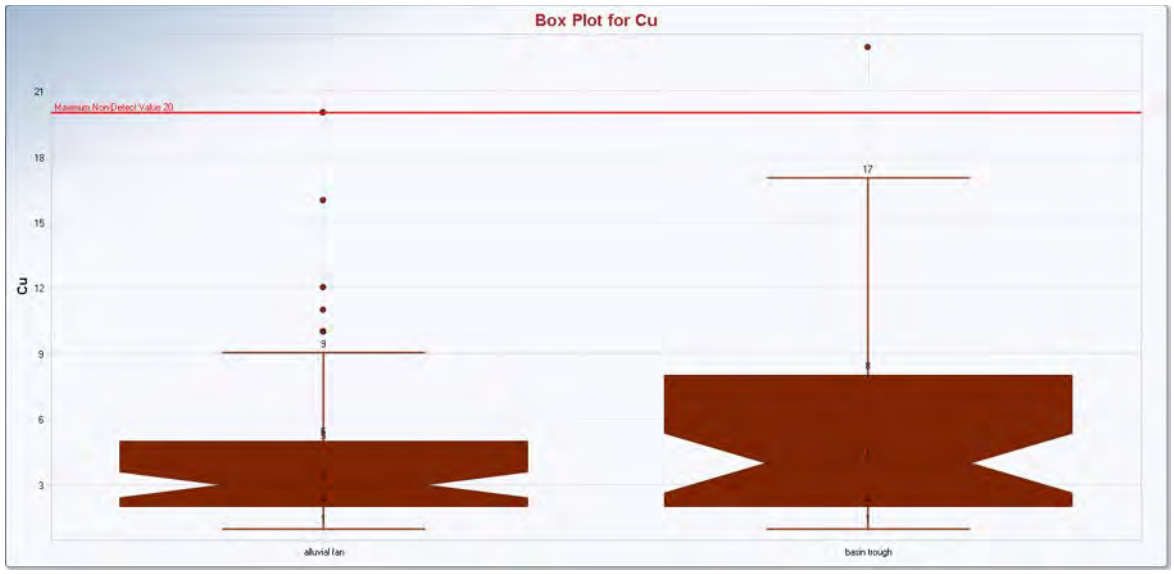

Figure 6-4. Box plots Comparing Cu in Two Zones: Alluvial Fan versus Basin Trough

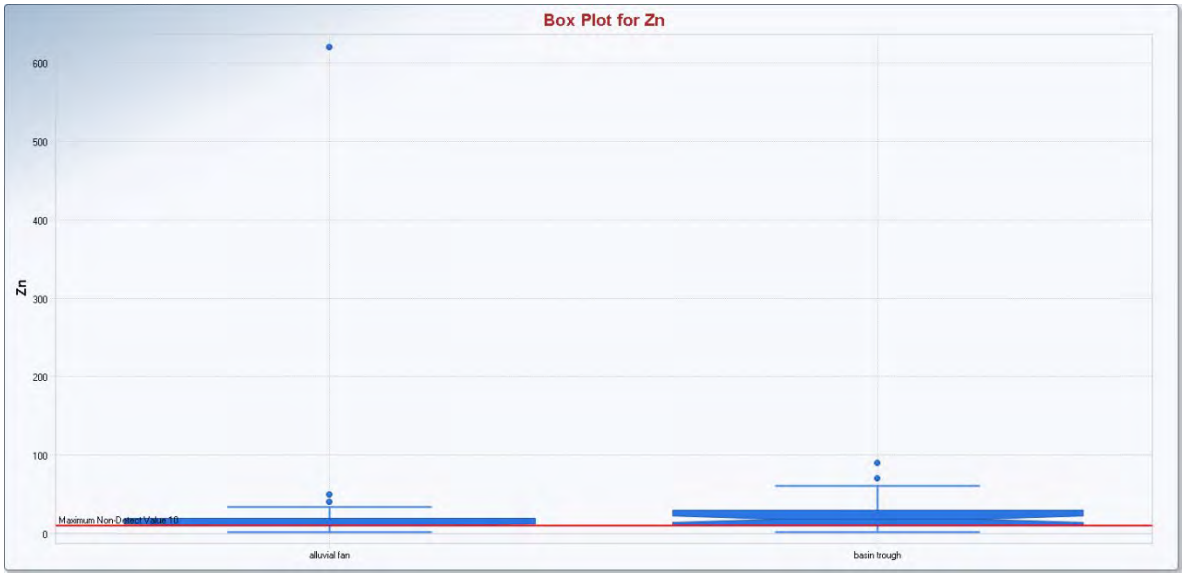

Figure 6-5. Box Plots Comparing Zn in Two Zones: Alluvial Fan versus Basin Trough

Table 6-8. Gehan Test Comparing the Location Parameters of Copper (Cu) in Two Zones
H_0: Cu concentrations in two zones, Alluvial Fan and Basin Trough, are comparable

Selected Null Hypothesis	Sample 1 Mean/Median = Sample 2 Mean/Median (Two Sided Alternative)
Alternative Hypothesis	Sample 1 Mean/Median <> Sample 2 Mean/Median

Sample 1 Data: Cu(alluvial fan)
Sample 2 Data: Cu(basin trough)

Raw Statistics

	Sample 1	Sample 2
Number of Valid Data	65	49
Number of Missing Observations	3	1
Number of Non-Detects	17	14
Number of Detect Data	48	35
Minimum Non-Detect	1	1
Maximum Non-Detect	20	15
Percent Non-detects	26.15%	28.57%
Minimum Detect	1	1
Maximum Detect	20	23
Mean of Detects	4.146	5.229
Median of Detects	2	3
SD of Detects	4.005	5.214

Sample 1 vs Sample 2 Gehan Test

H0: Mean of Sample 1 = Mean of background

Gehan z Test Value	-1.372
Lower Critical z (0.025)	-1.96
Upper Critical z (0.975)	1.96
P-Value	0.17

Conclusion with Alpha = 0.05
Do Not Reject H0, Conclude Sample 1 = Sample 2
P-Value >= alpha (0.05)

Conclusion: Based upon the box plots shown in Figure 6-3 and the Gehan test summarized in Table 6-8, the null hypothesis is not rejected, and it is concluded that the mean/median Cu concentrations in groundwater from the two zones are comparable.

Table 6-9. Tarone-Ware Comparing Location Parameters of Zinc Concentrations
H_0: Zn concentrations in groundwaters of Alluvial Fan = groundwaters of Basin Trough

Selected Null Hypothesis	Sample 1 Mean/Median = Sample 2 Mean/Median (Two Sided Alternative)
Alternative Hypothesis	Sample 1 Mean/Median <> Sample 2 Mean/Median

Sample 1 Data: Zn(alluvial fan)
Sample 2 Data: Zn(basin trough)

Raw Statistics

	Sample 1	Sample 2
Number of Valid Data	67	50
Number of Missing Observations	1	0
Number of Non-Detects	16	4
Number of Detects	51	46
Minimum Non-Detect	3	3
Maximum Non-Detect	10	10
Percent Non-detects	23.88%	8.00%
Minimum Detect	5	3
Maximum Detect	620	90
Mean of Detects	27.88	23.13
Median of Detects	11	20
SD of Detects	85.02	19.03

Sample 1 vs Sample 2 Tarone-Ware Test

H0: Mean/Median of Sample 1 = Mean/Median of Sample 2

TW Statistic	-2.113
Lower TW Critical Value(0.025)	-1.96
Upper TW Critical Value (0.975)	1.96
P-Value	0.0346

Conclusion with Alpha = 0.05

Reject H0, Conclude Sample 1 <> Sample 2

P-Value < alpha (0.05)

Conclusion: Based upon the box plots shown in Figure 6-5 and the T-W test results summarized in Table 6-9, the null hypothesis is rejected, and it is concluded that the Z_n concentrations in groundwaters of two zones are not comparable (*p*-value = 0.0346).

CHAPTER 7

Outlier Tests for Data Sets with and without Nondetect Values

Due to resource constraints, it is not possible (nor needed) to sample an entire population (e.g., reference area) of interest under investigation; only parts of the population are sampled to collect a random data set representing the population of interest. Statistical methods are then used on sampled data sets to draw conclusions about the populations under investigation. In practice, a sampled data set can consist of some wrong/incorrect values, which often result from transcription errors, data-coding errors, or instrument breakdown errors. Such wrong values could be outlying with respect to the rest of the data set; these outliers need to be fixed and corrected or when correction is not possible, removed before performing a statistical method. However, a sampled data set can also consist of some correct measurements that are extremely large or small relative to the majority of the data. If the sampling design was representative and competently executed, then these outlying measurements are truly reflective of the background population, which may indeed be skewed or even multimodal.

In practice, the boundaries of an environmental population (background) of interest may not be well-defined and the selected population actually may consist of areas (concentrations) not belonging to the dominant population of interest (e.g., reference area). Therefore, a sampled data set may consist of outlying observations coming from population(s) not belonging to the dominant background population of interest. Statistical tests based on parametric methods generally are more sensitive to the existence of outliers than are those based on nonparametric distribution-free methods. It is well-known (e.g., Rousseeuw and Leroy 1987; Barnett and Lewis 1994; Singh and Nocerino 1995) that the presence of outliers in a data set distorts the computations of all classical statistics (e.g., sample mean, *sd*, upper limits, hypotheses test statistics, GOF statistics, OLS regression estimates, covariance matrices, and also outlier test statistics themselves) of interest. Outliers also lead to both Types I and Type II errors by distorting the test statistics used for hypotheses testing. Statistics computed using a data set with outliers lack statistical power to address the objective/issue of interest (e.g., use of a BTV to identify contaminated locations). The use of such distorted statistics (e.g., two-sample tests, UCL95, UTL95-95) may lead to incorrect cleanup decisions which may not be cost-effective or protective of human health and the environment.

It is also well-known that classical outlier tests such as the Rosner Test suffer from masking effects (Huber 1981; Rousseeuw and Leroy 1987; Barnett and Lewis 1994; Singh and Nocerino 1995, and Marona, Martin, and Yohai 2006); this is especially true when outliers are present in clusters of data points and /or the data set represents multiple populations. Masking means that the presence of some outliers hides the presence of other intermediate outliers. The use of robust and resistant outlier identification methods is recommended in the presence of multiple outliers. Several modern robust outlier identification methods exist in the statistical literature cited above. However, robust outlier identification procedures are beyond the scope of the ProUCL software and this technical guidance document.

7.1 Outliers in Environmental Data Sets

In addition to representing contaminated locations, outliers in an environmental data set occur due to non-random, random and seasonal fluctuations in the environment. Outliers tests identify statistical outliers present in a data set. The variabilities of data sets originating from environmental applications are much

higher than the variabilties of data sets collected from other applications such as the biological and manufacturing processes, therefore, in environmental applications, not all outliers identified by a statistcial test may represent real physical outliers. Typically, extreme statistical outliers in a data set represent non-random situations potentially representing impacted locations; extreme outliers should not be included in statistical evaluations. Mild and intermediate statistical outliers may be present due to random natural fluctuations and variability in the environment; those outlying observations may be retained in statistical evaluations such as estimating BTVs. Based upon site CSM and expert knowledge, the project team should make these determinations.

The use of graphical displays is very helpful in distingushing between extreme statistical outliers (real physical outliers) and intermediate statistical outliers. It is suggested that outlier tests be supplemented with exploratory graphical displays such as Q-Q plots and box plots (Johnson and Wichern 2002; Hoaglin, Moseteller and Tukey 1983). ProUCL has several of these graphical methods which can be used to identify multiple outliers potentially present in a data set. Graphical displays provide additional insight into a data set that cannot be revealed by tests statistics (e.g., Rosner test, Dixon test, S-W test). Graphical displays help identify observations that are much larger or smaller than the bulk (majority) of the data. Based upon historical and current site and regional information, graphical displays, outlier test results, and investigation of suspect data, the project team and the decision makers should decide about the proper disposition of outliers to include or not to include them in the computation of the various decision-making statistics such as UCL95 and UTL95-95. Performing statistical analyses twice on the same data set, once using the full data set with outliers and once using the data set without high/extreme outliers coming from the far tails, helps the project team in determining the proper disposition of those outliers. Several examples illustrating these issues have been discussed in this technical guidance document (e.g., Chapters 2 through 5).

Note 1: In practice, extreme outliers represent: 1) nonrepresentative sampling, 2) gross measurement errors, 3) highly skewed distributions, or 4) observations coming from population(s) different from the dominant population of interest. On a normal exploratory Q-Q plot, observations well-separated (sticking out, significantly higher than the majority of the data) from the majority of observations may represent extreme physical outliers.

Note 2 (about Normality): Rosner and Dixon outlier tests require normality of a data set without the **suspected outliers.** Literature about these outlier tests is somewhat confusing and users tend to believe that the original data (with outliers) should follow a normal distribution. A data set with outliers very seldom follow a normal distribution as the presence of outliers tends to destroy the normality of a data set.

Note 3 (Outlier tests on Log-tranformed data): Statistical literature is abundant with methods applicable to normally distributed data sets. From theoretical point of view, one can use methods applicable to a normally distributed data set on log-transformed data sets following a lognormal distribution. Based upon this scenario, the use of a lognormal distribution is quite common on environmental data sets without realizing the problems and issues associated with its use (e.g., as described in Chapters 2-5 of this documents). While performing outlier tests on a background data set, in addition to accommodating contamination (extreme elevated outliers), the use of those outlier tests (e.g., Rosner test) may incorrectly identify the lower background level concentrations as outliers. Without looking into these issues carefully, some environmental documents (e.g., EPA 2009e, Helsel (2005, 2012)) suggest the use of statistical methods on log-transformed data sets. These documents suggest the use of outlier (e.g., Rosner test) tests on log-

tranformed data without realizing the pitfalls associated with its use. Based upon a real data set, an Example 7-0 illustrating these issues is provided below in Section 7.2.

Note 4: Methods incorporated in ProUCL can be used on any data set with or without NDs, and with or without the outliers. In the past, some practitioners have mis-stated that ProUCL software is restricted and can be used only on data sets without outliers. Just like any other software, it is not a requirement to exclude outliers before using any of the statistical methods incorporated in ProUCL. However, it is the intent of the developers of the ProUCL software to inform the users on how the inclusion of outliers can yield distorted UCL95; UPLs, UTLs, as well as other statistics. The outlying observations should be investigated separately to determine the reasons for their occurrences (e.g., nonrepresentative sampling, errors or contaminated locations). It is suggested that statistics be computed with and without the outliers followed by evaluation of the potential impact of outliers on the decision-making processes.

7.2 Outliers and Normality

The presence of outliers in a data set destroys the normality of the data set (Wilks 1963; Barnett and Lewis 1994; Singh and Nocerino 1995). It is highly likely that a data set which contains outliers will not follow a normal distribution unless the outliers are present in clusters. The classical outlier tests, Dixon and Rosner tests, assume that the data set *without* the suspected outliers follow a normal distribution; that is for both Rosner and Dixon tests, the data set representing the main body of the data obtained after removing the outliers, and not the original data set with outliers needs to follow a normal distribution. There appears to be some confusion among some practitioners (Helsel and Gilroy 2012) who mistakenly assume that one can perform Dixon and Rosner tests only when the data set, including outliers, follows a normal distribution, which is only rarely true.

A Q-Q plot is a more reliable guide as to whether the bulk of the data, without outliers, may follow an approximate normal distribution. Outliers are not known in advance. ProUCL has normal Q-Q plots which can be used to get an idea about the number of outliers or mixture populations potentially present in a data set. This can help a user to determine the suspected number of outliers needed to perform the Rosner test. Since the Dixon and Rosner tests may not identify all potential outliers present in a data set, the data set obtained, even without the identified outliers, may not follow a normal distribution.

The following example illustrates an issue to be careful of when applying outlier tests to log-transformed data sets.

Example 7-1. Rosner Test on Log-transformed Data Set.

A background data set for total polycyclic aromatic hydrocarbons (tPAH) from a Superfund site was used by the consultanats for the responsible party (RP) to establish BTVs. Based upon the log-transformed data, they failed to identiy potential outliers present in the upper end of the distribution and determined observations in the lower end of the data set as outliers which probably represent real background level concentrations. An exploratory Q-Q plot based upon the tPAH data set is shown in Figure 7-1. From this figure, it is noted that there are at least 3 high observations which may represent outliers. The Rosner test for 3 outliers was performed on raw and log-transformed data, those results are presented in Tables 7-1 and Table7-2. Rosner test performed on raw background data set identified higher observations: 42258, 47505,

and 55075 (Table 7-1) as outliers, whereas Rosner test performed on log-transformed data identified lower values 146, 153, and 222.1 (Table 7-2).

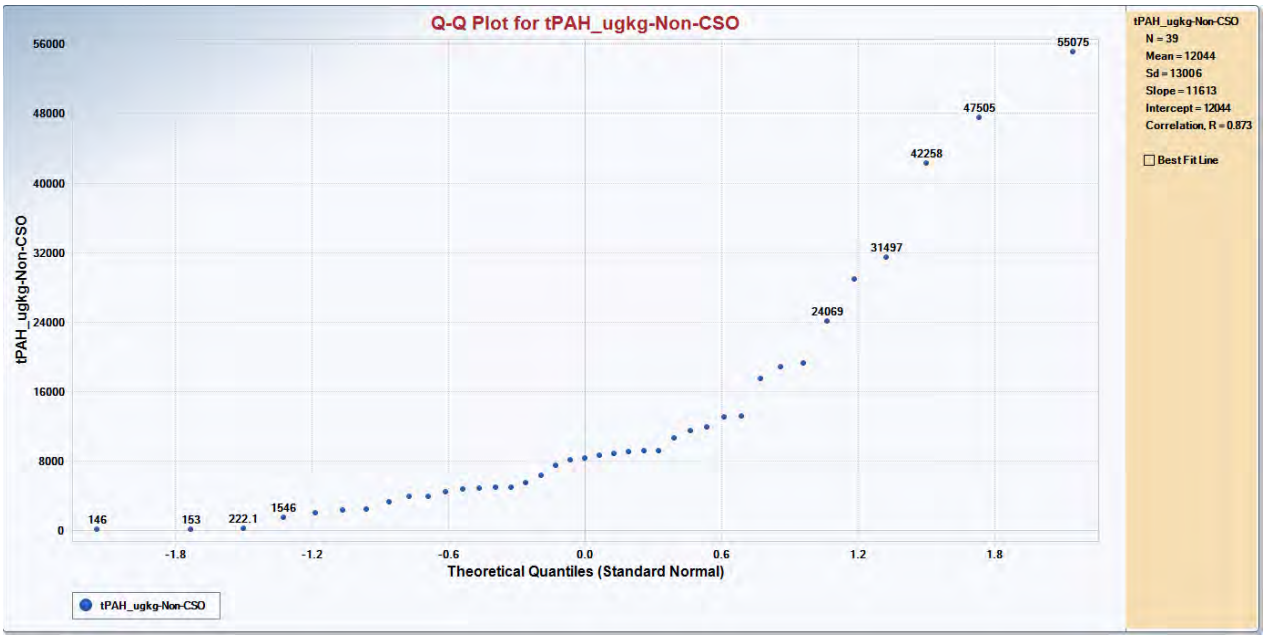

Figure 7-1. Q-Q Plot of tPAH Data

Table 7-1. Rosner Outlier Test Results on Raw tPAH Data Set

Rosner's Outlier Test for tPAH_ugkg-Non-CSO

Mean	**12044**
Standard Deviation	**13006**
Number of data	**39**
Number of suspected outliers	**3**

#	Mean	sd	Potential outlier	Obs. Number	Test value	Critical value (5%)	Critical value (1%)
1	12044	12839	55075	39	3.352	3.03	3.37
2	10912	11062	47505	38	3.308	3.01	3.36
3	9923	9358	42258	37	3.455	3	3.34

For 5% significance level, there are 3 Potential Outliers

Potential outliers are:

55075, 47505, 42258

For 1% Significance Level, there are 3 Potential Outliers

Potential outliers are:

55075, 47505, 42258

Table 7-2. Rosner Outlier Test Results on Log-transformed tPAH Data Set

Rosner's Outlier Test for ln(tPAH-ug/kg)-Non-CSO							
	Mean	8.776					
	Standard Deviation	1.368					
	Number of data	39					
	Number of suspected outliers	3					
#	Mean	sd	Potential outlier	Obs. Number	Test value	Critical value (5%)	Critical value (1%)
1	8.776	1.351	4.984	1	2.807	3.03	3.37
2	8.875	1.235	5.031	2	3.114	3.01	3.36
3	8.979	1.07	5.403	3	3.343	3	3.34
For 5% significance level, there are 3 Potential Outliers							
Potential outliers are:							
4.984, 5.031, 5.403							
For 1% Significance Level, there are 3 Potential Outliers							
Potential outliers are:							
4.984, 5.031, 5.403							

7.3 Outlier Tests for Data Sets without Nondetect Observations

A couple of classical outlier tests discussed in the environmental literature (EPA 2006b, and Gilbert 1987) and included in ProUCL software are described as follows. It is noted that these classical tests 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 (e.g., various AOCs of a site) may be present in a data set. Such scenarios can be revealed by using exploratory graphical displays including Q-Q and box plots.

7.3.1 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. Initially, this test was derived for manual computations. This test is described here for historical reasons. 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, one may want to perform a test for normality on the data without the suspected outlier. However, since the Dixon test may not identify all potential outliers present in a data set, the data set obtained after excluding the identified outliers may still not follow a normal distribution. This does not imply that the identified extreme value does not represent an outlier.

7.3.1.1 Directions for the Dixon's Test

Steps described below are provided for interested users, as ProUCL performs all of the operations described as follows:

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.

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. Since $X_{(1)}$ is the sample minimum, it need not be investigated if flagged as an outlier unless some potential error in sampling, sample handling and preservation or analysis is suspected.

STEP 4: $X_{(n)}$, the sample maximum, 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.

7.3.2 Rosner's Test

An outlier test developed by Rosner (1975, 1983) can be used to identify up to 10 outliers in data sets of sizes ≥ 25. The details of the test can be found in Gilbert (1987). Like the Dixon test, the critical values associated with the Rosner test are computed using the normal distribution of the data set *without the k (≤10) suspected outliers.* The assumption here is that the data set <u>*without*</u> the suspected outliers follows a normal distribution, as a data set with outliers tends not to follow a normal distribution. A graphical display, such as a Q-Q plot, can be used to identify suspected outliers needed to perform the Rosner test. Like the Dixon test, the Rosner test also suffers from masking.

7.3.2.1 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 *sd*.

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 *(without outliers)*. A data set with outliers seldom passes the normality test.

STEP 2: Compute the sample mean, $\bar{x}$, and the sample *sd*, 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 *sd*, 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, we have:

$$\left[\bar{x}^{(0)}, s^{(0)}, y^{(0)}\right]; \left[\bar{x}^{(1)}, s^{(1)}, y^{(1)}\right]; \ldots, \left[\bar{x}^{(r_0-1)}, s^{(r_0-1)}, y^{(r_0-1)}\right] \text{where}$$

$$\bar{x}^{(i)} = \frac{1}{n-1}\sum_{j=1}^{n-i} x_j \ , s^{(i)} = \sqrt{\frac{1}{n-i}\sum_{j=1}^{n-i}\left(x_j - \bar{x}^{(i)}\right)^2}, \text{ and } y^{(i)} \text{ is the farthest value } \bar{x}^{(i)}.$$

The above formulae for $\bar{x}^{(i)}$ and $s^{(i)}$ assume that the data have been re-numbered 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.

7.4 Outlier Tests for Data Sets with Nondetect Observations

In environmental studies, identification of detected high outliers, coming from the right tail of the data distribution and potentially representing impacted locations, is important as locations represented by those extreme high values may require further investigation. Therefore, for the purpose of the identification of high outliers, one may replace the NDs by their respective DLs, DL/2, or may just ignore them (especially when elevated DLs are associated with NDs and/or when the number of detected values is large) from any of the outlier test (e.g., Rosner test) computations, including the graphical displays such as Q-Q plots. Both of these procedures, ignoring NDs with elevated DLs or replacing them by DL/2, for identification of outliers are available in ProUCL for data sets containing NDs. Like uncensored full data sets, outlier tests on data sets with NDs should be supplemented with graphical displays. ProUCL can be used to generate Q-Q plots and box plots for data sets with ND observations.

Notes: 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 should not matter how the ND observations are treated. To compute

test statistics (e.g., Gehan test) and decision statistics (e.g., UCL95, UTL95-95), one should follow the procedures as described in Chapters 4 through 6.

Example 7-1. Consider a lead data set of size 10 collected from a Superfund site. The site data set appears to have some outliers. Since the data set is of small size, only the Dixon test can be used to identify outliers. The normal Q-Q plot of the lead data is shown in Figure 7-2 below. Figure 7-2 immediately suggests that the data set has some outliers. The Dixon test cannot directly identify all outliers present in a data set, only robust methods can identify multiple outliers. Multiple outliers may be identified one at a time iteratively by using the Dixon test on data sets after removing outliers identified in previous iterations. However, due to masking, the iterative process based upon the Dixon test may or may not be able to identify multiple outliers.

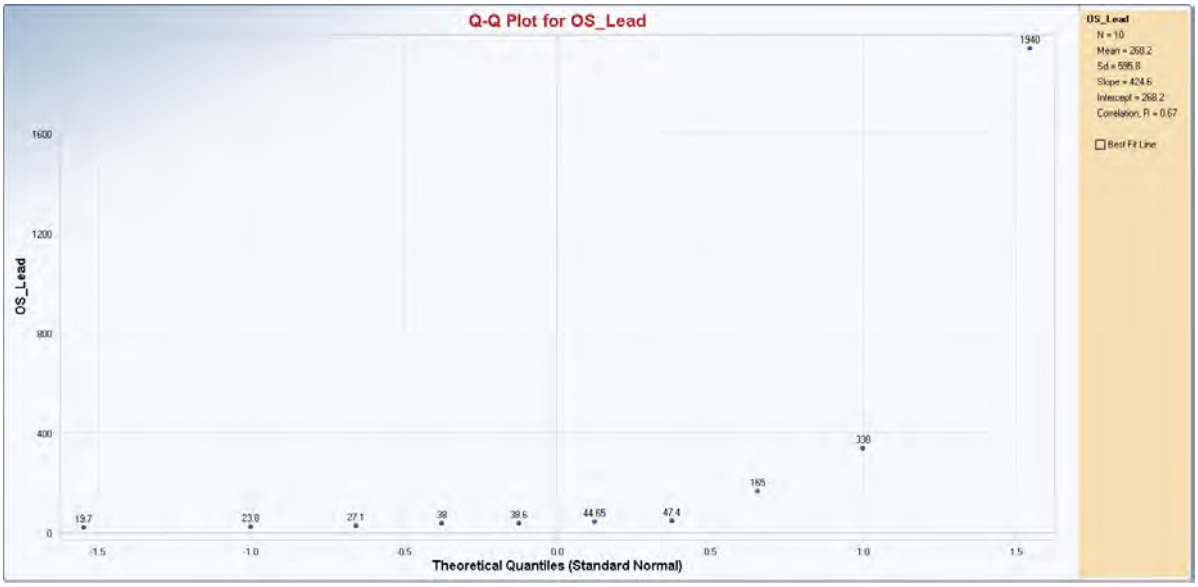

Figure 7-2. Normal Q-Q Plot Identifying Outliers

Table 7-3. Dixon Outlier Test Results for Site Lead Data Set

Dixon's Outlier Test for OS_Lead
Number of Observations = 10
10% critical value: 0.409
5% critical value: 0.477
1% critical value: 0.597
1. Observation Value 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. Observation Value 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.

Example 7-2. Consider She's (1997) pyrene data set of size n=56 with 11 NDs. The Rosner test results on data without the 11 NDs are summarized in Table 7-4, and the normal Q-Q plot without NDs is shown in Figure 7-3 below.

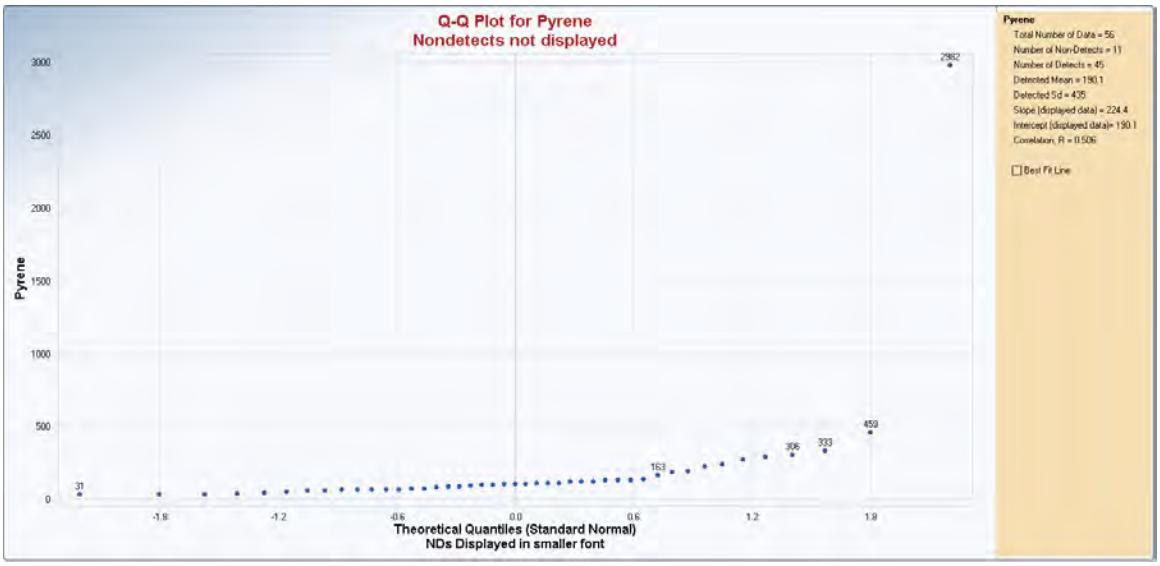

Figure 7-3. Normal Q-Q Plot of Pyrene Data Set Excluding NDs

Table 7-4. Rosner Test Results on Pyrene Data Set Excluding NDs

Rosner's Outlier Test for 10 Outliers in Pyrene

Total N	**56**
Number NDs	**11**
Number Detects	**45**
Mean of Detects	**190.1**
SD of Detects	**435**
Number of data	**45**
Number of suspected outliers	**10**
s not included in the following:	

#	Mean	sd	Potential outlier	Obs. Number	Test value	Critical value (5%)	Critical value (1%)
1	190.1	430.1	2982	45	6.491	3.09	3.44
2	126.6	90.7	459	44	3.665	3.08	3.43
3	118.9	75.7	333	43	2.828	3.07	3.41
4	113.8	68.74	306	42	2.796	3.06	3.4
5	109.1	62.43	289	41	2.881	3.05	3.39
6	104.6	56.1	273	40	3.001	3.038	3.378
7	100.3	49.65	238	39	2.773	3.026	3.366
8	96.68	44.78	222	38	2.798	3.014	3.354
9	93.3	40.17	190	37	2.408	3.002	3.342
10	90.61	37.21	187	36	2.59	2.99	3.33

For 5% significance level, there are 2 Potential Outliers

2982, 459

For 1% Significance Level, there are 2 Potential Outliers

2982, 459

Example 7-3. Consider the aluminum data set of size 28 collected from a Superfund site. The normal Q-Q plot is shown in Figure 7-4 below. Figure 7-4 suggests that there are 4 outliers (at least the

observation=30,000) present in the data set. The Rosner test results are shown in Table 7-5. Due to masking, the Rosner test could not even identify the outlying observation of 30,000.

