EPA United States
Environmental Protection
Agency

ProUCL Version 5.0.00
Technical Guide

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

RESEARCH AND DEVELOPMENT

ProUCL Version 5.0.00
Technical Guide

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

Prepared for:

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

Prepared by:

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

[1]Lockheed Martin IS&GS-CIVIL
2890 Woodbridge Ave
Edison NJ 08837

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

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

i

129cmb07

NOTICE

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

- ProUCL software was developed by Lockheed Martin, IS&GS - CIVIL under a contract with the EPA and is made available through the EPA Technical Support Center in Atlanta, Georgia.

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

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

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

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

Minimum Hardware Requirements

ProUCL 5.0.00 will function but will run slowly and page a lot.

- Intel Pentium 1.0 GHz
- 45 MB of hard drive space
- 512 MB of memory (RAM)
- CD-ROM drive or internet connection
- Windows XP (with SP3), Vista (with SP1 or later), and Windows 7.

ProUCL 5.0.00 will function but some titles and some Graphical User Interfaces (GUIs) will need to be scrolled. Definition without color will be marginal.

- 800 by 600 Pixels
- Basic Color is preferred

Preferred Hardware Requirements

- 1 gigahertz (GHz) or faster Processor.
- 1 gigabyte (GB) of memory (RAM)
- 1024 by 768 Pixels or greater color display

Software Requirements

ProUCL 5.0.00 has been developed in the Microsoft .NET Framework 4.0 using the C# programming language. To properly run ProUCL 5.0.00 software, the computer using the program must have the .NET Framework 4.0 pre-installed. The downloadable .NET Framework 4.0 files can be obtained from one of the following websites:

- http://msdn.microsoft.com/netframework/downloads/updates/default.aspx
  http://www.microsoft.com/en-us/download/details.aspx?id=17851
  Quicker site for 32 Bit Operating systems

- http://www.microsoft.com/en-us/download/details.aspx?id=24872
  Use this site if you have a 64 Bit operating system

Installation Instructions when Downloading from the EPA Web Site

- Download the file SETUP.EXE from the EPA Web site and save to a temporary location.

- Run the SETUP.EXE program. This will create a ProUCL directory and two folders: 1) The USER GUIDE (this document), and 2) DATA (example data sets).

- 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.

Installation Instructions when Copying from a CD

- Create a folder named **ProUCL 5.0** on a local hard drive of the machine you wish to install ProUCL 5.0.

- Extract the zipped file **ProUCL.zip** to the folder you have just created.

- Run **ProUCL.exe**.

Note: If you have extension turned off, the program will show with the name **ProUCL** in your directory and have an Icon with the label **ProUCL**.

Creating a Shortcut for ProUCL 5.0 on Desktop

- 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 5.0.

Caution: It should be noted that since all files in your ProUCL directory are needed to execute the ProUCL software, one needs you generate a shortcut using the process described above. Specifically, 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 5.0.00

Software ProUCL version 5.0.00 (ProUCL 5.0), its earlier versions: ProUCL version 3.00.01, 4.00.02, 4.00.04, 4.00.05, 4.1.00, and 4.1.01, associated Facts Sheet, User Guides and Technical Guides (e.g., EPA 2010a, 2010b) can be downloaded from the following EPA website:

http://www.epa.gov/osp/hstl/tsc/software.htm
http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

Material for a couple of ProUCL webinars offered in March 2011, and relevant literature used in the development of ProUCL 5.0 can also be downloaded from the above EPA website.

Contact Information for all Versions of ProUCL

The ProUCL software is 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 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

The main objective of the ProUCL software funded by the USEPA is to compute rigorous statistics to help decision makers and project teams in making correct decisions at a polluted site which are cost-effective, and protective of human health and the environment. The ProUCL software is based upon the philosophy that rigorous statistical methods can be used to compute correct estimates of population parameters and decision making statistics including: the upper confidence limit (UCL) of the mean, the upper tolerance limit (UTL), and the upper prediction limit (UPL) to help decision makers and project teams in making correct decisions. A few commonly used text book type methods (e.g., CLT, Student's t-UCL) alone cannot address all scenarios and situations occurring in the various environmental studies. Since many environmental decisions are based upon a 95% UCL (UCL95) of the population mean, it is important to compute correct UCLs of practical merit. The use and applicability of a statistical method (e.g., student's t-UCL, Central Limit Theorem (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, distribution, skewness, and sample size. It is anticipated that the availability of the statistical methods in the ProUCL software covering a wide range of environmental data sets will help the decision makers in making more informative and correct decisions at the various Superfund and RCRA sites.

It is noted that for moderately skewed to highly skewed environmental data sets, UCLs based on the CLT and the Student's t-statistic fail to provide the desired coverage (e.g., 0.95) to the population mean even when the sample sizes are as large as 100 or more. The sample size requirements associated with the CLT increases with skewness. It will be naive and incorrect to state that a CLT or Student's statistic based UCLs are adequate to estimate EPC terms based upon skewed data sets. These facts have been described in the published documents summarizing simulation experiments conducted on positively skewed data sets to evaluate the performances of the various UCL computation methods. The use of a parametric lognormal distribution on a lognormally distributed data set yields unstable impractically large UCLs values, especially when the standard deviation (*sd*) of the log-transformed data becomes greater than 1.0 and the data set is of small size less than 30-50. Many environmental data sets can be modeled by a gamma as well as a lognormal distribution. The use of a gamma distribution on gamma distributed data sets tends to yield UCL values of practical merit. Therefore, the use of gamma distribution based decision statistics such as UCLs, UPLs, and UTLs cannot be dismissed by stating that it is easier (than a gamma model) to use a lognormal model to compute these upper limits.

The suggestions made in ProUCL are based upon the extensive experience of the developers in environmental statistical methods, published environmental literature, and procedures described in various EPA guidance documents. The inclusion of outliers in the computation of the various decision statistics tends to yield inflated values of those decision statistics, which can lead to incorrect decisions. Often inflated statistics computed using a few outliers tend to represent those outliers rather than representing the main dominant population of interest (e.g., reference area). It is suggested to identify outliers, observations coming from population(s) other than the main dominant population, before computing the decision statistics needed to address project objectives. The project team may want to perform the statistical evaluations twice, once with outliers and once without outliers. This exercise will help the project team in computing correct and defensible decision statistics needed to make cleanup and remediation decisions at polluted sites.

The initial development during 1999-2000 and all subsequent upgrades and enhancements of the ProUCL software have been funded by USEPA through its Office of Research and Development (ORD). Initially

ProUCL was developed as a research tool for USEPA scientists and researchers of the Technical Support Center and ORD-NERL, EPA Las Vegas. Background evaluations, groundwater monitoring, exposure and risk management and cleanup decisions in support of the Comprehensive Environmental Recovery, Compensation, and Liability Act (CERCLA) and Resource Conservation and Recovery Act (RCRA) site projects of USEPA are often derived based upon the various test statistics (e.g., Shapiro-Wilk test, t-test, Wilcoxon-Mann-Whitney (WMW) test, analysis of variance [ANOVA], Mann-Kendall [MK] test) and decision statistics including UCLs of mean, UPLs, and UTLs. To address the statistical needs of the environmental projects of the USEPA, over the years ProUCL software has been upgraded and enhanced to include many graphical tools and statistical methods described in the various EPA guidance documents including: EPA 1989a, 1989b, 1991, 1992a, 1992b, 2000 (MARSSIM), 2002a, 2002b, 2002c, 2006a, 2006b, and 2009. Several statistically rigorous methods (e.g., for data sets with NDs) not easily available in the existing guidance documents and in the environmental literature are also available in ProUCL version 5.0.00 (ProUCL 5.0).

ProUCL 5.0 has graphical, estimation, and hypotheses testing methods for uncensored-full data sets and for left-censored data sets consisting of NDs observations with multiple detection limits (DLs) or reporting limits (RLs). In addition to computing general statistics, ProUCL 5.0 has goodness-of-fit (GOF) tests for normal, lognormal and gamma distributions, parametric and nonparametric methods including bootstrap methods for skewed data sets to compute various decision making statistics such as UCLs of mean (EPA 2002a), percentiles, UPLs for a certain number of future observations (e.g., k with k=1, 2, 3,...), UPLs for mean of future k ($\geq$1) observations, and UTLs (e.g., EPA 1992b, 2002b, and 2009). Many positively skewed environmental data sets can be modeled by a lognormal as well as a gamma model. It is well-known that for moderately skewed to highly skewed data sets, the use of a lognormal distribution tends to yield inflated and unrealistically large values of the decision statistics especially when the sample size is small (e.g., <20-30). For gamma distributed skewed uncensored and left-censored data sets, ProUCL software computes decision statistics including UCLs, percentiles, UPLs for future k ($\geq$1) observations, UTLs, and upper simultaneous limits (USLs).

For data sets with NDs, ProUCL has several estimation methods including the Kaplan-Meier (KM) method, regression on order statistics (ROS) methods and substitution methods (e.g., replacing NDs by DL, DL/2). ProUCL 5.0 can be used to compute upper limits which adjust for data skewness; specifically, for skewed data sets, ProUCL 5.0 computes upper limits using KM estimates in gamma (lognormal) UCL and UTL equations provided the detected observations in the left-censored data set follow a gamma (lognormal) distribution. Some poor performing commonly used and cited methods such as the DL/2 substitution method and H-statistic based UCL computation method have been incorporated in ProUCL for historical reasons, and research and comparison purposes.

The Sample Sizes module of ProUCL can be used to develop data quality objectives (DQOs) based sampling designs and to perform power evaluations needed to address statistical issues associated with the various polluted sites projects. ProUCL provides user friendly options to enter the desired values for the decision parameters such as Type I and Type II error rates, and other DQOs used to determine the minimum sample sizes needed to address project objectives. The Sample Sizes module can compute DQOs based minimum sample sizes needed: to estimate the population mean; to perform single and two-sample hypotheses testing approaches; and in acceptance sampling to accept or reject a batch of discrete items such as a lot of drums consisting of hazardous waste. Both parametric (e.g., t-test) and nonparametric (e.g., Sign test, WMW test, test for proportions) sample size determination methods are available in ProUCL.

ProUCL has exploratory graphical methods for both uncensored data sets and for left-censored data sets consisting of ND observations. Graphical methods in ProUCL include histograms, multiple quantile-quantile (Q-Q) plots, and side-by-side box plots. The use of graphical displays provides additional insight

about the information contained in a data set that may not otherwise be revealed by the use of estimates (e.g., 95% upper limits) and test statistics (e.g., two-sample t-test, WMW test). In addition to providing information about the data distributions (e.g., normal or gamma), Q-Q plots are also useful in identifying outliers and the presence of mixture populations (e.g., data from several populations) potentially present in a data set. Side-by-side box plots and multiple Q-Q plots are useful to visually compare two or more data sets, such as: site-versus-background constituent concentrations, surface-versus-subsurface concentrations, and constituent concentrations of several groundwater monitoring wells (MWs). ProUCL also has a couple of classical outlier test procedures, such as the Dixon test and the Rosner test which can be used on uncensored data sets as well as on left-censored data sets consisting of ND observations.

ProUCL has parametric and nonparametric single-sample and two-sample hypotheses testing approaches for uncensored as well as left-censored data sets. Single-sample hypotheses tests: Student's t-test, Sign test, Wilcoxon Signed Rank test, and the Proportion test are used to compare site mean/median concentrations (or some other threshold such as an upper percentile) with some average cleanup standard, C_s (or a not-to-exceed compliance limit, A_0) to verify the attainment of cleanup levels (EPA, 1989a; MARSSIM, 2000; EPA 2006a) at remediated site areas of concern. Single-sample tests such as the Sign test and Proportion test, and upper limits including UTLs and UPLs are also used to perform intra-well comparisons. Several two-sample hypotheses tests as described in EPA guidance documents (e.g., EPA 2002b, 2006b, 2009) are also available in the ProUCL software. The two-sample hypotheses testing approaches in ProUCL include: Student's t-test, WMW test, Gehan test and Tarone-Ware test. The two-sample tests are used to compare concentrations of two populations such as site versus background, surface versus subsurface soils, and upgradient versus downgradient wells.

The Oneway Analysis of Variance (ANOVA) module in ProUCL has both classical and nonparametric Kruskal-Wallis (K-W) tests. Oneway ANOVA is used to compare means (or medians) of multiple groups such as comparing mean concentrations of several areas of concern and to perform inter-well comparisons. In groundwater (GW) monitoring applications, the ordinary least squares (OLS) of regression, trend tests, and time series plots are used to identify upwards or downwards trends potentially present in constituent concentrations identified in GW monitoring wells over a certain period of time. The Trend Analysis module performs Mann-Kendall trend test and Theil-Sen trend test on data sets with missing values; and generates trend graphs displaying a parametric OLS regression line and nonparametric Theil-Sen trend line. The Time Series Plots option can be used to compare multiple time-series data sets.

The use of the incremental sampling methodology (ISM) has been recommended (ITRC, 2012) to collect ISM soil samples needed to compute mean concentrations of the decision units (DUs) and sampling units (SUs) requiring characterization and remediation activities. At many polluted sites, a large amount of discrete onsite and/or offsite background data are already available which cannot be directly compared with newly collected ISM data. In order to provide a tool to compare the existing discrete background data with actual field onsite or background ISM data, a Monte Carlo Background Incremental Sample Simulator (BISS) module has been incorporated in ProUCL 5.0 (currently blocked from general use) which may be used on a <u>large</u> existing discrete background data set. The BISS module simulates incremental sampling methodology based equivalent background incremental samples. The availability of a <u>large</u> discrete background data set collected from areas with geological conditions comparable to the DU(s) of interest is a pre-requisite for successful application of this module. The BISS module has been temporarily blocked for use in ProUCL 5.0 as this module is awaiting adequate guidance and instructions for its intended use on discrete background data sets.

ProUCL 5.0 is a user friendly freeware package providing statistical and graphical tools needed to address statistical issues described in the various EPA guidance documents. ProUCL 5.0 can process many

constituents (variables) simultaneously to: perform various tests (e.g., ANOVA and trend test statistics) and compute decision statistics including UCLs of mean, UPLs, and UTLs – a capability not available in several commercial software packages such as Minitab 16 and NADA for R (Helsel, 2013). ProUCL 5.0 also has the capability of processing data by group variables. ProUCL 5.0 is easy to use and it does not require any programming skills as needed when using other software packages such as Minitab, SAS, and programs written in R script.