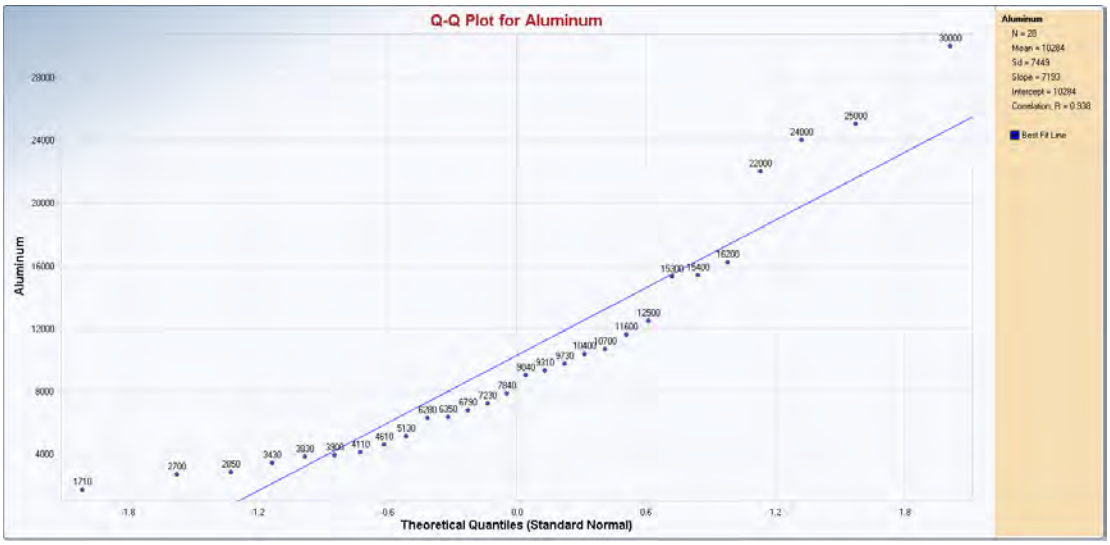

Figure 7-4. Normal Q-Q Plot of Aluminum Concentrations

Table 7-5. Rosner Test Results on Pyrene Data Set Excluding NDs

Rosner's Outlier Test for Aluminum

Mean	**10284**
Standard Deviation	**7449**
Number of data	**28**
Number of suspected outliers	**10**

#	Mean	sd	Potential outlier	Obs. Number	Test value	Critical value (5%)	Critical value (1%)
1	10284	7315	30000	26	2.695	2.88	3.2
2	9553	6490	25000	25	2.38	2.86	3.18
3	8959	5822	24000	27	2.584	2.84	3.16
4	8358	5050	22000	28	2.702	2.82	3.14
5	7789	4264	16200	10	1.973	2.8	3.11
6	7423	3956	15400	13	2.016	2.776	3.082
7	7061	3637	15300	6	2.265	2.752	3.054
8	6669	3215	12500	14	1.814	2.728	3.026
9	6377	3000	11600	21	1.741	2.704	2.998
10	6102	2812	10700	12	1.635	2.68	2.97

For 5% Significance Level, there is no Potential Outlier

For 1% Significance Level, there is no Potential Outlier

As mentioned earlier, there are robust outlier identification methods which can be used to identify multiple outliers/multiple populations present in a data set. Several of those methods are incorporated in Scout 2008 (EPA 2009d). A couple of formal (with test statistics) robust graphs based upon the PROP influence function and MCD method (Singh and Nocerino 1995) are shown in Figures 7-5 and 7-6. The details of these methods are beyond the scope of ProUCL. The two graphs suggest that there are several outliers present including the elevated value of 30,000. All observations exceeding the horizontal lines displayed at critical values of the Largest Mahalanobis Distance (MD) (Wilks 1963; Barnett and Lewis 1994) represent outliers.

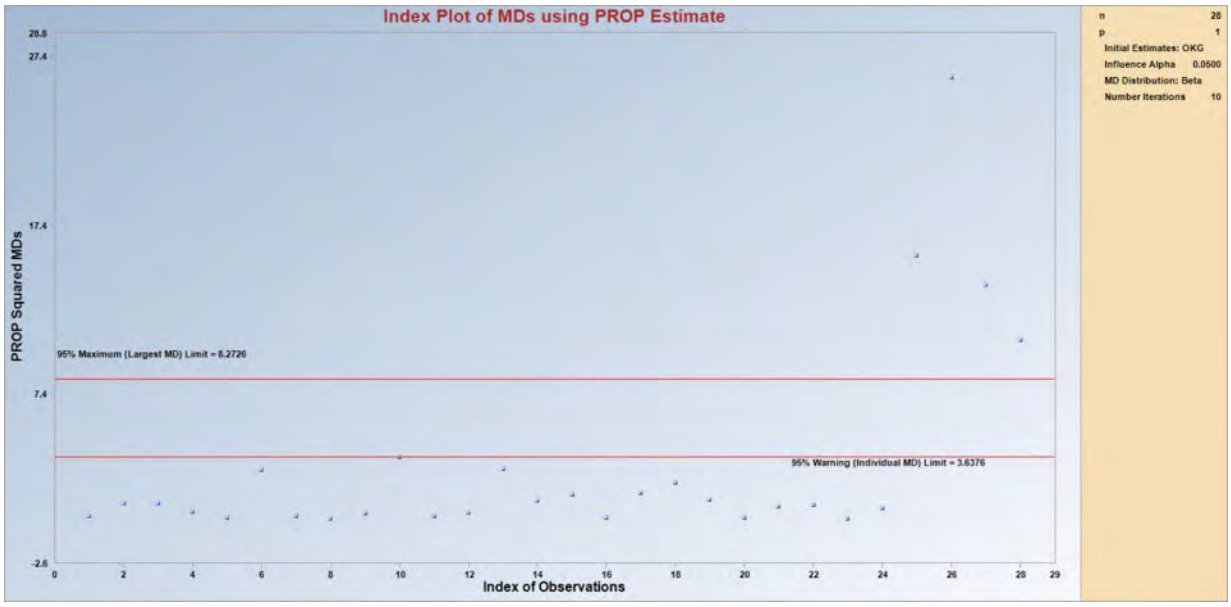

Figure 7-5. Robust Index Plot of MDs Based Upon the PROP Influence Function

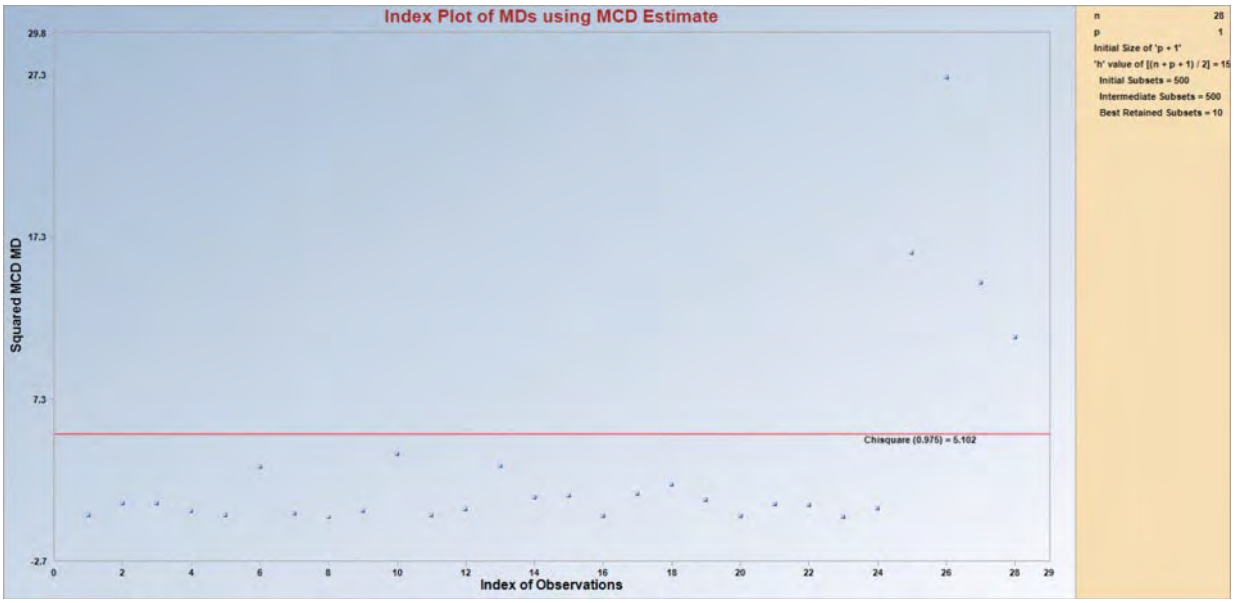

Figure 7-6. Robust Index Plot of MDs Based upon the MCD Method

CHAPTER 8

Determining Minimum Sample Sizes for User Specified Decision Parameters and Power Assessment

This chapter describes mathematical formulae used to determine data quality objectives (DQOs)-based minimum sample sizes required by estimation, and hypothesis testing approaches used to address statistical issues for environmental projects (EPA 2006a, 2006b). The sample size determination formulae for estimation of the unknown population parameters (e.g., mean, percentiles) depend upon the pre-specified values of the decision parameters: CC, *(1-α),* and the allowable error margin, Δ, between the estimate and the unknown true population parameter. For example, if the environmental problem requires the calculation of the minimum number of samples required to estimate the true unknown population mean, Δ would represent the maximum allowable difference between the estimate of the sample mean and the unknown population mean. Similarly, for hypotheses testing approaches, sample size determination formulae depend upon the pre-specified values of the decision parameters chosen while defining and describing the DQOs associated with an environmental project. The decision parameters associated with hypotheses testing approaches include the Type I false positive error rate, α; and the Type II false negative error rate, β=1-power; and the allowable width, Δ, of the gray region. For values of the parameter of interest (e.g., mean, proportion) lying in the gray region, the consequences of committing the two types of errors described in Chapter 6 are not significant from both the human health and the cost effectiveness points of view.

Even though the same symbol, Δ, has been used to denote the allowable error margin in an estimate (e.g., of mean) and the width of the gray region associated with the various hypothesis testing approaches, there are differences in the meanings of the error margin and width of the gray region. A brief description of these terminology is provided in this chapter. The user is advised to consult the already existing EPA guidance documents (EPA 2006a, 2006b; MARSSIM 2000) for the detailed description of the terms with interpretation used in this chapter. Both parametric (assuming normality) and nonparametric (distribution free) DQOs-based sample size determination formulae as described in EPA guidance documents (MARSSIM 2000; EPA 2002c, 2006a, 2006b, and 2009) are available in the ProUCL software. These formulae yield minimum sample sizes needed to perform statistical methods meeting pre-specified DQOs. The **Stats/ Sample Sizes** module of ProUCL has the minimum sample size determination methods for most of the parametric and nonparametric one-sided and two-sided hypotheses testing approaches available in ProUCL.

ProUCL includes the DQOs-based parametric minimum sample size formula to estimate the population mean, assuming that the sample mean follows a normal distribution or assuming that the criteria is met due to the CLT]. ProUCL outputs a non-negative integer as the minimum sample size. This minimum sample size is calculated by rounding the value, obtained by using a sample size formula, upward. For all sample size determination formulae incorporated in ProUCL, it is implicitly assumed that samples (e.g., soil, groundwater, sediment samples) are randomly collected from the same statistical population (e.g., AOC or MW), and therefore the sampled data (e.g., analytical results) represent independently and identically distributed (*i.i.d*) observations from a single statistical population. During the development of the **Stats/Sample Sizes** module of ProUCL, emphasis was given to assure that the module is user friendly with a straight forward unambiguous mechanism (e.g., graphics user interface [GUIs]) to input desired decision

parameters (e.g., α, β error rates, width, Δ of the gray region) needed to compute the minimum sample size for a selected statistical application.

Most of the sample size formulae available in the literature and incorporated in ProUCL) require an estimate (e.g., preliminary from other sites and pilot studies or based upon actual collected data) of the population variability. In practice, the population variance, σ^2, is unknown, and is estimated by the sample variance, s^2. During the planning stage, an estimate of the population variance is usually computed using: 1) historical information when available, 2) data collected from a pilot study when possible, or 3) information from a similar site. If historical, similar site or pilot data are not available, the minimum sample size can be computed for a range of values of the variance, and an appropriate and practical sample size from both a defensible decision making and budget point of view is selected.

New in ProUCL 5.0 and higher: The **Sample Size** module in ProUCL can be used at two different stages of a project. As mentioned above, most of the sample size formulae require some estimate of the population standard deviation (variability). Depending upon the project stage, a standard deviation: 1) represents a preliminary estimate of the population (e.g., study area) variability needed to compute the minimum sample size during the planning stage; or 2) represents the sample standard deviation computed using the data collected without considering the DQOs process, which is used to assess the power of the test based upon the collected data. During the power assessment stage, if the computed sample size is larger than the size of the already collected data set, it can be inferred that the size of the collected data set is not large enough to achieve the desired power. The formulae to compute the sample sizes during the planning stage and after performing a statistical test are the same except that the estimates of standard deviations are computed/estimated differently.

These two stages are briefly described as follows:

Planning stage before collecting data: Sample size formulae are commonly used during the planning stage of a project to determine the minimum sample sizes needed to address project objectives (estimation, hypothesis testing) with specified values of the decision parameters (e.g., Type I and II errors, width of gray region). During the planning stage, since the data are not collected *a priori*, a preliminary rough estimate of the population standard deviation, to be expected in sampled data, is obtained from other similar sites, pilot studies, or expert opinions. An estimate of the expected standard deviation along with the specified values of the other decision parameters are used to compute the minimum sample sizes needed to address the project objectives during the sampling planning stage. The project team is expected to collect the number of samples thus obtained. The detailed discussion of the sample size determination approaches during the planning stage can be found in EPA 2006a and MARSSIM 2000.

Power assessment stage after performing a statistical method: Often, in practice, environmental samples/data sets are collected without taking the DQOs process into consideration. Under this scenario, the project team performs statistical tests on the already collected data set. However, once a statistical test (e.g., WMW test) has been performed, the project team can assess the power associated with the test in retrospect. That is for specified DQOs and decision errors (Type I error and power of the test =1-Type II error) and using the sample standard deviation computed based upon the already collected data, the minimum sample size needed to perform the test for specified values of the decision parameters is computed.

If the computed sample size obtained using the sample variance is less than the size of the already collected data set used to perform the test, it may be determined that the power of the test has been achieved. However, if the sample size of the collected data is less than the minimum sample size computed in retrospect, the user may want to collect additional samples to assure that the test achieves the desired power.

It should be pointed out that there could be differences in the sample sizes computed in the two different stages due to the differences in the values of the estimated variability. Specifically, the preliminary estimate of the variance computed using information from similar sites could be significantly different from the variance computed using the available data already collected from the study area under investigation which will yield different values of the sample size.

Sample size determination methods in ProUCL can be used for both stages. The only difference will be in the input value of the standard deviation/variance. It is the users' responsibility to input a correct value for the standard deviation during the two stages.

8.1 Sample Size Determination to Estimate the Population Mean

In exposure and risk assessment studies, a UCL95 of the population mean is used to estimate the EPC term. Listed below are several variations of methods available in the literature to compute the minimum sample size, *n*, needed to estimate the population mean with specified confidence coefficient (CC), *(1 - α)*, and allowable/tolerable error margin (allowable absolute difference between the estimate and the parameter), *Δ* in an estimate of the mean.

8.1.1 Sample Size Formula to Estimate Mean without Considering Type II (β) Error Rate

The sample size can be computed using the following normal distribution based equation (when population variance is known),

$$n = \sigma^2 z^2_{1-(\alpha/2)}/\Delta^2, \tag{8-1}$$

or by using the following approximate standard normal distribution based equation (when population variance is not known),

$$n = s^2 z^2_{1-(\alpha/2)}/\Delta^2 \tag{8-2}$$

or, alternatively, by using the t- distribution based equation (when population variance is not known):

$$n = s^2 t^2_{(n-1),(1-\alpha/2)}/\Delta^2 \tag{8-3}$$

Here Δ represents the allowable error margin (±) in the mean estimate. The computed sample size assures that the sample mean will be within ± Δ units of the true population mean with probability *(1-α)*.

Throughout this chapter, z_v represents that value from a standard normal distribution (SND) for which the proportion of the distribution to the left of this value (z_v) is v*;* and $t_{(n-1),\,v}$ represents that value from a t-distribution with *(n-1)* degrees of freedom for which the proportion of the distribution to the left of this value is v.

Note: The sample size formulae described above are for estimating the population mean (and not for the median) and are based upon the underlying assumption that the distribution of the sample mean follows a normal distribution (which can be assumed due to the CLT). ProUCL does not compute minimum sample sizes required to estimate the population median. While estimating the mean, the symbol Δ represents the allowable error margin (+/-) in the mean estimate. For example for Δ = 10, the sample size is computed to assure that the error in the estimate will be within ±10 units of the true unknown population mean with specified CC of *(1-α)*.

For estimation of the mean, the most commonly used formula to compute the sample size, *n*, is given by (8-2) above; however, under normal theory, the use of t-distribution based formula (8-3) is more appropriate to compute *n*. It is noted that the difference between the sample sizes obtained using (8-2) or (8-3) is not significant. They usually differ by only 2 to 3 samples (Blackwood 1991; Singh, Singh, and Engelhardt 1999). It is a common practice to address this difference by using the following adjusted formula (Kupper and Hafner 1989; Bain and Engelhardt 1991) to compute the minimum sample size needed to estimate the mean for specified CC, *(1 - α),* and margin of error, Δ.

$$n = s^2 z^2_{1-(\alpha/2)}/\Delta^2 + z^2_{1-(\alpha/2)}/2 \qquad (8\text{-}4)$$

To be able to use a normal (instead of t-critical value) distribution based critical value, as used in (8-4), a similar adjustment factor is used in other sample size formulae described in the following sections (e.g., two-sample t-test, WRS test). This adjustment is also used in various sample size formulae described in EPA guidance documents (MARSSIM 2000; EPA 2002c, 2006a, 2006b). ProUCL uses equation (8-4) to compute sample sizes needed to estimate the population mean for specified values of CC, *(1- α)*, and error margin, Δ. An example illustrating the sample size determination to estimate the mean is given as follows.

Table 8-1. Sample Size for Estimation of the Mean (CC = *0.95*, *s* = *25*, error margin, Δ = 10)

	Sample Size for Estimation of Mean
	Based on Specified Values of Decision Parameters/DQOs (Data Quality Objectives)
Date/Time of Computation	2/26/2010 12:12:37 PM
User Selected Options	
Confidence Coefficient	95%
Allowable Error Margin	10
Estimate of Standard Deviation	25
	Approximate Minimum Sample Size
95% Confidence Coefficient:	26

8.1.2 Sample Size Formula to Estimate Mean with Consideration to Both Type I (α) and Type II (β) Error Rates

This scenario corresponds to the single-sample hypothesis testing approach. For specified decision error rates, *α* and *β*, and width, *Δ*, of the gray region, ProUCL can be used to compute the minimum sample size based upon the assumption of normality. ProUCL also has nonparametric minimum sample size determination formulae to perform Sign and WSR tests. The nonparametric Sign test and WSR test are used to perform single sample hypothesis tests for the population location parameter (mean or median).

A brief description of the standard terminology used in the sample size calculations associated with hypothesis testing approaches is described first as follows.

α = False Rejection Rate (Type I Decision Error), i.e., the probability of rejecting the null hypothesis when in fact the null hypothesis is true

β = False Acceptance Rate (Type II Decision Error), i.e., the probability of not rejecting the null hypothesis when in fact the null hypothesis is false

$z_{1-\alpha}$ = a value from a standard normal distribution for which the proportion of the distribution to the left of this value is $1 - \alpha$

$z_{1-\beta}$ = a value from a standard normal distribution for which the proportion of thedistribution to the left of this value is $1 - \beta$

Δ = width of the gray region (specified by the user); in a gray region, decisions are "too close to call", a gray region is that area where the consequences of making a decision error (Type I or Type II) are relatively minor.

The user is advised to note the difference between the gray region (associated with hypothesis testing approaches) and error margin (associated with estimation approaches).

Example illustrating the above terminology: Let the null and alternative hypotheses be: H_0: $\mu \leq C_s$, and H_A: $\mu > C_s$. The width, Δ, of the gray region for this one sided alternative hypothesis is $\Delta = \mu_1 - C_s$, where C_s is the cleanup standard specified in the null hypothesis, and μ_1 ($>C_s$) represents an alternative value belonging to the parameter value set determined by the alternative hypothesis. Note that the gray region lies to the right (e.g., see Figure 8-1) of the cleanup standard, C_s, and for all values of μ in the interval, $(C_s, \mu_1]$, with length of the interval = width of gray region= $\Delta = \mu_1 - C_s$. The consequences of making an incorrect decision (e.g., accepting the null hypothesis when in fact it is false) will be minor.

8.2 Sample Sizes for Single-Sample Tests

8.2.1 Sample Size for Single-Sample t-test (Assuming Normality)

This section describes formulae to determine the minimum number of samples, *n*, needed to conduct a single-sample t-test, for 1-sided as well as two-sided alternatives, with pre-specified decision error rates and width of the gray region. This hypothesis test is used when the objective is to determine whether the mean concentration of an AOC exceeds an action level (AL); or to verify the attainment of a cleanup standard, C_s (EPA 1989a). In the following, *s* represents an estimate (e.g., an initial guess, historical estimate, or based upon expert knowledge) of the population *sd*, σ.

Three cases/forms of hypothesis testing as incorporated in ProUCL are described as follows:

8.2.1.1 Case I (Right-Sided Alternative Hypothesis, Form 1)

H_0: site mean, $\mu \leq$ AL or a C_s vs.

H_A: site mean, μ > AL or a C_s

Gray Region: Range of the mean concentrations where the consequences of deciding that the site mean is less than the AL when in fact it is greater (that is a dirty site is declared clean) are not significant. The upper bound of the gray region, Δ, is defined as the alternative mean concentration level, μ_1 (> C_s), where the human health and environmental consequences of concluding that the site is clean (when in fact it is not clean) are relatively significant. The false acceptance error rate, β, is associated with this upper bound (μ_1) of the gray region: $\Delta=\mu_1 - C_s$. These are illustrated in Figure 8-1 below (EPA 2006a). A similar explanation of the gray region applies to other single-sample Form 1 right-sided alternative hypotheses tests (e.g., Sign test, WSR test) considered later in this chapter.

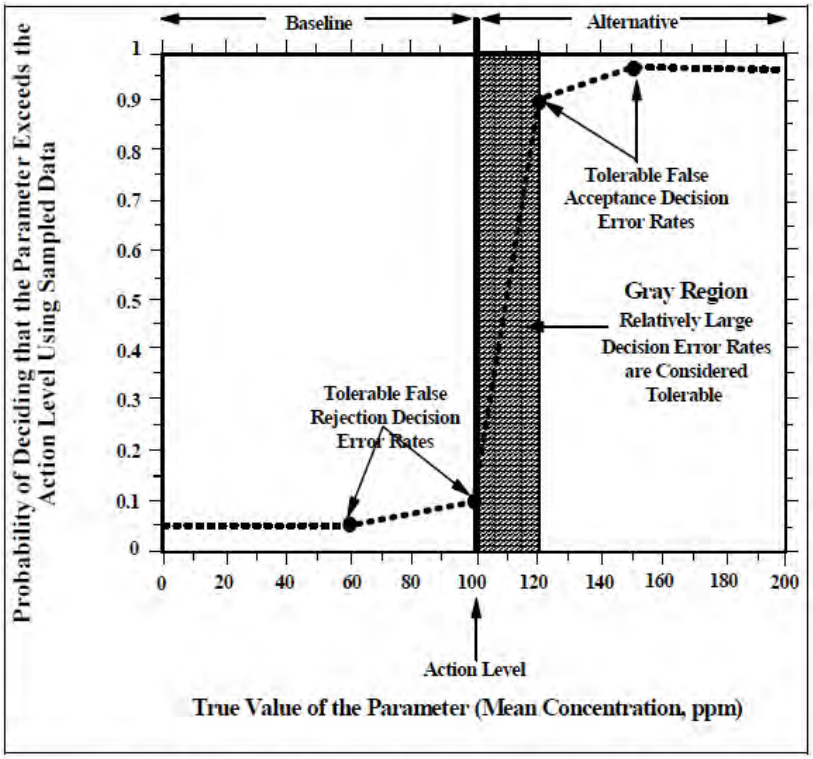

Figure 8-1. Gray Region for Right-Sided (Form 1) Alternative Hypothesis Tests (EPA 2006a)

8.2.1.2 Case II (Left-Sided Alternative Hypothesis, Form 2)

H_0: site mean, $\mu \geq$ AL or C_s vs.

H_A: site mean, μ <AL or C_s

Gray Region: Range of true mean concentrations where the consequences of deciding that the site mean is greater than or equal to the cleanup standard or action level, AL, when in fact it is smaller (that is a clean site is declared dirty) are not considered significant. The lower bound of the gray region is defined as the alternative mean concentration, μ_1 (< C_s), where the consequences of concluding that the site is dirty (when

in fact it is not dirty) would be costly requiring unnecessary cleaning of a site. The false acceptance rate, β, is associated with that lower bound (μ_1) of the gray region, $\Delta = C_s - \mu_1$. These are illustrated in Figure 8-2.

A similar explanation of the gray region applies to other single-sample left-sided (left-tailed) alternative hypotheses tests including the Sign test and WSR test.

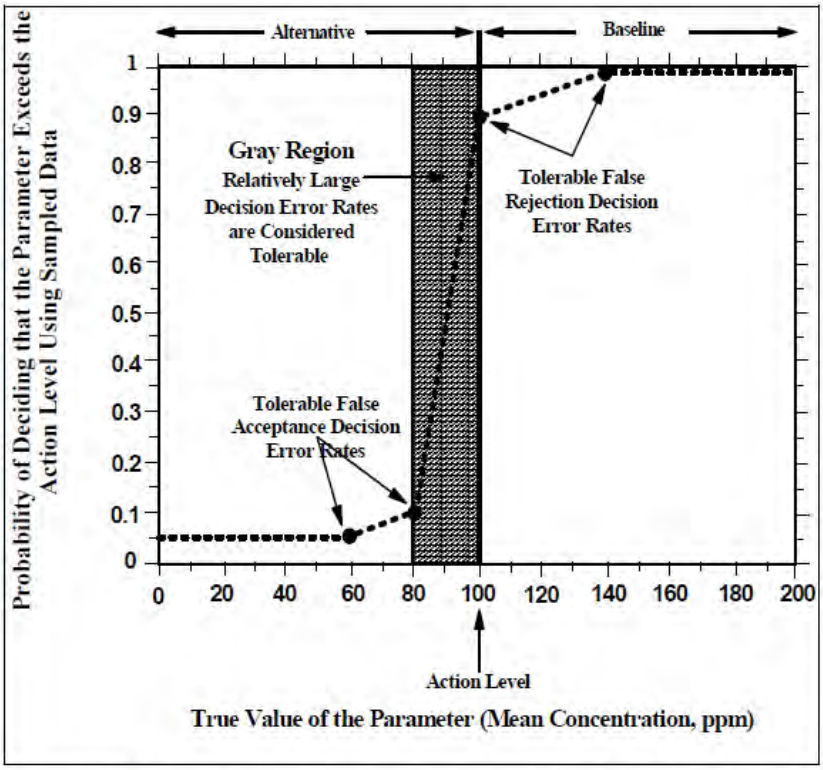

Figure 8-2. Gray Region for Left-Sided (Form 2) Alternative Hypothesis Tests (EPA 2006a)

The minimum sample size, *n*, needed to perform the single-sample one-sided t-test (both Forms 1 and 2 described above) is given by

$$n = \left(z_{1-\alpha} + z_{1-\beta}\right)^2 \left(\frac{s}{\Delta}\right)^2 + \frac{z_{1-\alpha}^2}{2} \qquad (8\text{-}5)$$

8.2.1.3 Case III (Two-Sided Alternative Hypothesis)

H_0: site mean, $\mu = C_s$; vs.

H_A: site mean, $\mu \neq C_s$

The minimum sample size for specified performance (decision) parameters is given by:

$$n = \left(z_{1-\alpha} + z_{1-\beta}\right)^2 \left(\frac{s}{\Delta}\right)^2 + \frac{z_{1-\alpha/2}^2}{2} \qquad (8\text{-}6)$$

Δ = width of the gray region, $\Delta = abs\ (C_s - \mu_1)$, *abs* represents the absolute value operation.

In this case, the gray region represents a two-sided region symmetrically placed around the mean concentration level equal to C_s, or AL; consequences of committing the two types of errors in this gray region would be minor (not significant). A similar explanation of the gray region applies to other single-sample two-sided (two-tailed) alternative hypotheses tests such as the Sign test and WSR test.

In equations (8-5) and (8-6), the computation of the estimated variance, s^2 depends upon the project stage. Specifically,

> s^2 = a preliminary estimate of the population variance (e.g., estimated from similar sites, pilot studies, expert opinions) which is used during the planning stage; or
>
> s^2 = actual sample variance of the collected data to be used when assessing the power of the test in retrospect based upon collected data.

Note: ProUCL outputs the estimated variance based upon the collected data on single sample t-test output sheet; ProUCL sample size GUI draws users' attention to input an appropriate estimate of variance, the user should input an appropriate value depending upon the project stage/data availability.

The following example: "Sample Sizes for Single-sample t-Test" discussed in *Guidance on Systematic Planning Using the Data Quality Objective Process* (EPA 2006a, page 49) is used here to illustrate the sample size determination for a single-sample t-test. For specified values of the decision parameters, the minimum number of samples is given by $n \geq 8.04$. For a one-sided alternative hypothesis, ProUCL computes the minimum sample size to be 9 (rounding up), and a sample size of 11 is computed for a two-sided alternative hypothesis.

Table 8-2. Sample Size for Single-Sample t-Test Sample Sizes (α = 0.05, β = 0.2, s = 10.41, Δ = 10)

	Sample Sizes for Single Sample t Test
	Based on Specified Values of Decision Parameters/DQOs (Data Quality Objectives)
Date/Time of Computation	2/26/2010 12:41:58 PM
User Selected Options	
False Rejection Rate [Alpha]	0.05
False Acceptance Rate [Beta]	0.2
Width of Gray Region [Delta]	10
Estimate of Standard Deviation	10.41
	Approximate Minimum Sample Size
Single Sided Alternative Hypothesis:	9
Two Sided Alternative Hypothesis:	11

8.2.2 Single Sample Proportion Test

This section describes formulae used to determine the minimum number of samples, n, needed to compare an upper percentile or proportion, P, with a specified proportion, P_0 (e.g., proportion of exceedances, proportion of defective items/drums, proportion of observations above the specified AL), for user selected

decision parameters. The details are given in EPA guidance document (2006a). Sample size formulae for three forms of the hypotheses testing approach are described as follows.

8.2.2.1 Case I (Right-Sided Alternative Hypothesis, Form 1)

H_0: population proportion ≤ specified value (P_0)vs.

H_A: population proportion > specified value (P_0)

Gray Region: Range of true proportions where the consequences of deciding that the site proportion, P, is less than the specified proportion, P_0, when in fact it is greater (that is a dirty site is declared clean) are not significant. The upper bound of the gray region, Δ, is defined as the alternative proportion, P_1 ($> P_0$), where the human health and environmental consequences of concluding that the site is clean (when in fact it is not clean) are relatively significant. The false acceptance error rate, β, is associated with this upper bound (P_1) of the gray region ($\Delta = P_1 - P_0$).

8.2.2.2 Case II (Left-Sided Alternative Hypothesis, Form 2)

H_0: population proportion ≥ specified value (P_0)vs. H_A: population proportion < specified value (P0)

Gray Region: Range of true proportions where the consequences of deciding that the site proportion, P, is greater than or equal to the specified proportion, P_0, when in fact it is smaller (a clean site is declared dirty) are not considered significant. The lower bound of the gray region is defined as the alternative proportion, P_1 ($< P_0$), where the consequences of concluding that the site is dirty (when in fact it is not dirty) would be costly requiring unnecessary cleaning of a clean site. The false acceptance rate, β, is associated with that lower bound (P_1) of the gray region ($\Delta = P_0 - P_1$).

The minimum sample size, *n*, for the single-sample proportion test (for both cases I and II) is given by

$$n = \left(\frac{z_{1-\alpha}\sqrt{P_0(1-P_0)} + z_{1-\beta}\sqrt{P_1(1-P_1)}}{P_1 - P_0} \right)^2 \qquad (8\text{-}7)$$

8.2.2.3 Case III (Two-Sided Alternative Hypothesis)

H_0: population proportion = specified value (P_0) vs.

H_A: population proportion ≠ specified value (P_0)

The following procedure is used to determine the minimum sample size needed to conduct a two-sided proportion test.

$a = \left(\frac{z_{1-\alpha/2}\sqrt{P_0(1-P_0)} + z_{1-\beta}\sqrt{P_1(1-P_1)}}{P_1 - P_0} \right)^2$ for right-sided alternative;

when $P_1 = P_0 + \Delta$; and

$b = \left(\frac{z_{1-\alpha/2}\sqrt{P_0(1-P_0)}+z_{1-\beta}\sqrt{P_1(1-P_1)}}{P_1-P_0}\right)^2$ for left-sided alternative;

when $P_1 = P_0 - \Delta$

P_0 = specified proportion

P_1 = outer bound of the gray region.

Δ = width of the gray region = $|P_0 - P_1|$=abs ($P_0 - P_1$)

The sample size, *n*, for two-sided proportion test (Case III) is given by

$$n = max(\alpha, \beta) \quad (8\text{-}8)$$

An example illustrating the single-sample proportion test is considered next. This example: "Sample Sizes for Single-sample Proportion Test" is also discussed in EPA 2006a (page 59). For this example, for the specified decision parameters, the number of samples is given by $n \geq 365$. However, ProUCL computes the sample size to be 419 for the right-sided alternative hypothesis, 368 for the left-sided alternative hypothesis, and 528 for the two-sided alternative hypothesis.

Table 8-3. Output for Single-Sample Proportion Test Sample Size (α = 0.05, β = 0.2, P_0 = 0.2, Δ = 0.05)

	Sample Sizes for Single Sample Proportion Test
	Based on Specified Values of Decision Parameters/DQOs (Data Quality Objectives)
Date/Time of Computation	2/26/2010 12:50:52 PM
User Selected Options	
False Rejection Rate [Alpha]	0.05
False Acceptance Rate [Beta]	0.2
Width of Gray Region [Delta]	0.05
Proportion/Action Level [P0]	0.2
	Approximate Minimum Sample Size
Right Sided Alternative Hypothesis:	419
Left Sided Alternative Hypothesis:	368
Two Sided Alternative Hypothesis:	max(471, 528)

Notes: The correct use of the **Sample Size** module, to determine the minimum sample size needed to perform a proportion test, requires that the users have some familiarity with the single-sample hypothesis test for proportions. Specifically the user should input feasible values for the specified proportion, P_0, and width, Δ, of the gray region. The following example shows the output screen when unfeasible values are selected for these parameters.

Table 8-4. Output - Single-sample Proportion Test Sample Sizes (α = 0.05, β = 0.2, P_0 = 0.7, Δ = 0.8)

	Sample Sizes for Single Sample Proportion Test
	Based on Specified Values of Decision Parameters/DQOs (Data Quality Objectives)
Date/Time of Computation	2/26/2010 12:55:51 PM
User Selected Options	
False Rejection Rate [Alpha]	0.05
False Acceptance Rate [Beta]	0.2
Width of Gray Region [Delta]	0.8
Proportion/Action Level [P0]	0.7
	Approximate Minimum Sample Size
Right Sided Alternative Hypothesis:	Not Feasible - Please check your Decision Parameters/DQOs
Left Sided Alternative Hypothesis:	Not Feasible - Please check your Decision Parameters/DQOs
Two Sided Alternative Hypothesis:	Not Feasible - Please check your Decision Parameters/DQOs

8.2.3 Nonparametric Single-sample Sign Test (does not require normality)

The purpose of the single-sample nonparametric Sign test is to test a hypothesis involving the true location parameter (mean or median) of a population against an AL or C_s without assuming normality of the underlying population. The details of sample size determinations for nonparametric tests can be found in Conover (1999).

8.2.3.1 Case I (Right-Sided Alternative Hypothesis)

H_0: population location parameter ≤ specified value, C_s vs.

H_A: population location parameter > specified value, C_s

A description of the gray region associated with the right-sided Sign test is given in Section 8.2.1.1.

8.2.3.2 Case II (Left-Sided Alternative Hypothesis)

H_0: population location parameter ≥ specified value, C_s vs.

H_A: population location parameter < specified value, C_s

A description of the gray region associated with this left-sided Sign test is given in Section 8.2.1.2.

The minimum sample size, *n*, for the single-sample one-sided (both left-sided and right-sided) Sign test is given by the following equation:

$$n = \frac{\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{4(Sign\,P-0.5)^2}, \text{ where} \qquad (8\text{-}9)$$

$$Sign\,P = \Phi\left(\frac{\Delta}{sd}\right) \qquad (8\text{-}10)$$

Δ = width of the gray region

sd = an estimate of the population (e.g., reference area, AOC, survey unit) standard deviation

Some guidance on the selection of an estimate of the population *sd*, σ, is given in Section 8.1.1 above.

$\Phi(x)$ = Cumulative probability distribution representing the probability that a standard normal variate, Z, takes on a value $\leq x$.

8.2.3.3 Case III (Two-Sided Alternative Hypothesis)

H_0: population location parameter = specified value, C_s vs. H_A: population location parameter $\neq$ specified value, C_s

A description of the gray region associated with the two-sided Sign test can be found in Section 8.1.2.3.

The minimum sample size, *n*, for a two-sided Sign test is given by the following equation:

$$n = \frac{\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2}{4(Sign\ P-0.5)^2}$$

In the following example, ProUCL computes the sample size to be 35 for a single-sided alternative hypothesis and 43 for a two-sided alternative hypothesis for default values of the decision parameters.

Note: Like the parametric t-test, the computation of the standard deviation (*sd*) depends upon the project stage. Specifically,

sd^2 (used to compute P in equation (8-10)) = a preliminary estimate of the population variance (e.g., estimated from similar sites, pilot studies, expert opinion) which is used during the planning stage; and

sd^2 (*used to compute P*) = sample variance computed using the actual collected data to be used when assessing the power of the test in retrospect based upon the collected data.

ProUCL outputs the sample variance based upon the collected data on the Sign test output sheet; and ProUCL sample size GUI draws user's attention to input an appropriate estimate, *sd^2*, the user should input an appropriate value depending upon the project stage/data availability.

Table 8-5. Output for Single-Sample Sign Test Sample Sizes (α = 0.05, β = 0.1, sd = 3, Δ = 2)

	Sample Sizes for Single Sample Sign Test
	Based on Specified Values of Decision Parameters/DQOs (Data Quality Objectives)
Date/Time of Computation	2/26/2010 12:15:27 PM
User Selected Options	
False Rejection Rate [Alpha]	0.05
False Acceptance Rate [Beta]	0.1
Width of Gray Region [Delta]	2
Estimate of Standard Deviation	3
	Approximate Minimum Sample Size
Single Sided Alternative Hypothesis:	35
Two Sided Alternative Hypothesis:	43

8.2.4 Nonparametric Single Sample Wilcoxon Sign Rank (WSR) Test

The purpose of the single WSR test is similar to that of the Sign test described above. This test is used to compare the true location parameter (mean or median) of a population against an AL or C_s without assuming normality of the underlying population. The details of this test can be found in Conover (1999) and EPA (2006a).

8.2.4.1 Case I (Right-Sided Alternative Hypothesis)

H_0: population location parameter ≤ specified value, C_s vs.

H_A: population location parameter > specified value, C_s

A description of the gray region associated with this right-sided test is given in Section 8.1.2.1.

8.2.4.2 Case II (Left-Sided Alternative Hypothesis)

H_0: population location parameter ≥ specified value, C_s vs.

H_A: population location parameter < specified value, C_s

A description of the gray region associated with this left-sided (left-tailed) test is given in Section 8.1.2.2.

The minimum sample size, n, needed to perform the single-sample one-sided (both left-sided and right-sided) WSR test is given as follows.

$$n = 1.16\left(\frac{sd^2(z_{1-\alpha}+z_{1-\beta})^2}{\Delta^2} + \frac{z_{1-\alpha}^2}{2}\right) \quad (8\text{-}11)$$

Where:

sd^2 = a preliminary estimate of the population variance which is used during the planning stage; and

sd^2 = actual sample variance computed using the collected data to be used when assessing the power of the test in retrospect based upon collected data

Note: ProUCL sample size GUI draws user's attention to input an appropriate estimate, sd^2; the user should input an appropriate value depending upon the project stage/data availability.

8.2.4.3 Case III (Two-Sided Alternative Hypothesis)

H_0: population location parameter = specified value, C_s vs.

H_A: population location parameter ≠ specified value, C_s

A description of the gray region associated with the two-sided WSR test is given in Section 8.1.2.3.

The sample size, *n*, needed to perform the single-sample two-sided WSR test is given by:

$$n = 1.16\left(\frac{sd^2\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2}{\Delta^2}+\frac{z_{1-\alpha/2}^2}{2}\right) \quad (8\text{-}12)$$

Where:

sd^2 = a preliminary estimate of the population variance (e.g., estimated from similar sites) which is used during the planning stage; and

sd^2 = sample variance computed using actual collected data to be used to assess the power of the test in retrospect.

Note: ProUCL sample size GUI draws user's attention to input an appropriate estimate, sd^2, the user should input an appropriate value depending upon the project stage/data availability.

The following example: "Sample Sizes for Single-sample Wilcoxon Signed Rank Test" is discussed in the EPA 2006a (page 65). ProUCL computes the sample size to be 10 for a one-sided alternative hypothesis, and 14 for a two-sided alternative hypothesis.

Table 8-6. Output for Single-sample WSR Test Sample Sizes (α = 0.1, β = 0.2, sd = 130, Δ = 100)

	Sample Sizes for Single Sample Wilcoxon Signed Rank Test
	Based on Specified Values of Decision Parameters/DQOs (Data Quality Objectives)
Date/Time of Computation	2/26/2010 1:13:58 PM
User Selected Options	
False Rejection Rate [Alpha]	0.1
False Acceptance Rate [Beta]	0.2
Width of Gray Region [Delta]	100
Estimate of Standard Deviation	130

	Approximate Minimum Sample Size
Single Sided Alternative Hypothesis:	10
Two Sided Alternative Hypothesis:	14

8.3 Sample Sizes for Two-Sample Tests for Independent Sample

This section describes minimum sample size determination formulae needed to compute sample sizes (same number of samples (n=m) from two populations) to compare the location parameters of two populations (e.g., reference area vs. survey unit, two AOC, two MW) for specified values of the decision parameters. ProUCL computes sample sizes for one-sided as well as two-sided alternative hypotheses. The sample size formulae described in this section assume that samples are collected following the simple random or systematic random sampling (e.g., EPA 2006a) approaches. It is also assumed that samples are collected randomly from two independently distributed populations (e.g., two different uncorrelated AOCs); and samples (analytical results) collected from each of population represent independently and identically distributed observations from their respective populations.

8.3.1 Parametric Two-sample t-test (Assuming Normality)

The details of the two-sample t-test can be found in Chapter 6 of this ProUCL Technical Guide.

8.3.1.1 Case I (Right-Sided Alternative Hypothesis)

H_0: site mean, $\mu_1 \leq$ background mean, μ_2 vs.

H_A: site mean, $\mu_1 >$ background mean, μ_2

Gray Region: Range of true concentrations where the consequences of deciding the site mean is less than or equal to the background mean (when in fact it is greater), that is, a dirty site is declared clean, are relatively minor. The upper bound of the gray region is defined as the alternative site mean concentration level, μ_1 ($> \mu_2$), where the human health, and environmental consequences of concluding that the site is clean (or comparable to background) are relatively significant. The false acceptance rate, β, is associated with the upper bound of the gray region, Δ.

8.3.1.2 Case II (Left-Sided Alternative Hypothesis)

H_0: site mean, $\mu_1 \geq$ background mean, μ_2 vs.

H_A: site mean, $\mu_1 <$ background mean, μ_2

Gray Region: Range of true mean values where consequences of deciding the site mean is greater than or equal to the background mean (when in fact it is smaller); that is, a clean site is declared a dirty site, are considered relatively minor. The lower bound, μ_1 ($< \mu_2$) of the gray region, is defined as the concentration where consequences of concluding that the site is dirty would be too costly, potentially requiring unnecessary cleanup. The false acceptance rate is associated with the lower bound of the gray region.

The minimum sample sizes (equal sample sizes for both populations) for the two-sample one-sided t-test (both cases I and II described above) are given by:

$$m = n = 2\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \left(\frac{s_p}{\Delta}\right)^2 + \frac{z_{1-\alpha}^2}{4} \qquad (8\text{-}13)$$

The decision parameters used in equations (8-13) and (8-14) have been defined earlier in Section 8.1.1.2.

Δ = width (e.g., difference between two means) of the gray region

S_p = a preliminary estimate of the common population standard deviation, σ, of the two populations (discussed in Chapter 6). Some guidance on the selection of an estimate of the population *sd*, σ, is given above in Section 8.1.2.

S_p = pooled standard deviation computed using the actual collected data to be used when assessing the power of the test in retrospect.

8.3.1.3 Case III (Two-Sided Alternative Hypothesis)

H_0: site mean, $\mu_1 =$ background mean, μ_2 vs.

H_A: site mean, $\mu_1 \neq$ background mean, μ_2

The minimum sample sizes for specified decision parameters are given by:

$$m = n = 2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(\frac{s_p}{\Delta}\right)^2 + \frac{z_{1-\alpha/2}^2}{4} \qquad (8\text{-}14)$$

The following example: "Sample Sizes for Two-sample t Test" is discussed in the EPA 2006a guidance document (page 68). According to this example, for the specified decision parameters, the minimum number of samples from each population comes out to be $m = n \geq 4.94$. ProUCL computes minimum sample sizes for the two populations to be 5 (rounding up) for the single sided alternative hypotheses and 7 for the two-sided alternative hypothesis.

Note: S_p represents the pooled estimate of the populations under comparison. During the planning stage, the user inputs a preliminary estimate of variance while computing the minimum sample sizes; and while

assessing the power associated with the t-test, the user inputs the pooled standard deviation, S_p, computed using the actual collected data.

> S_p = a preliminary estimate of the common population standard deviation (e.g., estimated from similar sites, pilot studies, expert opinion) which is used during the planning stage; and
>
> S_p = pooled standard deviation computed using the collected data to be used when assessing the power of the test in retrospect.

ProUCL outputs the pooled standard deviation, S_p, based upon the collected data on the two sample t-test output sheet; ProUCL sample size GUI draws user's attention to input an appropriate estimate of the standard deviation, the user should input an appropriate value depending upon the project stage/data availability.

Table 8-7. Output for Two-Sample t-Test Sample Sizes (α = 0.05, β = 0.2, s_p = 1.467, Δ = 2.5)

	Sample Sizes for Two Sample t Test
	Based on Specified Values of Decision Parameters/DQOs (Data Quality Objectives)
Date/Time of Computation	2/26/2010 1:17:57 PM
User Selected Options	
False Rejection Rate [Alpha]	0.05
False Acceptance Rate [Beta]	0.2
Width of Gray Region [Delta]	2.5
Estimate of Pooled SD	1.467
	Approximate Minimum Sample Size
Single Sided Alternative Hypothesis:	5
Two Sided Alternative Hypothesis:	7

8.3.2 Wilcoxon-Mann-Whitney (WMW) Test (Nonparametric Test)

The details of the two-sample nonparametric WMW can be found in Chapter 6; this test is also known as the two-sample WRS test.

8.3.2.1 Case I (Right-Sided Alternative Hypothesis)

H_0: site median ≤ background median vs.

H_A: site median > background median

The gray region for the WMW Right-Sided alternative hypothesis is similar to that of the two-sample t- test described in Section 8.1.3.1.

8.3.2.2 Case II (Left-Sided Alternative Hypothesis)

H_0: site median ≥ background median vs.

H_A: site median < background median

The gray region for the WMW left-sided alternative hypothesis is similar to that of two-sample t-test described in Section 8.1.3.2.

The sample sizes *n* and *m*, for one-sided two-sample WMW tests are given by

$$m = n = 1.16\left(2\left(z_{1-\alpha} + z_{1-\beta}\right)^2\left(\frac{sd}{\Delta}\right)^2 + \frac{z_{1-\alpha}^2}{4}\right) \quad (8\text{-}15)$$

Here:

sd^2 =a preliminary estimate of the common variance, σ^2 (obtained from similar sites, expert opinions), of the two populations and to be used during the planning stage; and

sd^2 = pooled variance computed using the collected data to be used when assessing the power of the test in retrospect.

Note: ProUCL outputs the pooled variance based upon the collected data; ProUCL sample size GUI draws user's attention to input an appropriate estimate of sd^2. The user should input an appropriate value depending upon the project stage/data availability.

8.3.2.3 Case III (Two-Sided Alternative Hypothesis)

H_0: site median = background median vs.

H_A: site median ≠ background median

The sample sizes (equal number of samples from the two populations) for the two-sided alternative hypothesis for specified decision parameters are given by:

$$m = n = 1.16\left(2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2\left(\frac{sd}{\Delta}\right)^2 + \frac{z_{1-\alpha/2}^2}{4}\right) \quad (8\text{-}16)$$

Here:

sd^2 =a preliminary estimate of the common variance, σ^2 (obtained from similar sites, expert opinions), of the two populations and to be used during the planning stage; and

sd^2 = pooled variance computed using the collected data to be used when assessing the power of the test in retrospect.

Note: ProUCL sample size GUI draws user's attention to input an appropriate estimate of sd^2. The user should input an appropriate value depending upon the project stage/data availability.

In the following example, ProUCL computes (default option) the sample size to be 46 for the single-sided alternative hypothesis and 56 for the two-sided alternative hypothesis when the user selects the default values of the decision parameters.

Table 8-8. Output for Two-sample WMW Test Sample Sizes (α = 0.05, β = 0.1, s = 3, Δ = 2)

	Sample Sizes for Two Sample Wilcoxon-Mann-Whitney Test
	Based on Specified Values of Decision Parameters/DQOs (Data Quality Objectives)
Date/Time of Computation	2/26/2010 12:18:47 PM
User Selected Options	
False Rejection Rate [Alpha]	0.05
False Acceptance Rate [Beta]	0.1
Width of Gray Region [Delta]	2
Estimate of Standard Deviation	3
	Approximate Minimum Sample Size
Single Sided Alternative Hypothesis:	46
Two Sided Alternative Hypothesis:	56

8.3.3 Sample Size forWMW Test Suggested by Noether (1987)

For the two-sample WRS test (WMW test), the MARSSIM guidance document (EPA 2000) uses the following combined sample size formula suggested by Noether (1987). The combined sample size, *N=(m+n)* equation for the one-sided alternative hypothesis defined in Case I (Section 8.3.2.1) and Case II (Section 8.3.2.2*)* above is given as follows:

$$N = m + n = \frac{\left(z_{1-\alpha}+z_{1-\beta}\right)^2}{3(P-0.5)^2}, \text{ where}$$

$$P = \Phi\left(\frac{\Delta}{\sqrt{2}sd}\right)$$

Δ = Width of the gray region

sd = an estimate of the common standard deviation of the two populations.

P = $\Phi(x)$ = Cumulative probability distribution representing the probability that a standard normal variate, Z, takes on a value $\leq x$.

Some guidance on the selection of an estimate of the population standard deviation, σ, is given in Section 1.1.1. More details can be found in EPA 2006a. The combined sample size, *N=(n+m)* for the two-sided alternative hypothesis (Case III, Section 8.3.2.3) is given as follows:

$$N = m + n = \frac{\left(z_{1-\alpha/2}+z_{1-\beta}\right)^2}{3(P-0.5)^2}$$

Note: In practice the sample sizes obtained using equations described in Sections 8.3.2.1 through 8.3.2.3 are slightly higher than those obtained using Noether's equations described in this Section, 8.3.3. This could

be the reason that the MARSSIM guidance document suggests increasing the sample size obtained using Noether equations by 20%; ProUCL does not increase the calculated sample size by 20%.

Example: An example illustrating these sample size calculations is discussed as follows. In the following example, ProUCL computes the sample size to be 46 for the single sided alternative hypothesis and 56 for the two sided alternative hypothesis when the user selects the default values of the decision parameters.

Using Noether's formula (as used in MARSSIM document), the combined sample size, N= $m + n$ (assuming $m = n$) is 87 for the single sided alternative hypothesis, and 107 for the two sided alternative hypothesis.

Table 8-9. Output for Two-Sample WMW Test Sample Sizes ($\alpha = 0.05$, $\beta = 0.1$, $s = 3$, $\Delta = 2$)

	Sample Sizes for Two Sample Wilcoxon-Mann-Whitney Test
	Based on Specified Values of Decision Parameters/DQOs (Data Quality Objectives)
Date/Time of Computation	7/23/2010 11:58:40 AM
User Selected Options	
False Rejection Rate [Alpha]	0.05
False Acceptance Rate [Beta]	0.1
Width of Gray Region [Delta]	2
Estimate of Standard Deviation	3
	Approximate Minimum Sample Size
Single Sided Alternative Hypothesis:	46
Two Sided Alternative Hypothesis:	56
MARSSIM WRS Test (Noether, 1987)	
	Approximate Minimum Sample Size
Single Sided Alternative Hypothesis:	87
Two Sided Alternative Hypothesis:	107

8.4 Acceptance Sampling for Discrete Objects

ProUCL can be used to determine the minimum number of discrete items that should be sampled, from a lot consisting of n discrete items, to accept or reject the lot (drums containing hazardous waste) based upon the number of defective items (e.g., mean contamination above an action level, not satisfying a characteristic of interest) found in the sampled items. This acceptance sampling approach is specifically useful when the sampling is destructive, that is an item needs to be destroyed (e.g., drums need to be sectioned) to determine if the item is defective or not. The number of items that need to be sampled is determined for the allowable number of defective items, d= 0, 1, 2, …, n. The sample size determination is not straight forward as it involves the use of the beta and hypergeometric distributions. Several researchers (Scheffe and Tukey 1944; Laga and Likes 1975; Hahn and Meeker 1991) have developed statistical methods and algorithms to compute the minimum number of discrete objects that should be sampled to meet specified (desirable) decision parameters. These methods are based upon nonparametric tolerance limits. That is, computing a sample size so that the associated UTL will not exceed the acceptance threshold of the characteristic of interest. The details of the terminology and algorithms used for acceptance sampling of lots (e.g., a batch of drums containing hazardous waste) can be found in the RCRA guidance document (EPA 2002c).

In acceptance sampling, sample sizes based upon the specified values of decision parameters can be computed using the exact beta distribution (Laga and Likes 1975) or the approximate chi-square distribution (Scheffe and Tukey 1944). Exact as well as approximate algorithms have been incorporated in ProUCL 4.1 and higher versions of ProUCL. It is noted that the approximate and exact results are often in complete agreement for most values of the decision parameters. A brief description now follows.

8.4.1 Acceptance Sampling Based upon Chi-square Distribution

The sample size, *n*, for acceptance sampling using the approximate chi-square distribution is given by:

$$n = \frac{m-1}{2} + \left(\frac{(1+p)}{4(1-p)}\right)\chi^2_{\alpha}(2m) \qquad (8\text{-}17)$$

Where:

m = number of non-conforming defective items (always ≥ 1, *m* = 1 implies '0' exceedance rule)

p = 1 – proportion

proportion = pre-specified proportion of non-conforming items

α = *1* – confidence coefficient, and

$\chi^2_{\alpha,2m}$= the cumulative percentage point of a chi-square distribution with *2m df*; the area to the left of $\chi^2_{\alpha,2m}$ is α.

8.4.2 Acceptance Sampling Based upon Binomial/Beta Distribution

Let *x* be a random variable with arbitrary continuous probability density function *f(x)*. Let $x_1 < x_2 < \ldots < x_n$ be an ordered sample size *n* from this distribution.

For a pre-assigned proportion, *p*, and confidence coefficient, *(1–α)*, let the following probability statement given by equation (8-10) be true.

$$P\left\{\int_{x_r}^{x_{n+1-s}} f(x)dx > p\right\} = 1 - \alpha \qquad (8\text{-}18)$$

The statement given by (8-18) implies that the interval (x_r, x_{n+1-s}) contains at least a proportion, *p*, of the distribution with the probability, *(1 – α)*. The interval, (x_r, x_{n+1-s}), whose endpoints are the r^{th} smallest and s^{th} largest observations in a sample size of *n*, is a nonparametric *100p%* tolerance interval with a confidence coefficient of *(1 – α)*, and x_r and x_{n+1-s} are the lower and upper tolerance limits respectively.

The variable $z = \int_{x_r}^{x_{n+1-s}} f(x)dx$ has the following beta probability density function:

$$g(z) = \begin{cases} \frac{1}{B(n-m,\ m)} \cdot z^{n-m} \cdot (1-z)^{m-1}, & 0 < z < 1 \\ 0 & \text{otherwise} \end{cases} \qquad (8\text{-}19)$$

Where

$m = r + s$ and B (p, q) denotes the well known beta function.