Methods incorporated in ProUCL 5.0 have been tested and verified extensively by the developers and the various researchers, scientists, and users. The results obtained by ProUCL are in agreement with the results obtained by using other software packages including Minitab, SAS, and programs written in R Script. ProUCL 5.0 computes decision statistics (e.g., UPL, UTL) based upon the KM method in a straight forward manner without flipping the data and re-flipping the computed statistics for left-censored data sets; these operations are not easy for a typical user to understand and perform. This can unnecessarily become tedious when computing decision statistics for multiple variables/analytes. Moreover, unlike survival analysis, it is important to compute an accurate estimate of the *sd* which is needed to compute decision making statistics including UPLs and UTLs. For left-censored data sets, ProUCL computes a KM estimate of *sd* directly. These issues are elaborated by examples discussed in this Technical Guide and in the accompanying ProUCL 5.0 User Guide.

x

ACRONYMS and ABBREVIATIONS

ACL	alternative compliance or concentration limit
A-D, AD	Anderson-Darling test
AL	Action limit
AM	arithmetic mean
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
CDF, cdf	cumulative distribution function
CERCLA	Comprehensive Environmental Recovery, Compensation, and Liability Act
CL	compliance limit
CLT	central limit theorem
COC	contaminant/constituent of concern
COPC	contaminant/constituent of potential concern
C_s	cleanup standards
CSM	conceptual site model
CV	coefficient of variation
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
EU	exposure units
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	a positive integer representing future or next k observations
K	shape parameter of a gamma distribution
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
Log	refers to log base e, natural logarithm
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
MAD	median absolute deviation
MARSSIM	Multi-Agency Radiation Survey and Site Investigation Manual
MCL	maximum concentration limit, maximum compliance limit
MD	Mahalanobis distance
MDC	minimum detectable concentration
MDD	minimum detectable difference
MDL	method detection limit
MK, M-K	Mann-Kendall
ML	maximum likelihood
MLE	maximum likelihood estimate
MLE (t)	UCL based upon ML estimates using Student's t-distribution critical value
MS	mean sum of squares
MSE	mean sum of squares error
MSR	mean sum of squares regression
MV	minimum variance
n	number of observations/measurements in a sample
N	number of observations/measurements in a population
MVUE	minimum variance unbiased estimate
MW	monitoring well
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
OU	operating unit
PCA	principal component analysis
PDF, pdf	probability density function
pdf	files in pdf format
PLE	product limit estimate
PRG	preliminary remediation goals
PROP	proposed influence function
QA	quality assurance
QC	quality
Q-Q	quantile-quantile
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
SE	standard error
SND	standard normal deviate
SNV	standard normal variate
S_p	pooled standard deviation
SSE	sum of squares error
SSL	soil screening levels
SST	sum of squares total
SSR	sum of squares regression

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
U.S. EPA, USEPA	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
WMW	Wilcoxon-Mann-Whitney
WRS	Wilcoxon Rank Sum
WSR	Wilcoxon Signed Rank
X_p	p^{th} percentile of a distribution
$<$	less than
$>$	greater than
$\geq$	greater than or equal to
$\leq$	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.

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 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.

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

Kolmogorov-Smirnov (KS) test: The Kolmogorov-Smirnov test is used to decide if a 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.

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

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

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

p-value: In statistical hypothesis testing, the p-value 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).

pth Percentile or pth 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 then 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: This is similar to a stable UCL.

Robustness: Robustness is used to compare statistical tests. A robust test is the one with good performance (that is not unduly affected by outliers 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: A sample here represents a random sample (data set) obtained from the population of interest (e.g., a site area, a reference area, or a monitoring well). The sample is supposed to be a representative sample of the population under study. The sample is used to draw inferences about the population parameter(s).

Shapiro-Wilk (SW) test: 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 characteristic under study (e.g., lead concentrations). It can also be measured in terms of the standard deviation of log-transformed data. The greater is 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 merits, 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 is 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.

Uncensored Data Set: A data set without any censored observations is called an uncensored data set.

Unreliable UCL, Unstable UCL, Unrealistic UCL: The UCL of a population mean is unstable, unrealistic, or unreliable if it is orders of magnitude higher than the other UCLs of population mean. It represents an impractically large value that cannot be achieved in practice. For example, the use of Land's H-statistic often results in 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 95th percentile.

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

xBar: arithmetic average of computed using the sampled data values

ACKNOWLEDGEMENTS

Table of Contents

INTRODUCTION

OVERVIEW OF ProUCL VERSION 5.0.00 SOFTWARE

The main objective of the ProUCL software funded by the USEPA is to compute rigorous decision statistics to help the decision makers in making correct decisions which are cost-effective, and protective of human health and the environment. The ProUCL software is based upon the philosophy that rigorous statistical methods can be used to compute the correct estimates of population parameters (e.g., site mean, background percentiles) and decision making statistics including the upper confidence limit of (UCL) the mean, the upper tolerance limit (UTL), and the upper prediction limit (UPL) to help the decision makers and project teams in making correct decisions. The use and applicability of a statistical method (e.g., student's t-UCL, Central Limit Theorem (CLT)-UCL, adjusted gamma-UCL, Chebyshev UCL, bootstrap-t UCL) depend upon data size, data variability, 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 text book methods described in most of the statistical text books (e.g., Hogg and Craig, 1995) based upon the Student's t-statistic and the CLT alone cannot address all scenarios and situations commonly occurring in the various environmental studies. It is naive and incorrect to state or assume that Student's t-statistic and/or CLT based UCLs of mean will provide the desired coverage (e.g., 0.95) to the population mean irrespective of the skewness of the data set/population under consideration. These issues have been discussed in detail in Chapters 2 and 4 of this Technical guide. Several examples have been discussed throughout this guidance document and also in the accompanying ProUCL 5.0 User Guide to elaborate on these issues.

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 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, UPL, and UTLs cannot be dismissed just because it is easier to use a lognormal model to compute these upper limits or incorrectly assuming that the two distributions behave in a similar manner. The advantages of computing the gamma distribution based decision statistics are discussed in Chapters 2-5 of this guidance document.

Since many environmental decisions are made based upon a 95% UCL of the population mean, it is important to compute correct UCLs and other decision making statistics of practical merit. In an effort to compute correct UCLs of the population mean and other decision making statistics, in addition to computing the Student's t statistic and the CLT based statistics (e.g., UCLs, UPLs), significant effort has been made to incorporate rigorous statistical methods based UCLs (and other limits) in the ProUCL software covering a wide-range of data skewness and sample sizes (e.g., Singh, Singh, and Engelhardt, 1997; Singh, Singh, and Iaci, 2002; and Singh, Singh, 2003). It is anticipated that the availability of the statistical methods in the ProUCL software covering a wide range of environmental data sets will help the decision makers in making more informative and correct decisions at the various polluted sites.

It is noted that even for skewed data sets, practitioners tend to use the CLT or Student's t-statistic based UCLs of mean based upon 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 (standard deviation of the log-transformed data) starts exceeding 1. The large sample requirement associated with the use of the CLT 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 the data skewness; and for skewed data sets, even samples of size greater than (>)100 may not be large enough for the sample mean to follow an approximate normal distribution. For moderately skewed to highly skewed environmental data sets, as expected, UCLs based on the CLT and the Student's t- statistic fail to provide the desired coverage to the population mean even when the sample sizes are as large as 100 or more. These facts have been verified in the published simulation experiments conducted on positively skewed data sets (e.g., Singh, Singh, and Engelhardt, 1997; Singh, Singh, and Iaci, 2002; and Singh and Singh, 2003).

The initial development and all subsequent upgrades and enhancement of the ProUCL software have been funded by the USEPA through its Office of Research and Development (ORD). Initially ProUCL was developed as a research tool for scientists and researchers of the Technical Support Center and ORD-NERL, EPA Las Vegas. During 1999-2001, the initial intent and objectives of developing the ProUCL software (Version 1.0 and Version 2.0) were to provide a statistical research tool to EPA scientists which can be used to compute theoretically sound 95% upper confidence limits (UCL95s) of the mean routinely used in exposure assessment, risk management and cleanup decisions made at various CERCLA and RCRA sites (EPA 1992a, 2002a). During 2002, the peer-reviewed ProUCL version 2.1 (with Chebyshev inequality based UCLs) was released for public use. Several researchers have developed rigorous parametric and nonparametric statistical methods (e.g., Johnson, 1978; Grice and Bain, 1980; Efron (1981, 1982); Efron and Tibshirani, 1993; Hall (1988, 1992); Sutton, 1993; Chen, 1995; Singh, Singh, and Engelhardt, 1997; Singh, Singh, and Iaci, 2002) to compute upper limits (e.g., UCLs) which adjust for data skewness. Since Student's t-UCL, CLT-UCL, and percentile bootstrap UCL fail to provide the desired coverage to the population mean of skewed distributions, several parametric (e.g., gamma distribution based) and nonparametric (e.g., BCA bootstrap and bootstrap-t, Chebyshev UCL) UCL computation methods which adjust for data skewness were incorporated in ProUCL versions 3.0 and 3.00.02 during 2003-2004. ProUCL version 3.00.02 also had graphical quantile-quantile (Q-Q) plots and GOF tests for normal, lognormal, and gamma distributions; capabilities to statistically analyze multiple variables simultaneously were also incorporated in ProUCL 3.00.02 (EPA 2004).

It is important to compute decision statistics (e.g., UCLs, UTLs) which are cost-effective and protective of human health and the environment (balancing between Type I and Type II errors), therefore, one cannot dismiss the use of the better [better than t-UCL, CLT-UCL, ROS and KM percentile bootstrap UCL, KM-UCL (t)] performing UCL computation methods including gamma UCLs and the various bootstrap UCLs which adjust for data skewness. During 2004-2007, ProUCL was upgraded to versions 4.00.02, and 4.00.04. These upgrades included exploratory graphical (e.g., Q-Q plots, box plots) and statistical (e.g., maximum likelihood estimation [MLE], KM, and ROS) methods for left-censored data sets consisting of nondetect (NDs) observations with multiple DLs or RLs. For uncensored and left-censored data sets, these upgrades provide statistical methods to compute upper limits: percentiles, UPLs and UTLs needed to estimate site-specific background level constituent concentrations or background threshold values (BTVs). To address statistical needs of background evaluation projects (e.g., MARSSIM, 2000; EPA 2002b), several single-sample and two-sample hypotheses testing approaches were also included in these ProUCL upgrades.

During 2008-2010, ProUCL was upgraded to ProUCL 4.00.05. The upgraded ProUCL was enhanced by including methods to compute gamma distribution based UPLs and UTLs (Krishnamoorthy, Mathew, and Mukherjee, 2008). The Sample Size module to compute DQOs based minimum sample sizes needed to

address statistical issues associated with the various environmental projects (e.g., MARSSIM, 2000; EPA [2002c, 2006a, 2006b]) was also incorporated in ProUCL 4.00.05.

During 2009-2011, ProUCL 4.00.05 was upgraded to ProUCL 4.1 and 4.1.01. ProUCL 4.1 (2010) and 4.1.01 (2011) retain all capabilities of the previous versions of ProUCL software. Two new modules: Oneway ANOVA and Trend Analysis were included in ProUCL 4.1. The Oneway ANOVA module has both parametric and nonparametric ANOVA tests to perform inter-well comparisons. The Trend Analysis module can be used to determine potential upward or downward trends present in constituent concentrations identified in GW monitoring wells (MWs). The Trend Analysis module can compute Mann-Kendall (MK) and Theil-Sen (T-S) trend statistics to determine upward or downward trends potentially present in analyte concentrations. ProUCL 4.1 also has the Ordinary Least Squares (OLS) Regression module. In ProUCL 4.1, some modifications were made in decision tables used to make recommendations regarding the use of UCL95 to estimate EPC terms. Specifically, based upon the recent experience, developers of ProUCL re-iterated that the use of a lognormal distribution to estimate EPC terms and BTVs should be avoided, as the use of lognormal distribution tends to yield unrealistic and unstable values of the decision making statistics including UCL, UPL, and UTL; this is especially true when the sample size is <20-30 and the data set is moderately skewed to highly skewed. During March 2011, a couple of webinars were presented describing the capabilities and use of the methods available in ProUCL 4.1.

ProUCL version 5.0.00 represents an upgrade of ProUCL 4.1.01 (EPA, June 2011) which represents an upgrade of ProUCL 4.1.00 (EPA 2010). For uncensored and left-censored data sets, ProUCL 5.0 consists of all statistical and graphical methods that are available in the previous versions of the ProUCL software package except for a couple of poor performing and restricted (e.g., can be used only when a single detection limit is present) estimation methods such as the MLE and winsorization methods for left-censored data sets. ProUCL has GOF tests for normal, lognormal, and gamma distributions for uncensored and left-censored data sets with NDs. ProUCL 5.0 has the extended version of the Shapiro-Wilk (S-W) test to perform normal and lognormal GOF tests for data sets of sizes up to 2000 (Royston [1982, 1982a]). In addition to normal and lognormal distribution based decision statistics, ProUCL software computes UCLs, UPLs, and UTLs based upon the gamma distribution.

Several enhancements have been made in the UCLs and BTVs modules of the ProUCL 5.0 software. A new statistic, an upper simultaneous limit (Singh and Nocerino, 2002; Wilks, 1963) has been incorporated in the Upper limits/BTVs module of ProUCL 5.0.00 for data sets consisting of NDs with multiple DLs, a two-sample hypothesis test, the Tarone-Ware (T-W; Tarone and Ware, 1978) test has been incorporated in ProUCL 5.0. Nonparametric tolerance limits have been enhanced, and for specific values of confidence coefficients, coverage probability, and sample size, ProUCL 5.0 outputs the confidence coefficient actually achieved by a UTL. The Trend Analysis and OLS Regression modules can handle missing events to compute trend test statistics and generate trend graphs. Some new methods using KM estimates in gamma (and lognormal) distribution based UCL, UPL, and UTL equations have been incorporated to compute the decision statistics for data sets consisting of nondetect observations. To facilitate the computation of UCLs from ISM based samples (ITRC, 2012); the minimum sample size requirement has been lowered to 3, so that one can compute the UCL95 based upon ISM data sets of sizes $\geq$3.

All known bugs, typographical errors, and discrepancies found by the developers and the various users of the ProUCL software package have been addressed in the ProUCL version 5.0.00. Specifically, a discrepancy found in the estimate of mean based upon the KM method has been fixed in ProUCL 5.0. Some changes have been made in the decision logic used in GOF and UCL modules. In practice, based upon a given data set, it is well known that the two statistical tests (e.g., Theil-Sen and OLS trend tests)

can lead to different conclusions. To streamline the decision logic associated with the computation of the various UCLs, the decision tables in ProUCL 5.0 have been updated. Specifically, for each distribution if at least one of the two GOF tests (e.g., Shapiro-Wilk or Lilliefors test for normality) determines that the hypothesized distribution holds, then ProUCL concludes that the data set follows the hypothesized distribution, and decision statistics are computed accordingly. Additionally, for gamma distributed data sets, ProUCL 5.0 suggests the use of the: adjusted gamma UCL for samples of sizes ≤ 50 (instead of 40 suggested in previous versions); and approximate gamma UCL for samples of sizes >50.

Also, 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 (Anderson-Darling and Kolmogorov Smirnov tests) obtained using the two gamma deviate generation algorithms: Whitaker (1974) and Marsaglia and Tsang (2000). For values of $k \leq 0.2$, the critical values of the two gamma GOF tests: Anderson-Darling (A-D) and Kolmogorov-Smirnov (K-S) tests have been updated using the currently available more accurate 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 values of the shape parameter, *k=0.025, 0.05, 0.1*, and *0.2*, the critical value tables for these two tests have been updated by incorporating the newly generated critical values for the three significance levels: 0.05, 0.1, and 0.01. The updated tables are provided in Appendix A. It should be noted that for *k=0.2*, the older and the newly generated critical values are in general agreement.

ProUCL 5.0 also has a new Background Incremental Sample Simulator (BISS) module (temporarily blocked for general public use) which can be used on a large existing discrete background data set to simulate background incremental samples (BIS). The availability of a large discrete data set collected from areas with geological formations and conditions comparable to the DUs (background or onsite) of interest is a requirement for successful application of this module. The simulated BISS data can be compared with the actual field ISM (ITRC, 2012) data collected from the various DUs using other modules of ProUCL 5.0. The values of the BISS data are not directly available to users; however, the simulated BISS data can be accessed by the various modules of ProUCL 5.0 to perform desired statistical evaluations. For example, the simulated background BISS data can be merged with the actual field ISM data after comparing the two data sets using a two-sample t-test; the simulated BISS or the merged data can be used to compute a UCL of the mean or a UTL.

Note: The ISM methodology used to develop the BISS module is a relatively new approach; methods incorporated in this BISS module require further investigation. The BISS module has been temporarily blocked for use in ProUCL 5.0 as this module is awaiting adequate guidance and instructions for its intended use on discrete background data sets.