The probability $P(z \geq p)$ can be expressed in terms of binomial distribution as follows:

$$P(z \geq p) = \sum_{t=0}^{n-m} \binom{n}{t} p^t (1-p)^{n-t} \quad (8\text{-}20)$$

For given values of m, p and α, the minimum sample size, n, for acceptance sampling is obtained by solving the inequality:

$$P(z \geq p) \geq 1 - \alpha \quad (8\text{-}21)$$

$$\sum_{t=0}^{n-m} \binom{n}{t} p^t (1-p)^{n-t} \geq 1 - \alpha \quad (8\text{-}22)$$

Where:

m = number of non-conforming items (always greater than 1)

p = 1 – proportion

proportion = pre-specified proportion of non-conforming items; and

α = 1 – confidence coefficient.

An example output generated by ProUCL is given as follows.

Table 8-10. Output Screen for Sample Sizes for Acceptance Sampling (default options)

	Acceptance Sampling for Pre-specified Proportion of Non-conforming Items
	Based on Specified Values of Decision Parameters/DQOs
Date/Time of Computation	2/26/2010 12:20:34 PM
User Selected Options	
Confidence Coefficient	0.95
Pre-specified proportion of non-conforming items in the lot	0.05
Number of allowable non-conforming items in the lot	0

	Approximate Minimum Sample Size
Exact Binomial/Beta Distribution	59
Approximate Chisquare Distribution (Tukey-Scheffe)	59

CHAPTER 9

Oneway Analysis of Variance Module

Both parametric and nonparametric Oneway Analysis of Variance (ANOVA) methods are available in ProUCL under the **Statistical Tests** module. A brief description of Oneway ANOVA is described in this chapter.

9.1 Oneway Analysis of Variance (ANOVA)

In addition to the two-sample hypothesis tests, ProUCL software has Oneway ANOVA to compare the location (mean, median) parameters of more than two populations (groups, treatments, monitoring wells). Both classical and nonparametric ANOVA are available in ProUCL. Classical Oneway ANOVA assumes the normality of all data sets collected from the various populations under comparison; classical ANOVA also assumes the homoscedasticity of the populations that are being compared. Homoscedasticity means that the variances (spread) of the populations under comparisons are comparable. Classical Oneway ANOVA represents a generalization of the two-sample t-test (Chapter 6). ProUCL has GOF tests to evaluate the normality of the data sets but a formal F-test to compare the variances of more than two populations has not been incorporated in ProUCL. The users may want to use graphical displays such as side-by-side box plots to compare the spreads present in data sets collected from the populations that are being compared. A nonparametric Oneway ANOVA test: Kruskal-Wallis (K-W) test is also available in ProUCL. The K-W test represents a generalization of the two-sample WMW test described in Chapter 6. The K-W test does not require the normality of the data sets collected from the various populations/groups. However, for each group, the distribution of the characteristic of interest should be continuous and those distributions should have comparable shapes and variabilities.

9.1.1 General Oneway ANOVA Terminology

Statistical terminology used in Oneway ANOVA is described as follows:

g	number of groups, populations, treatments under comparison
i	an index used for the i^{th} group, $i = 1, 2, ..., g$
n_i	number of observations in the i^{th} group
j	an index used for the j^{th} observation in a group; for the i^{th}, $j = 1, 2, ..., n_i$
x_{ij}	the j^{th} observation of the response variable in the i^{th} group
n	total number of observations= $n_1 + n_2 + \ldots + n_g$
$\sum_{j=1}^{n_i} x_{j,i}$	sum of all observations in the i^{th} group
$\bar{x}_i$	mean of the observations collected from the i^{th} group
$\bar{x}$	mean of all, n_t (the observations)
μ_i	true (unknown) mean of the i^{th} group

In Oneway ANOVA, the null hypothesis, $H_{0,}$ is stated as: the g groups under comparison have equal means (medians) and that any differences in the sample means/medians are due to chance. The alternative hypothesis, H_A is stated as: the means/medians of the g groups are not equal.

The decision to reject or accept the null hypothesis is based upon a test statistic computed using the available data collected from the g groups.

9.2 Classical Oneway ANOVA Model

The ANOVA model is represented by a regression model in which the predictor variables are the treatment or group variables. The Oneway ANOVA model is given as follows:

$$x_{i,j} = \mu_i + e_{i,j} \tag{9-1}$$

Where μ_i is the population mean (or median) of the i^{th} group, and errors, $e_{i,j}$, are assumed to be independently and normally distributed with mean = 0 and with a constant variance, σ^2. All observations in a given group have the same expectation (mean) and all observations have the same variance regardless of the group. The details of Oneway ANOVA can be found in most statistical books including the text by Kunter *et al*. (2004).

The null and the alternative hypotheses for Oneway ANOVA are given as follows:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_i = \cdots = \mu_g$$
$$H_A: At\ least\ one\ of\ the\ means\ (or\ medians)\ is\ not\ equal\ to\ others$$

Based upon the available data collected from the g groups, the following statistics are computed. ProUCL summarizes these results in an ANOVA Table.

Sum of Squares Between Groups is given by:

$$SS_{Between\ Groups} = \sum_{i=1}^{g} n_i(\bar{x}_i - \bar{x})^2 \tag{9-2}$$

Sum of Squares Within Groups is given by:

$$SS_{Within\ Groups} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(x_{j,i} - \bar{x}_i\right)^2 \tag{9-3}$$

Total Sum of Squares is given by:

$$SS_{Total} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(x_{j,i} - \bar{x}_i\right)^2 \tag{9-4}$$

Between Groups Degrees of Freedom (*df*):	*g-1*
Within Groups *df*:	*n-g*
Total *df*:	*n-1*

Mean Squares Between Groups is given by:

$$MS_{Between\ Groups} = \frac{SS_{Between\ Groups}}{g-1} \quad (9\text{-}5)$$

Mean Squares Within Groups:

$$MS_{Within\ Groups} = \frac{SS_{Within\ Groups}}{n-g} \quad (9\text{-}6)$$

Scale estimate is given by:

$$S = \sqrt{MS_{Within\ Groups}} \quad (9\text{-}7)$$

R^2 is given by:

$$R^2 = 1 - \frac{SS_{Within\ Groups}}{SS_{Total}} \quad (9\text{-}8)$$

Decision statistic, F, is given by:

$$F\ Statistic = \frac{MS_{Between\ Groups}}{MS_{Within\ Groups}} \quad (9\text{-}9)$$

Under the null hypothesis, the F-statistic given in equation (9-9) follows the $F_{(g-1),\ (n-g)}$ distribution with *(g-1)* and *(n-g)* degrees of freedom, provided the data sets collected from the g groups follow normal distributions. ProUCL software computes *p*-values using the F distribution, $F_{(g-1),\ (n-g)}$.

Conclusion: The null hypothesis is rejected for all levels of significance, $\alpha \geq p$-value.

9.3 Nonparametric Oneway ANOVA (Kruskal-Wallis Test)

Nonparametric Oneway ANOVA or the K-W test (Kruskal and Wallis 1952, Hollander and Wolfe 1999) represents a generalization of the two-sample WMW, test which is used to compare the equality of medians of two groups. Like the WMW test, analysis for the K-W test is also conducted on ranked data, therefore, the distributions of the *g* groups under comparisons do not have to follow a known statistical distribution (e.g., normal). However, distributions of the *g* groups should be continuous with comparable shapes and variabilities. Also the *g* groups should represent independently distributed populations.

The null and alternative hypotheses are defined in terms of medians, m_i of the *g* groups:

$$H_0: m_1 = m_2 = \ldots = m_i = \ldots = m_g \quad (9\text{-}10)$$
$$H_A: At\ least\ one\ of\ the\ g\ medians\ is\ not\ equal\ to\ others$$

While performing the K-W test, all *n* observations in the *g* groups are arranged in ascending order with the smallest observation receiving the smallest rank and the largest observation getting the highest rank. All tied observations receive the average rank of those tied observations.

K-W Test on Data Sets with NDs: It should be noted that the K-W test may be used on data sets with NDs provided all NDs are below the largest detected value. All NDs are considered as tied observations

irrespective of reporting limits (RLs) and receive the same rank. However, the performance of the K-W test on data sets with NDs is not well studied; therefore, it is suggested that the conclusion derived using the K-W test statistics be supplemented with graphical displays such as side-by-side box plots. Side-by-side box plots can also be used as an exploratory tool to compare the variabilities of the *g* populations based upon the *g* data sets collected from those populations.

The K-W ANOVA table displays the following information and statistics:

Mean Rank of the i^{th} Group, $\bar{R}_i$: Average of the ranks (in the combined data set of size, *n*) of the n_i observations in the i^{th} group.

Overall Mean Rank, $\bar{R}$: Average of the ranks of all *n* observations.

Z-value of each group are computed using the following equation (Standardized Z):

$$Z_i = \frac{\bar{R}_i - \bar{R}}{\sqrt{\frac{\left((n+1)\left(\frac{n}{n_i}-1\right)\right)}{12}}} \tag{9-11}$$

n total number of observations = $n_1 + n_2 + \ldots + n_g$

n_i observation in the i^{th} group

g number of groups

Z_i given by (9-11) represents standardized normal deviates. The Z_i can be used to determine the significance of the difference between the average rank of the i^{th} group and the overall average rank, R, of the combined data set of sized n.

Kruskal-Wallis H-Statistic (without ties) is given by:

$$H = \frac{12\sum_{i=1}^{g} n_i(\bar{R}_i - \bar{R})^2}{n(n+1)} \tag{9-12}$$

K-W H-Statistic adjusted for ties is given by:

$$H_{adj-ties} = \frac{H}{1-\left[\frac{\sum_{i=1}^{g} t_i^3 - t_i}{n^3 - n}\right]} \tag{9-13}$$

Where t_i = number of tied values in i^{th} group

For large values of *n*, the H-statistic given above follows an approximate chi-square distribution with *(g-1)* degrees of freedom. *P*-values associated with the H-statistic given by (9-12) and (9-13) are computed by using a chi-square distribution with *(g-1)* degrees of freedom. The *p*-values based upon a chi-square approximation test are fairly accurate when the number of observations, *n*, is large such as ≥ 30.

Conclusion: The null hypothesis is rejected in favor of the alternative hypothesis for all levels of significance, $\alpha \geq p$-value.

Table 9-1. Classical Oneway ANOVA Results Comparing Petal Widths of 3 Iris Species from Fisher's Famous Iris Data Set (Fisher 1936).

	Classical Oneway ANOVA
Date/Time of Computation	3/2/2013 1:25:29 PM
From File	FULLIRIS.xls
Full Precision	OFF

pt-width

Group	Obs	Mean	SD	Variance
1	50	0.246	0.105	0.0111
2	50	1.326	0.198	0.0391
3	50	2.026	0.275	0.0754
Grand Statistics (All data)	150	1.199	0.762	0.581

Classical One-Way Analysis of Variance Table

Source	SS	DOF	MS	V.R.(F Stat)	P-Value
Between Groups	80.41	2	40.21	960	0
Within Groups	6.157	147	0.0419		
Total	86.57	149			

Pooled Standard Deviation	0.205
R-Sq	0.929

Note: A p-value <= 0.05 (or some other selected level) suggests that there are significant differences in mean/median characteristics of the various groups at 0.05 or other selected level of significance

A p-value > 0.05 (or other selected level) suggests that mean/median characteristics of the various groups are comparable.

Table 9-2 (Iris Data). The K-W Oneway ANOVA Results Comparing Petal Widths of 3 iris Species.

	Nonparametric Oneway ANOVA (Kruskal-Wallis Test)							
Date/Time of Computation	3/2/2013 1:29:12 PM							
From File	FULLIRIS.xls							
Full Precision	OFF							
pt-width								
Group	Obs	Median	Ave Rank	Z				
1	50	0.2	25.5	-9.967				
2	50	1.3	76.48	0.195				
3	50	2	124.5	9.771				
Overall	150	1.3	75.5					
K-W (H-Stat)	DOF	P-Value	(Approx. Chisquare)					
129.9	2	0						
131.2	2	0	(Adjusted for Ties)					
Note: A p-value <= 0.05 (or some other selected level) suggests that there are significant differences in								
mean/median characteristics of the various groups at 0.05 or other selected level of significance								
A p-value > 0.05 (or other selected level) suggests that mean/median characteristics of the various groups are comparable								

CHAPTER 10

Ordinary Least Squares Regression and Trend Analysis

Trend tests and ordinary least squares (OLS) regression methods are used to determine trends (e.g., decreasing, increasing) in time series data sets. Typically, OLS regression is used to determine linear relationships between a dependent response variable and one or more predictor (independent) variables (Draper and Smith 1998); however statistical inference on the slope of the OLS line can also be used to determine trends in the time series data used to estimate an OLS line. A couple of nonparametric statistical tests, the Mann-Kendall (M-K) test and the Theil-Sen (T-S) test to perform trend analysis have also been incorporated in ProUCL since version 5.0. Methods to perform trend analysis and OLS Regression with graphical displays are available under the **Statistical Tests** module of ProUCL. In environmental monitoring studies, OLS regression and trend tests can be used on time series data sets to determine potential trends in constituents' concentrations over a defined period of time. Specifically, the OLS regression with time or a simple index variable as the predictor variable can be used to determine a potential increasing or decreasing trend in mean concentrations of an analyte over a period of time. A significant positive (negative) slope of the regression line obtained using the time series data set with predictor variable as a time variable suggests an upward (downward) trend. A brief description of the classical OLS regression as function of the time variable, *T* (*t)*, is described as follows. It should however be noted that the OLS regression and associated graphical displays can be used to determine a linear relation for any pair of dependent variable, Y, and independent variable, X. The independent variable does not have to be a time variable.

10.1 Ordinary Least Squares Regression

The linear regression model for a response variable, Y and a predictor (independent) variable, *t* is given as follows:

$$Y = b_0 + b_1 t + e; \qquad (10\text{-}1)$$

$$E[Y] = b_0 + b_1 t \; = \; mean\, response\, at\, t$$

In (10-1), variable *e* is a random variable representing random measurement error in the response variable, Y (concentrations). The error variable, *e*, is assumed to follow a normal distribution, *N (0, σ2)*, with mean 0 and unknown variance, σ^2. Let (t_i, y_i); $i: =1, 2, \ldots n$ represent the paired data set of size *n*, where y_i is the measured response when the predictor variable, $t = t_i$. It is noted that multiple observations may be collected at one or more values of the prediction variable, *t*. Using the regression model (10-1) on this data set, we have:

$$y_i = b_0 + b_1 t_i + e; \qquad (10\text{-}2)$$

$$E[y_i] = b_0 + b_1 t_i = mean\, response\, when\, t \; = \; t_i$$

For each fixed value, t_i of the predictor variable, t, the random error, e_i is normally distributed with $N(0,\sigma^2)$. Random errors, e_i, are independently distributed. Without the random error, e, all points will lie exactly on the population regression line estimated by the OLS line. The OLS estimates of the intercept, b_0 and slope, b_1 are obtained by minimizing the residual sum of squares. The details of deriving the OLS estimates, $\hat{b}_0$ *and* $\hat{b}_1$ of the intercept and slope can be found in Draper and Smith (1998).

The OLS regression method can be used to determine increasing or decreasing trends in the response variable Y (e.g., constituent concentrations in a MW) over a time period (e.g., quarters during a 5 year time period). A positive statistically significant slope estimate suggests an upward trend and a statistically significant negative slope estimate suggests a downward or decreasing trend in the mean constituent concentrations. The significance of the slope estimate is determined based upon the normal assumption of the distribution of error terms, e_i, and therefore, of responses, y_i, $i:=1,2,\ldots,n$

ProUCL computes OLS estimates of parameters b_0 and b_1; performs inference about the slope and intercept estimates, and outputs the regression ANOVA table including the coefficient of determination, R^2, and estimate of the error variance, σ^2. Note that R^2 represents the square of the Pearson correlation coefficient between the dependent response variable, y, and the independent predictor variable, t. ProUCL also computes confidence intervals and prediction intervals around the OLS regression line; and can be used to generate scatter plots of n pairs, (t, y), displaying the OLS regression line, confidence interval for mean responses, and prediction interval band for individual observations (e.g., future observations).

General OLS terminology and sum of squares computed using the collected data are described as follows:

$$S_{ty} = \sum_{i=1}^{n} t_i y_i - (\sum_{i=1}^{n} t_i \sum_{i=1}^{n} y_i)/n \; ; \; and \qquad (10\text{-}3)$$

$$S_{tt} = \sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2 /n \; ; \; and \; S_{yy} = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 /n$$

The OLS estimates of slope and intercept are given as follows:

$$\hat{b}_1 = S_{ty}/S_{tt}; \; and$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{t}$$

$$\bar{t} = \sum_{i=1}^{n} t_i/n \qquad (10\text{-}4)$$

The estimated OLS regression line is given by: $\hat{y} = \hat{b}_0 + \hat{b}_1 t$ and error estimates also called residuals are given by $\hat{e}_i = y_i - \hat{y}_i; \; i = 1,2,\ldots,n$. It should be noted that for each i, $\hat{y}_i$ represents the mean response at value, t_i of the predictor variable, t, for $i:=1,2,\ldots,n$.

The residual sum of squares is given by:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (10\text{-}5)$$

Estimate of the error variance, σ^2, and variances of the OLS estimates, $\hat{b}_0$ *and* $\hat{b}_1$ are given as follows:

$$\hat{\sigma}^2 = MSE = SSE/(n-2)$$

$$Var(\hat{b}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{t}^2}{S_{tt}}\right)$$

$$Var(\hat{b}_1) = \sigma^2/S_{tt} \qquad (10\text{-}6)$$

Estimates of the variances of the OLS estimates $\hat{b}_0$ *and* $\hat{b}_1$ are obtained by replacing σ^2 by its estimate, mean sum of squares error (MSE), given in (10-6). Standard errors (SEs) of the OLS estimates: $\hat{b}_0$ *and* $\hat{b}_1$ are their respective standard deviations. ProUCL tests the significance of slope and intercept of the regression line given by (10-1). Details for testing the significance of the slope are given as follows. It should be noted that the parametric OLS regression line given by (10-4) estimates the change in *the mean* concentration over time.

Testing Significance of the Slope, b_1: Under normality and independence of random errors, e_i, in responses, y_i, the test statistic given by (10-7) follows a Student's t-distribution with *(n-2)* degrees of freedom. One can perform any of the 3 hypothesis forms including: 1) H_0: $b_1 = 0$ vs. the alternative hypothesis, H_1: $b_1 \neq 0$; 2) H_0: $b_1 = 0$ vs. the alternative, H_1:$b_1 > 0$; and 3) or H_0: $b_1 = 0$ vs. the alternative, H_1: $b_1 < 0$. Under the null hypothesis, the test statistic is obtained by dividing the regression estimate by its SE:

$$t = \hat{b}_1/SE(\hat{b}_1) \qquad (10\text{-}7)$$

Under normality of the responses, y_i (and the random errors, e_i), the test statistic given in (10-7) follows a Student's t-distribution with (*n*-2) degrees of freedom (*df*). A similar process is used to perform inference about the intercept, b_0 of the regression line. The test statistic associated with the OLS estimate of the intercept, $\hat{b}_0$ also follows a Student's t-distribution with (*n*-2) degrees of freedom.

P-values: ProUCL computes and outputs t-distribution based *p*-values associated with the two-sided alternative hypothesis, H_1: $b_1 \neq 0$. The *p*-values are displayed on the output sheet as well as on the regression graph generated by ProUCL.

Note: ProUCL displays residuals including standardized residuals on the OLS output sheet. Those residuals can be imported (copying and pasting) in an excel data file to assess the normality of those OLS residuals. The parametric trend evaluations based upon the OLS slope (significance, confidence interval) are valid provided the OLS residuals are normally distributed. Therefore, it is suggested that the user assesses the normality of OLS residuals before drawing trend conclusions using a parametric test based upon the OLS slope estimate. When the assumptions are not met, one can use graphical displays and nonparametric trend tests, M-K and T-S tests, to determine potential trends in time series data set.

10.1.1 Regression ANOVA Table

The following statistics are displayed on the regression ANOVA table.

Sum of Squares Regression (SSR): SSR represents that part of the variation in the response variable, Y, which is explained by the regression model, and is given by:

$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad (10\text{-}8)$$

Sum of Squares Error (SSE): SSE represents that part of the variation in the response variable, Y, which is attributed to random measurement errors, and is given by:

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Sum of Squares Total (SST): SST is the total variation present in the response variable, Y and is equal to the sum of SSR and SSE.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = SSR + SSE \quad (10\text{-}9)$$

Regression Degrees of Freedom (*df*): 1 (1 predictor variable)

Error *df*: *n-2;* and Total *df*: *n-1*

Mean Sum of Squares (MS) Regression (MSR): is given by SSR divided by the regression *df* which is equal to 1 in the present scenario with only one predictor variable.

$$MSR = SSR$$

Mean Sum of Squares Error (MSE): is given by SSE divided by the error degrees of freedom

$$MSE = \frac{SSE}{n-2}$$

MSE represents an unbiased estimate of the error variance, σ^2. In regression terminology, σ is called the scale parameter, and $\sqrt{MSE}$ is called the scale estimate.

F-statistic: is computed as the ratio of MSR to MSE, and follows an F distribution with *1* and *(n-2)* degrees of freedom (*df)*.

$$F = \frac{MSR}{MSE} \quad (10\text{-}10)$$

P-value: The overall *p-value* associated with the regression model is computed using the $F_{1,(n-2)}$ distribution of the test- statistic given by equation (10-10).

R^2: represents the variation explained in the response variable, Y, by the regression model, and is given by:

$$R^2 = 1 - \frac{SSE}{SST} \quad (10\text{-}11)$$

Adjusted R square (Adjusted R^2): The adjusted R^2 is considered a better measure of the variation explained in the response variable, Y, and is given by:

$$R^2_{adjusted} = 1 - \left(\frac{n-1}{n-2}\right)\frac{SSE}{SST}$$

10.1.2 Confidence Interval and Prediction Interval around the Regression Line

ProUCL also computes confidence and prediction intervals around the regression line and displays these intervals along with the regression line on the scatter plot of the paired data used in the OLS regression. ProUCL generates, when selected, a summary table displaying these intervals and residuals.

Confidence Interval (LCL, UCL): represents a band within which the estimated mean responses, $\hat{y}_i$, are expected to fall with specified confidence coefficient, *(1-α)*. Upper and lower confidence limits (LCL and UCL) are computed for each mean response estimate, $\hat{y}_i$, observed at value, t_i, of the predictor variable, *t*. These confidence limits are given by:

$$\hat{y}_i \pm t_{((1-\alpha/2),(n-2))} sd[\hat{y}_i] \qquad (10\text{-}12)$$

Where the estimated standard deviation, $sd(\hat{y}_i)$, of the mean response, $\hat{y}_i$, is given by:

$$sd[\hat{y}_i] = \sqrt{MSE\left(\frac{1}{n} + \frac{(t_i - \bar{t})^2}{S_{tt}}\right)} \; ; \; i = 1,2,\ldots,n$$

A confidence band can be generated by computing the confidence limits given by (10-12) for each value, t_i of the predictor variable, *t; i:=1,2,…n*.

Prediction Limits (LPL, UPL): represents a band within which a predicted response (and not the mean response), $\hat{y}_0$, for a specified new value, t_0 ,of the predictor variable, *t*, is expected to fall. Since the variances of the individual predicted responses are higher than the variances of the mean responses, a prediction band around the OLS line is wider than the confidence band. The LPL and UPL comprising the prediction band are given by:

$$\hat{y}_0 \pm t_{((1-\alpha/2),(n-2))} sd(\hat{y}_0); \; with \; \hat{y}_0 = \hat{b}_0 + \hat{b}_1 x_0 \qquad (10\text{-}13)$$

Where the estimated standard deviation, $sd(\hat{y}_0)$, of a new response, $\hat{y}_0$,(or the individual response for existing observations) is given by:

$$sd(\hat{y}_0) = \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(t_0 - \bar{t})^2}{S_{tt}}\right)};$$

Like the confidence band, a prediction band around the OLS line can be generated by computing the prediction limits given by (10-13) for each value, t_i , of the predictor variable, *t*, and also other values of *t* (within the experiment range) for which the response, *y*, was not observed.

Notes: Unlike M-K and T-S trend tests, multiple observations may be collected at one or more values of the predictor variable. Specifically, OLS can be performed on data sets with multiple measurements collected at one or more values of the predictor variable (e.g., sampling time variable, t).

Example 10-1. Consider the time series data set for sulfate as described in RCRA Guidance (EPA 2009e). The OLS graph with relevant test statistics is shown in Figure 10-1 below. The positive slope estimate, 33.12, is significant with a *p-value* of 0 suggesting that there is an upward trend in sulfate concentrations.

Figure 10-1. OLS Regression of Sulfate as a Function of Time

Table 10-1. Regression Results from Example 10-1

Number Reported (x-values)			22			
Dependendant Variable			Sulfate			
Independent Variable			Date			
Regression Estimates and Inference Table						
Paramater	Estimates	Std. Error	T-values	p-values		
intercept	-2503	410.7	-6.095	5.8853E-6		
Date	33.12	4.422	7.49	3.1763E-7		
OLS ANOVA Table						
Source of Variation		**SS**	**DOF**	**MS**	**F-Value**	**P-Value**
Regression		126230	1	126230	56.1	0.0000
Error		45003	20	2250		
Total		171233	21			
R Square			0.737			
Adjusted R Square			0.724			
Sqrt(MSE) = Scale			47.44			

10.2 Trend Analysis

Time Series Data Set: When the predictor variable, *t*, represents a time variable (or an index variable), the data set $(t_i, y_i);\ i:=1,2,....n$ is called a time series data set, provided values of the variable, *t*, satisfy: $t_1 < t_2 < t_3 <\ldots< t_n$.

The **Trend Analysis** module of ProUCL includes two trend tests, the M-K) test and the T-S test. The trend tests in ProUCL are performed on time series data sets. Both M-K and T-S tests in ProUCL can handle missing values. Like all other methods, these tests can be performed by a group variable - performing the selected trend test for each group in the data set. A detailed description of these tests is described in the following sections.

Notes: The two trend tests are meant to identify trends in time series data (data collected over a certain period of time such as daily, monthly, quarterly, etc) with distinct values of the time variable (time of sampling events); that is only one measurement is reported (collected) at each sampling event time. If multiple measurements are collected at a sampling event, the user may want to use the average (or median, mode, minimum or maximum) of those measurements resulting in a time series with one measurement per sampling time event. When multiple observations are present for a sampling event, ProUCL computes the average of those observations. Trend tests in ProUCL software assume that the user has entered data in chronological order. If the data are not entered properly in chronological order, the graphical trend displays may be meaningless. T-S tests takes sampling events into consideration; however, those sampling events do not have to be performed at regular intervals. When sampling events are not provided, the user can assign numeric values in chronological order for sampled observations. At present ProUCL does not does not read dates (years, quarters etc.). If dates are provided, the user needs to assign numeric values in chronological order.

Handling Nondetects: The trend module in ProUCL does not recognize a nondetect column consisting of zeros and ones. For data sets consisting of nondetects with varying DLs, one can replace all NDs with half of the lowest DL (DL/2) or by replacing all NDs by a single value lower than the lowest DL. When multiple DLs are present in a data set, the use of substitution methods should be avoided. Replacing NDs by their respective DLs or by their DL/2 values is like performing trend test on DLs or on DL/2s, especially when the percentage of NDs present in the data set is high.

10.2.1 Mann–Kendall Test

The M-K trend test is a nonparametric test which is used on a time series data set, $(t_i, y_i);\ i:=1,2,....n$ as described earlier. As a nonparametric procedure, the M-K test does not require the underlying data to follow a specific distribution. The M-K test can be used to determine increasing or decreasing trends in measurement values of the response variable, *y*, observed during a certain time period. If an increasing trend in measurements exists, then the measurement taken first from any randomly selected pair of measurements should, on average, have a lower response (concentration) than the measurement collected at a later point.

The M-K statistic, *S*, is computed by examining all possible distinct pairs of measurements in the time series data set and scoring each pair as follows. It should be noted that for a measurement data set of size, *n,* there are $n(n-1)/2$ distinct pairs, (y_j, y_i) with $j>i$, which are being compared.

- If an earlier measurement, y_i, is less in magnitude than a later measurement, y_j, then that pair is assigned a score of 1;
- If an earlier measurement value is greater in magnitude than a later value, the pair is assigned a score of –1; and
- Pairs with identical ($y_i = y_j$) measurements values are assigned a score of 0.

The M-K test statistic, *S*, equals the sum of scores assigned to all pairs. The following conclusions are derived based upon the values of the M-K statistic, *S*.

- A positive value of *S* implies that a majority of the differences between earlier and later measurements are positive suggesting the presence of a potential upward and increasing trend over time.
- A negative value for *S* implies that a majority of the differences between earlier and later measurements are negative suggesting the presence of a potential downward/decreasing trend.
- A value of *S* close to zero indicates a roughly equal number of positive and negative scores assigned to all possible distinct pairs, (y_j, y_i) with $j>i$, suggesting that the data do not exhibit any evidence of an increasing or decreasing trend.

When no trend is present in time series measurements, positive differences in randomly selected pairs of measurements should balance negative differences. In other words, the expected value of the test statistic *S,* E[*S*], should be close to '0' when the measurement data set does not exhibit any evidence of a trend. To account for randomness and inherent variability in measurements, the statistical significance of the M-K test statistic is determined. The larger the absolute value of *S*, the stronger the evidence for a real increasing or decreasing trend. The M-K test in ProUCL can be used to test the following hypotheses:

Null Hypothesis, H_0: Data set does not exhibit sufficient evidence of any trends (stationary measurements) vs.

- H_A: Data set exhibits an upward trend (not necessarily linear); or
- H_A: Data set exhibits a downward trend(not necessarily linear); or
- H_A: Data set exhibits a trend (two-sided alternative - (not necessarily linear)).

Under the null hypothesis of no trend, it is expected that the mean value of S =0; that is E[*S*] =0.

Notes: The M-K test in ProUCL can be used for testing a two-sided alternative, H_A, stated above. For a two-sided alternative hypothesis, the *p*-values (exact as well as approximate) reported by ProUCL need to be doubled.

10.2.1.1 Large Sample Approximation for M-K Test

When the sample size *n* is large, the exact critical values for the statistic *S* are not readily available. However, as a sum of identically-distributed random quantities, the distribution of *S* tends to approximately follow a normal distribution by the CLT. The exact *p*-values for the M-K test are available for sample sizes

up to 22 and have been incorporated in ProUCL. For samples of sizes larger than 22, a normal approximation to *S* is used. In this case, a standardized *S*-statistic, denoted by Z is computed by using the expected mean value and *sd* of the test statistic, *S*.

The *sd* of *S, sd(S)* is computed using the following equation:

$$sd(S) = \sqrt{\frac{1}{18}\left[n(n-1)(2n+5) - \sum_{j=1}^{g} t_j(t_j-1)(2t_j+5)\right]} \quad (10\text{-}14)$$

Where n is the sample size, g represents the number of groups of ties (if any) in the data set, and t_j is the number of tied observations in the j^{th} group of ties. If no ties or NDs are present, the equation reduces to the simpler form:

$$sd(S) = \sqrt{\frac{1}{18}[n(n-1)(2n+5)]} \quad (10\text{-}15)$$

The standardized *S* statistic denoted by Z for an increasing (or decreasing) trend is given as follows:

$$\begin{aligned} Z &= \tfrac{(S-1)}{sd(S)}\ if\ S > 0; \\ Z &= 0\ if\ S = 0;\ and \\ Z &= \tfrac{(S+1)}{sd(S)}\ if\ S < 0 \end{aligned} \quad (10\text{-}16)$$

Like the *S* statistic, the sign of Z determines the direction of a potential trend in the data set. A positive value of Z suggests an upward (increasing) trend and a negative value of Z suggests a downward or decreasing trend. The statistical significance of a trend is determined by comparing Z with the critical value, z_α, of the standard normal distribution; where z_α represents that value such that the area to the right of z_α under the standard normal curve is α.

10.2.1.2 Step-by-Step Procedure to perform the Mann-Kendall Test

The M-K test does not require the availability of an event or a time variable. However, if graphical trend displays (e.g., T-S line) are desired, the user should provide the values for a time variable. When a time or an event variable is not provided, ProUCL generates an index variable and displays the time-series graph using the index variable.

Step 1. Order the measurement data: $y_1, y_2, \ldots, y_n$ by sampling event or time of collection. If the numerical values of data collection times (event variable) are not known, the user should enter data values according to the order they were collected. Next, compute all possible differences between pairs of measurements, ($y_j - y_i$) for $j > i$. For each pair, compute the *sign* of the difference, defined by:

$$sgn(y_j - y_i) = \begin{cases} 1\ if\ (y_j - y_i) > 0 \\ 0\ if\ (y_j - y_i) = 0 \\ -1\ if\ (y_j - y_i) < 0 \end{cases} \quad (10\text{-}17)$$

Step 2. Compute the M-K test statistic, *S*, given by the following equation:

$$S = \sum_{i=1}^{n} \sum_{j=i+1}^{n} sgn\,(y_j - y_i) \qquad (10\text{-}18)$$

In the above equation the summation starts with a comparison of the very first sampling event against each of the subsequent measurements. Then the second event is compared with each of the samples taken after it (*i.e.*, the third, fourth, and so on). Following this pattern is probably the most convenient way to ensure that all distinct pairs have been considered in computing *S*. For a sample of size *n*, there will be *n*(*n*-1)/2 distinct pairs, *(i, j)* with *j*>*i*.

Step 3. For *n<23,* the tabulated critical levels, α_{cp} (tabulated *p*-values) given in Hollander and Wolfe (1999), have been incorporated in ProUCL. If $S > 0$ and $\alpha > \alpha_{cp}$, conclude there is statistically significant evidence of an increasing trend at the α significance level. If $S < 0$ and $\alpha > \alpha_{cp}$, conclude there is statistically significant evidence of a decreasing trend. If $\alpha \leq \alpha_{cp}$, conclude that data do not exhibit sufficient evidence of any significant trend at the α level of significance.

Specifically, the M-K test in ProUCL tests for one-sided alternative hypothesis as follows:

H_0: no trend vs. H_A: upward trend

or

H_0: no trend vs. H_A: downward trend

ProUCL computes tabulated *p*-values (for sample sizes <23) based upon the sign of the M-K statistic, *S*, as follows:

If *S>0*, the tabulated *p*-value (α_{cp}) is computed for *H_0: no trend*, vs. *H_A: upward trend*

If *S<0*, the tabulated *p*-value (α_{cp}) is computed for *H_0: no trend* vs. H_A*: downward trend*

If the *p*-value is larger than the specified α (e.g., 0.05), the null hypothesis of no trend is not rejected.

Step 4. For $n > 22$, large sample normal approximation is used for *S*, and a standardized *S* is computed. Under the null hypothesis of no trend, E(*S*) =0, and the *sd* is computed using equations (10-14) or (10-15). When ties are present, *sd(S)* is computed by adjusting for ties as given in (10-14). Standardized *S*, denoted by Z is computed using equation (10-16).

Step 5. For a given significance level *(α),* the critical value z_α is determined from the standard normal distribution.

If Z >0, a critical value and *p*-value are computed for *H_0: no trend*, vs. *H_A: upward trend*.

If Z<0, a critical value and *p*-value are computed for *H_0: no trend* vs. *H_A: downward trend*

If the *p-value* is larger than the specified α (e.g., 0.05), the null hypothesis of no trend is not rejected.

Specifically, compare *Z* against this critical value, z_α. If Z>0 and $Z > z_\alpha$, conclude there is a statistically significant evidence of an increasing trend at an α-level of significance. If Z<0 and $Z < -z_\alpha$, conclude there is statistically significant evidence of a decreasing trend. If neither exists, conclude that the data do not

exhibit sufficient evidence of any significant trend. For large samples, ProUCL computes the *p*-value associated with Z.

Notes: As mentioned, the M-K test in ProUCL can be used for testing a two-sided alternative, H_A stated above. For a two-sided alternative hypothesis, *p*-values (both exact and approximate) reported by ProUCL need to be doubled.

Example 10-2. Consider a nitrate concentration data set collected over a period of time. The objective is to determine if there is a downward trend in nitrate concentrations. No sampling time event values were provided. The M-K test has been used to establish a potential trend in nitrate concentrations. However, if the user also wants to see a trend graph, ProUCL generates an index variable and displays the trend graph along with OLS line and the T-S nonparametric line (based upon the index variable) as shown in Figure 10-2 below. Figure 10-2 displays all the statistics of interest.

Table 10-2. M-K Trend Statistics.

	Mann-Kendall Trend Test Analysis
User Selected Options	
Date/Time of Computation	3/2/2013 3:45:38 PM
From File	Trend-data forNitrate_a.xls
Full Precision	OFF
Confidence Coefficient	0.95
Level of Significance	0.05

Nitrate	
General Statistics	
Number Values	204
Number Values Missing	2
Number Values Reported (n)	202
Minimum	9.312
Maximum	19.96
Mean	14.29
Geometric Mean	14.2
Median	13.96
Standard Deviation	1.688
Mann-Kendall Test	
Test Value (S)	-4684
Critical Value (0.05)	-1.645
Standard Deviation of S	960.5
Standardized Value of S	-4.876
Approximate p-value	5.4240E-7
Statistically significant evidence of a decreasing trend at the specified level of significance.	

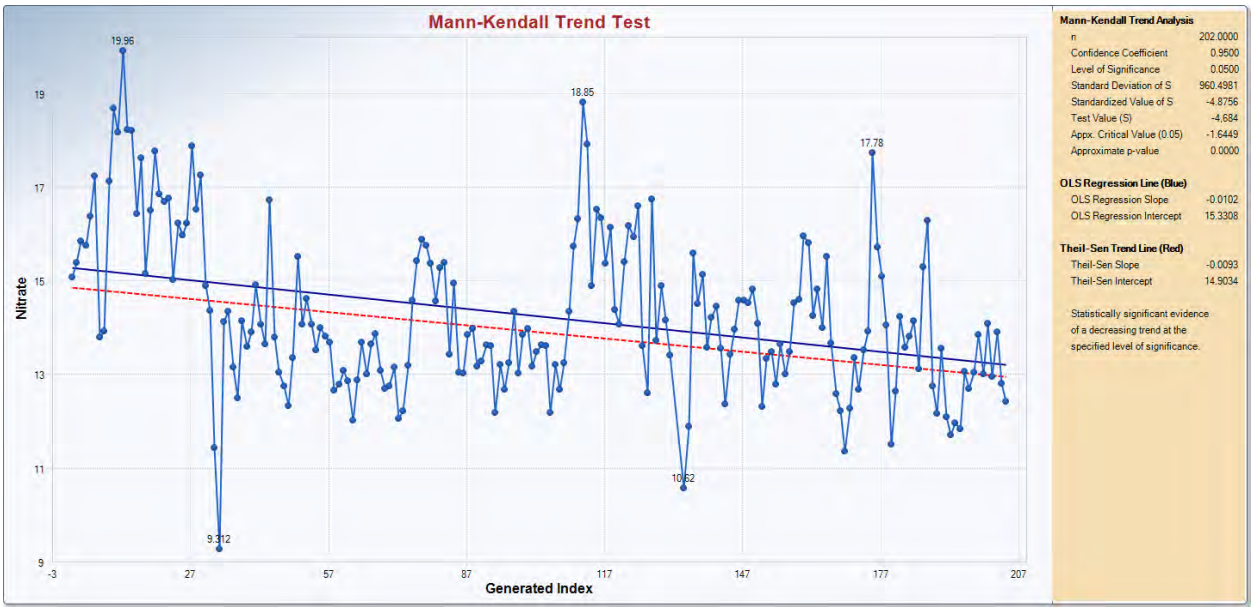

Figure 10-2. Trend Graph with M-K Test Results and OLS Line and Nonparametric Theil-Sen Line

10.2.2 Theil - Sen Line Test

The details of T-S test can be found in Hollander and Wolfe (1999). The T-S test represents a nonparametric version of the parametric OLS regression analysis and requires the values of the time variable at which the response measurements were collected. The T-S procedure does not require normally distributed trend residuals and responses as required by the OLS regression procedure. It is also not critical that the residuals be homoscedastic (having equal variance over time). For large samples, even a relatively mild to modest slope of the T-S trend line can be statistically significantly different from zero. It is best to first identify whether or not a significant trend (slope) exists, and then determine how steeply the concentration levels are increasing (or decreasing) over time for a significant trend.

New since ProUCL 5.1: ProUCL computes y-hat values and residuals based upon the Theil-Sen nonparametric regression line. ProUCL outputs the slope and intercept of the T-S trend line, which can be used to compute residuals associated with the T-S regression line.

Unlike the M-K test, actual concentration values are used in the computation of the slope estimate associated with the T-S trend test. The test is based upon the idea that if a simple *slope estimate* is computed for every pair (*n(n-1)/2* pairs in all) of *distinct* measurements in the sample (known as the set of *pairwise slopes*), the average of this set of *n(n-1)/2* slopes would approximate the true unknown slope. Since the T-S test is a nonparametric test, instead of taking an *arithmetic average* of the pairwise slopes, the *median* slope value is used as an estimate of the unknown population slope. By taking the median pairwise slope instead of the mean, extreme pairwise slopes - perhaps due to one or more outliers or other errors - are ignored and have little or negligible impact on the final slope estimator.

The T-S trend line is also nonparametric because the median pairwise slope is combined with the median concentration value and the median of the time values to construct the final trend line. Therefore, the T-S line estimates the change in *median* concentration over time and not the *mean* as in linear OLS regression; the parametric OLS regression line described in Section 10.1 estimates the change in *the mean* concentration over time (when the dependent variable represents the time variable).

Averaging of Multiple Measurements at Sampling Events: In practice, when multiple observations are collected/reported at one or more sampling events (times), one or more pairwise slopes may become infinite, resulting in a failure to compute the T-S test statistic. In such cases, the user may want to pre-process the data before using the T-S test. Specifically, to assure that only one measurement is available at each sampling event, the user pre-processes the time series data by computing average, median, mode, minimum, or maximum of the multiple observations collected at those sampling events. The T-S test in ProUCL provides the option of averaging multiple measurements collected at the various sampling events. This option also computes M-K test and OLS regression statistics using the averages of multiple measurements collected at the various sampling event.

Note: The OLS regression and M-K test can be performed on data sets with multiple measurements taken at the various sampling time events. However, often it is desirable to use the averages (or median) of measurements taken at the various sampling events to determine potential trends present in a time-series data set.

10.2.2.1 Step-by-Step Procedure to Compute Theil-Sen Slope

Step 1. Order the data set by sampling event or time of collection of those measurements. Let $y_1, y_2, \ldots, y_n$ represent ordered measurement values. Consider all possible distinct pairs of measurements, (y_i, y_j) for $j > i$. For each pair, compute the simple pairwise slope estimate given by:

$$m_{ij} = \frac{(y_j - y_i)}{j - i} \; for \; j > i$$

For a time-series data set of size n, there are $N=n(n-1)/2$ such pairwise slope estimates, m_{ij}. If a given observation is a ND, one may use half of the DL or the RL as its estimated concentration. Alternatively, depending upon the distribution of detected values (also called the censored data set), the users may want to use imputed estimates of ND values obtained using the GROS or LROS method.

Step 2. Order the N pairwise slope estimates, m_{ij} from the smallest to the largest and re-label them as $m(1)$, $m(2),\ldots, m(N)$. Determine the T-S estimate of slope, Q, as the median value of this set of N ordered slopes. Computation of the median slope depends on whether N is even or odd. The median slope is computed using the following algorithm:

$$Q = \begin{cases} m_{([N+1]/2)} \; if \; N = odd \\ \left(m_{([N/2])} + m_{([N/2]/2)}\right)/_2 \; if \; N = even \end{cases} \qquad (10\text{-}19)$$

Step 3. Arrange the n measurements in ascending order from smallest to the largest value: $y(1), y(2),\ldots, y(n)$. Determine the median measurement using the following algorithm:

$$\bar{y} = \begin{cases} y_{([n+1]/2)} \; if \; n = odd \\ \left(y_{([n/2])} + y_{([n/2]/2)}\right)/_2 \; if \; n = even \end{cases} \quad (10\text{-}20)$$

Similarly, compute the median time, $\tilde{t}$ of the *n* ordered sampling times: t_1, t_2, to t_n by using the same median computation algorithm as used in (10-19) and (10-20).

Step 4. Compute the T-S trend line using the following equation:

$$y = \tilde{y} + Q(t - \tilde{t}) = (\tilde{y} - Q\tilde{t}) + Qt$$

10.2.2.2 Large Sample Inference for Theil – Sen Test Based upon Normal Approximation

As described in Step 2 above, order the *N* pairwise slope estimates, m_{ij} in ascending order from smallest to the largest: *m*(1), *m*(2),…, *m*(N). Compute *S* given in (10-18) and its *sd* given below:

$$sd(S) = \sqrt{\frac{1}{18}\left[n(n-1)(2n+5) - \sum_{j=1}^{g} t_j(t_j-1)(2t_j+5)\right]} \quad (10\text{-}21)$$

ProUCL can be used to test the following hypotheses:

H_0: Data set does not exhibit sufficient evidence of any trends (stationary measurements) vs.

H_A: Data set exhibits a trend (two-sided alternative)

H_A: Data set exhibits an upward trend; or

H_A: Data set exhibits a downward trend.

Case I. Testing for the null hypothesis, *H_0: Time series data set does not exhibit any trend*, vs. the two-sided alternative hypothesis, *H_A: Data Set exhibits a trend*.

Compute the critical value, C_α using the following equation:

$$C_\alpha = Z\alpha_{/2} sd(S)$$

Compute M_1 and M_2 as:

$$M_1 = \left[\frac{N - C_\alpha}{2}\right]; \quad \text{and} \quad M_2 = \left[\frac{N + C_\alpha}{2}\right]$$

Obtain the M_1^{th} largest and M_2^{th} largest slopes, $(m_{(M_1)})$ and $(m_{(M_2)})$, from the set consisting of all *n(n-1)/2* slopes. Then the probability of the T-S slope, Q, lying between these two slopes is given by the statement:

$$P(m_{(M_1)} < Q < m_{(M_2)}) = 1 - \alpha$$

On ProUCL output, $(m_{(M_1)})$ is labeled as LCL and $(m_{(M_2)})$ is labeled as UCL.

Conclusion: If 0 belongs to the interval, $(m_{(M_1)}, m_{(M_2)})$, conclude that T-S test slope is insignificant; that is, conclude that there is no significant trend present in the time series data set.

Cases II and III: Test for an upward (downward) trend with Null hypothesis, H_0: *Time series data set does not exhibit any trend*, vs. the alternative hypothesis, H_A: *data set exhibits an upward (downward) trend*.

For specified level of significance, α, compute the following:

$$C_\alpha = Z_\alpha * sd(S)$$

$$M_1 = \left[\frac{N - C_\alpha}{2}\right] \text{ and } M_2 = \left[\frac{N + C_\alpha}{2}\right]$$

Obtain the M_1^{th} largest and M_2^{th} largest slopes, $(m_{(M_1)})$ and $(m_{(M_2)})$ from the set consisting of all *n(n-1)/2* slopes.

Conclusion:

If $(m_{(M_1)}) > 0$, then the data set exhibits a significant upward trend.

If $(m_{(M_2)}) < 0$, then the data set exhibits a significant downward trend.

Example 10-3. Time series data (time event, concentration) were collected from several groundwater MWs on a Superfund site. The objective is to determine potential trends present in concentration data collected quarterly from those wells over a period of time. Some missing sampling events (quarters) are also present. ProUCL handles the missing values, computes trend test statistics and generates a time series graph along with the OLS and T-S lines.

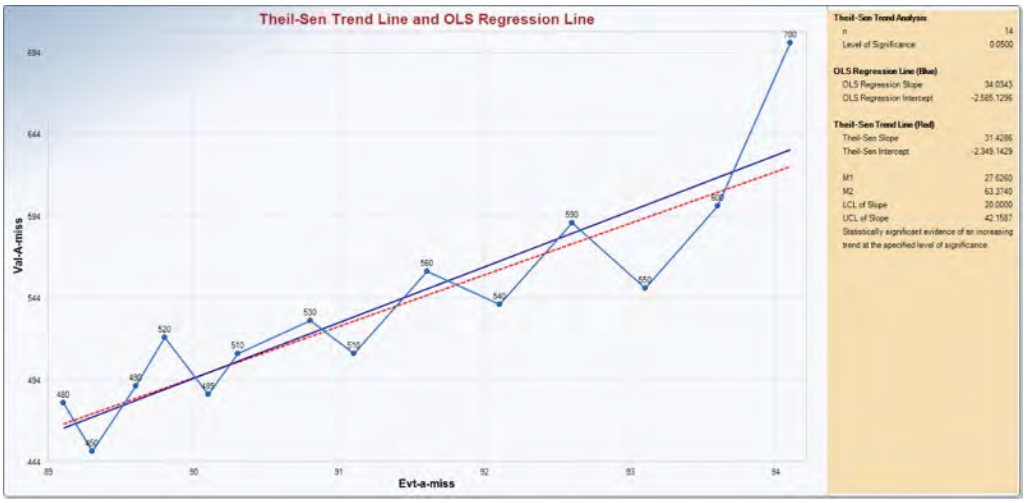

Figure 10-3. Time Series Plot and OLS and Theil-Sen Results with Missing Values

Table 10-3. The Excel output sheet, generated by ProUCL and showing all relevant results.

Val-A-miss	
General Statistics	
Number of Events	14
Number Values Observations	16
Number Values Missing	2
Number Values Reported (n)	14
Minimum	450
Maximum	700
Mean	536.8
Geometric Mean	533.6
Median	525
Standard Deviation	62.99

Approximate inference for Theil-Sen Trend Test	
Mann-Kendall Statistic (S)	72
Standard Deviation of S	18.24
Standardized Value of S	3.893
Approximate p-value	4.9562E-5
Number of Slopes	91
Theil-Sen Slope	31.43
Theil-Sen Intercept	-2349
M1'	30.5
One-sided 95% lower limit of Slope	21.36
95% LCL of Slope (0.025)	20
95% UCL of Slope (0.975)	42.16
Statistically significant evidence of an increasing trend at the specified level of significance.	

Notes: As with other statistical tests (e.g., Shapiro-Wilk and Lilliefors GOF tests for normality), it is very likely, that based upon a given data set, the three trend tests described here will lead to different trend conclusions. It is important that the user verifies the underlying assumptions required by these tests (e.g., normality of OLS residuals). A parametric OLS slope test is preferred when the underlying assumptions are met. Conclusions derived using nonparametric tests supplemented with graphical displays are preferred when OLS residuals are not normally distributed. These tests can also yield different results when the data set consists of missing values and/or there are gaps in the time series data set. It should be pointed out that an OLS line (therefore slope) can become significant even by the inclusion of an extreme value (e.g., collected after skipping of several intermediate sampling events) extending the domain of the sampling events time interval. For example, a perfect OLS line can be generated using two points at two extreme ends resulting in a significant slope; whereas nonparametric trend tests are not as influenced by such irregularities in the data collection and sampling events. In such circumstances, the user should draw a conclusion based upon the site CSM, expert and historical site knowledge and expert opinions.

10.3 Multiple Time Series Plots

The **Time Series Plot** option of the **Trend Analysis** module can generate time series plots for multiple groups/wells comparing concentration levels of those groups over a period of time. Time series plots are also useful for comparing concentrations of a MW during multiple periods (every 2 years, 5 years, ...) collected quarterly, semi-annually. This option can also handle missing sampling events. However, the number of observations in each group should be the same, sharing the same time event variable (if provided). An example time series plot comparing concentrations of three MWs during the same period of time is shown as follows.

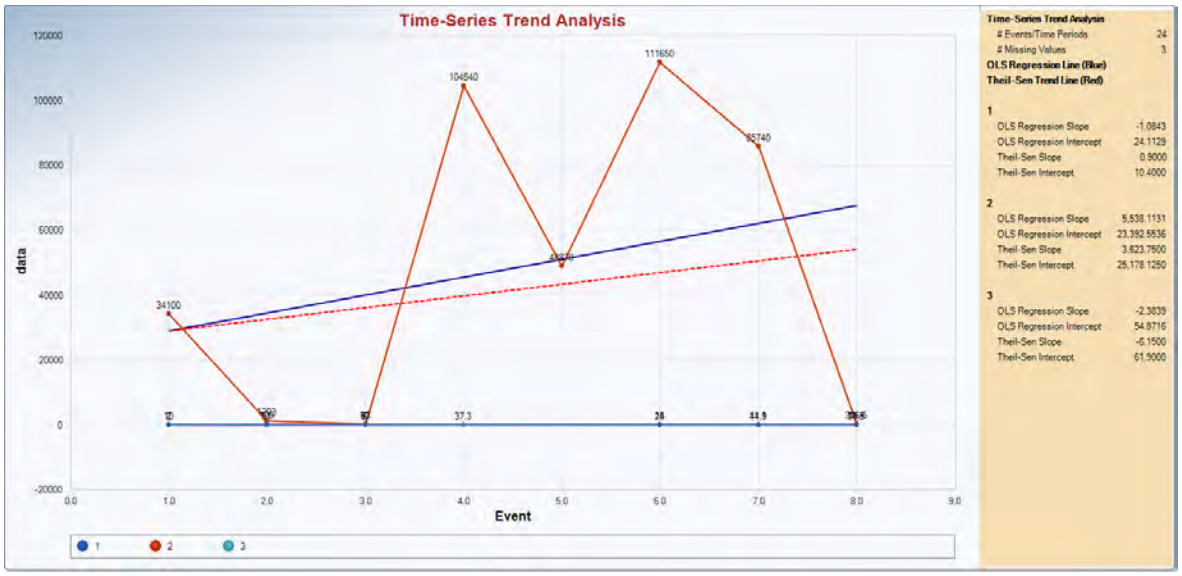

Figure 10-4. Time Series Plot Comparing Concentrations of Multiple Wells over a Period of Time

This option is specifically useful when the user wants to compare the concentrations of multiple groups (wells) and the exact sampling event dates are not available (data only option). The user may just want to graphically compare the time-series data collected from multiple groups/wells during several quarters (every year, every 5 years, …). Each group (e.g., well) defined by a group variable must have the same number of observations and should share the same sampling event values (when available). That is the number of sampling events and values (e.g., quarter ID, year ID, etc.) for each group (well) must be the same for this option to work. However, the exact sampling dates (not needed to use this option) in the various quarters (years) do not have to be the same as long as the values of the sampling quarters (1,3,5,6,7,9, etc.) used in generating the time-series plots for the various groups (wells) match. Using the geological and hydrological information, this kind of comparison may help the project team in identifying non-compliance wells (e.g., with upward trends in constituent concentrations) and associated reasons.

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 Engelhardt, M. 1991. *Statistical Analysis of Reliability and Life Testing Models*, Theory and Methods. 2nd Edition. Dekker, New York.

Bain, L.J. and Engelhardt, M. 1992. *Introduction to probability and Mathematical Statistics*. Second Edition. Duxbury Press, California.

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.

Best, D.J. 1983. A note on gamma variate generators with shape parameters less than unity**.** Computing, 30(2):185-188, 1983.

Blackwood, L. G. 1991. *Assurance Levels of Standard Sample Size Formulas*, Environmental Science and Technology, Vol. 25, No. 8, pp. 1366-1367.

Blom, G. 1958. Statistical Estimates and Transformed Beta Variables. John Wiley and Sons, New York.

Bowman, K. O. and Shenton, L.R. 1988. *Properties of Estimators for the Gamma Distribution*, Volume 89. Marcel Dekker, Inc., New York.

Bradu, D. and Mundlak, Y. 1970. *Estimation in Lognormal Linear Models*. Journal of the American Statistical Association, 65, 198-211.

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.

Conover W.J.. 1999. *Practical Nonparametric Statistics*, 3rd Edition, John Wiley & Sons, New York.

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.

David, H.A. and Nagaraja, H.N. 2003. *Order Statistics*. Third Edition. John Wiley.

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.

Draper, N.R. and Smith, H. 1998. Applied Regression Analysis (3rd Edition). New York: John Wiley & Sons.

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.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** (2): 179–188.

Flagg, K., M. Fitzgerald, M. Higgs, and P. Black. 2017. *UCL Estimators and considerations of sampling distribution, bias, and variability* [Unpublished manuscript]. Neptune and Company, Inc., June 2017.

Fleischhauer, H. and Korte, N. 1990. *Formation of Cleanup Standards Trace Elements with Probability Plot.* Environmental Management, Vol. 14, No. 1. 95-105.

Frost, J. 2018. *Central Limit Theorem Explained*. Retrieved from https://statisticsbyjim.com/basics/central-limit-theorem/.

Gehan, E.A. 1965. A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Sample. Biometrika 52, 203-223.

Gerlach, R. W., and J. M. Nocerino. 2003. Guidance for Obtaining Representative Laboratory Analytical Subsamples from Particulate Laboratory Samples. EPA/600/R-03/027. www.epa.gov/esd/tsc/images/particulate.pdf.

Gibbons. 1994. Statistical Methods for Groundwater Monitoring. John Wiley &Sons.

Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, New York.

Gilliespie, B.W., Chen, Q., Reichert H., Franzblau A., Hedgeman E., Lepkowski J., Adriaens P., Demond A., Luksemburg W., and Garabrant DH. 2010. *Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator*. Epidemiology, Vol. 21, No. 4.

Gleit, A. 1985. *Estimation for Small Normal Data Sets with Detection Limits*. Environmental Science and Technolog*y, 19,* pp. 1206-1213, 1985.

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.

Gu, M.G., and Zhang, C.H. 1993. Asymptotic properties of self-consistent estimators based on doubly censored data. Annals of Statistics. Vol. 21, 611-624.

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.

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.

Hawkins, D. M., and Wixley, R. A. J. 1986. *A Note on the Transformation of Chi-Squared Variables to Normality.* The American Statistician, 40, 296–298.

Hayes, A. F. 2005. *Statistical Methods for Communication Science,* Lawrence Erlbaum Associates, Publishers.

Helsel, D.R. 2005. *Nondetects and Data Analysis*. Statistics for Censored Environmental Data. John Wiley and Sons, NY.

Helsel, D.R. 2102a. *Practical Stats Webinar on ProUCL v4.* The Unofficial User Guide; October 15, 2012.

Helsel, D.R. 2012b. Statistics for Censored Environmental Data Using Minitab and R. Second Edition. John Wiley and Sons, NY.

Helsel, D.R. 2013. Nondetects and Data Analysis for Environmental Data, NADA in R

Helsel, D.R. and E. J. Gilroy. 2012. *The Unofficial Users Guide to ProUCL4.* Amazon, Kindle Edition.

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.

Holgresson, M. and Jorner U. 1978. *Decomposition of a Mixture into Normal Components: a Review.* Journal of Bio-Medicine. Vol. 9. 367-392.

Hollander M & Wolfe DA (1999). *Nonparametric Statistical Methods (2nd Edition)*. New York: John Wiley & Sons.

Hogg, R.V. and Craig, A. 1995. *Introduction to Mathematical Statistics*; 5th edition. Macmillan.

Huber, P.J. 1981, *Robust Statistics*, John Wiley and Sons, NY.

Hyndman, R. J. and Fan, Y. 1996. *Sample quantiles in statistical packages*, American Statistician, **50**, 361–365.