Software ProUCL version 5.0, its earlier versions: ProUCL version 3.00.02, 4.00.02, 4.00.04, 4.1.00, and 4.1.01, associated Facts Sheet, User Guides and Technical Guides (e.g., EPA [2004, 2007, 2009a, 2009b, 2010a, 2010b]) can be downloaded from the EPA website:

http://www.epa.gov/osp/hstl/tsc/software.htm
http://www.epa.gov/osp/hstl/tsc/softwaredocs.htm

ProUCL 5.0 is a user-friendly freeware package providing statistical and graphical tools needed to address statistical issues described in several EPA guidance documents. Considerable effort has been made to provide a detailed technical guide to help practitioners understand statistical methods needed to address statistical needs of their environmental projects. ProUCL generates detailed output sheets and

graphical displays for each method which can be used to educate students learning environmental statistical methods. Like previous versions, ProUCL 5.0 can process many variables simultaneously to compute various tests (e.g., ANOVA and trend test statistics) and decision statistics including UCL of mean, UPLs, and UTLs, a capability not available in other software packages such as Minitab 16 and NADA for R (Helsel, 2013). Without the availability of this option, the user has to compute decision and test statistics for one variable at a time which becomes cumbersome when dealing with a large number of variables. ProUCL 5.0 also has the capability of processing data by groups. ProUCL 5.0 is easy to use; it does not require any programming skills as needed when using programs written in R Script.

The Need for ProUCL Software

EPA guidance documents (e.g., EPA [1989a, 1989b, 1992a, 1992b, 1994, 1996, 2000, 2002a, 2002b, 2002c, 2006a, 2006b, 2009a, and 2009b]) describe statistical methods including: DQOs based sample size determination procedures, methods to compute decision statistics: UCL95, UPL, and UTLs, parametric and nonparametric hypotheses testing approaches, Oneway ANOVA, OLS regression, and trend determination approaches. Specifically, EPA guidance documents (e.g., EPA [2002c, 2006a, 2006b; and MARSSIM, 2000]) describe DQOs based parametric and nonparametric minimum sample size determination procedures needed: to compute decision statistics (e.g., UCL95); to perform site versus background comparisons (e.g., t-test, proportion test, WMW test); and to determine the number of discrete items (e.g., drums filled with hazardous material) that need to be sampled to meet the DQOs (e.g., specified proportion, p_0 of defective items, allowable error margin in an estimate of mean). Statistical methods are used to compute test statistics (e.g., S-W test, t-test, WMW test, T-S trend statistic) and decision statistics (e.g., 95% UCL, 95% UPL, UTL95-95) needed to address statistical issues associated with CERCLA and RCRA site projects. For example, exposure and risk management and cleanup decisions in support of EPA projects are often made based upon the mean concentrations of the contaminants/constituents of potential concern (COPCs). Site-specific BTVs are used in site versus background evaluation studies. A UCL95 is used to estimate the EPC terms (EPA1992a, 2002a); and upper limits such as upper percentiles, UPLs, or UTLs are used to estimate BTVs or not-to-exceed values (EPA 1992b, 2002b, and 2009). The estimated BTVs are also used: to identify the COPCs; to identify the site areas of concern (AOCs); to perform intra-well comparisons to identify MWs not meeting specified standards; and to compare onsite constituent concentrations with site-specific background level constituent concentrations. Oneway ANOVA is used to perform inter-well comparisons, OLS regression and trend tests are often used to determine potential trends present in constituent concentrations identified in groundwater monitoring wells (MWs). Most of the methods described in this paragraph are available in the ProUCL 5.0 software package.

It is noted that not much guidance is available in the guidance documents cited above to compute rigorous UCLs, UPLs, and UTLs for moderately skewed to highly skewed uncensored and left-censored data sets consisting of NDs with multiple DLs, a common occurrence in environmental data sets. Several parametric and nonparametric methods are available in the statistical literature (Singh, Singh, and Engelhardt, 1997; Singh, Singh, and Iaci, 2002; Krishnamoorthy et al. 2008; Singh, Maichle, and Lee, 2006) to compute UCLs and other upper limits which adjust for data skewness. During the years, as new methods became available to address statistical issues related to the environmental projects, those methods were incorporated in ProUCL software so that environmental scientists and decision makers can make more accurate and informative decisions based upon those rigorous statistical methods. Until 2006, not much guidance was provided on how to compute UCL95 of mean and other upper limits (e.g., UPLs and UTLs) based upon data sets consisting of NDs with multiple DLs. For data sets with NDs, Singh, Maichle, and Lee (EPA 2006) conducted an extensive simulation study to compare the performances of the various estimation methods (in terms of bias in the mean estimate) and UCL computation methods (in

terms of coverage provided by a UCL). They demonstrated that the nonparametric KM method performs well in terms of bias in estimates of mean. They also concluded that UCLs computed using the Student's t-statistic and percentile bootstrap method using the KM estimates do not provide the desired coverage to the population mean of skewed data sets. They demonstrated that the depending upon sample size and data skewness, UCLs computed using KM estimates and: the BCA bootstrap method (mildly skewed data sets); the bootstrap-t method, and the Chebyshev inequality (moderately to highly skewed data sets) provide better coverage (closer to the specified 95% coverage) to the population mean than the various other UCL computation methods. Based upon their findings, during 2006-2007, several UCL and other upper limits computation methods based upon KM and ROS estimates were incorporated in the ProUCL 4.0 software. It is noted that since the inclusion of the KM method in ProUCL 4.0 (2007), the use of the KM method based upper limits has become popular in many environmental applications to estimate EPC terms and background threshold values (BTVs). The KM method is also described in the latest version of the unified RCRA guidance document (EPA 2009).

It is not easy to justify distributional assumptions of data sets consisting of both detects and NDs with multiple DLs. Therefore, based upon the published literature and recent experience, parametric UCL computation methods such as the MLE methods for normal and lognormal distributions are excluded from ProUCL 5.0. Additionally, the winsorization method (Gilbert, 1987) has also been excluded from ProUCL 5.0 due to its poor performance. ProUCL software is also used for teaching environmental statistics courses therefore, in addition to statistical and graphical methods routinely used to address statistical needs of environmental projects, due to their popularity some poor performing methods such as the substitution DL/2 method and Land's (1975) H-statistic based UCL computation method have been retained in ProUCL version 5.0.00 for research and comparison purposes.

Methods incorporated in ProUCL 5.0 and in its earlier versions have been tested and verified extensively by the developers and various researchers, scientists, and users. Specifically, the results obtained by ProUCL 5.0 are in agreement with the results obtained by using other software packages including Minitab, SAS, and programs available in R-Script (not all methods are available in these software packages). Additionally, ProUCL 5.0 outputs several intermediate results (e.g., khat and biased corrected kstar estimates of the gamma shape parameter, k) and critical values (e.g., K factor used to compute UTLs, d2max needed to compute USL) needed to compute the various decision statistics of interest, which may help the interested users to verify statistical results computed by the ProUCL software. ProUCL is a user friendly software which can be used to: process multiple variables (analytes) simultaneously (e.g., perform ANOVA on many variables); process grouped data; to generate and display multiple plots (Q-Q plots) on the same graphical display. No programming skills are needed to use ProUCL software. ProUCL provides warning messages and makes suggestions to help a typical user in selecting the most appropriate decision statistic (e.g., UCL).

Note: The availability of intermediate results and critical values can be used to compute lower limits and two-sided intervals which are not as yet available in the ProUCL software.

For left-censored data sets, ProUCL 5.0 computes decision statistics (e.g., UCL, UPL, and UTL) based upon KM estimates computed in a straight forward manner without flipping the data and re-flipping the decision statistics; these operations are not easy for a typical user to understand and perform and can become quite tedious when multiple analytes need to be processed. Moreover, in environmental applications it is important to compute accurate estimates of standard deviations which are needed to compute the decision making statistics including UPLs and UTLs. Decision statistics (UPL, UTL) based upon a KM estimate of the of standard deviation computed using indirect methods can be different from the statistics computed using an estimate of *sd* obtained using the KM method directly, especially when

one is dealing with skewed data set or using a log-transformation. These issues are elaborated by examples discussed in this Guide and the accompanying ProUCL 5.0 User Guide.

For uncensored data sets, researchers (e.g., Johnson (1978), Chen (1995), Efron and Tibshirani (1993), Hall [1988, 1992], more references in Chapters 2 and 3) had developed parametric (e.g., gamma distribution based) and nonparametric (bootstrap-t and Hall's bootstrap method, modified-t) methods to compute decision statistics which adjust for data skewness. For uncensored positively skewed data sets, Singh, Singh, and Iaci (2002) and Singh and Singh (2003) performed simulation experiments to compare the performances (in terms of coverage probabilities) of the various UCL computation methods described in the literature. They demonstrated that for skewed data sets, UCLs based upon Student's t statistic, central limit theorem (CLT), and percentile bootstrap method tend to underestimate the population mean (EPC term). It is reasonable to state and assume the findings of the simulation studies performed on uncensored skewed data sets to compare the performances of the various UCL computation methods can be extended to skewed left-censored data sets. Based upon the findings of those studies performed on uncensored data sets and also using the findings summarized in Singh, Maichle, and Lee (2006), it is concluded that t-statistic, CLT, and the percentile bootstrap method based UCLs computed using KM estimates (and also ROS estimates) underestimate the population mean of moderately skewed to highly skewed data sets. Interested users may want to verify these statements via simulation experiments or otherwise. Like uncensored skewed data sets, for left-censored data sets, ProUCL 5.0 offers several parametric and nonparametric methods to compute UCLs and other limits which adjust for data skewness.

In earlier versions of the ProUCL software (e.g., ProUCL 4.00.02), for left-censored data sets, KM estimates were used in the normal distribution based equations to compute the various upper limits. However, normal distribution based upper limits (e.g., t-UCL) using KM estimates (or any other estimates such as ROS estimates) fail to provide the specified coverage to the parameters (e.g., mean, percentiles) of populations with skewed distributions (Singh, Singh, and Iaci, 2002, Johnson, 1978, Chen 1995). Also, the nonparametric UCL computation methods (e.g., percentile bootstrap) do not provide the desired coverage to the population means of skewed distributions (e.g., Hall [1988, 1992], Efron and Tibshirani, 1993). For an example, the use of t-UCL or the percentile bootstrap UCL method on robust ROS estimates or on KM estimates underestimates the population mean for moderately skewed to highly skewed data sets. Chapters 3 and 5 of this document describe parametric and nonparametric KM method based upper limits computation methods (and available in ProUCL 5.0) which adjust for data skewness.

The KM method yields good estimates of the population mean and standard deviation (Singh, Maichle, and Lee, 2006); however upper limits computed using the KM or ROS estimates in normal equations or in the percentile bootstrap method do not account for skewness present in the data set. Appropriate UCL computation methods which account for data skewness should be used on KM or ROS estimates. For left-censored data sets, ProUCL 5.0 computes upper limits using KM estimates in gamma (lognormal) UCL, UPL, and UTL equations (e.g., also suggested in EPA 2009) provided the detected observations in the left-censored data set follow a gamma (lognormal) distribution.

Recently, the use of the ISM methodology has been recommended (ISM ITRC, 2012) to collect soil samples needed to estimate mean concentrations of the DUs requiring characterization and remediation activities. ProUCL can be used to compute UCLs based upon ISM data as described and recommended in the ITRC ISM Tech Reg Guide (2012). At many sites, large amounts of discrete background data are already available which are not directly comparable to the actual field ISM data (onsite or background). To compare the existing discrete background data with field ISM data, the BISS module (blocked for general use in ProUCL version 5.0 awaiting guidance and instructions for its intended use) of ProUCL 5.0 can be used on a large (e.g., consisting of at least 30 observations) existing discrete background data set. The BISS module simulates incremental sampling methodology based equivalent incremental background

samples; and each simulated BISS sample represents an estimate of the mean of the population represented by the discrete background data set. The availability of a <u>large </u>discrete background data set collected from areas with geological conditions comparable to the DU(s) of interest (onsite DUs) is a requirement for successful application of this module. The user cannot see the simulated BISS data; however the simulated BISS data can be accessed by the various other modules of ProUCL 5.0 to perform desired statistical evaluations. For example, the simulated BISS data can be merged with the actual field ISM background data after comparing the two data sets using a two-sample t-test. The actual field ISM or the merged ISM and BISS data can be accessed by the various modules of ProUCL to compute a UCL of mean or a UTL.

ProUCL 5.0 Capabilities

A summary of statistical methods available in the ProUCL software is provided as follows.

<u>Assumptions:</u> Like most statistical methods, statistical methods to compute 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). If multiple populations (e.g., background and site data mixed together) are present in a data set, it is recommended to separate them out first by using the population partitioning techniques (e.g., Singh, Singh, and Flatman 1994), and then compute appropriate decision statistics (e.g., 95% UCLs) 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 5.0 and this guidance document. Parametric estimation and hypotheses testing methods (e.g., t-test, UCLs, UTLs) are based upon distributional (e.g., normal distribution, gamma) assumptions. ProUCL has GOF tests for normal, gamma, and lognormal distributions.

<u>Multiple Constituents/Variables:</u> Environmental scientists need to evaluate many constituents in their decision making processes (exposure and risk assessment). 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 test statistics (e.g., ANOVA test, trend test) for those variables/analytes.

<u>Analysis by a Group Variable:</u> 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 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 useful option to perform various statistical tests and methods including graphical displays separately for each of the group (samples from different populations) that may be present in a data set. For an example, the same data set may consist of analytical data from the various groups or populations representing site, background, two or more AOCs, surface, subsurface, monitoring wells. 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.

<u>Exploratory Graphical Displays for Uncensored and Left-Censored Data Sets:</u> Graphical methods

included in the Graph module of ProUCL include: Q-Q plots (data in same column), multiple Q-Q plots (data in different columns), box plots, multiple box plots, and histograms. These graphs can also be generated for data sets consisting of ND observations. Additionally, the OLS Regression and Trend Analysis module can be used to generate graphs displaying parametric OLS regression lines with confidence intervals and prediction intervals around the regression lines and nonparametric Theil-Sen trend lines. The Trend Analysis module can generate trend graphs for data sets without a sampling event variable, and also generate time series graphs for data sets with a sampling event (time) variable. ProUCL 5.0 accepts only numerical values for the event variable. Graphical displays of a data set are useful to gain added insight contained in a data set that may not otherwise be clear by looking at test statistics such as t-test, Dixon test or T-S test. 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 nondetect 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 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 outliers and multiple populations that may be present in a data set. On a Q-Q plot, observations well-separated from the majority of the data may represent potential outliers coming from a population different from the main dominant population (e.g., background population). In a Q-Q plot, jumps and breaks of significant magnitude suggest the presence of observations coming from multiple populations (onsite and offsite areas). ProUCL can also be used to display box plots with horizontal lines displayed at pre-specified compliance limits or computed upper limits (e.g., UPL, UTL) superimposed on the same graph. This kind of graph provides a visual comparison of site data with compliance limits and/or BTV estimates.

Outlier Tests: ProUCL also has a couple of classical outlier test procedures (EPA 2006b, 2009), such as the Dixon test and the Rosner test. The details of these outlier tests are described in Chapter 7. These outlier tests often suffer from masking effects in the presence of multiple outliers. It is suggested that the classical outlier procedures should always be accompanied by graphical displays including box plots and Q-Q plots. Description and use of the robust and resistant (to masking) outlier procedures (Rousseeuw and Leroy, 1987; Singh and Nocerino, 1995) are beyond the scope of ProUCL 5.0. Interested users are encouraged to try the Scout 2008 software package (EPA 2009) to use the robust outlier identification methods especially when dealing with multivariate data sets consisting of data for several variables/analytes.

Outliers represent observations coming from populations different from the main 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 dominant population and not to compute distorted statistics by accommodating a few low probability outliers (e.g., by using a lognormal distribution). 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 the higher order statistics.

Goodness-of-Fit Tests: In addition to computing simple summary statistics for data sets with and without NDs, ProUCL 5.0 has GOF tests for normal, lognormal and gamma distributions. To test for normality (lognormality) of a data set, ProUCL has the Lilliefors test and the extended S-W test for samples of sizes up to 2000 (Royston, 1982, 1982a). For the gamma distribution, two GOF tests: the Anderson-Darling test (1954) and Kolmogorov Smirnov test (Schneider, 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 (Anderson-Darling and Kolmogorov Smirnov tests) obtained using the two gamma deviate generation algorithms: Whitaker (1974) and Marsaglia and Tsang (2000). For values of $k \leq 0.2$, the critical values of the two gamma GOF tests: Anderson-Darling (A-D) and Kolmogorov-Smirnov (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 values of the shape parameter, $k=0.025, 0.05, 0.1$, and 0.2, the critical value tables for these two GOF tests 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. It should be noted that for $k=0.2$, the older (generated in 2002) and the newly generated critical values are in general agreement.

ProUCL also generates GOF Q-Q plots for normal, lognormal, and gamma distribution 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 this guidance document. For data sets consisting of 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 the detected values. Typically, 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 methods listed above. Using the "Imputed NDs using ROS Methods" option of the "Stats/Sample Sizes" module of ProUCL 5.0, additional columns can be generated to store 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: Sample Sizes module in ProUCL can be used to develop DQOs based sampling designs needed to address statistical issues associated with the various polluted sites projects. ProUCL 5.0 provides user friendly options to enter 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 the selected statistical applications including: estimation of mean, single and two-sample hypothesis testing approaches, and acceptance sampling. Both parametric (e.g., for t-tests) and nonparametric (e.g., Sign test, WRS test) sample size determination methods as described in EPA (2002c, 2006a, 2006b) and MARSSIM (2000) guidance documents are available in ProUCL version 5.0. 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 of 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.

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

10

[1988,1992]; Singh, Singh, and Iaci 2002; Singh and Singh, 2003, Singh, Maichle and Lee 2006) as the interval estimates based upon these bootstrap methods fail to provide the specified coverage (e.g., UCL 95 does not provide adequate 95% coverage to population mean) to the 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 based upon data sets consisting of NDs and outliers. Due to the reasons described above, whenever applicable, ProUCL 5.0 provides cautionary notes and warning messages regarding the use of bootstrap-t and Halls bootstrap UCL methods.

- For nonparametric uncensored and left-censored data sets with NDs, depending upon data variability and skewness, ProUCL recommends the use of BCA bootstrap, bootstrap-t, or Chebyshev inequality based methods to compute decision statistics.

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 without NDs, 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 normal approximation are computed using the continuity correction factors. The mathematical details of the various Single-sample and Two-Sample hypotheses testing approaches are described in Chapter 6 of this document.

- Single-sample: parametric (Student's t-test) and nonparametric (Sign test, WSR test, tests for proportions and percentiles) hypotheses testing approaches are available in ProUCL. The single-sample hypotheses tests are used when the 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. Specifically, a t-test (or a sign test) may be used to verify the attainment of cleanup levels at an AOC) after a remediation activity has taken place; and a test for proportion may be used to verify if the proportion of exceedances of an action level (or a compliance limit) by sample 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. Specifically, a t-test or a Wilcoxon Signed Rank (WSR) test are used to compare the measures of location and central tendencies (e.g., mean, median) of a site area (e.g., AOC) to a cleanup standard, C_s or action level also representing a measure of central tendency (e.g., mean, median); whereas, a proportion test compares if the proportion of site observations from an AOC exceeding a compliance limit (CL) exceeds a specified proportion, P_0 (e.g., 5%, 10%). The percentile test compares a specified percentile (e.g., 95th) of the site data to a pre-specified upper threshold (e.g., action level).

- Two-sample: 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, compare concentrations of GW monitoring wells (MWs). It should be noted that as cited in the literature, some of the hypotheses testing approaches (e.g., nonparametric two-sample WMW) deal with the single detection limit scenario. 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 the multiple Q-Q plots and side-by-side box plots. Gehan test or Tarone-Ware (new in ProUCL 5.0) should be used in cases where multiple detection limits are present.

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 the various decision making statistics such as UCLs of mean (EPA 2002a), percentiles, UPLs for future k (≥1) observations, UTLs (e.g., EPA 1992b, EPA 2009) and upper simultaneous limits (USLs) (Singh and Nocerino, [1995, 2002]) based upon uncensored full data sets and left-censored data sets consisting of 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 5.0 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 KM method (1958), ROS methods (Helsel, 2005) and substitution methods such as replacing NDs by DL or DL/2 (Gilbert, 1987, EPA 2006b). Substitution DL/2 method has been incorporated in ProUCL for research and comparison purposes as requested by EPA scientists.

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 MVUEs of parameters of a lognormal distribution). Nonparametric UCL computation methods for data sets without NDs include: CLT-based UCL, Modified-t-statistic (adjusted for skewness)-based UCL, Adjusted-CLT (adjusted for Skewness)-based UCL, Chebyshev inequality based-UCL (using sample mean and standard deviation), Jackknife method-based UCL, 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.

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 normal, gamma, and lognormal distributions. Nonparametric upper limits are typically based upon order statistics of a data set such as a background or a reference area 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 confidence coefficient and coverage probability, ProUCL 5.0 outputs the actual confidence coefficient 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 this document.

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 mean and *sd* computed using logROS (LROS, robust ROS), Gamma ROS (GROS), KM, and DL/2 methods. For nonparametric data sets, to adjust for skewness, ProUCL uses bootstrap methods and Chebyshev inequality to compute UCLs and other limits using estimates of mean and standard deviation obtained using methods listed above. ProUCL also 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 to

compute UCLs, UTLs, and other upper limits. Based upon a Monte Carlo study performed by Singh, Maichle, and Lee (EPA, 2006), ProUCL recommends the use of the Kaplan-Meier (1958) estimates in bootstrap and Chebyshev inequality to compute the various decision statistics (e.g.,UCL95, UPL, UTL) of interest. ProUCL 5.0 suggests the use of KM-Gamma upper limits when the detected data follow a gamma distribution. ProUCL computes KM estimates directly using left-censored data sets without flipping data and re-flipping decision statistics. The KM method incorporated in ProUCL computes both *sd* and standard error (*SE*) of the mean. 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. The inclusion of the substitution method 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 description of the various other upper limits: UPLs, UTLs, and USLs for data sets with NDs are given in Chapter 5.

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; and performing inter-well comparisons comparing concentrations of 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: M-K trend test and Theil-Sen trend test. Using this option, one can generate trend graphs and time-series graphs displaying Theil-Sen trend line and all other statistics of interest with associated p-values.

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 the various GW monitoring wells over a certain period of time (EPA 2009). The details of Oneway ANOVA are given in Chapter 9, and OLS regression line and Trend tests methods are described in Chapter 10.

BISS Module: At many sites, a large amount of discrete onsite and background data are already available which are not directly comparable to actual field ISM data. In order to provide a tool to compare the existing discrete data with ISM data, the BISS module of ProUCL 5.0 may be used on a large existing discrete data set. The ISM methodology used to develop the BISS module is a relatively new approach; methods incorporated in this BISS module require further investigation. The BISS module has been temporarily blocked for use in ProUCL 5.0 as this module is awaiting adequate guidance for its intended use on discrete background data sets.

Recommendations and Suggestions in ProUCL: Not much guidance is available in the environmental literature including the available guidance documents to compute rigorous UCLs, UPLs, and UTLs for moderately skewed to highly skewed uncensored and left-censored data sets consisting of NDs with multiple DLs, a common occurrence in environmental data sets. For uncensored positively skewed data sets, Singh, Singh, and Iaci (2002) and Singh and Singh (2003) performed extensive simulation experiments to compare the performances (in terms of coverage probabilities) of several UCL computation methods described in statistical and environmental literature. They noted that the optimal choice of a decision statistic (e.g., UCL 95) depends upon the sample size, data distribution and data skewness. Until 2006, not much guidance was available on how to compute UCL95 of mean and other upper limits (e.g., UPLs and UTLs) based upon skewed data sets consisting of NDs with multiple DLs. For data sets with NDs, Singh, Maichle, and Lee (EPA 2006) conducted a similar simulation study to compare the performances of the various estimation methods (in terms of bias in the mean estimate); and

of some the 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 estimate of mean; and for skewed data sets, t-statistic, CLT, and the percentile bootstrap method based UCLs computed using KM estimates (and ROS estimates) underestimate the population mean. Based upon the 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 to compare the performances of the various UCL computation methods can be extended to skewed left-censored data sets.

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, based upon the sample size, data distribution, and data skewness, some modules of ProUCL make suggestions about using a decision statistic to estimate population parameters of interest (e.g., EPC). The recommendations made in ProUCL are based upon the extensive experience of the developers in environmental statistical methods, published literature (e.g., Efron and Tibshirani, 1993; Hall, 1988; Singh, Singh, and Engelhardt 1997; Singh, Singh, and Iaci 2002; and Singh, Maichle, and Lee 2006) and procedures described in the various EPA guidance documents (EPA [1992a, 1992b 2002a, 2002b, 2006b, 2009, 2009a, 2009b]). Based upon the conceptual site model (CSM), expert site and regional knowledge, the project team should make the final decision regarding using or not using the suggestions made by ProUCL. If deemed necessary, the project team may want to consult a statistician.

Even though, ProUCL 5.0 has been developed using limited government funding, for data sets with and without NDs, ProUCL 5.0 provides many statistical and graphical methods described in the EPA documents cited above. However, one may not compare the availability of methods in ProUCL 5.0 with methods available in the commercial software packages such as SAS and Minitab 16. For example, trend tests correcting for seasonal/spatial variations are not available in the ProUCL software. For those methods the user is referred to the commercial software packages. As mentioned earlier, it is recommended to supplement test results (e.g., two-sample test) with graphical displays (e.g., Q-Q plots, side-by-side box plots); especially when data sets consist of NDs and outliers. With the inclusion of BISS module, Oneway ANOVA, Regression and Trend tests, and the user-friendly DQOs based Sample Size determination modules, ProUCL represents a comprehensive statistical software package equipped with statistical methods and graphical tools needed to address many environmental sampling and statistical issues as described in the various CERCLA (EPA 1989a, 1992a, 2002a, 2002b, 2006a, 2006b), MARSSIM (EPA 2000), and RCRA (EPA 1989b, 1992b, 2002c, 2009) guidance documents.

Finally, the users and practitioners 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 (2012) will fail to provide desired coverage to the 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.

14

ProUCL 5.0 User Guide

In addition to this Technical Guide, a User Guide also accompanies the ProUCL 5.0 software, providing details of using the statistical and graphical methods incorporated in ProUCL 5.0.00. The User Guide provides details about the input and output operations that can be performed using ProUCL 5.0. The User guide also provides details about saving edited input files, output Excel-type spreadsheets and graphical displays generated by ProUCL 5.0.

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 remediation decisions that affect human health and the environment. For example, a background data set of size less than 6 is not large enough to characterize background population, to compute background threshold values (BTV) estimates, or to perform background versus site comparisons. Several EPA guidance documents (e.g., MARSSIM 2000; EPA [2006a, 2006b]) describe data quality objectives (DQOs) and minimum sample size computations needed to address statistical issues associated with the various environmental applications. In order to obtain reliable results using statistical methods, an adequate amount of data should be collected using desired DQOs (confidence coefficient, decision error rates). The Sample Sizes module of ProUCL computes DQOs based minimum sample sizes needed to use the statistical methods described in the various guidance documents. In some cases, it may not be possible (e.g., due to resource constraints) to collect DQOs based number of samples; under these circumstances one can use the Sample Sizes module to assess the power of the test statistic used in retrospect. Some suggestions about the minimum sample size requirements needed to use statistical methods to estimate environmental parameters of interest such as exposure point concentration (EPC) terms and BTVs, to compare site data with background data or with some pre-established screening levels (e.g., action levels [ALs], compliance limits [CLs]), are provided in this chapter. It is noted that similar minimum sample size suggestions made by ProUCL (EPA 2007, 2009a, 2009b) have been made in some other guidance documents including the RCRA Guidance Document (EPA 2009).

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 with. 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 onsite observations are compared with those threshold values. It should be noted that 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 to perform the comparisons described above.

1.1   Background Data Sets

Based upon the conceptual site model (CSM), the project team familiar with the site 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. An appropriate random sample of independent observations (e.g., *i.i.d*) should be collected from the background area. A defensible background data set represents a "single" population possibly without any outliers. In a background data set, in addition to reporting

and/or laboratory errors, statistical outliers may also be present. A few elevated statistical outliers present in a background data set may actually represent potentially contaminated locations belonging to an impacted site areas and/or possibly from other polluted site(s); those elevated outliers may not be coming from the main dominant background population under evaluation. Since the presence of outliers in a data set tends to yield distorted (incorrect and misleading) values of the decision making statistics (e.g., UCLs, UPLs and UTLs), elevated outliers should not be included in background data sets and estimation of BTVs. The objective here is to compute background statistics based upon the majority of the data set representing the main dominant background population, and not to accommodate a few low probability high outliers (e.g., coming from extreme tails of the data distribution) that may also be present in the background data set. The occurrence of elevated outliers is common 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. Since both of these classical tests suffer from masking effects (e.g., some extreme outliers may mask the occurrence of other intermediate outliers), it is suggested that these classical outlier tests be supplemented with graphical displays such as a box plot and a Q-Q plot. The use of exploratory graphical displays helps in determining the number of outliers potentially present in a data set. The use of graphical displays also helps in identifying extreme high outliers as well as intermediate and mild outliers. The use of robust and resistant outlier identification procedures (Singh and Nocerino, 1995, Rousseeuw and Leroy, 1987) is recommended when multiple outliers are present in a data set. Those methods are beyond the scope of ProUCL 5.0. However, several robust outlier identification methods are available in the Scout 2008 version 1.0 software package (EPA 2009).

An appropriate background data set of a reasonable size (preferably computed using DQOs processes) is needed to represent a background area and to compute upper limits (e.g., estimates of BTVs) based upon background data sets and also to compare site and background data sets using hypotheses testing approaches. At the minimum, a background data set should have at least 10 (more observations are preferable) observations to perform background evaluations.

1.2   Site Data Sets

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

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 action levels 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, it is recommended to use UTL or upper simultaneous limit (USL) to control the false pos1itive error rate (Type I Error Rate). Alternatively, one can use hypothesis testing approaches 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 the BTV (e.g., 95th 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.

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

- two-sample hypotheses 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 10 observations from each population under comparison.

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) thus obtained is questionable. In these situations it is suggested to supplement the test statistics decisions by graphical displays.

1.3 Discrete Samples or Composite Samples?

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

- When using a single-sample hypothesis testing approach, site data can be obtained by collecting all discrete or all composite samples. The hypothesis testing approach is used when many (e.g., $\geq$

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 the ProUCL 5.0 Tech Guide.

- 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 5.0 are described in Chapter 6 of the ProUCL 5.0 Tech Guide.

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 to represent: 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, an average value, μ_0, and a not-to-exceed value, A_0, represent two significantly different parameters, and different statistical methods and limits are used to compare the site data with these two very different threshold values. Statistical limits, such as an UCL of the population mean, an 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: an EPC term (μ_0) and a BTV (A_0). A new upper limit, USL has been included in ProUCL 5.0 which may be used to estimate a BTV based upon a well-established background data set 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. Specifically, the differences between UCLs and UPLs (or upper percentiles), and UCLs and UTLs should be clearly understood and acknowledged. A UCL with a 95% confidence limit (UCL95) of the mean represents an estimate of the population mean (measure of the central tendency), 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 $\leq$ UCL95 of Mean $\leq$ Upper 95th Percentile $\leq$ UPL95 of a Single Observation $\leq$ UTL95-95

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). The differences among the various upper limits discussed above are illustrated by the following example.