Interstate Technology Regulatory Council (ITRC). 2012. *Incremental Sampling Methodology.* Technical and Regulatory Guidance, 2012.

Interstate Technology Regulatory Council (ITRC). 2013 *Groundwater Statistics and Monitoring Compliance.* Technical and Regulatory Guidance, December 2013.

Interstate Technology Regulatory Council (ITRC). 2015. *Decision Making at Contaminated Sites.* Technical and Regulatory Guidance, January 2015.

Interstate Technology Regulatory Council (ITRC). 2020. *Incremental Sampling Methodology (ISM) Update*, October 2020.

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 D. Wichern. 2002. *Applied Multivariate Statistical Analysis*. 6th Edition. 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.

Krishnamoorthy, K., Mathew, T., and Mukherjee, S. 2008. Normal distribution based methods for a Gamma distribution: Prediction and Tolerance Interval and stress-strength reliability. Technometrics, 50, 69-78.

Kroese, D.P., Taimre, T., and Botev Z.I. 2011. *Handbook of Monte Carlo Methods.* John Wiley & Sons.

Kruskal, W. H., and Wallis, A. 1952. *Use of ranks in one-criterion variance analysis.* Journal of the American Statistical Association, 47, 583-621.

Kupper, L. L. and Hafner, K. B. 1989, *How Appropriate Are Popular Sample Size Formulas?* The American Statistician, Vol. 43, No. 2, pp. 101-105

Kunter, M. J., C. J. Nachtsheim, J. Neter, and Li W. 2004. *Applied Linear Statistical Methods.* Fifth Edition. McGraw-Hill/Irwin.

Laga, J., and Likes, J. 1975, *Sample Sizes for Distribution-Free Tolerance Intervals* Statistical Papers. Vol. 16, No. 1. 39-56

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.

Levene, Howard. 1960. *Robust tests for equality of variances.* In Olkin, Harold, et alia. Stanford University Press. pp.278–292.

Lilliefors, H.W. 1967. *On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown.* Journal of the American Statistical Association, 62, 399-404.

Looney and Gulledge. 1985. Use of the Correlation Coefficient with Normal Probability Plots. The American Statistician, 75-79.

Manly, B.F.J. 1997. Randomization, Bootstrap, and Monte Carlo Methods in Biology. Second Edition. Chapman Hall, London.

Maronna, R.A., Martin, R.D., and Yohai, V.J. 2006, *Robust Statistics: Theory and Methods*, John Wiley and Sons, Hoboken, NJ.

Marsaglia, G. and Tsang, W. 2000. *A simple method for generating gamma variables.* ACM Transactions on Mathematical Software, 26(3):363-372.

Millard, S. P. and Deverel, S. J. 1988. Nonparametric statistical methods for comparing two sites based on data sets with multiple nondetect limits. Water Resources Research, 24, pp. 2087-2098.

Millard, S.P. and Neerchal, M.K. 2002. *Environmental Stats for S-PLUS*. Second Edition. Springer.

Minitab version 16. 2012. *Statistical Software.*

Molin, P., and Abdi H. 1998. *New Tables and numerical approximations for the Kolmogorov-Smirnov/Lilliefors/ Van Soest's test of normality*. In Encyclopedia of Measurement and Statistics, Neil Salkind (Editor, 2007). Sage Publication Inc. Thousand Oaks (CA).

Natrella, M.G. 1963. *Experimental Statistics*. National Bureau of Standards, Hand Book No. 91, U.S. Government Printing Office, Washington, DC.

Noether, G.E. 1987 *Sample Size Determination for some Common Nonparametric Tests*, Journal American Statistical Assoc., 82, 645-647

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.

R Core Team, 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rosner, B. 1975. *On the detection of many outliers*. Technometrics, 17, 221-227.

Rosner, B. 1983. Percentage points for a generalized ESD many-outlier procedure. Technometrics, 25, 165-172.

Rousseeuw, P.J. and Leroy, A.M. 1987. *Robust Regression and Outlier Detection*. John Wiley.

Royston, P. 1982a. *Algorithm AS 181: The W test for Normality*. Applied Statistics, 31, 176–180.

Royston, P. 1982. An extension of Shapiro and Wilk's W test for normality to large samples. Applied Statistics, 31, 115–124.

Shacklette, H.T, and Boerngen, J.G. 1984. Element Concentrations in Soils and Other Surficial Materials in the Conterminous United States, U.S. Geological Survey Professional Paper 1270.

Scheffe, H., and Tukey, J.W. 1944. *A formula for Sample Sizes for Population Tolerance Limits*. The Annals of Mathematical Statistics. Vol 15, 217.

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.

Scheffe, H., and Tukey, J.W. 1944. *A formula for Sample Sizes for Population Tolerance Limits*. The Annals of Mathematical Statistics. Vol 15, 217.

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, B. E. 1978. *Kolmogorov-Smirnov Test Statistic for the Gamma Distribution with Unknown Parameters*, Dissertation, Department of Statistics, Temple University, Philadelphia, PA.

Schneider, H. 1986. *Truncated and Censored Samples from Normal Populations*. Vol. 70, Marcel Dekker Inc., New York, 1986.

She, N. 1997. *Analyzing Censored Water Quality Data Using a Nonparametric Approach.* Journal of the American Water Resources Association 33, pp. 615-624.

Shea, B. 1988. Algorithm AS 239: Chi-square and Incomplete Gamma Integrals. Applied Statistics, 37: 466-473.

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.

Sinclair, A.J. 1976. *Applications of Probability Graphs in Mineral Exploration.* Association of Exploration Geochemists, Rexdale Ontario, p 95.

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. 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., and Lee, S. 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/600/R-06/022, March 2006. http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

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. and Nocerino, J.M. 1997. *Robust Intervals for Some Environmental Applications."* The Journal of Chemometrics and Intelligent Laboratory Systems, Vol 37, 55-69.

Singh, A. and Nocerino, J.M. 2002. Robust Estimation of the Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations, Vol. 60, pp 69-86.

Singh, A.K. and Ananda. M. 2002. Rank kriging for characterization of mercury contamination at the East Fork Poplar Creek, Oak Ridge, Tennessee. Environmetrics, Vol. 13, pp. 679-691.

Singh, A. and Singh, A.K. 2007. *ProUCL Version 4 Technical Guide (Draft)*. Publication EPA/600/R-07/041. January, 2007. http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

Singh, A. and Singh, A.K. 2009. *ProUCL Version 4.00.04 Technical Guide (Draft)*. Publication EPA/600/R-07/041. February, 2009. http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

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., Singh A.K., and Engelhardt, M. 1999, Some Practical Aspects of sample Size and Power Computations for Estimating the Mean of Positively Skewed Distributions in Environmental Applications. Office of Research and Development. EPA/006/s-99/006. November 1999.
http://www.epa.gov/esd/tsc/images/325cmb99rpt.pdf

Singh, A., Singh, A.K., and Flatman, G. 1994. *Estimation of Background Levels of Contaminants*. Math Geology, Vol. 26, No, 3, 361-388.

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.

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.

Tarone, R. and Ware, J. 1978. On Distribution-free Tests for Equality of Survival Distributions. Biometrika, 64, 156-160.

Thom, H.C.S. 1968. *Direct and Inverse Tables of the Gamma Distribution*. Silver Spring, MD; Environmental Data Service.

U.S. Environmental Protection Agency (EPA). 1989a. 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). 1989b. *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). 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). 1992a. *Supplemental Guidance to RAGS: Calculating the Concentration Term*. Publication EPA 9285.7-081, May 1992.

U.S. Environmental Protection Agency (EPA). 1992b. *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). 1994. *Statistical Methods for Evaluating the Attainment of Cleanup Standards*, EPA 230-R-94-004, Washington, DC.

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). MARSSIM. 2000. U.S. Nuclear Regulatory Commission, *et al. 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). 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). 2002c. RCRA Waste Sampling, Draft Technical Guidance – Planning, Implementation and Assessment. EPA 530-D-02-002, 2002.

U.S. Environmental Protection Agency (EPA). 2004. *ProUCL Version 3.1, Statistical Software*. National Exposure Research Lab, EPA, Las Vegas Nevada, October 2004.
http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

U.S. Environmental Protection Agency (EPA). 2006a, *Guidance on Systematic Planning Using the Data Quality Objective Process*, EPA QA/G-4, EPA/240/B-06/001. Office of Environmental Information, Washington, DC. Download from: http://www.epa.gov/quality/qs-docs/g4-final.pdf

U.S. Environmental Protection Agency (EPA). 2006b. *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

U.S. Environmental Protection Agency (EPA). 2007. *ProUCL Version 4.0 Technical Guide*. EPA 600-R-07-041, January 2007.

U.S. Environmental Protection Agency (EPA). 2009a. ProUCL Version 4.00.05 User Guide (Draft). Statistical Software for Environmental Applications for Data Sets with and without nondetect observations. National Exposure Research Lab, EPA, Las Vegas. EPA/600/R-07/038, February 2009. Down load from: http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

U.S. Environmental Protection Agency (EPA). 2009b. ProUCL Version 4.00.05 Technical Guide (Draft). Statistical Software for Environmental Applications for Data Sets with and without nondetect observations. National Exposure Research Lab, EPA, Las Vegas. EPA/600/R-07/038, February 2009. Down load from: http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

U.S. Environmental Protection Agency (EPA). 2009c. ProUCL4.00.05 Facts Sheet. Statistical Software for Environmental Applications for Data Sets with and without nondetect observations. National Exposure Research Lab, EPA, Las Vegas, Nevada, 2009.http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

U.S. Environmental Protection Agency (EPA). 2009d. *Scout 2008 – A Robust Statistical Package*, Office of Research and Development, February 2009. http://archive.epa.gov/esd/archive-scout/web/html/

U.S. Environmental Protection Agency (EPA). 2009e. Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities – Unified Guidance. EPA 530-R-09-007, 2009.

U.S. Environmental Protection Agency (EPA). 2010a. *A Quick Guide to the Procedures in Scout (Draft),* Office of Research and Development, April 2010. http://archive.epa.gov/esd/archive-scout/web/html/

U.S. Environmental Protection Agency (EPA). 2010b. *ProUCL Version 4.00.05 User Guide (Draft)*. EPA/600/R-07/041, May 2010. http://www.epa.gov/osp/hstl/tsc/software.htm

U.S. Environmental Protection Agency (EPA). 2010c. ProUCL Version 4.00.05 Technical Guide (Draft). EPA/600/R-07/041, May, 2010. http://www.epa.gov/osp/hstl/tsc/software.htm

U.S. Environmental Protection Agency (EPA). 2010d. *ProUCL 4.00.05, Statistical Software for Environmental Applications for Data Sets with and without nondetect observations*. National Exposure Research Lab, EPA, Las Vegas Nevada, May 2010. http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

U.S. Environmental Protection Agency (EPA). 2010e. *Scout 2008 User Guide (Draft)* EPA/600/R-08/038, Office of Research and Development, April 2010.http://archive.epa.gov/esd/archive-scout/web/html/

U.S. Environmental Protection Agency (EPA). 2011. *ProUCL 4.1.00, Statistical Software for Environmental Applications for Data Sets with and without nondetect observations*. National Exposure Research Lab, EPA, Las Vegas Nevada, June 2011. http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

U.S. Environmental Protection Agency (EPA). 2013a. *ProUCL 5.0.00 Technical Guide (Draft)* EPA/600/R-07/041. September 2013. Office of Research and Development. http://www.epa.gov/esd/tsc/TSC_form.htm

U.S. Environmental Protection Agency (EPA). 2013b. *ProUCL 5.0.00 User Guide (Draft)* EPA/600/R-07/041. September 2013. Office of Research and Development. http://www.epa.gov/esd/tsc/TSC_form.htm

U.S. Environmental Protection Agency (EPA). 2014. *ProUCL 5.0.00 Statistical Software for Environmental Applications for Datasets with and without Nondetect Observations,* Office of Research and Development, August 2014. http://www.epa.gov/esd/tsc/TSC_form.htm

Wald, A. 1943. *An Extension of Wilks' Method for Setting Tolerance Intervals.* Annals of Mathematical Statistics. Vol. 14, 44-55.

Whittaker, J. 1974. Generating Gamma and Beta Random Variables with Non-integral Shape Parameters. Applied Statistics, 23, No. 2, 210-214.

Wilks, S.S. 1941. *Determination of Sample Sizes for Setting Tolerance Limits.* Annals of Mathematical Statistics, Vol. 12, 91-96.

Wilks, S.S. 1963. *Multivariate statistical outliers*. Sankhya A, 25: 407-426.

Wilson, E.B., and Hilferty, M.M. 1931, "*The Distribution of Chi-Squares*," Proceedings of the National Academy of Sciences, 17, 684–688.

Wong, A. 1993. A Note on Inference for the Mean Parameter of the Gamma Distribution. Statistics Probability Letters, Vol. 17, 61-66.

APPENDIX A

Simulated Critical Values for Gamma GOF Tests, the Anderson-Darling Test and the Kolmogorov-Smirnov Test

Critical Values of Gamma GOF Test Statistics

For values of the gamma distribution shape parameter, $k \leq 0.2$, critical values of the two gamma empirical distribution tests (EDF) GOF tests: Anderson-Darling (A-D) and Kolmogorov Smirnov (K-S) tests incorporated in ProUCL 4.1 and earlier versions have been updated in ProUCL 5.0. Critical values incorporated in earlier versions of ProUCL were simulated using the gamma deviate generation algorithm (Whittaker 1974) available at the time and with the source code provided in the book *Numerical Recipes in C, the Art of Scientific Computing* (Press *et al.* 1990). It is noted that the gamma deviate generation

algorithm available at the time was not very efficient, especially for smaller values of the shape parameter, $k \leq 0.1$. For small values of the shape parameter, k, significant discrepancies were found in the critical values of the two gamma GOF test statistics obtained using the two gamma deviate generation algorithms: Whitaker (1974) and Marsaglia and Tsang (2000).

Even though, discrepancies were identified in critical values of the two GOF tests for value of $k \leq 0.1$, for comparison purposes, critical values of the two tests have also been re-generated for $k=0.2$. For values of $k \leq 0.2$, critical values for the two gammas EDF GOF tests have been re-generated and tables of critical values of the two gamma GOF tests have been updated in this Appendix A. Specifically, for values of the shape parameter, k (e.g., $k \leq 0.2$), critical values of the two gamma GOF tests have been generated using the more efficient gamma deviate generation algorithm as described in Marsaglia and Tsang (2000) and Best (1983). Detailed description about the implementation of Marsaglia and Tsang's algorithm to generate gamma deviates can be found in Kroese, Taimre, and Botev (2011). It is noted that for values of $k > 0.1$, the simulated critical values obtained using Marsaglia and Tsang's algorithm (2000) are in general agreement with the critical values of the two GOF test statistics incorporated in ProUCL 4.1 for the various values of the sample size considered. Therefore, those critical values for values of $k > 0.2$ have not been updated in tables as summarized in this Appendix A. The developers double checked the critical values of the two GOF tests by using MatLab to generate gamma deviates. Critical values obtained using MatLab code are in general agreement with the newly simulated critical values incorporated in critical value tables summarized in this appendix.

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. Here the notation n=5(25)1 means that n takes values starting at 5 all the way up to 25 at increments of 1 each; n=30(50)5 means that n takes values starting at 30 all the way up to 50 at increments of 5 each, and so on. Random deviates of sample size n were generated from a gamma, (k, θ), population. The considered values of the shape parameter, k, are: 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.

Step 1. Generate a random sample of the specified size, n, from a gamma, G (k, 1), distribution. For values of $k>0.2$, the algorithm as outlined in Whittaker (1974) was used to generate the gamma deviates; and for values of $k \leq 0.2$, Marsaglia and Tsang's algorithm (2000) has been used to generate 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; Schneider and Clickner 1976) using the incomplete gamma function (details can be found in Chapter 2 of this document).

Step 3. Repeat Steps 1 and 2, a large number (iterations) of times. For values of $k>0.2$, 20,000 iterations were used to compute critical values. However, since generation of gamma deviates are quite unstable for

smaller values of k ($\leq$0.1), 500,000 iterations have been used to obtain the newly generated critical values of the two test statistics based upon Marsaglia and Tsang's algorithm.

Step 4. Arrange the resulting test statistics in ascending order. Compute the 90%, 95%, and 99% percentiles of the K-S test statistic and the A-D test statistic.

The resulting raw 10%, 5%, and 1% critical values for the two tests are summarized in Tables 1 through 6 as follows. The critical values as summarized in Tables 1-6 are in agreement (up to 3 significant digits) with all available exact or asymptotic critical values (note that critical values of the two GOF tests are not available for values of k<1). 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. 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.

Note on computation of the gamma distribution based decision statistics and critical values: While computing the various decision statistics (e.g., UCL and BTVs), ProUCL uses biased corrected estimates, kstar, $\hat{k}^*$, and theta star, $\hat{\theta}^*$ (described in Section 2.3.3) of the shape, k, and scale, θ, parameters of the gamma distribution. It is noted that the critical values for the two gamma GOF tests reported in the literature (D'Agostino and Stephens 1986; Schneider and Clickner 1976; Schneider 1978) were computed using the MLE estimates, $\hat{k}$ and $\hat{\theta}$, of the two gamma parameters, k and θ. Therefore, the critical values of A-D and K-S tests incorporated in ProUCL have also been computed using the MLE estimates: khat, $\hat{k}$, and theta hat, $\hat{\theta}$, of the two gamma parameters, k and θ.

Table A-1. Critical Values for A-D Test Statistic for Significance Level = 0.10

n\k	0.025	0.05	0.1	0.2	0.5	1	2	5	10	20	50
5	0.919726	0.802558	0.715363	0.655580	0.612	0.599	0.594	0.591	0.589	0.589	0.588
6	0.923855	0.819622	0.735533	0.670716	0.625	0.61	0.603	0.599	0.599	0.598	0.598
7	0.924777	0.829767	0.746369	0.684718	0.635	0.618	0.609	0.607	0.606	0.604	0.605
8	0.928382	0.834365	0.758146	0.694671	0.641	0.624	0.616	0.612	0.61	0.609	0.608
9	0.928959	0.840361	0.765446	0.701756	0.648	0.629	0.62	0.614	0.613	0.613	0.612
10	0.930055	0.847992	0.771909	0.707396	0.652	0.632	0.623	0.618	0.616	0.615	0.614
15	0.934218	0.864609	0.792009	0.727067	0.663	0.642	0.63	0.624	0.622	0.621	0.621
16	0.934888	0.866151	0.795984	0.727392	0.665	0.642	0.632	0.626	0.624	0.622	0.621
17	0.935586	0.866978	0.796929	0.729339	0.666	0.644	0.632	0.626	0.623	0.623	0.622
18	0.936246	0.869658	0.799900	0.731904	0.668	0.643	0.634	0.626	0.623	0.624	0.623
19	0.937456	0.870368	0.800417	0.732093	0.67	0.645	0.633	0.626	0.625	0.624	0.624
20	0.937518	0.871858	0.801716	0.733548	0.669	0.645	0.633	0.627	0.626	0.624	0.624
21	0.937751	0.874119	0.803861	0.735995	0.671	0.646	0.634	0.628	0.626	0.626	0.624
22	0.938503	0.874483	0.804803	0.736736	0.67	0.646	0.636	0.628	0.627	0.625	0.625
23	0.938587	0.875008	0.805412	0.737239	0.671	0.645	0.635	0.629	0.627	0.625	0.625
24	0.939277	0.875990	0.806629	0.738236	0.672	0.647	0.635	0.628	0.627	0.626	0.625
25	0.940150	0.876204	0.807918	0.738591	0.673	0.648	0.636	0.629	0.627	0.626	0.625
30	0.941743	0.882689	0.811964	0.741572	0.674	0.65	0.637	0.629	0.628	0.627	0.626
35	0.943737	0.885557	0.814862	0.743736	0.676	0.65	0.638	0.631	0.629	0.628	0.627
40	0.945107	0.885878	0.817072	0.747438	0.677	0.651	0.637	0.631	0.629	0.628	0.628
45	0.947909	0.887142	0.817778	0.748890	0.677	0.651	0.639	0.632	0.63	0.628	0.629
50	0.947922	0.887286	0.818568	0.749399	0.677	0.652	0.64	0.632	0.63	0.629	0.629
60	0.948128	0.890153	0.820774	0.749930	0.679	0.652	0.64	0.632	0.631	0.629	0.629
70	0.948223	0.891061	0.822280	0.750605	0.679	0.653	0.641	0.633	0.63	0.63	0.63
80	0.949613	0.891764	0.823067	0.751452	0.68	0.654	0.641	0.633	0.631	0.63	0.629
90	0.951013	0.892197	0.823429	0.752461	0.68	0.654	0.642	0.634	0.631	0.629	0.63
100	0.951781	0.892833	0.824216	0.752765	0.681	0.654	0.642	0.633	0.631	0.63	0.63
200	0.952429	0.893123	0.826133	0.753696	0.682	0.654	0.642	0.634	0.631	0.631	0.63
300	0.953464	0.893406	0.826715	0.754433	0.682	0.655	0.641	0.634	0.633	0.631	0.63
400	0.955133	0.898383	0.827845	0.755130	0.683	0.655	0.641	0.635	0.633	0.631	0.631
500	0.956040	0.898554	0.827995	0.755946	0.683	0.655	0.643	0.635	0.632	0.631	0.631
1000	0.957279	0.898937	0.828584	0.757750	0.684	0.655	0.643	0.635	0.632	0.631	0.63

Table A-2. Critical Values for K-S Test Statistic for Significance Level = 0.10

n\k	0.025	0.050	0.10	0.2	0.50	1.0	2.0	5.0	10.0	20.0	50.0
5	0.382954	0.377607	0.370075	0.358618	0.346	0.339	0.336	0.334	0.333	0.333	0.333
6	0.359913	0.352996	0.343783	0.332729	0.319	0.313	0.31	0.307	0.307	0.307	0.307
7	0.336053	0.329477	0.321855	0.312905	0.301	0.294	0.29	0.288	0.288	0.287	0.287
8	0.315927	0.312018	0.305500	0.295750	0.284	0.278	0.274	0.272	0.271	0.271	0.271
9	0.300867	0.296565	0.290030	0.280550	0.27	0.264	0.26	0.258	0.257	0.257	0.257
10	0.286755	0.283476	0.276246	0.268807	0.257	0.251	0.248	0.246	0.245	0.245	0.245
15	0.238755	0.237248	0.231259	0.223045	0.214	0.209	0.206	0.204	0.204	0.203	0.203
16	0.232063	0.228963	0.224049	0.216626	0.208	0.203	0.2	0.198	0.198	0.197	0.197
17	0.225072	0.222829	0.218089	0.211438	0.202	0.197	0.194	0.193	0.192	0.192	0.192
18	0.218863	0.216723	0.212018	0.205572	0.197	0.192	0.189	0.188	0.187	0.187	0.187
19	0.213757	0.211493	0.206688	0.201002	0.192	0.187	0.184	0.183	0.182	0.182	0.182
20	0.209044	0.205869	0.202242	0.196004	0.187	0.183	0.18	0.179	0.178	0.178	0.178
21	0.204615	0.201904	0.197476	0.191444	0.183	0.179	0.176	0.175	0.174	0.174	0.174
22	0.199688	0.197629	0.193503	0.187686	0.179	0.175	0.172	0.171	0.17	0.17	0.17
23	0.195776	0.193173	0.188985	0.182952	0.175	0.171	0.169	0.167	0.167	0.166	0.166
24	0.192131	0.189663	0.185566	0.179881	0.172	0.168	0.165	0.164	0.163	0.163	0.163
25	0.188048	0.185450	0.181905	0.176186	0.169	0.165	0.162	0.161	0.16	0.16	0.16
30	0.172990	0.169910	0.166986	0.161481	0.155	0.151	0.149	0.147	0.147	0.147	0.147
35	0.160170	0.158322	0.155010	0.150173	0.144	0.14	0.138	0.137	0.136	0.136	0.136
40	0.150448	0.148475	0.145216	0.140819	0.135	0.132	0.13	0.128	0.128	0.128	0.128
45	0.142187	0.140171	0.137475	0.133398	0.127	0.124	0.122	0.121	0.121	0.121	0.121
50	0.135132	0.133619	0.130496	0.126836	0.121	0.118	0.116	0.115	0.115	0.115	0.115
60	0.123535	0.122107	0.119488	0.116212	0.111	0.108	0.107	0.106	0.105	0.105	0.105
70	0.114659	0.113414	0.110949	0.107529	0.103	0.1	0.099	0.098	0.098	0.097	0.097
80	0.107576	0.106191	0.104090	0.100923	0.096	0.094	0.093	0.092	0.092	0.091	0.091
90	0.101373	0.100267	0.097963	0.095191	0.091	0.089	0.088	0.087	0.086	0.086	0.086
100	0.096533	0.095061	0.093359	0.090566	0.086	0.084	0.083	0.082	0.082	0.082	0.082
200	0.068958	0.067898	0.066258	0.064542	0.062	0.06	0.059	0.059	0.058	0.058	0.058
300	0.056122	0.055572	0.054295	0.052716	0.05	0.049	0.048	0.048	0.048	0.048	0.048
400	0.048635	0.048048	0.047103	0.045745	0.044	0.043	0.042	0.042	0.042	0.041	0.041
500	0.043530	0.042949	0.042053	0.040913	0.039	0.038	0.038	0.037	0.037	0.037	0.037
1000	0.030869	0.030621	0.029802	0.028999	0.028	0.027	0.027	0.026	0.026	0.026	0.026

Table A-3. Critical Values for A-D Test Statistic for Significance Level = 0.05

n\k	0.025	0.05	0.1	0.2	0.5	1	2	5	10	20	50
5	1.151052	0.993916	0.867326	0.775584	0.711	0.691	0.684	0.681	0.679	0.679	0.678
6	1.163733	1.015175	0.892648	0.801734	0.736	0.715	0.704	0.698	0.698	0.697	0.697
7	1.164504	1.027713	0.910212	0.822761	0.752	0.728	0.715	0.71	0.708	0.707	0.708
8	1.164753	1.033965	0.926242	0.835780	0.762	0.736	0.724	0.719	0.715	0.716	0.715
9	1.165715	1.039023	0.936047	0.847305	0.771	0.743	0.73	0.723	0.722	0.721	0.721
10	1.165767	1.051305	0.945231	0.855135	0.777	0.748	0.736	0.729	0.725	0.725	0.724
15	1.166499	1.072701	0.971851	0.883252	0.793	0.763	0.747	0.739	0.737	0.735	0.734
16	1.166685	1.072764	0.976822	0.883572	0.796	0.763	0.75	0.741	0.739	0.737	0.735
17	1.168544	1.074729	0.979261	0.885946	0.798	0.766	0.749	0.742	0.739	0.738	0.737
18	1.168987	1.076805	0.982322	0.889231	0.8	0.767	0.753	0.743	0.739	0.739	0.738
19	1.169801	1.078026	0.983408	0.891016	0.803	0.769	0.752	0.742	0.741	0.74	0.74
20	1.169916	1.080724	0.985352	0.892498	0.803	0.768	0.752	0.745	0.742	0.741	0.739
21	1.170231	1.082101	0.988749	0.895978	0.805	0.77	0.754	0.745	0.743	0.743	0.741
22	1.170651	1.083139	0.989794	0.896739	0.804	0.771	0.756	0.746	0.744	0.74	0.743
23	1.170815	1.084161	0.990147	0.897642	0.805	0.769	0.755	0.747	0.744	0.742	0.741
24	1.171897	1.085896	0.991640	0.898680	0.806	0.772	0.755	0.746	0.744	0.742	0.742
25	1.173062	1.086184	0.991848	0.899874	0.807	0.773	0.756	0.747	0.745	0.743	0.742
30	1.174361	1.095072	1.000576	0.903940	0.809	0.775	0.758	0.746	0.745	0.744	0.744
35	1.174900	1.095964	1.000838	0.907253	0.812	0.776	0.76	0.75	0.748	0.747	0.745
40	1.177053	1.097870	1.004925	0.909633	0.813	0.779	0.759	0.751	0.748	0.747	0.746
45	1.178564	1.099630	1.006416	0.911353	0.813	0.777	0.761	0.753	0.748	0.748	0.747
50	1.178640	1.100960	1.007896	0.912084	0.814	0.78	0.763	0.754	0.75	0.748	0.748
60	1.179045	1.103255	1.009514	0.914286	0.816	0.779	0.763	0.753	0.751	0.749	0.748
70	1.179960	1.105666	1.013808	0.914724	0.817	0.78	0.763	0.754	0.751	0.749	0.749
80	1.180934	1.106509	1.014011	0.914808	0.819	0.782	0.763	0.754	0.75	0.751	0.748
90	1.183445	1.106661	1.015090	0.915898	0.818	0.783	0.765	0.755	0.752	0.75	0.751
100	1.183507	1.107269	1.015433	0.917512	0.818	0.783	0.765	0.754	0.752	0.75	0.75
200	1.184370	1.108491	1.018998	0.920264	0.821	0.784	0.766	0.756	0.751	0.751	0.75
300	1.186474	1.112771	1.019934	0.920502	0.822	0.784	0.766	0.757	0.755	0.751	0.752
400	1.186711	1.113282	1.020022	0.920551	0.823	0.785	0.766	0.757	0.754	0.751	0.752
500	1.186903	1.114064	1.020267	0.921806	0.822	0.785	0.767	0.756	0.753	0.752	0.752
1000	1.188089	1.114697	1.020335	0.923848	0.824	0.785	0.768	0.757	0.753	0.752	0.75