Example 1.1. Consider a background real data set collected from a Superfund site (EPA 2002b). The data set has several inorganic COPC, including aluminum, arsenic, chromium, iron, and lead. Iron concentrations follow a normal distribution. Some upper limits for the iron data set are summarized as follows. However, the various upper limits do follow the order as described above.

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

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

A brief discussion about the differences between the applications and uses of the various statistical limits is provided below.

- 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 based upon 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 an UTL represents an upper limit to be used for point-by-point individual site observation comparisons. UPLs and UTLs are computed based upon background data sets, and 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 least10) 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 5.0.

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 as 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) that site location (or a monitoring well) exhibiting constituent concentration in excess of the BTV. If all collocated (or collected during the same time period) sample observations collected 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 cleanup.

When BTV constituent concentrations are not known or pre-established, one has to collect or extract a background data set of an appropriate size that can be considered representing 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, enough background observations (minimum of 10) should be collected, perhaps using an appropriate DQOs process as described in EPA (2006a) and MARSSIM (2000). Several statistical limits listed above are used to estimate the 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 contaminants/constituents of potential concern (COPCs), 2) identify the potentially polluted site areas of concern (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 various onsite locations representing the site area under investigation, then the use of hypotheses testing approaches (both single-sample and a two-sample) is preferred. The use of hypothesis testing approaches 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 extract, 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) about the number of site observations that should be collected and compared with the "pre-established" standards before coming to a conclusion about the status (clean or polluted) of the site AOCs. As mentioned earlier, when the number of available site samples is less than 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 (e.g., the average value, μ_0, or a not-to-exceed value, A_0), represented by the known threshold value, one can use single-sample 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 Technical Guide.

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 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 all NDs (e.g., associated detection limits) fall below the specified threshold value, C_s. These tests are used to compare the site location (e.g., median, mean) with some specified 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 (or equivalently a UTL 95-95 UTL 95-90) may be used to compare site proportion (or site percentile) with the specified threshold or action level, A_0.

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) to perform two-sample hypothesis test approaches are described in EPA (2002b, 2006a, 2006b), MARSSIM (2000) and also in Chapter 6 of the ProUCL 5.0 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 5.0 are listed as follows:

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

The Gehan and Tarone-Ware tests are meant to be used on left-censored data sets with multiple detection limits (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 be all discrete or all composite (obtained using the same design and pattern), and be collected from the same medium (soil) at similar depths (e.g., all surface samples or all subsurface samples) and time (e.g., during the same quarter in groundwater applications) using comparable (preferably same) analytical methods. Good sample collection methods and sampling strategies are given in EPA (1996, 2003) guidance documents.

Notes: ProUCL 5.0 (and previous versions) has been developed using limited government funding. ProUCL 5.0 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 5.0 with methods incorporated in commercial software packages such as SAS and Minitab 16. 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 may 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 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) need to be collected from background areas (e.g., site-specific, general reference area, or historical data). The project team and site experts should decide what represents a site population and what represents a background population. The project team should determine the population area and boundaries based upon all current and 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 these statistical tests, one needs to determine the appropriate sample sizes that need to be collected from the populations (e.g., site and background) under investigation using appropriate DQOs

processes (EPA [2006a, 2006b]; MARSSIM, 2000). ProUCL has the Sample Sizes module which can be used to develop DQOs based sampling designs needed to address statistical issues associated with the various 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 5.0. ProUCL 5.0 also has sample size determination methods for acceptance sampling of lots of discrete objects such as a lot of drums containing hazardous waste (e.g., RCRA applications, EPA 2002c).

However, due to budget constraints, it may not be possible to collect the same number of samples as determined by using a DQOs process. For example, the data might have already been collected (often is the case) without using a DQOs process, or due to resource constraints, it may not be possible to collect as many samples as determined by using a DQOs based sample size formula. In practice, the project team and the decision makers may decide not to collect enough background samples. It is suggested to collect at least10 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, though 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 DQOs 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 (e.g., using Sample Size module of ProUCL) 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.

Notes: From a mathematical point of view, the statistical methods incorporated in ProUCL and described in this guidance document to estimate EPC terms and BTVs, and compare 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. It is recommended not to use those statistics to draw cleanup and remediation decisions potentially impacting human health and the environment. The minimum sample size recommendation (at least 10 observations) may be used only when data sets of size determined by a DQOs process (EPA, 2006) cannot be collected. Some of the recent guidance documents (e.g., EPA 2009) are also suggesting collecting a minimum of about 10 samples in the circumstance that data cannot be collected using a DQOs based process.

- To allow the users to compute decision statistics based upon composite data collected using the Incremental Sampling Methodology (ITRC, 2012), ProUCL 5.0 will compute decision statistics (e.g., UCLs, UPLs, UTLs) based upon samples of sizes as small as 3. The user is referred to the ITRC ISM Tech Reg Guide (2012) to determine which UCL (e.g., Student's t-UCL or Chebyshev UCL) should be used to estimate the EPC term.

1.7.1 Sample Sizes for Bootstrap Methods

Several nonparametric methods including bootstrap methods to compute UCL, UTL, and other limits for both full-uncensored data sets and left-censored data sets with NDs are available in ProUCL 5.0. 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 (with replacement) 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. Moreover, bootstrapping a large data set of size greater than 500 or 1000 will be time consuming.

1.8   Statistical Analyses by a Group ID

The analyses of data categorized by a group ID variable such as: 1) Surface vs. Subsurface; 2) AOC1 vs. AOC2; 3) Site vs. Background; and 4) Upgradient vs. Downgradient monitoring wells are common in environmental and various other applications. ProUCL 5.0 offers this option for data sets with and without NDs. The Group Option provides a useful tool to perform various statistical tests, methods and generate graphical displays separately for each of the group (samples from different populations) that may be present in a data set. The graphical displays (e.g., box plots, (quantile-quantile) Q-Q 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 users' responsibility to provide adequate amount of data to perform the group operations. For an 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, standard deviation [*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. ProUCL generates a warning message (colored orange) in the lower Log Panel of the ProUCL 5.0 screen.

1.9   Statistical Analyses for Many Constituents/Variables

ProUCL software can process multiple analytes/variables simultaneously in a user friendly manner – an option not available in other software packages such as Minitab 16 (2012), NADA for R (Helsel, 2013). This option is very 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 those 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 tend to use the maximum detected value as an estimate of the EPC term. This is especially true when the sample size is small such as < 5, or when a UCL95 exceeds the maximum detected values (EPA, 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). The use of higher order statistics to estimate the UTLs depends upon the sample size. For an example, 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 values (EPA, 1992a). 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. ProUCL computes 95% UCLs of mean using several methods based upon normal, gamma, lognormal, and non-discernible distributions. In the past (e.g., EPA 1992), a lognormal distribution was used as the default distribution to model positively skewed environmental data sets; and 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 (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 skewed data sets of smaller sizes (e.g., < 30, < 50,...), H-UCL often exceeds the maximum detected value. Since the use of a lognormal distribution has been quite common (e.g., suggested as a default model in a 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 a potential outlier 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. Singh and Singh (2003) studied the performance of the max test (using the maximum observed value to estimate the EPC) via Monte Carlo simulation experiments. They noted that for skewed data sets of small sizes (e.g., < 10-20), even the max test does not provide the specified 95% coverage to the population mean, and for larger data sets it overestimates the EPC term, which may lead to unnecessary further remediation.

Today, several methods, some of which are described in EPA (2002a), are available in the various versions of ProUCL (e.g., ProUCL 3.00.02 [EPA 2004], ProUCL 4.0 [EPA 2007], ProUCL 4.00.05[EPA 2009, 2010]) to estimate the EPC terms. For data sets with NDs, ProUCL 5.0 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   *Chebyshev Inequality Based UCL95*

ProUCL 5.0 (and its earlier versions) displays a warning message when the suggested 95% UCL (e.g., Hall's or bootstrap-t UCL with outliers) of the mean exceeds the detected maximum concentration.
When a 95% UCL does exceed the maximum observed value, ProUCL recommends the use of an alternative UCL computation method based upon the Chebyshev inequality. One may use a 97.5% or 99% Chebyshev UCL to estimate the mean of a highly skewed population. The use of the Chebyshev inequality to compute UCLs tends to yield more conservative (but stable) UCLs than other methods available in ProUCL software. In such cases, when the sample size is large (and other UCL methods such as the bootstrap-t method yield unrealistically high values due to presence of outlier(s)), one may want to use a 95% Chebyshev UCL or a Chebyshev UCL with lower confidence coefficient such as 90% as an estimate of the population mean, especially when the sample size is large (e.g., >100, 150). The detailed recommendations (as functions of sample size and skewness) for the use of those UCLs are summarized in various versions of ProUCL Technical Guides (EPA, 2004, 2007, 2009, and 2010d).

Notes: It is recommended not to use the maximum observed value to estimate the EPC term representing the average exposure contracted by an individual over an EA. For the sake of interested users, ProUCL displays a warning message when the recommended 95% UCL (e.g., Hall's bootstrap UCL) of the mean exceeds the observed maximum concentration. For such scenarios (when a 95% UCL does exceed the maximum observed value), an alternative 95% UCL computation method based upon Chebyshev inequality is recommended by the ProUCL software.

1.11   Samples with Nondetect Observations

ND observations are inevitable in most environmental data sets. Singh, Maichle, and Lee (EPA, 2006) studied the performances (in terms of coverages) of the various UCL95 computation methods including the simple substitution methods (such as the DL/2 and DL methods) for data sets with ND observations. They concluded that the UCLs obtained using the substitution methods, including the replacement of NDs by respective 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 to compute UCL95 based upon data sets with ND observations.

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

Based upon the results of the report by Singh, Maichle, and Lee (EPA, 2006), it is recommended to avoid the use of the DL/2 method to perform a GOF test, and to compute the summary statistics and various other limits (e.g., UCL, UPL, UTLs) often used to estimate the EPC terms and BTVs. Until recently, the DL/2 method has been the most commonly used method to compute the various statistics of interest for data sets with 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 Kaplan-Meier [KM] estimates) including Chebyshev inequality and bootstrap methods with better performance are available that can be used to

compute the various upper limits of interest. Several of those parametric and nonparametric methods are available in ProUCL 4.0 and higher versions. It should be noted that 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 DL/2 method in ProUCL for comparison and research purposes.

Notes: Even though the DL/2 method (to compute UCLs, UPLs, and for goodness-of-fit [GOF] tests) has been incorporated in ProUCL, its use is not recommended due to its poor performance. The DL/2 method has been retained in ProUCL 5.0 for historical and comparison purposes. NERL-EPA, Las Vegas strongly recommends avoiding the use of DL/2 method even when the % of NDs is as low as 5% to 10%.

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 reporting limit (RL) or the maximum RL/2. Statistics (e.g., UCL95) computed based upon only a few detected values (e.g., < 4) cannot be considered reliable enough to estimate the EPC terms having 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 the EPC terms and other upper limits. Specifically, it is suggested that for data sets consisting of less than 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 together should decide on a site-specific basis on how to estimate the average exposure (EPC term) for the constituent and area under consideration. For such data sets with low detection frequencies, other measures such as the median or mode represents better estimates (with lesser uncertainty) of the population measure of central tendency.

Additionally, it is also suggested that when most (e.g., > 95%) of the observations for a constituent lie below the DLs, the sample median or the sample mode (rather than the sample average) may be used as an estimate the EPC term. 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 5.0

In addition to performing background versus site comparisons for CERCLA and RCRA sites, and estimating the EPC terms in exposure and risk evaluation studies, the statistical methods as incorporated in ProUCL 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 the COPCs during the screening phase of a cleanup project to be conducted at a contaminated site. The screening for the COPCs is performed prior to any characterization and remediation activities that may have to be conducted at the site. 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 to obtain reliable 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 by a MW concentration may be considered an indication of contamination in that MW. In such individual concentration comparisons, the presence of contamination (determined by an exceedance) may have to be confirmed by re-sampling from that MW. If concentrations of constituents in the original sample and re-sample(s) exceed the MCL or BTV, then that MW may require further scrutiny, perhaps triggering remediation remedies as determined by the project team. If the concentration data from a MW for about 4 to 5 continuous quarters (or some other designated time period determined by the project team) are below the MCL or BTV level, then that MW may be considered as complying with (achieving) the pre-established or estimated standards.

1.13.3 Verification of the Attainment of Cleanup Standards, C_s

Hypothesis testing approaches are used to verify the attainment of the cleanup standard, C_s, at polluted 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 DQOs 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 site 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, μ_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 be considered as representing impacted locations with elevated concentrations (hot spots).

1.14   Some General Issues and Recommendations made by ProUCL

Some general issues regarding the handling of multiple detection limits and field duplicates by ProUCL and recommendations made about various substitution and regression on order statistics (ROS) methods for data sets with NDs are described in the following sections.

1.14.1 Multiple Detection Limits

ProUCL 5.0 does not make distinctions between method detection limits (MDLs), adjusted MDLs, sample quantitation limits (SQLs), or DLs. Multiple DLs in ProUCL mean different values of the DL. An indicator variable with of 0 (=nondetect) and 1(= detect) is assigned to each variable consisting of NDs. All ND observations in ProUCL are indentified by the value '0' of the indicator variable used in ProUCL to distinguish between detected (=1) and nondetected (=0) observations. It is the users' responsibility to supply correct numerical values for NDs (should be entered as the reported detection limit or RL values) and not as qualifiers (e.g., J, U, B, UJ, ...) for ND observations in the data set.

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

For data sets with NDs, ProUCL 5.0 can compute point estimates of population mean and standard deviation using the KM and ROS methods (and also using DL/2 method). The DL/2 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 the various 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 EPC terms than other UCL methods based upon Student's t-statistic on KM estimates, percentile bootstrap method on KM or ROS estimates.

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

Several ProUCL 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 ProUCL v4 software and its associated guidance documents (EPA 2007, 2009a, 2009b, 2010c, and 2010d) in the Unofficial ProUCL v4 User Guide 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 comments made in the Unofficial ProUCL v4 User Guide to assure that: 1) rigorous statistical methods are used to compute the 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 Practical Stats 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 5.0 Technical and User Guides. The detailed responses to the comments made about the ProUCL software in the Unofficial ProUCL v4 User Guide are provided elsewhere.

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 may not compare ProUCL with the commercial software packages which are expensive and not as easy to use as the ProUCL software to address 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 the 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 and 2010d]). Even though for uncensored data sets, programs to compute gamma distribution based UCLs and UPLs are available in R Script, programs to compute a 95% UCL of mean based upon a gamma distribution on KM estimates are not easily available in commercial software packages and in R script.

- 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 (2010) 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) has all of these 9 types which can be computed 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 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 (2012) 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) are given by:

$$\text{mean} = \mu_1 = \exp(\mu + 0.5\sigma^2); \text{ and median} = M = \exp(\mu)$$

  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 pth 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.

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 1-2. Probabilities $p(\bar{x} < \mu_1)$ **Computed for Lognormal Distributions with μ=4 and Varying Values of σ**

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

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

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. 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 to gain information (e.g., outliers, multiple populations, data distribution, patterns, and skewness) present in 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 available in the various commercial software packages to get an idea about the data distribution (skewed or symmetric), to identify outliers, and to compare multiple groups (main objectives of box plots in ProUCL).

  o As mentioned earlier, it is a matter of choice of using percentiles/quartiles to construct a box plot. There is no 'best' method to construct a box plot. Many software packages use one method (e.g., out of 9 described 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 identify outliers, Q-Q plots are also used to assess data distributions. Multiple Q-Q plots are useful to perform 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 to generate 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 also, 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 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 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.

  o However, 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 $\mu_1 - \mu_2$ 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]).

- Helsel (2012a) suggested 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, 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 to use data mining exploratory methods on multivariate data sets with NDs. Additional discussion on this topic is considered in Chapter 4.

  o 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 Gehan and Tarone-Ware tests in ProUCL5.0 which are better suited for data sets with multiple detection limits.

- <u>Outliers:</u> 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; and therefore may require further investigation.

  o 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 part 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 2009) software package.

  o 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. 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 potentially representing 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.

- 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 (2012 a) 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).

- 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 an 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 do 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 tends not to accommodate contamination (elevated values) as can be seen in Examples 2-1and 2-2 of Chapter 2. 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.

o   In the context of computing a UCL95 of mean, Helsel and Gilroy (2012) and Helsel (2012a) state that GROS and LROS are probably never better than KM. It should be noted that these three estimation methods compute estimates of mean and standard deviation and not the upper limits used to estimate EPC terms and BTVs. The use of KM method does yield good estimates of mean and standard deviation as noted by Singh, Maichle, and Lee (2006). 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.

o   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 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)) have developed parametric (e.g., gamma) and nonparametric (e.g., bootstrap-t and Hall's bootstrap method, modified-t, and Chebyshev inequality) methods to compute confidence intervals and upper limits which adjust for data skewness.

o   Analytically, it is not feasible to compare the various estimation and UCL computation methods for skewed data sets consisting of nondetect 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 to compare the performances of the various UCL computation methods can be extended to skewed left-censored data sets.

o   Like uncensored skewed data sets, for left-censored data sets, ProUCL 5.0 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 bootstrap-t method, and in 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 this ProUCL Technical Guide), 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 data skewness and data distribution, for gamma distributed data sets, the use of Gamma distribution was suggested to compute the UCL of mean.

Upper limits (e.g., UCLs, UPLs, UTLs) computed using the Student's t statistic and percentile bootstrap method (Helsel, 2012, 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 have been incorporated in ProUCL 5.0 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 USEPA is to compute rigorous decision statistics to help the decision makers and project teams in making correct decisions which are 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 (2012):

"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 left-censored data set consisting of nondetect (NDs) observations with one reporting limit of 20 illustrating this issue is described as follows.

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

36

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 method. The statement made in Helsel's text holds when all observations reported as detects are greater than the single reporting limit which is seldom true in environmental data sets consisting of analytical concentrations.

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 cleanup decisions are made based upon the mean concentrations of the contaminants/constituents of potential concern (COPCs). To address the uncertainty associated with the sample mean, a 95% upper confidence limit (UCL95) is used to estimate the unknown population mean, μ_1. A UCL95 is routinely used to estimate the exposure point concentration (EPC) term (EPA, 1992; EPA, 2002a). A UCL95 of mean represents that limit such that one can be 95% confident that the population mean, μ_1, will be less than that limit with 95% confidence. From a risk point of view, a 95% UCL of mean represents a number that is health protective when used to compute risk and health hazards. It is therefore important to compute a reliable, defensible (from human health point of view) and cost-effective estimate of the exposure point concentration (EPC) term. 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.

Recently, the use of a Student's t-UCL95 and Chebyshev inequality based UCL95 has been recommended (ITRC, 2012) to estimate EPC terms using incremental sampling methodology (ISM) based soil samples collected from the various decision units (DUs). In order to facilitate the computation of ISM data based estimates of the EPC term, ProUCL5.0 can compute a 95% UCL of the mean based upon data sets of sizes as small as 3. Additionally, the UCL module of ProUCL can be used on datasets with NDs collected using the ISM approach. However, it is advised that the users do not compute the decision making statistics (e.g., UCLs, upper prediction limits [UPLs], upper tolerance limits [UTLs]) based upon discrete data sets consisting of less than 8-10 observations.

For uncensored data sets without nondetect (ND) observations, theoretical details of the various UCL computation methods based upon Student's t- and percentile bootstrap as wells as more complicated bootstrap-t and gamma distribution 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; and an H-statistic based Land's (Land, 1971, 1975) H-UCL is used to estimate the EPC term. 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 (by increasing variability) to increase by a large amount (Singh, Singh, and Iaci, 2002). Realizing that the use of the H-statistic can result in an

unreasonably large UCL, it has been recommended (EPA, 1992a) to use the maximum value as an estimate of EPC term in cases when the H-UCL exceeds the largest value in the data set. For uncensored data sets without any NDs, ProUCL makes recommendations on how to compute an appropriate UCL95. Those recommendations are based upon the findings of simulation studies summarized in Singh and Singh (2003) and Singh, Singh, and Iaci (2002).

It is noted that 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 if an environmental data set follows a gamma distribution. For data sets following a gamma distribution, the EPC term should be estimated using an adjusted gamma (when $n<50$) or approximate gamma (when $n\geq 50$) UCL95 of the mean. A gamma distribution based UCL95 of the mean approximately provides the specified 95% coverage to the population mean, $\mu_1 = k\theta$ of a gamma distribution, $G(k, \theta)$ with k and θ respectively representing the shape and scale parameters. For highly skewed gamma distributed data sets with values of the shape parameter, $k < 0.1$, a 95% UCL may be computed using the bootstrap-t-method or Hall's bootstrap method when the sample size, n, is less small, such as <15 to 20, and for larger samples 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.

It is noted that <u>unlike</u> the percentile bootstrap and bias-corrected accelerated bootstrap (BCA) methods, bootstrap-t and Hall's bootstrap methods (Efron and Tibshirani, 1993) account for data skewness and their use is recommended on skewed data sets to compute UCLs of the mean. However, it should be noted that 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 based upon these two methods are much higher than the rest of the UCL computation methods, then this 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. In case these two bootstrap methods yield erratic and inflated UCLs, the UCL of the mean may be computed using the Chebyshev inequality.

ProUCL 5.0 has graphical (e.g., quantile-quantile [Q-Q] plots) and formal goodness-of-fit (GOF) tests for normal, lognormal, and a 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 distribution 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. Singh, Singh, and Engelhardt (1997, 1999); Singh, Singh, and Iaci (2002); and Singh and Singh (2003) evaluated the performances of several parametric, nonparametric and bootstrap UCL computation methods; some of those methods that have been included in the ProUCL software.

ProUCL computes various summary statistics for raw, as well as, log-transformed data sets with and without nondetect observations. In this Technical Guide and in ProUCL software, log-transformation (*log*) stands for the natural logarithm (*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 various population parameters of normal, lognormal, and gamma distributions. Nonparametric UCL computation methods in ProUCL include: Jackknife, central limit theorem (CLT),

adjusted-CLT, Chebyshev inequality, and bootstrap methods. It is well known that the Jackknife method (with sample mean as an estimator) and Student's t-method yield identical UCL values. Moreover, it is noted that UCLs based upon the standard bootstrap and the percentile bootstrap methods do not perform well (e.g., do not provide the specified coverage for the mean) for skewed data sets.

Note on Computing Lower Confidence Limits (LCLs) of Mean: In several environmental applications, one needs to compute a LCL of the unknown population mean. At present, ProUCL does not directly compute LCLs of mean. However, it should be pointed out that for data sets with and without nondetects, except for the bootstrap methods, gamma distribution, and H-statistic based LCL 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 mean incorporated in ProUCL. 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 2009c) 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, $y_i = \log(x_i)$; $i = 1, 2, ... , n$. Specifically, let

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

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

Similarly, let $\bar{x}$ and s_x be the sample mean and *sd* of the raw data, x_1, x_2, .. , 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 5.0. These include the normal, lognormal, and the gamma distributions. Shapiro-Wilk (for n up to 2000) and Lilliefors 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. The empirical distribution function (EDF) based methods: the K-S test and the A-D test are used to test for a gamma distribution. Extensive critical values for these two test statistics have been obtained via Monte Carlo simulation experiments (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 quantile-quantile (Q-Q) plots are also available to visually inspect data distributions. 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.

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, namely, 0.01, 0.05, and 0.1 (or confidence levels: 0.99, 0.95, and 0.90). For normal distributions, ProUCL outputs approximate p-values for the Shapiro-Wilk (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). 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 (e.g., 0.95 or greater) of the correlation coefficient of the linear pattern may suggest approximate normality (or lognormality) of the data set under study. However, it should be noted that on this graphical display, observations well-separated from the linear pattern displayed by the majority of data represent the outlying observations not belonging to the main dominant population (whose distribution one is assessing based upon a data set); and apparent jumps and breaks in the Q-Q plot suggest the presence of multiple populations. The correlation of the Q-Q plot based upon 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 possible to comprehend based upon statistics such as Shapiro-Wilk (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. The goodness-of-fit (GOF) test of a data set should always be judged based upon the formal (e.g., S-W statistic) as well as informal graphical displays. The normal Q-Q plot may be used as an aid to identify outliers or to identify multiple populations. 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 as graphical displays provide added insight about the data set and graphical displays do not get distorted by outliers and/or mixture populations. Therefore the final conclusion about the data distribution should be based upon the formal goodness-of-fit tests as wells as Q-Q plots. This statement is true for all GOF tests (e.g., 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 , [1982, 1982a]). For samples of 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 approximate p-values are displayed. 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 (Dudewicz and Misra, 1988; Conover, 1999). 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. The user should use this information to compute an appropriate UCL of the mean. The program prints the relevant statistics on the Q-Q plot of data.

- For convenience, normality, lognormality, or gamma distribution test results *for 0.05 level of significance (built-in)* are displayed on the UCL (and background statistics) output sheets. This helps the user in selecting the most appropriate UCL to estimate the EPC term. 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. The user should make a decision based upon the information provided by the associated Q-Q plot and the values of the GOF test statistics.

New in ProUCL 5.0: To streamline the decision process used to compute upper limits (e.g., UCL95), ProUCL 5.0 makes 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 leads 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 suggestions to use parametric or nonparametric decision statistics (e.g., UCL95) have been made based upon this conclusion.

2.2.2   Gamma Distribution

A continuous random variable, X (e.g., concentration of an analyte), is said to follow a gamma distribution, $G(k, \theta)$ with parameters $k > 0$ (shape parameter) and $\theta > 0$ (scale parameter), if its probability density function is given by the following equation:

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

(2-3)

Many positively skewed data sets follow a lognormal as well as a gamma distribution. It is observed that 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). These critical values for the two test statistics are given in Appendix A. The 1%, 5%, and 10% critical values of these two test statistics have been incorporated in ProUCL 5.0. The GOF tests for a gamma distribution depend upon the MLEs of the gamma parameters, k and θ, which should be computed first before performing the goodness-of-fit 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 these methods for a gamma distribution is provided as follows.

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

Let $x_1, x_2, ... , x_n$ be a random sample from the gamma distribution, $G(k, \theta)$; and let $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ 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 gamma distribution is obtained by plotting the scatter plot of pairs, $(x_{0i}, x_{(i)})$ $i := 1, 2, ..., n$. The gamma quantiles, x_{0i}, are given by the equation, $x_{0i} = z_{0i}\hat{\theta}/2$; $i := 1, 2, ...,$ n, where the quantiles z_{0i} (already ordered) are obtained by using the inverse chi-square distribution and are given as follows:

$$\int_{0}^{z_{0i}} f(\chi_{2\hat{k}}^2)d\chi_{2\hat{k}}^2 = (i-1/2)/n; \quad i := 1, 2, ..., n \tag{2-4}$$

In (2-4), $\chi_{2\hat{k}}^2$ represents a chi-square random variable with $2\hat{k}$ degrees of freedom (*df*). The program, PPCHI2 (Algorithm AS91) described in Best and Roberts (Applied Statistics [1975, Vol. 24, No. 3]) 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 such a Q-Q plot (e.g., with outliers and jumps) can still be high which does not signify that the data follow a gamma distribution. Therefore, 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 $z_i = F(x_i)$; $i := 1, 2, ..., n$. 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 $z_{(1)} \leq z_{(2)} \leq ... \leq z_{(n)}$. Let $\bar{z} = \left(\sum_{i=1}^{n} z_i\right)/n$ be the mean of the n, z_i; $i := 1, 2, ..., n$.

Compute the following two statistics:

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

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

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

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 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 ProUCL 4.1.001 and its previous versions 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: 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 (e.g., D'Agostino and Stephens [1986], Schneider and Clickner [1976] and 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.

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: Anderson-Darling and Kolmogorov Smirnov tests incorporated in ProUCL 4.1.01 and earlier versions have been updated in ProUCL 5.0. Critical values incorporated in ProUCL 4.1 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). 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 (e.g., ≤ 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 (e.g., $k \leq 0.1$), 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). It should be pointed out that from 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. However, for comparison purposes, for *k=0.2* both older and newly generated critical values of the two GOF have been included in tables presented in Appendix A. More details on this subject are provided in Appendix A.

2.3 Estimation of Parameters of the Three Distributions Incorporated in ProUCL

Let μ_1 and $\sigma_1{}^2$ 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, $\sigma_1{}^2$) with mean, μ_1, and variance, $\sigma_1{}^2$. The probability density function of a normal distribution is given by the following equation:

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

For normally distributed data sets, it is well known (Hogg and Craig, 1995) that the MVUEs of the mean, μ_1, and the variance, $\sigma_1{}^2$, are respectively given by the sample mean, $\bar{x}$, and sample variance, $s_x{}^2$. It is also well known that for normally distributed data sets, a UCL of the unknown mean, μ_1, based upon the Student's t-distribution is optimal. It was observed via Monte Carlo simulation experiments (Singh and Singh, 2003, Draft EPA Internal Report) that for normally distributed data sets, the modified-t-UCL and UCL based upon the bootstrap-t method provide the exact 95% coverage to the population mean. For normally distributed data sets, the UCLs based upon these three methods are in close agreement.

2.3.1 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 $\sigma_1{}^2$ represent the mean and variance of X. Some parameters of interest of a two-parameter lognormal distribution, LN(μ, σ^2), are given as follows:

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

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

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

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

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

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

Note that 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 larger is the *sd*, the larger are the *CV* and the skewness. For example, for a lognormal distribution: with $\sigma = 0.5,$ the skewness = 1.75; with $\sigma =1.0$, the skewness = 6.185; with $\sigma =1.5,$ the skewness = 33.468; and with $\sigma = 2.0$, the skewness = 414.36. The skewness of a lognormal distribution becomes unreasonably large as σ starts approaching and exceeding 1.5. Note that for a gamma distribution, the skewness is a function of the shape parameter, *k*. As *k* decreases, the skewness increases. It is observed (Singh, Singh, Engelhardt, 1997; 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 recommendations.

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
$0.5 \leq \sigma < 1.0$	Mild skewness to moderate skewness
$1.0 \leq \sigma < 1.5$	Moderate skewness to high skewness
$1.5 \leq \sigma < 2.0$	High skewness
$2.0 \leq \sigma < 3.0$	Very high skewness (moderate probability of outliers and/or multiple populations)
$\sigma \geq 3.0$	Extremely high skewness (high probability of outliers and/or multiple populations)

2.3.2.3   *MLEs of the Quantiles of a Lognormal Distribution*

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

The pth 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 pth quantile of the standard normal random variable, Z, with $P(Z \leq z_p) = p$, then the pth quantile of a lognormal distribution is given by $x_p = \exp(\mu + z_p\sigma)$. Thus the *MLE* of the pth 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 $LN(\mu, \sigma^2)$ 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).