Table A-4. Critical Values for K-S Test Statistic for Significance Level = 0.05

n\k	0.025	0.05	0.1	0.2	0.5	1	2	5	10	20	50
5	0.425015	0.416319	0.405292	0.388127	0.372	0.364	0.36	0.358	0.358	0.357	0.357
6	0.393430	0.384459	0.374897	0.364208	0.349	0.341	0.336	0.333	0.332	0.332	0.332
7	0.367179	0.361553	0.353471	0.342709	0.327	0.32	0.315	0.313	0.312	0.311	0.311
8	0.348874	0.342809	0.335397	0.323081	0.309	0.301	0.297	0.295	0.294	0.294	0.293
9	0.331231	0.325179	0.317725	0.308264	0.294	0.287	0.282	0.28	0.279	0.279	0.279
10	0.315236	0.311210	0.303682	0.294373	0.281	0.274	0.27	0.267	0.267	0.266	0.266
15	0.262979	0.260524	0.253994	0.245069	0.234	0.228	0.224	0.222	0.222	0.221	0.221
16	0.255659	0.251621	0.246493	0.238415	0.227	0.221	0.218	0.216	0.215	0.215	0.214
17	0.247795	0.244721	0.240192	0.231881	0.221	0.215	0.212	0.21	0.209	0.209	0.208
18	0.240719	0.237832	0.233566	0.226194	0.215	0.209	0.206	0.204	0.203	0.203	0.203
19	0.235887	0.232558	0.227223	0.220341	0.21	0.204	0.201	0.199	0.199	0.198	0.198
20	0.229517	0.227125	0.222103	0.214992	0.205	0.199	0.196	0.194	0.194	0.193	0.193
21	0.224925	0.221654	0.217434	0.209979	0.2	0.195	0.192	0.19	0.189	0.189	0.189
22	0.219973	0.217725	0.212415	0.205945	0.196	0.191	0.188	0.186	0.185	0.185	0.185
23	0.215140	0.212869	0.207622	0.201004	0.192	0.187	0.184	0.182	0.182	0.181	0.181
24	0.211022	0.208355	0.203870	0.197443	0.188	0.183	0.18	0.178	0.178	0.178	0.177
25	0.207233	0.204154	0.200009	0.193701	0.184	0.18	0.177	0.175	0.175	0.174	0.174
30	0.187026	0.187026	0.183312	0.177521	0.169	0.165	0.162	0.16	0.16	0.16	0.16
35	0.176132	0.174396	0.170208	0.165130	0.157	0.153	0.151	0.149	0.149	0.148	0.148
40	0.165449	0.163501	0.159727	0.154749	0.148	0.144	0.141	0.14	0.139	0.139	0.139
45	0.156286	0.154614	0.151477	0.146553	0.139	0.136	0.133	0.132	0.132	0.132	0.131
50	0.148646	0.146991	0.143731	0.139040	0.132	0.129	0.127	0.126	0.125	0.125	0.125
60	0.135915	0.134711	0.131391	0.127762	0.121	0.118	0.116	0.115	0.115	0.114	0.114
70	0.126014	0.124810	0.122186	0.118044	0.113	0.11	0.108	0.107	0.106	0.106	0.106
80	0.118350	0.116873	0.114417	0.111066	0.105	0.103	0.101	0.1	0.1	0.099	0.099
90	0.111619	0.110232	0.107708	0.104276	0.1	0.097	0.095	0.094	0.094	0.094	0.094
100	0.106157	0.104696	0.102748	0.099320	0.095	0.092	0.091	0.09	0.089	0.089	0.089
200	0.070489	0.074659	0.072990	0.070805	0.067	0.065	0.064	0.064	0.064	0.064	0.063
300	0.061746	0.061067	0.059533	0.057851	0.055	0.054	0.053	0.052	0.052	0.052	0.052
400	0.053335	0.052747	0.051917	0.050257	0.048	0.047	0.046	0.045	0.045	0.045	0.045
500	0.047696	0.047419	0.046238	0.044893	0.043	0.042	0.041	0.041	0.04	0.04	0.04
1000	0.034028	0.033719	0.032830	0.031659	0.03	0.03	0.029	0.029	0.029	0.029	0.029

Table A-5. Critical Values for A-D Test Statistic for Significance Level = 0.01

n\k	0.025	0.05	0.1	0.2	0.5	1	2	5	10	20	50
5	1.749166	1.518258	1.258545	1.068746	0.945	0.905	0.89	0.883	0.882	0.879	0.879
6	1.751877	1.543508	1.305996	1.123216	0.99	0.946	0.928	0.918	0.916	0.911	0.912
7	1.752404	1.556906	1.332339	1.162744	1.019	0.979	0.951	0.944	0.938	0.935	0.938
8	1.752700	1.561426	1.358108	1.187751	1.044	0.99	0.97	0.961	0.955	0.956	0.953
9	1.758051	1.567347	1.372050	1.210845	1.058	1.007	0.984	0.967	0.968	0.969	0.967
10	1.759366	1.575002	1.384541	1.218849	1.071	1.018	0.994	0.981	0.977	0.975	0.973
15	1.762174	1.593432	1.418705	1.263841	1.1	1.048	1.018	1.002	0.999	0.997	0.999
16	1.763292	1.596448	1.422813	1.273189	1.112	1.047	1.019	1.007	1.004	1	0.999
17	1.763403	1.599618	1.425118	1.273734	1.11	1.053	1.023	1.008	1.004	1.003	1
18	1.763822	1.599735	1.435826	1.274053	1.116	1.054	1.027	1.015	1.006	1.005	1.003
19	1.764890	1.603396	1.441772	1.278280	1.115	1.059	1.026	1.013	1.01	1.006	1.008
20	1.765012	1.604198	1.443435	1.279990	1.118	1.056	1.031	1.016	1.012	1.005	1.009
21	1.765021	1.604737	1.446116	1.281092	1.126	1.057	1.031	1.017	1.013	1.013	1.008
22	1.765611	1.605233	1.448791	1.284002	1.119	1.062	1.036	1.023	1.014	1.011	1.013
23	1.765703	1.609641	1.449964	1.288792	1.125	1.059	1.034	1.017	1.02	1.012	1.013
24	1.766530	1.609644	1.451442	1.289696	1.126	1.065	1.035	1.02	1.015	1.012	1.013
25	1.766655	1.609908	1.451659	1.290311	1.127	1.064	1.038	1.021	1.017	1.014	1.013
30	1.771265	1.617605	1.462230	1.295794	1.133	1.072	1.044	1.023	1.023	1.019	1.018
35	1.772614	1.620179	1.465890	1.296988	1.136	1.072	1.045	1.027	1.025	1.021	1.018
40	1.772920	1.622877	1.468763	1.304213	1.138	1.076	1.046	1.03	1.027	1.023	1.022
45	1.774318	1.624156	1.469148	1.308833	1.141	1.074	1.048	1.036	1.03	1.026	1.024
50	1.775401	1.630356	1.471192	1.311004	1.142	1.079	1.053	1.034	1.029	1.028	1.025
60	1.777021	1.630972	1.474981	1.312242	1.144	1.079	1.054	1.032	1.032	1.029	1.03
70	1.780583	1.634413	1.477148	1.313856	1.145	1.079	1.055	1.038	1.031	1.031	1.028
80	1.782174	1.636678	1.481082	1.315184	1.15	1.085	1.055	1.036	1.033	1.032	1.029
90	1.786462	1.637946	1.483922	1.316508	1.149	1.086	1.056	1.038	1.034	1.031	1.033
100	1.788600	1.639307	1.484231	1.318003	1.149	1.085	1.054	1.042	1.035	1.033	1.032
200	1.789565	1.640278	1.486139	1.318714	1.156	1.089	1.059	1.041	1.031	1.032	1.033
300	1.791785	1.640656	1.489654	1.322935	1.154	1.09	1.058	1.043	1.038	1.033	1.031
400	1.796178	1.641470	1.491079	1.323876	1.158	1.093	1.057	1.043	1.039	1.035	1.034
500	1.799037	1.642244	1.491158	1.328415	1.155	1.089	1.057	1.047	1.04	1.034	1.034
1000	1.810595	1.642639	1.492652	1.328852	1.157	1.092	1.06	1.043	1.035	1.036	1.031

Table A-6. Critical Values for K-S Test Statistic for Significance Level = 0.01

n\k	0.025	0.050	0.10	0.2	0.50	1.0	2.0	5.0	10.0	20.0	50.0
5	0.495311	0.482274	0.467859	0.449435	0.431	0.421	0.414	0.41	0.41	0.408	0.408
6	0.464286	0.454103	0.441814	0.423777	0.402	0.391	0.385	0.382	0.381	0.38	0.38
7	0.437809	0.426463	0.411589	0.398890	0.38	0.369	0.362	0.36	0.358	0.357	0.357
8	0.412467	0.404538	0.392838	0.379962	0.36	0.349	0.344	0.34	0.339	0.339	0.338
9	0.390183	0.383671	0.375103	0.361937	0.343	0.333	0.327	0.323	0.323	0.322	0.322
10	0.373002	0.368362	0.358647	0.348328	0.328	0.318	0.312	0.309	0.308	0.308	0.307
15	0.310445	0.307559	0.300791	0.289751	0.274	0.266	0.261	0.258	0.257	0.257	0.256
16	0.302682	0.298348	0.290148	0.280643	0.266	0.258	0.253	0.251	0.25	0.249	0.249
17	0.294519	0.289320	0.283394	0.274722	0.259	0.251	0.246	0.244	0.243	0.242	0.242
18	0.285220	0.280990	0.276126	0.265561	0.252	0.245	0.24	0.237	0.236	0.236	0.236
19	0.277810	0.275460	0.269173	0.260992	0.246	0.238	0.234	0.232	0.231	0.23	0.23
20	0.271994	0.268927	0.261936	0.253878	0.24	0.233	0.228	0.226	0.225	0.225	0.225
21	0.266096	0.262728	0.256686	0.247915	0.235	0.228	0.223	0.221	0.22	0.22	0.219
22	0.260430	0.256537	0.251727	0.242711	0.23	0.223	0.219	0.216	0.216	0.215	0.215
23	0.254210	0.252405	0.245607	0.236271	0.225	0.218	0.215	0.212	0.211	0.211	0.21
24	0.249574	0.246722	0.240947	0.233143	0.221	0.214	0.21	0.208	0.207	0.207	0.206
25	0.246298	0.242298	0.236164	0.228867	0.216	0.21	0.206	0.204	0.203	0.203	0.203
30	0.220685	0.222267	0.217254	0.209442	0.199	0.193	0.189	0.187	0.186	0.186	0.185
35	0.208407	0.206958	0.202296	0.194716	0.185	0.179	0.176	0.174	0.173	0.173	0.172
40	0.196230	0.193613	0.188617	0.182935	0.173	0.168	0.165	0.163	0.162	0.162	0.162
45	0.185995	0.183011	0.179728	0.173141	0.164	0.158	0.156	0.154	0.154	0.153	0.153
50	0.176191	0.173662	0.170513	0.163792	0.156	0.151	0.148	0.146	0.146	0.146	0.145
60	0.161519	0.158802	0.155658	0.150458	0.143	0.138	0.136	0.134	0.134	0.133	0.133
70	0.149283	0.148241	0.144542	0.139590	0.132	0.128	0.126	0.124	0.124	0.124	0.124
80	0.139831	0.138103	0.135441	0.131479	0.124	0.12	0.118	0.117	0.116	0.116	0.116
90	0.132254	0.130746	0.127231	0.123253	0.117	0.114	0.111	0.11	0.11	0.109	0.11
100	0.126224	0.123308	0.121414	0.117441	0.111	0.108	0.106	0.105	0.104	0.104	0.104
200	0.085150	0.088338	0.086339	0.083391	0.079	0.077	0.075	0.074	0.074	0.074	0.074
300	0.073232	0.072401	0.071096	0.068521	0.065	0.063	0.062	0.061	0.061	0.061	0.06
400	0.063283	0.062708	0.061239	0.059235	0.056	0.054	0.053	0.053	0.053	0.053	0.053
500	0.056181	0.056147	0.054822	0.053042	0.05	0.049	0.048	0.047	0.047	0.047	0.047
1000	0.040020	0.039807	0.038938	0.036987	0.036	0.035	0.034	0.034	0.033	0.033	0.033

APPENDIX B

Large Sample Size Requirements to use the Central Limit Theorem on Skewed Data Sets to Compute an Upper Confidence Limit of the Population Mean

As mentioned earlier, the main objective of the ProUCL software funded by the USEPA is to compute accurate and defensible decision statistics to help the decision makers in making reliable decisions which are cost-effective, and protective of human health and the environment. ProUCL software is based upon the philosophy that rigorous statistical methods can be used to compute the correct estimates of the population parameters (e.g., site mean, background percentiles) and decision making statistics including the upper confidence limit (UCL) of the population mean, the upper tolerance limit (UTL), and the upper prediction limit (UPL) to help decision makers and project teams in making decisions. The use and applicability of a statistical method (e.g., Student's t-UCL, CLT-UCL, adjusted gamma-UCL, Chebyshev UCL, bootstrap-t UCL) depend upon data size, data skewness, and data distribution. ProUCL computes decision statistics using several parametric and nonparametric methods covering a wide-range of data variability, skewness, and sample size. A couple of UCL computation methods described in the statistical text books (e.g., Hogg and Craig, 1995) based upon the Student's t-statistic and the Central Limit Theorem (CLT) alone cannot address all scenarios and situations commonly occurring in the various environmental studies.

Moreover, the properties of the CLT and Student's t-statistic are unknown when NDs with varying DLs are present in a data set - a common occurrence in data sets originating from environmental applications. The use of a parametric lognormal distribution on a lognormally distributed data set tends to yield unstable impractically large UCLs values, especially when the standard deviation (*sd)* of the log-transformed data is greater than 1.0 and the data set is of small size such as less than 30-50 (Hardin and Gilbert 1993; Singh, Singh, and Engelhardt, 1997). Many environmental data sets can be modeled by a gamma as well as a lognormal distribution. Generally, the use of a gamma distribution on gamma distributed data sets yields UCL values of practical merit (Singh, Singh, and Iaci 2002). Therefore, the use of gamma distribution-based decision statistics such as UCLs, upper prediction limits (UPLs), and UTLs should not be dismissed just because it is easier to use a lognormal model. The advantages of computing the gamma distribution-based decision statistics have been discussed in Chapters 2 through 5 of this technical guidance document.

Since many environmental decisions are made based upon a 95% UCL (UCL95) of the population mean, it is important to compute UCLs and other decision making statistics of practical merit. In an effort to compute correct and appropriate UCLs of the population mean and other decision making statistics, in addition to computing the Student's t statistic and the CLT based decision statistics (e.g., UCLs, UPLs), significant effort has been made to incorporate rigorous statistical methods based UCLs in ProUCL software covering a wide-range of data skewness and sample sizes (Singh, Singh, and Engelhardt 1997; Singh, Singh, and Iaci 2002). It is anticipated that the availability of the statistical limits in the ProUCL covering a wide range of environmental data sets will help decision makers in making more informative and defensible decisions at Superfund and RCRA sites.

It is noted that even for skewed data sets, practitioners tend to use the CLT or Student's t-statistic based UCLs of the mean for samples of sizes 25-30 (large sample rule-of-thumb to use CLT). However, this rule-of-thumb does not apply to moderately skewed to highly skewed data sets, specifically when σ (*sd* of the log-transformed data) starts exceeding 1. It should be noted that the large sample requirement depends upon the skewness of the data distribution under consideration. The large sample requirement for the sample mean to follow an approximate normal distribution increases with skewness. It is noted that for skewed data sets, even samples of size greater 100 may not be large enough for the sample mean to follow an approximate normal distribution (Figures B-1 through B-7 below) and the UCLs based upon the CLT and Student's t statistics fail to provide the desired 95% coverage of the population mean for samples of sizes as large as 100 as can be seen in Figures B-1 through B-7.

Noting that the Student's t-UCL and the CLT-UCL fail to provide the specified coverage of the population mean of skewed distributions, several researchers, including Chen (1995), Johnson (1978), Kleijnen, Kloppenburg, and Meeuwsen (1986), and Sutton (1993), proposed adjustments for data skewness in the Student's t statistic and the CLT. They suggested the use of a modified-t-statistic and skewness adjusted CLT for positively skewed distributions (for details see Chapter 2 of this Technical Guide). From statistical theory, the CLT yields UCL results slightly smaller than the Student's t-UCL and the adjusted CLT, and the Student's t-statistic yield UCLs smaller than the modified t-UCLs (details in Chapter 2 of this document). Therefore, only the modified t-UCL has been incorporated in the simulation results described in the following. Specifically, if a UCL95 based upon the modified t-statistic fails to provide the specified coverage to the population mean, then the other three UCL methods, Student's t-UCL, CLT-UCL, and the adjusted CLT-UCL, will also fail to provide the specified coverage of the population mean. The simulation graphs summarized in this appendix suggest that the skewness adjusted UCLs such as the Johnson's modified-t UCL (and therefore Student's t-UCL and CLT-UCL) do not provide the specified coverage to the population mean even for mildly to moderately skewed (σ in [0.5, 1.0]) data sets. The coverage of the population mean provided by these UCLs becomes worse (much smaller than the specified coverage) for highly skewed data sets.

The graphical displays, shown in Figures B-1 through B-7, cover mildly, moderately, and highly skewed data sets. Specifically, Figures B-1 through B-7 compare the UCL95 of the mean based upon parametric and nonparametric bootstrap methods and also UCLs computed using the modified-t UCL for mildly skewed (G(5,50), LN(5,0.05)); moderately skewed (G(2,50), LN(5,1)); and highly skewed (G(0.5, 50), G(1,50), and LN(5,1.5)) data distributions. From the simulation results presented in Figures B-1 through B-7, it is noted that for skewed distributions, as expected the UCLs based on the modified t-statistic (and therefore UCLs based upon the CLT and the Student's t-statistic) fail to provide the desired 95% coverage of the population mean of gamma distributions: G(0.5,50), G(1,50), G(2,50); and of lognormal distributions: LN(5,0.5), LN(5,1), LN(5,1.5) for samples of sizes as large as 100; and the large sample size requirement increases as the skewness increases.

The use of the CLT -UCL and Student's t-UCL underestimate the population mean/ EPC for most skewed data sets.

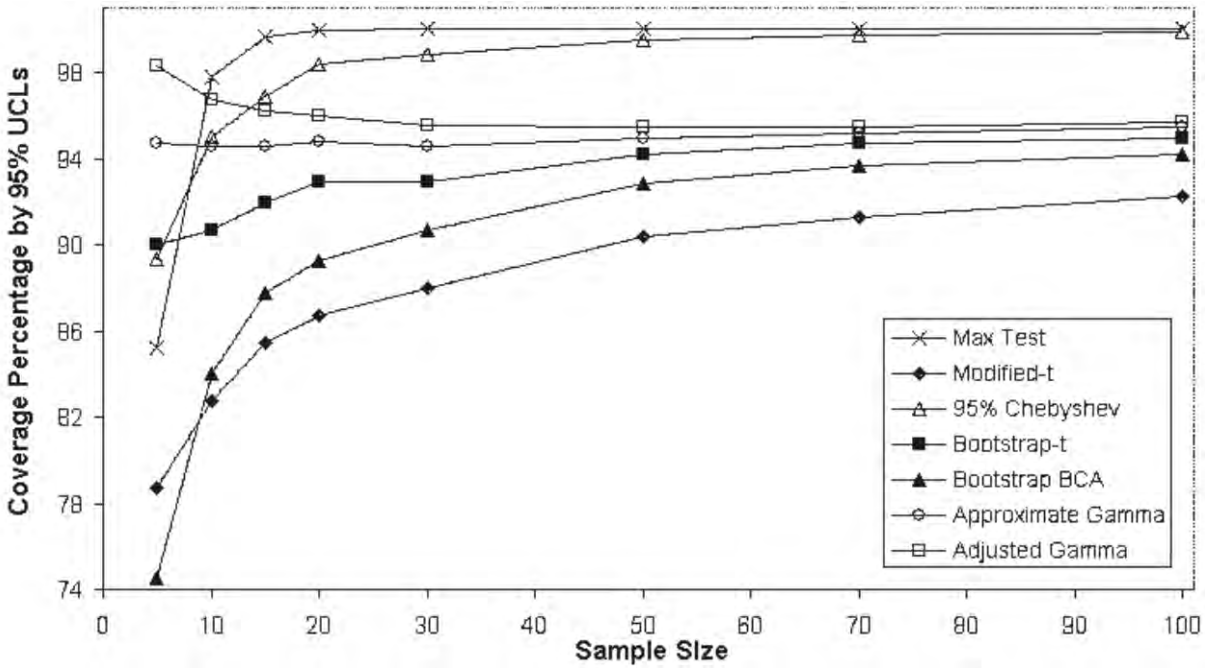

Figure B-1. Graphs of Coverage Probabilities by 95% UCLs of the mean of G (*k*=0.50, Θ=50)

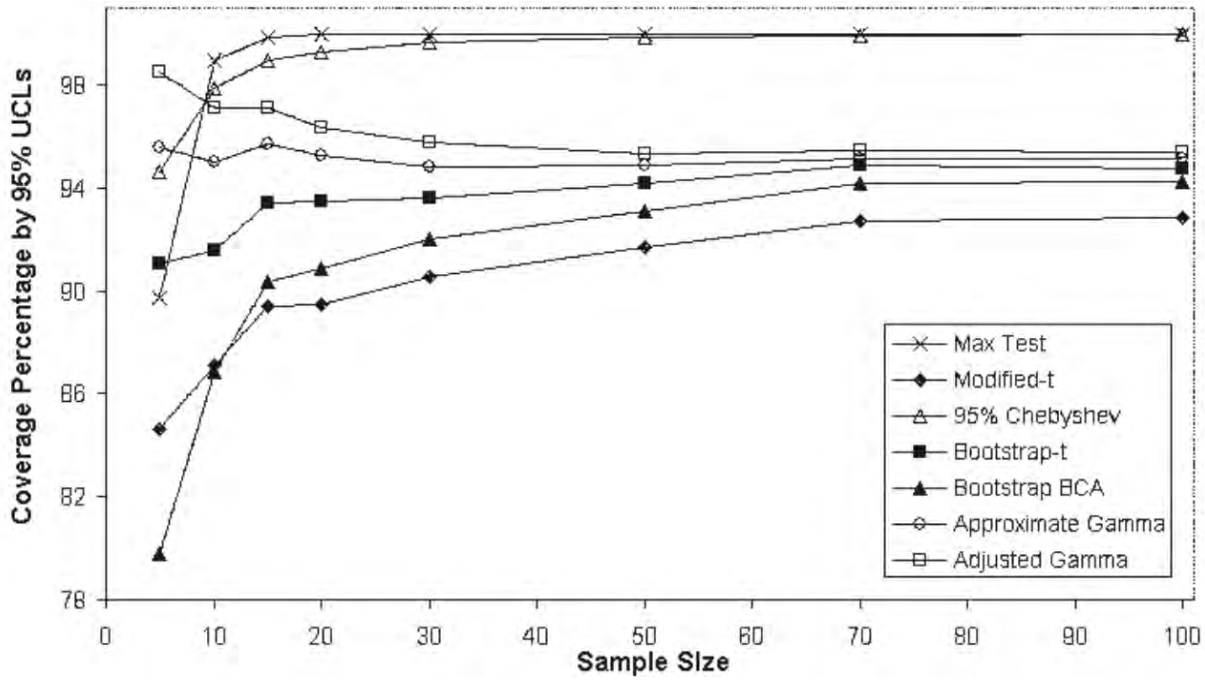

Figure B-2. Graphs of Coverage Probabilities by 95% UCLs of Mean of G(k=1.00, Θ=50)

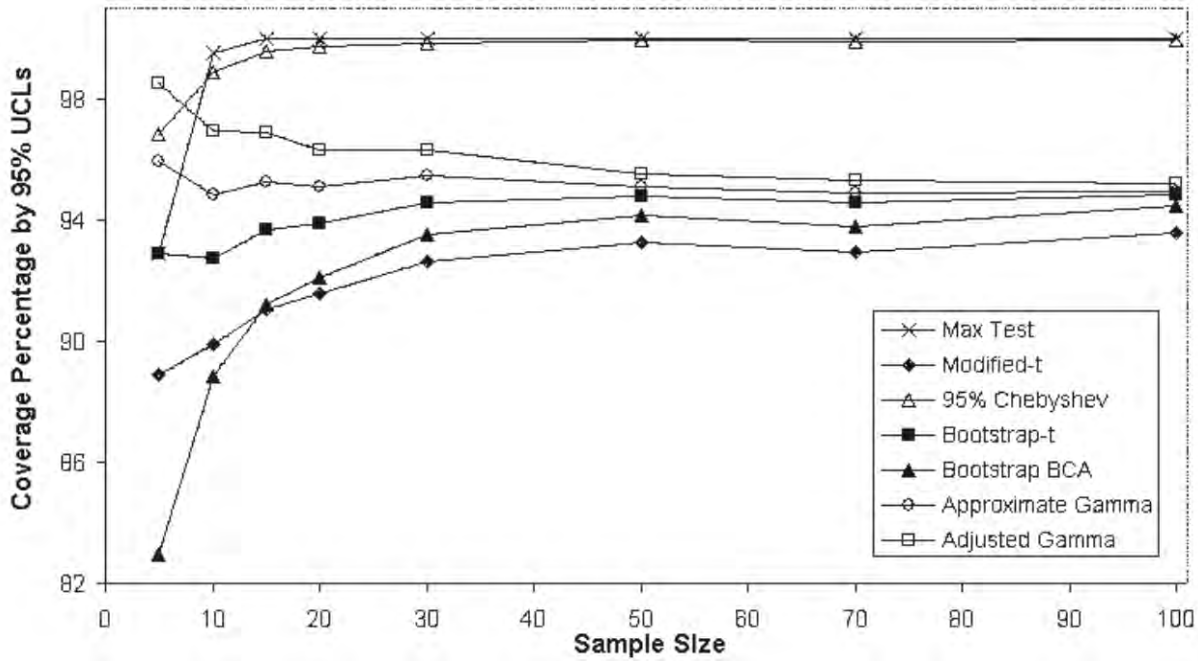

Figure B-3. Graphs of Coverage Probabilities by 95% UCLs of Mean of G(k=2.00, Θ=50)

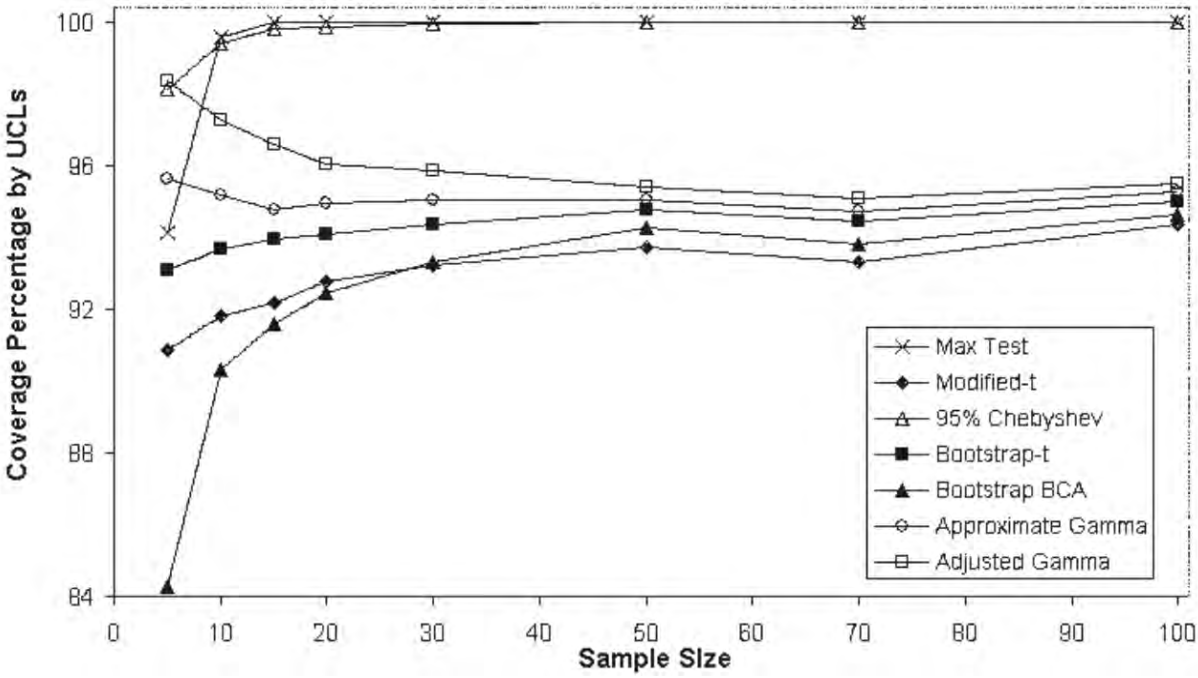

Figure B4. Graphs of Coverage Probabilities by 95% UCLs of Mean of G(k=5.00, Θ=50)

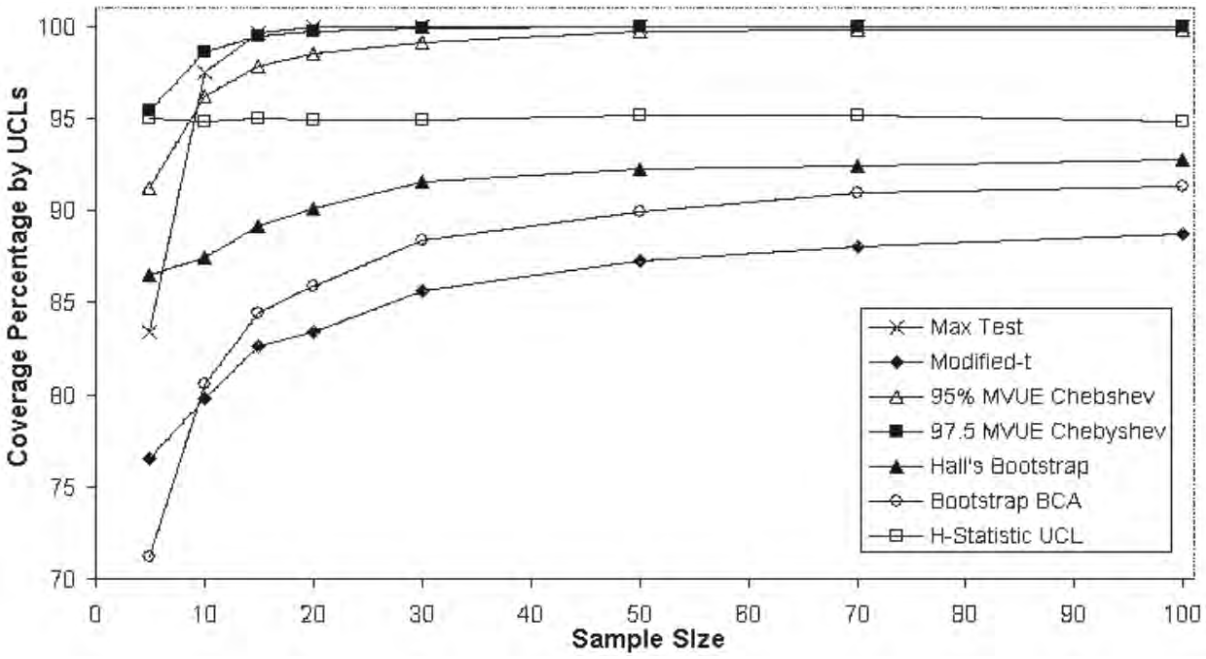

Figure B-5. Graphs of Coverage Probabilities by UCLs of Mean of LN(μ=5, σ=0.5)

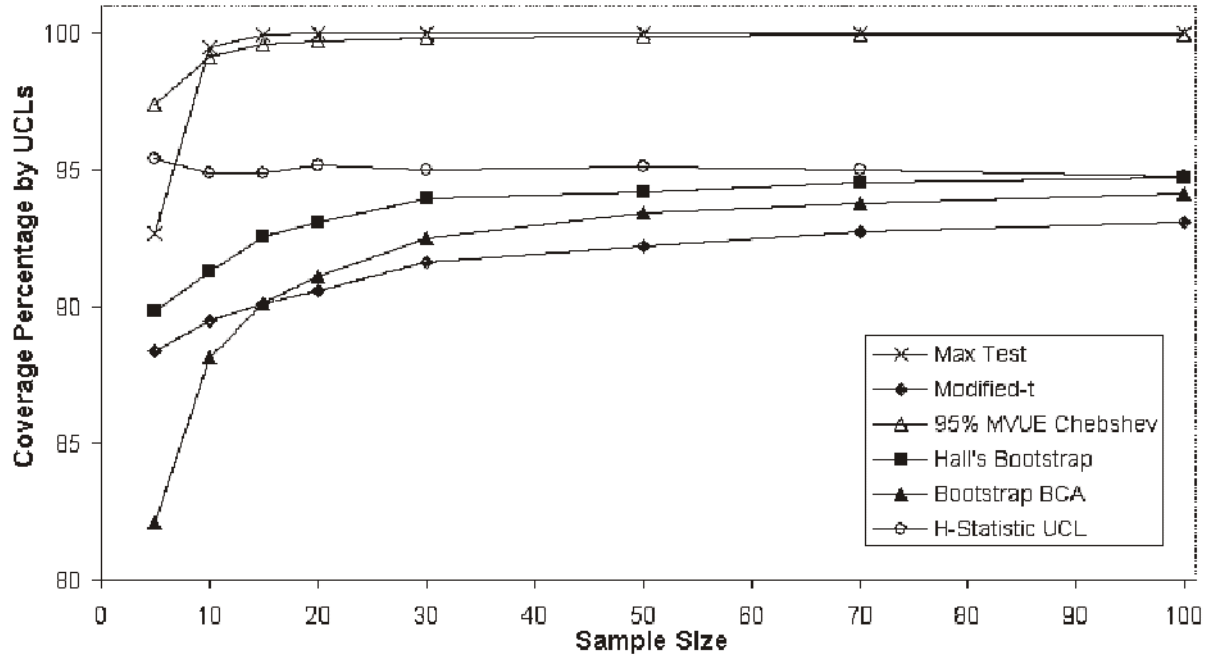

Figure B-6. Graphs of Coverage Probabilities by UCLs of Mean of LN(μ=5, σ=1.0)

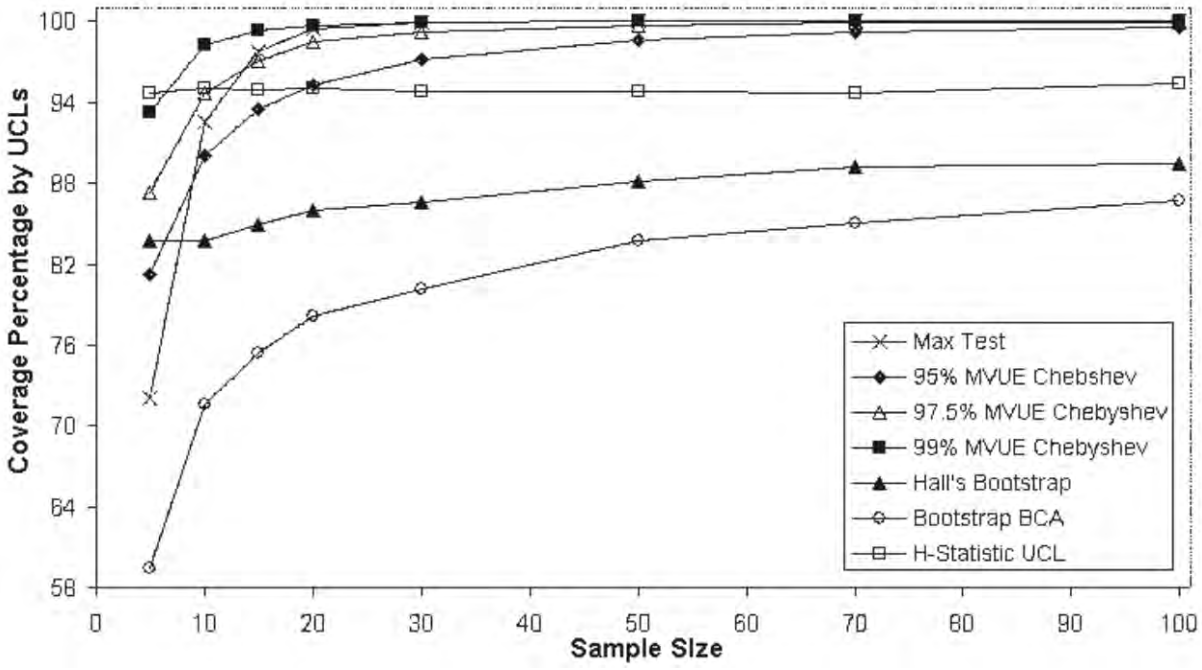

Figure B-7. Graphs of Coverage Probabilities by UCLs of Mean of LN(μ=5, σ=1.5)

APPENDIX C

UCL Recommendation Decision Logic

Decision Logic Flowcharts

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

Standard Deviation of Logged Data	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
$2.0 \le \sigma < 3.0$	Very high skewness
$\sigma \ge 3.0$	Extremely high skewness

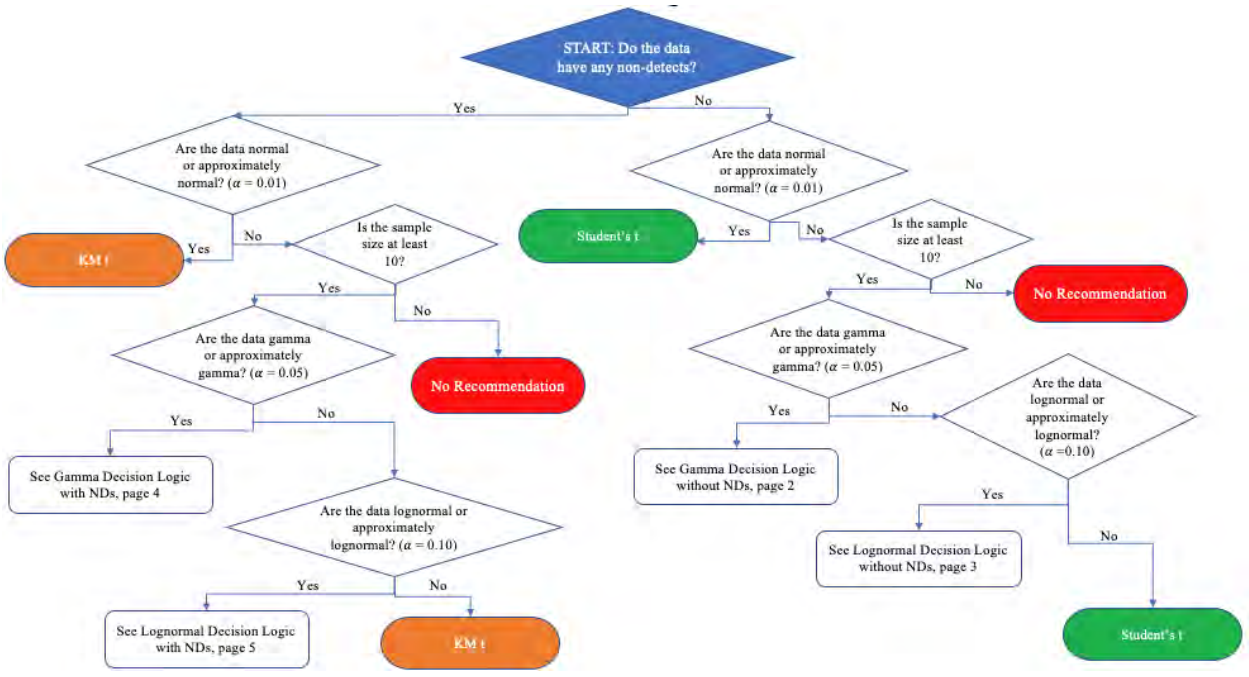

Figure C-1. General Decision Logic Framework, with and without NDs.

For highly skewed data sets with a CV exceeding 1.0, it is suggested that the user pre-process the data. It is very likely that the data include outliers and/or come from multiple populations. The population partitioning methods may be used to identify mixture populations present in the data set.

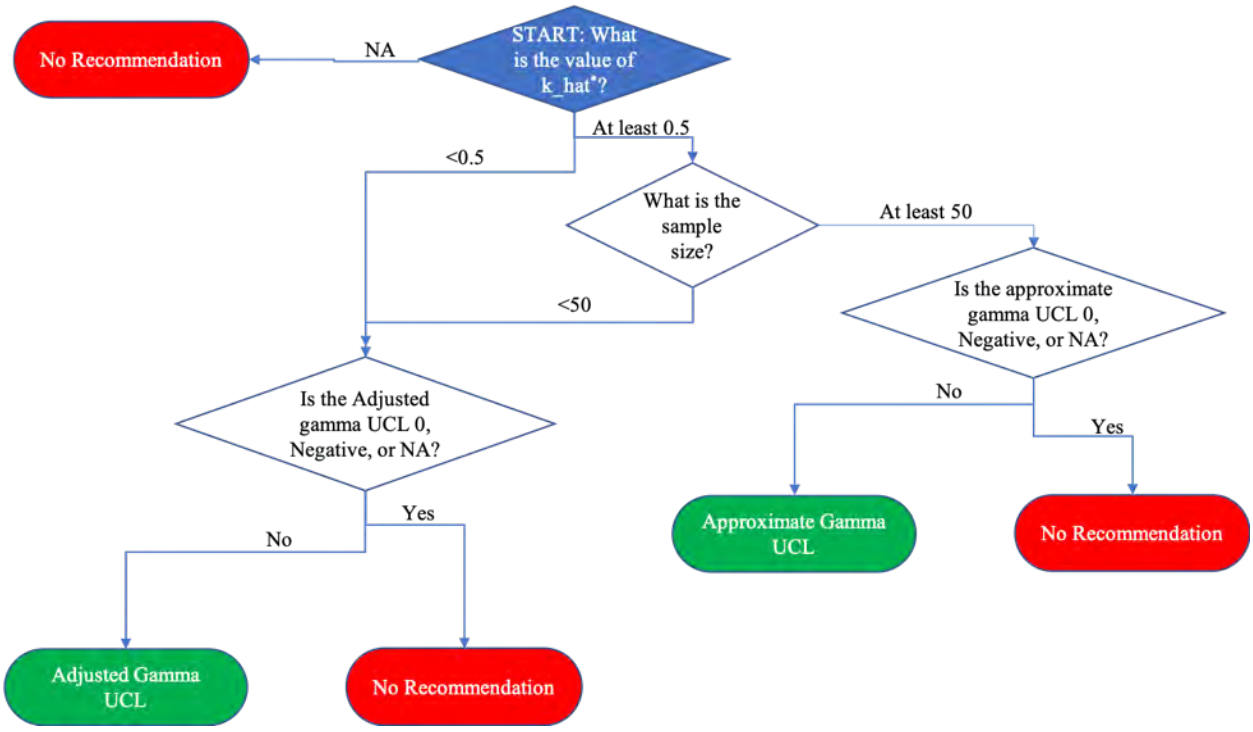

Figure C-2. Recommendations for UCLs of Data Following a Gamma Distribution, without NDs.

*k_hat is the MLE estimate of the shape parameter of the gamma distribution.

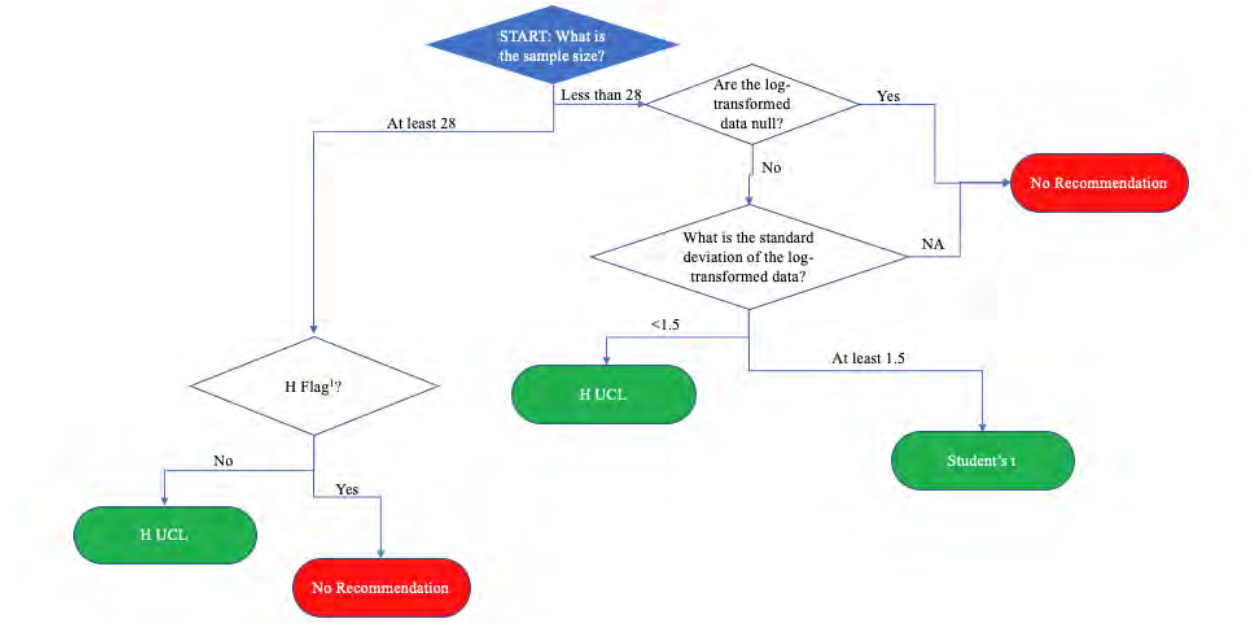

Figure C-3. Recommendations for UCLs of Data Following a Lognormal Distribution, without NDs.

* H Flag is "yes" if n < 3 or n > 1001, or if any of the values in the data are NA, negative, or 0. Otherwise "no".

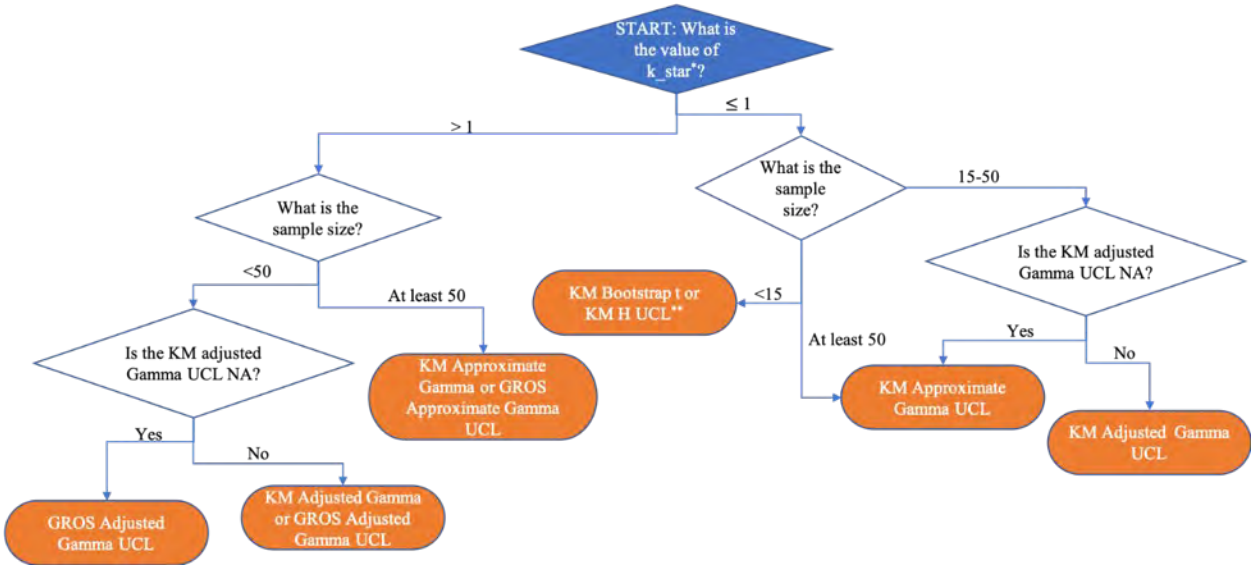

Figure C-4. Recommendations for UCLs of Data Following a Gamma Distribution, with NDs.

*k_star is the bias-corrected MLE estimate of the shape parameter of the gamma distribution, calculated using detects only.

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

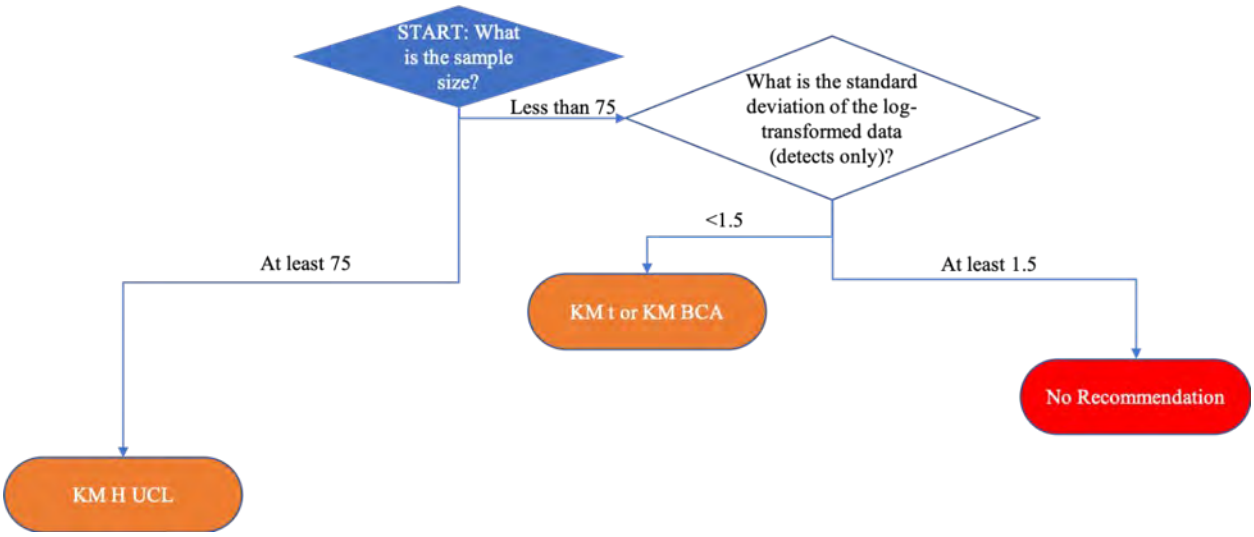

Figure C-5. Recommendations for UCLs of Data Following a Lognormal Distribution, with NDs.

Notes: Suggestions regarding the selection of a 95% UCL are provided to help the user to select the most appropriate 95% UCL. These suggestions are based upon the results of the simulation studies summarized in Singh, Singh, and Iaci (2002), Flagg et al (2017), Section 2.5.1.5, and Appendix D to this Technical Guide. For additional insight, the user may want to consult a statistician.

Optional Decision Logic Flowcharts for Lognormal Distributions without NDs

The recommendations in Figure C-3, Recommendations for UCLs of Data Following a Lognormal Distribution, without NDs, are based on the analysis presented in Appendix D to the Technical Guide. This is based on a machine learning algorithm (recursive partitioning, RPart) applied to a very large number of simulation results. One RPart tree was based on minimizing the average loss (risk) across the loss functions used. The other was based on minimizing the maximum loss (minimax risk) across loss functions.

The resulting decision trees were trimmed to control the complexity of the resulting decision rules to one decision point in sample size, N, and one decision point in sample log-scale standard deviation (log_SD). The decision trees can be expanded by reducing the penalty for complexity. This was done to add one additional decision level to each of the decision trees. The resulting expanded decision trees, in limited circumstances, also included recommending Hall's bootstrap and the Chebyshev 90% UCL. These decision trees are not used for ProUCL's UCL recommendations but are presented below to be used at the discretion of practitioners.

Minimum Average Risk Recommendations

The expanded minimum average risk RPart tree is presented in Figure C-6, below. Note that at each decision node in the plot, the left branch is taken when the node criterion is true, and the right branch is taken when it is false. The recommendation rules are summarized in Table C-2, below.

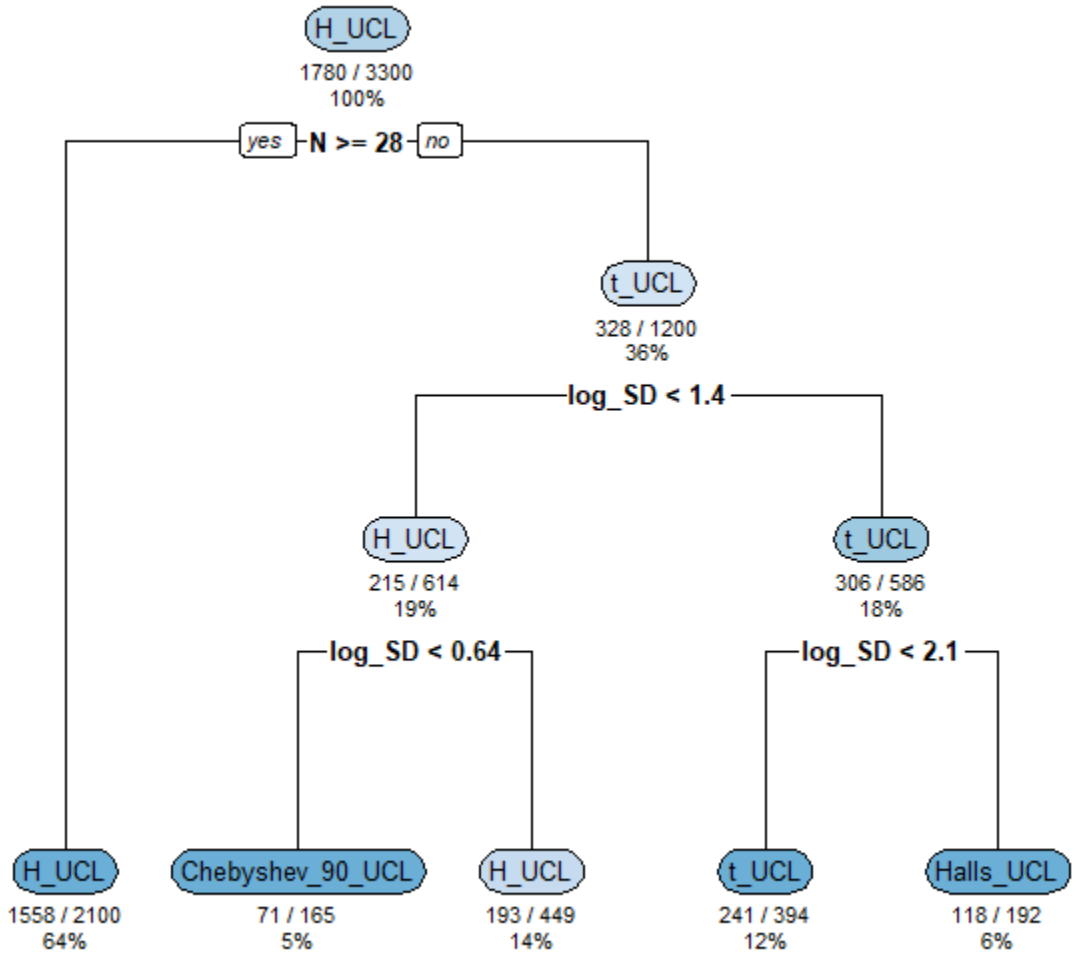

Figure C-6. Expanded minimum average risk recommendations for UCLs of Data Following a Lognormal Distribution, without NDs.

Table C-2. Expanded minimum average risk recommendations for UCLs of Data Following a Lognormal Distribution, without NDs.

Sample Size	Log standard deviation	Recommended UCL
< 28	< 0.64	Chebyshev 90% UCL
< 28	0.64 to 1.38	H-UCL
< 28	1.38 to 2.06	t-UCL
< 28	≥ 2.06	Halls_UCL
≥ 28	any	H-UCL

Minimax Risk Recommendations

The expanded minimax risk RPart tree is presented in Figure C-7, below. Note that at each decision node in the plot, the left branch is taken when the node criterion is true, and the right branch is taken when it is false. The recommendation rules are summarized in Table C-3, below.

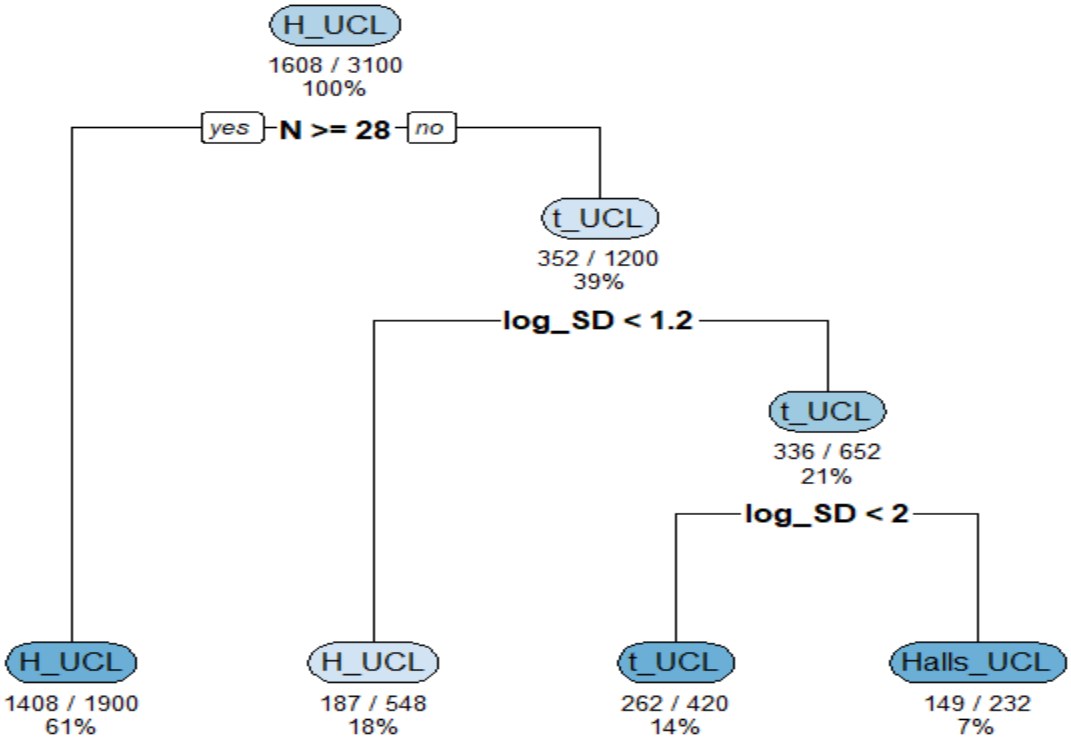

Figure C-7. Expanded minimax risk recommendations for UCLs of Data Following a Lognormal Distribution, without NDs.

Table C-2. Expanded minimax risk recommendations for UCLs of Data Following a Lognormal Distribution, without NDs.

Sample Size	Log standard deviation	Recommended UCL
< 28	< 1.2	H-UCL
< 28	1.2 to 2.0	t-UCL
< 28	≥ 2.0	Halls_UCL
≥ 28	any	H-UCL

APPENDIX D

Analysis of UCL Simulations at the Lognormal Distribution

Separate document