<u>Notes</u>: 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 80th percentile if and only if *σ > 1.69*, and, similarly, the mean, μ_1, will exceed the 95th percentile if and only if *σ > 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:

$$\hat{\mu}_1 = \exp(\bar{y})g_n(s_y^2/2) \tag{2-14}$$
$$\hat{\sigma}_1^2 = \exp(2\bar{y})[g_n(2s_y^2) - g_n((n-2)s_y^2/(n-1))] \tag{2-15}$$

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

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

The square root of the variance given by equation (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, $G(k, \theta)$, 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.

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

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

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 set the derivatives to zero. This results in the following two equations:

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

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

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

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

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:

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

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),

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

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:

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

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

For $k < 8$, one can use the following recurrence 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 lth 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 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 *MLE*s 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 bias in MLE mean estimate of the mean of a gamma distribution for various values of the shape parameter, k and sample size, *n*. It is noted that 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 estimate, $\hat{k}^*$ 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 estimate, $\hat{k}^*$ are lower than those computed using the MLE estimate, $\hat{k}$. 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 mean estimate obtained using the MLE estimate, $\hat{k}$ 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.

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,
- Jackknife UCL (yields the same result as Student's t-statistic UCL),
- Standard bootstrap UCL,
- Percentile bootstrap UCL,
- BCA bootstrap UCL,
- Bootstrap-t UCL, and
- Hall's bootstrap UCL.

50

For skewed data sets, Modified-t and adjusted CLT methods adjust for skewness. However, it is noted that (Singh, Singh, and Iaci, 2002) this adjustment is not adequate for moderately skewed to highly skewed data sets (levels of skewness described in Table 2-1).

Even though some UCL methods (e.g., *CLT*, *UCL* based upon Jackknife method, 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. Additionally, the inclusion of these methods also helps the user to make correct decisions. Based upon the 1) sample size, *n*, 2) data skewness, $\hat{\sigma}$, 3) and data distribution, ProUCL makes suggestions about using one or more *95% UCL* methods to estimate the EPC term. These suggestions are based upon the simulation results summarized in Singh, Singh, and Iaci (2002), Singh and Singh (2003) and professional experience of the developers of ProUCL software. When in doubt, the users may want to consult a statistician to select the most appropriate UCL95 to estimate an EPC term.

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 UCL95 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 Student's t-distribution with (*n* -1) *df*. Let $t_{\alpha,n-1}$ be the upper αth quantile of the Student's t-distribution with (*n* -1) *df*.

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

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

For a normally (when the skewness is about ~0) distributed 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 (e.g., |skewness|<0.5) data sets, where the skewness is defined in Table 2-1. Even for mildly to 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 (e.g., = 0.95) to the population mean. This is especially true when the sample size is smaller than 20-25 (Singh and Singh, (2003). The situation gets worse (coverage much smaller than 0.95) for higher values of the s*d*, $\hat{\sigma}$, or its *MLE*, s_y.

Notes: To streamline the decision process used to compute upper limits (e.g., UCL95), ProUCL 5.0 makes 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.

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% UCL of the mean and UPLs and UTLs based upon gamma distributed data sets (e.g., 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 (e.g., EPA, 2009; Helsel, 2012) 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: 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 5.0), and an approximate gamma UCL was recommended for data sets of sizes>40, whereas ProUCL 5.0 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 $2nk$ *df*. When the shape parameter, k, is known, a uniformly most powerful test of size of the null hypothesis, H$_0$: $\mu_1 \geq C_s$, against the alternative hypothesis, H$_A$: $\mu_1 < C_s$, is to reject H$_0$ if $\bar{x}/C_s < \chi^2_{2nk}(\alpha)/2nk$. The corresponding $(1 - \alpha)$ 100% uniformly most accurate UCL for the mean, μ_1, is then given by the probability statement.

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

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

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

It should be pointed out that the UCL given by equation (2-34) is an approximate UCL without guarantee that the confidence level of $(1 - \alpha)$ will be achieved by this UCL. Simulation studies conducted in Singh, Singh, and Iaci (2002) and in Singh and Singh (2003) 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 term.

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 $(1 - \alpha)*100$ UCL of the gamma mean, $\mu_1 = k\theta$, is given by the following equation:

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

Where β is given in 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	$\alpha = 0.05$ probability level, β	$\alpha = 0.1$ probability level, β	$\alpha = 0.01$ 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

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.

Notes: 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.

For all gamma distributed data sets for all values of k and n, 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 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, $\hat{k}$

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

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

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

Tables of H-statistic critical values can be found in Land (1975). Theoretically, when the population is lognormal, Land (1971) showed that the UCL given by equation (2-36) possesses optimal properties and is the uniformly most accurate unbiased confidence limit. However, in practice, the H-statistic based UCL can be quite disappointing and misleading, especially when the data set is 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 the *sd,* s_y, which in turn inflates the H-UCL. Furthermore, it is 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 (e.g., 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 (e.g., Example 2-2 below). It is noted that some recent documents (e.g., Helsel and Gilroy, 2012) incorrectly state that the two distributions are similar. Helsel (2012, 2012a) likes to 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. It is recommended to compute correct and defensible estimates and decision statistics taking the sample size and data skewness into consideration. 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 correct 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 (Singh and Singh, 2003). 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 (e.g., σ >1.0) lognormally distributed data sets. For such data sets, a Chebyshev inequality based UCL can be used to estimate the population mean. H-statistic often results in unstable values of UCL95 as shown in Examples 2-1 through 2-3.

54

Example 2-1. Consider the silver data set of size 56 (from NADA for R package [Helsel, 2013]). The normal GOF test graph is shown in Figure 2-1, it is noted that data set has an extreme outlier (an observation significantly different from the main body of the data set). The data set consists of 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 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
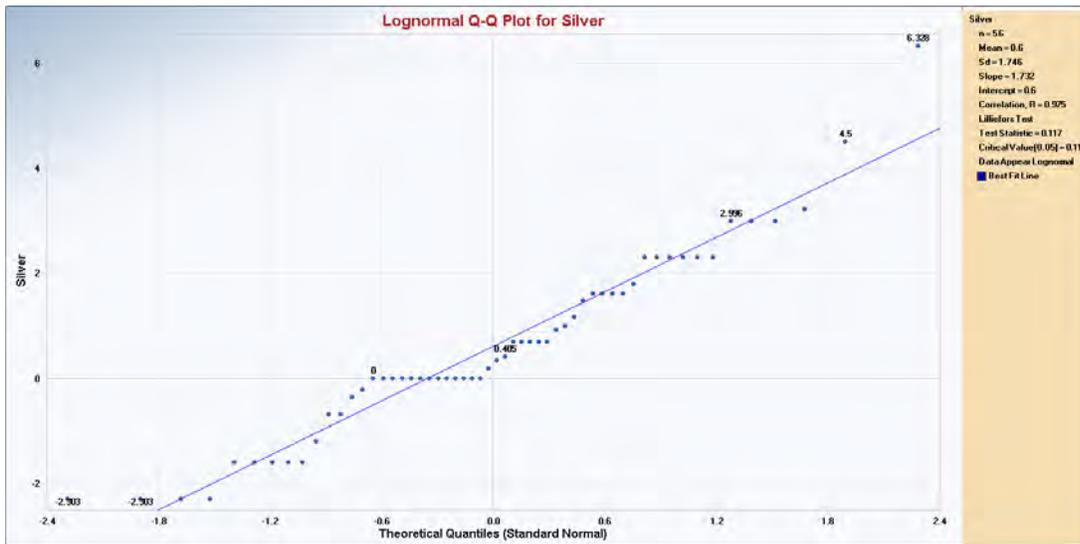
Figure 2-2. Lognormal Q-Q plot with GOF Test Statistics

55
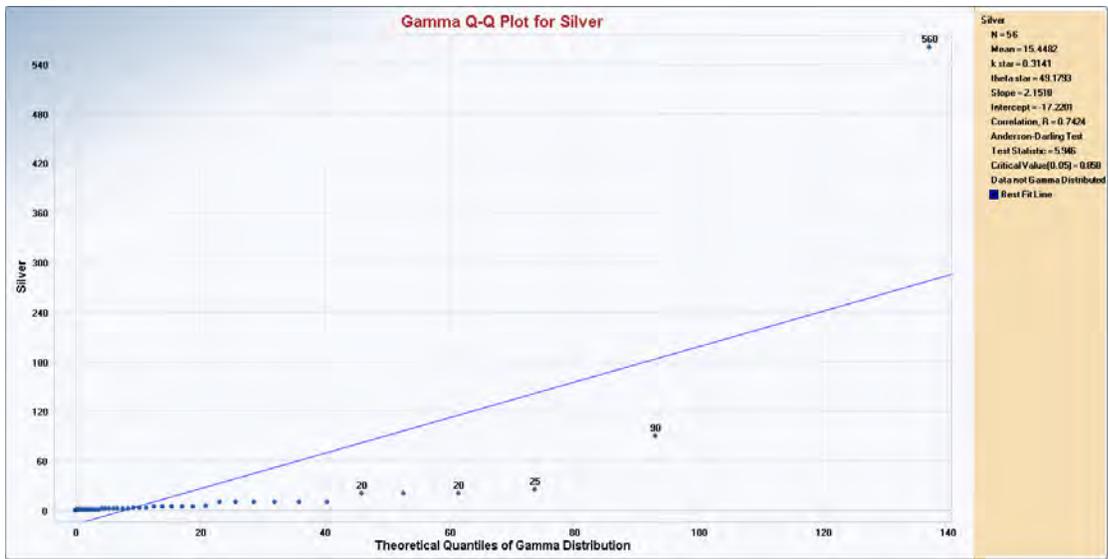
Figure 2-3. Gamma Q-Q plot with GOF Test Statistics
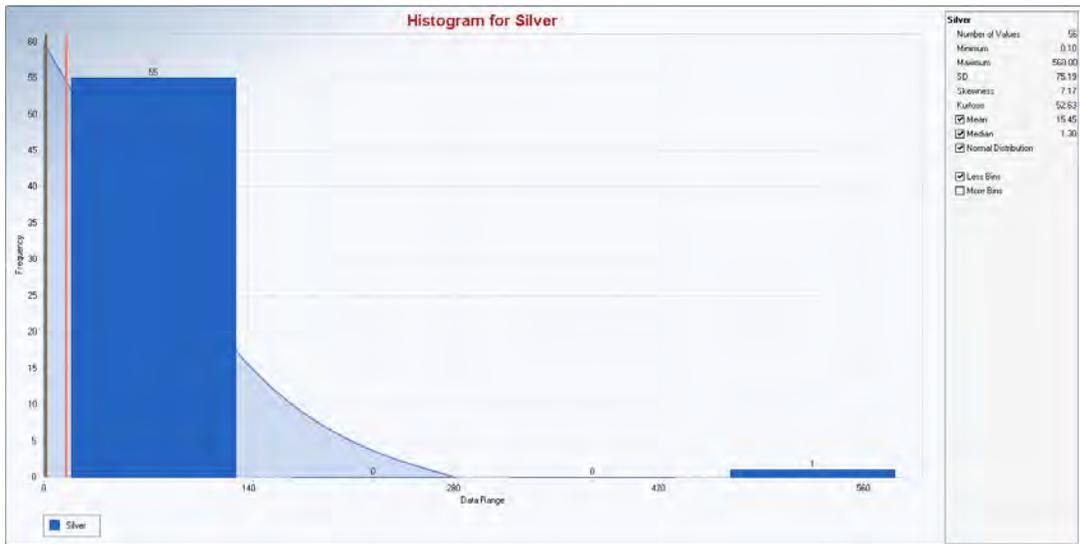
Figure 2-4. Histogram of Silver Data Set including outlier 560

56

Table2-3. Lognormal and Nonparamertic 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 sample mean is 15.45 and all lognormal distribution based UCL95s (e.g., H-UCL=18.54) are unrealistically low. In this case, the use of a lognormal distribution appears to underestimate the EPC term. The BCA bootstrap UCL95 is 52.45 and other nonparametric UCLs (e.g., percentile bootstrap UCL, Student's t-UCL) 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 Chebyshev UCL95 (Table 2-11) or BCA UCL95 to estimate the EPC term.

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 distorts the various decision making statistics.
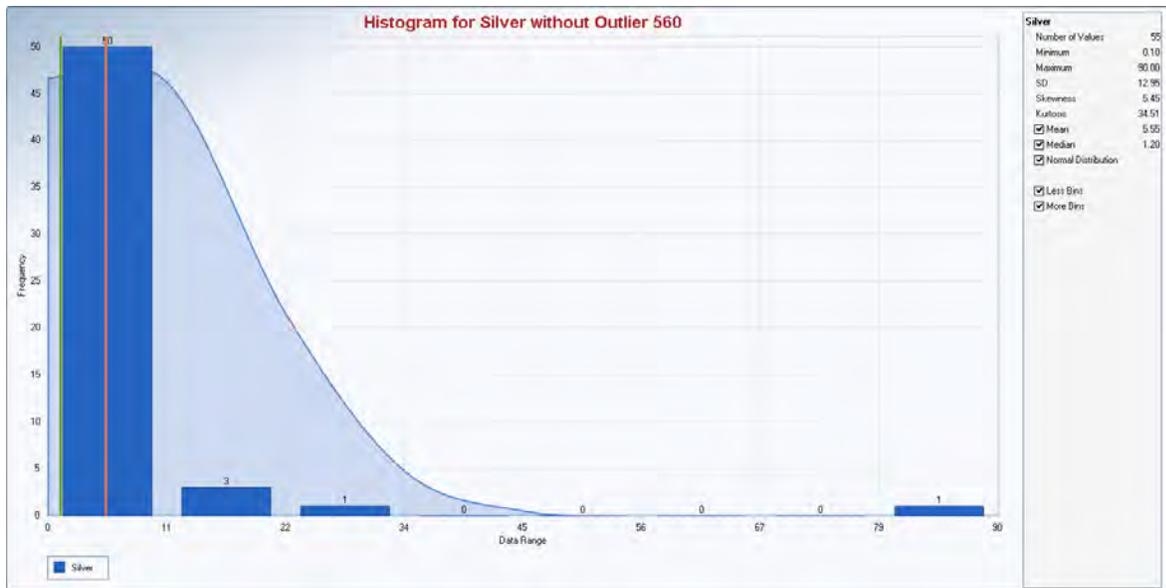
Figure 2-5. Histogram of Silver Data Set including outlier 560

Example 2-2: The positively skewed data set consisting of 25 observations, with values ranging from 0.35 to 170, follows a lognormal as well as 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 as well as a gamma distribution as shown in Figures 2-6a and 2-6b 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 in the following.

- The lognormal distribution based 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 correct decision statistics needed to make decisions which are cost-effective and protective of human health and the environment.

Table 2-4. Lognormal and Nonparamertic 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			
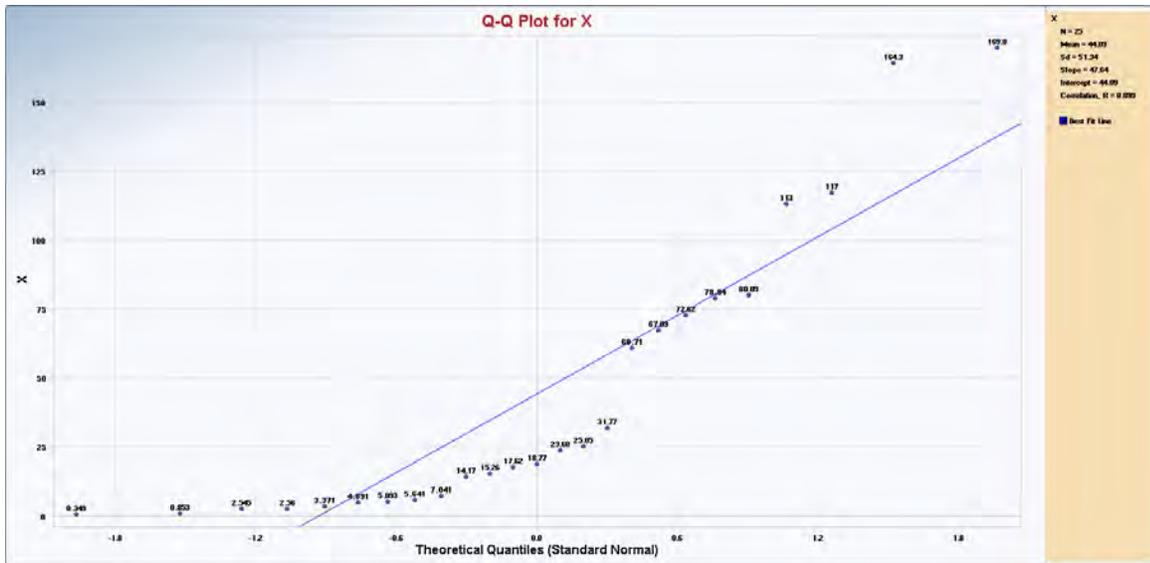
Figure 2-6. Normal Q-Q Plot of X
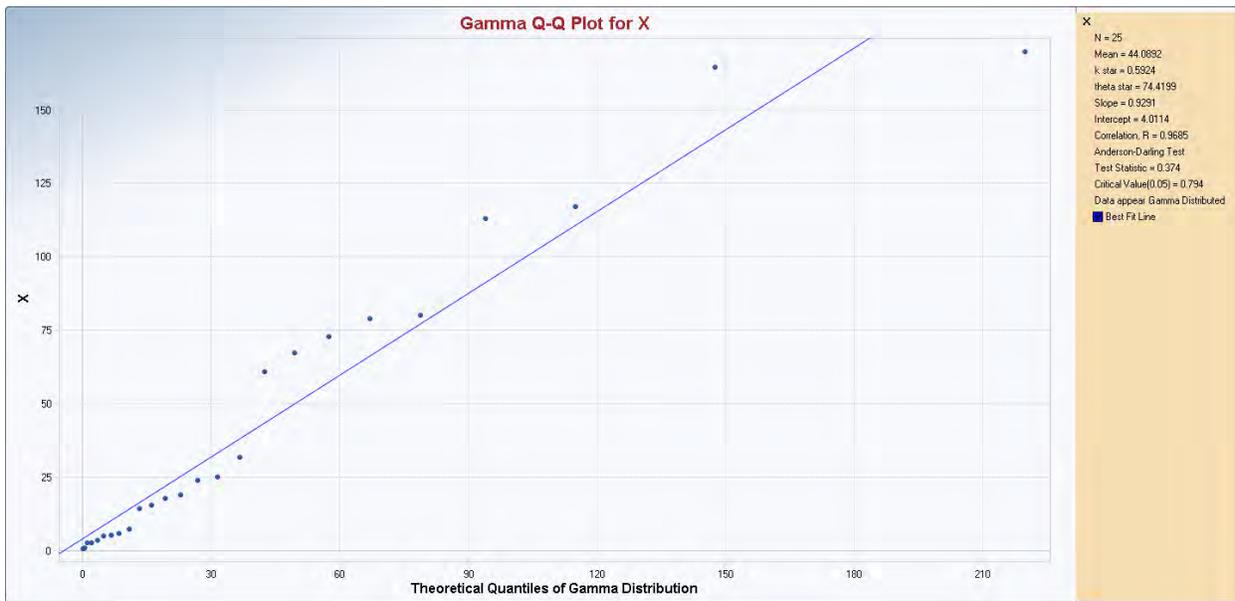
Figure 2-6a. Gamma Q-Q Plot of X

60
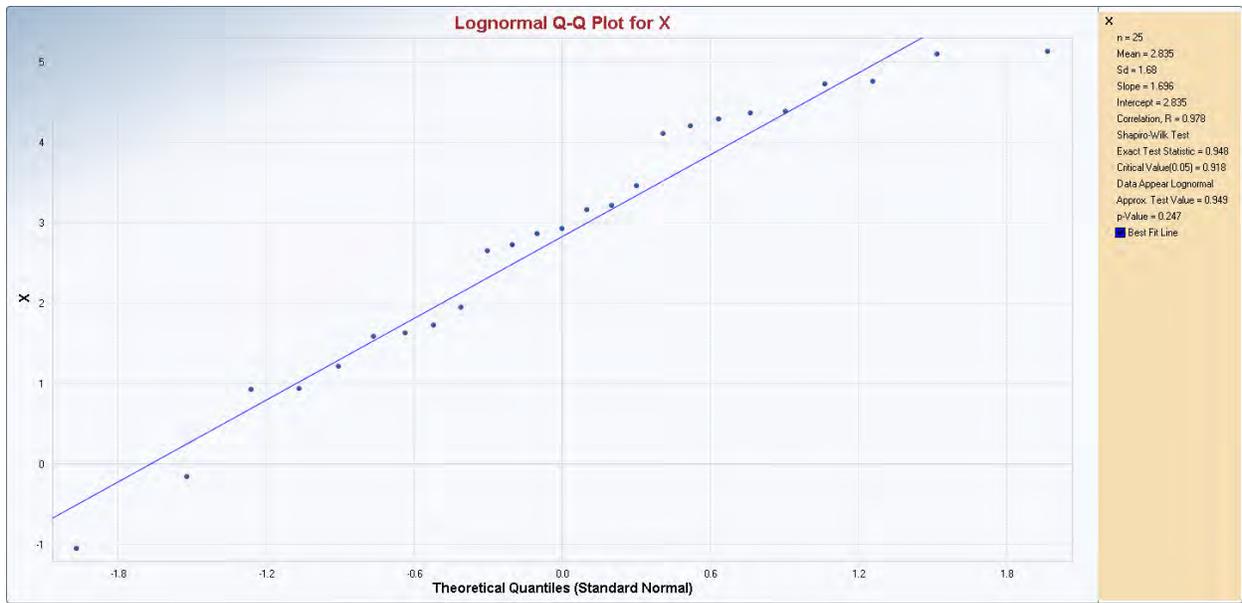
Figure 2-6b. Lognormal Q-Q Plot of X

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	

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			

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. The use of nonparametric methods is preferred to compute UCL95 for skewed data sets which do not follow a gamma distribution.

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 (e.g., 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:

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

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

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

This modification for a skewed distribution does not perform well even for mildly to moderately skewed data sets. Specifically, it is observed that the UCL given by equation (2-37) may not provide the desired coverage of the population mean, μ_1, when σ starts approaching and exceeding 0.75 (Singh, Singh, and Iaci, 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, 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 $(1 - \alpha)*100$ 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. More precisely, the sequence of random variables given by:

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

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). It should be noted that 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 $(1 - \alpha)*100$ UCL of the mean,

$$UCL = \bar{x} + z_\alpha s_x / \sqrt{n} \qquad\qquad (2\text{-}40)$$

An often cited and used rule of thumb for a sample size associated with a CLT based method is $n \geq 30$. 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 an adjustment for skewness.

63

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

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

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

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

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

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

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

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

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 5.0 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 2006, EPA 2009) is still commonly used to estimate EPC terms 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 and realistic estimates of the EPC terms. The use of the Chebyshev UCL has been recently adopted by the ITRC (2012) Technical Regulatory Guide to compute UCLs of the mean based upon data sets obtained using the Incremental Sampling Methodology (ISM) approach.

For moderately skewed data sets, the Chebyshev inequality yields conservative but realistic UCL95. 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; Singh and Singh, 2003). To address these issues, ProUCL computes and displays 97.5% or 99% Chebyshev UCLs. The user may want to consult a statistician to select the most appropriate UCL (e.g., 95% or 97.5% UCL) for highly skewed nonparametric data sets. 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 5.0 also outputs a UCL90 based upon the Chebyshev inequality.

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 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 $(1 - \alpha)*100$ UCL of the population mean, μ_1:

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

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

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

All versions of ProUCL computes the Chebyshev *(1 – α)*100 UCL* of the population mean using equation (2-46). This UCL is labeled as *Chebyshev (Mean, Sd)* on the output sheets generated by ProUCL. Since this Chebyshev method requires no distributional assumptions, this 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 > 1.5.

From the Monte-Carlo results discussed in Singh, Singh, and Iaci (2002) and Singh and Singh (2003), 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

65

Chebyshev 95% UCL tends to result in a higher (but stable) UCL of the mean of positively skewed gamma distributions.

Based upon the number of observations and data skewness, ProUCL makes recommendations to use a 95%, 97.5%, or 99% Chebyshev UCL. If these limits appear to be higher than expected, the project team collectively should make a decision about using an appropriate confidence coefficient (CC) to compute a Chebyshev inequality based upper limit. ProUCL can compute upper limits (e.g., UCLs, UTLs) for any user specified confidence coefficient level in the interval [0.5, 1). For convenience, ProUCL 5.0 also displays Chebyshev inequality based 90% UCL of the 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:

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

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 5.0 for historical reasons and research purposes. ProUCL 5.0 does not make any recommendations based upon this version of Chebyshev UCL.

Notes: As noted, 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 (e.g., 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, based upon the conceptual site model (CSM), expert site and regional knowledge, the project team should make the final decision regarding using or not using the suggestions made by ProUCL. If deemed necessary, the project team may want to consult a statistician.

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

Bootstrap and jackknife methods (e.g., Efron, 1981 and 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 (e.g., normal) and can be applied to a variety of situations. The bootstrap methods incorporated in ProUCL to compute the various 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).

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

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

For the jackknife method, n estimates of θ are computed by deleting one observation at a time (Dudewicz and Misra, 1988). For each index, i $(i=1,2,\ldots n)$, denote by $\hat{\theta}_{(i)}$, the estimate of θ (computed similarly as $\hat{\theta}$) when the ith observation is omitted from the original sample of size n, compute the arithmetic mean of these n jackknifed estimates given by:

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)} \tag{2-48}$$

A quantity known as the ith "pseudo-value" is given by:

$$J_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)} \tag{2-49}$$

Using equations (2-48) and (2-49), one computes the jackknife estimator of θ as follows:

$$J(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}J_i = n\hat{\theta} - (n-1)\tilde{\theta} \tag{2-50}$$

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

$$\hat{\sigma}_{J(\hat{\theta})} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(J_i - J(\hat{\theta})\right)^2} \tag{2-51}$$

Next, using the jackknife estimate, compute a t-type statistic given by

$$t = \frac{J(\hat{\theta}) - \theta}{\hat{\sigma}_{J(\hat{\theta})}} \tag{2-52}$$

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

$$UCL = J(\hat{\theta}) + t_{\alpha,n-1}\hat{\sigma}_{J(\hat{\theta})} \qquad\qquad (2\text{-}53)$$

If the sample size, n, is large, then the upper α^{th} t-quantile in the above equation can be replaced with the corresponding upper α^{th} standard normal quantile, z_α. Observe, also, that when $\hat{\theta}$ is the sample mean, $\bar{x}$, then the jackknife estimate is the same as the sample mean, $J(\bar{x}) = \bar{x}$, and the estimate of the standard error given by equation (2-51) simplifies to $s_x/n^{1/2}$, and the UCL in equation (2-53) reduces to the familiar t- statistic based UCL given by equation (2-32). ProUCL uses the jackknife estimate as the sample mean, that yields $J(\bar{x}) = \bar{x}$, which in turn translates equation (2-53) to Student's t- UCL given by equation (2-32). This method has been included in ProUCL to satisfy the curiosity of those users who do not recognize that the jackknife method (with sample mean as the estimator) yields a UCL of the population mean identical to the UCL based upon the Student's t- statistic as given by equation (2-32).

Notes: It is well known that the Jackknife method (with sample mean as an estimator) and Student's t-method yield identical UCLs. However, a typical user may be unaware of this fact, and some researchers may want to see these issues described and discussed at one place. Also, it has been suggested that a 95% UCL based upon the Jackknife method on the full data set obtained using the robust ROS (LROS) method may provide adequate coverage (e.g., Shumway, Kayhanian, and Azari, 2002) to the population mean of skewed distributions, which of course is not true since like Student's t-UCL, the Jackknife UCL does not account for data skewness. Finally, users are cautioned to note that for large data sets (e.g., sizes >10,000,...), the Jackknife method may take a long time (several hours) to compute a UCL of the mean.

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 thus obtained 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 $(x_{i1}, x_{i2}, ... , x_{in})$ represent the ith bootstrap sample of size n with replacement from the original data set, $(x_1, x_2, ..., x_n)$; 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 $\bar{x}_1, \bar{x}_2, ..., \bar{x}_N$. The bootstrap estimate of the population mean is the arithmetic mean, $\bar{x}_B$, of the N estimates $\bar{x}_i$: $i := 1, 2, ..., N$. The bootstrap estimate of the SE of the estimate, $\bar{x}$, is given by:

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

If some parameter, θ (say, the population median), other than the mean is of concern with an associated estimate (e.g., the sample median), then 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 $(1-\alpha)*100$ 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 $\bar{x}$. It is observed that the UCL obtained using the standard bootstrap method is quite similar to the UCL obtained using the Student's t-statistic 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 (e.g., > 500) data sets. The reason is that 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 can be considered as a good representative of the population itself. For data sets of larger sizes (e.g., > 500), methods based upon normal approximations should yield good estimates and results. Moreover, for larger data sets, bootstrap methods and specially the Jackknife method 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 $\bar{x}_{(1)} \le \bar{x}_{(2)} \le \cdots \le \bar{x}_{(N)}$. The (1 – α)*100 UCL of the population mean, μ_1, is given by the value that exceeds the (1 – α)*100 of the generated mean values. The 95% UCL of the mean is the 95th percentile of the generated means and is given by:

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

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

69

users of ProUCL and other software packages are cautioned about the suggested use of the percentile bootstrap method (e.g., Helsel, 2012) to compute 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 (e.g., Efron, 1981; Hall, 1988 and 1992; Efron and Tibshirani, 1993) have developed and proposed the use of skewness adjusted bootstrap methods. Simulations conducted by Singh and Singh (2003) verified 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 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). Singh and Singh (2003) investigated and compared its performance (in terms of coverage probabilities) with parametric methods and other bootstrap methods. They noted 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 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 $\alpha_2 \ast 100^{th}$ percentile computed using N bootstrap means $\bar{x}_i$; $i: = 1, 2, ..., N$. For example, when $N = 2000$, $\bar{x}^{(\alpha_2)} = (\alpha_2 N)^{th}$ 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:

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

$\Phi(z)$ is the standard normal cumulative distribution function and $z^{(1-\alpha)}$ is the $100(1-\alpha)^{th}$ percentile of a standard normal distribution. For example, $z^{(0.95)} = 1.645$, and $\Phi(1.645) = 0.95$. Also in (2-59), $\hat{z}_0$ (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 $\Phi^{-1}(x)$ is the inverse standard normal cumulative distribution function, e.g., $\Phi^{-1}(0.95) = 1.645$.

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

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 ith 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 to compute 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 ith 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, $t(1) \le t(2) \le … \le t(_N)$. The estimate of the lower α^{th} 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 0.05^{th} 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) - which is CORRECT.

Singh, Singh, and Iaci (2002) and Singh and Singh (2003) 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. They noted that for highly skewed ($k < 0.1$ or $\sigma > 2$) data sets of small sizes (e.g., $n < 10$ 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, $\bar{x}_i$ and $s_{x,i}$, and $\hat{k}_{3i}$, the sample mean, the sample standard deviation, and the sample skewness, respectively, are computed from the ith 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:

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

71

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

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

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

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

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 (e.g., $n = 100$). For lognormal data sets, the coverages provided by Hall's bootstrap and bootstrap-t methods do not increase much with the sample size, n. For highly skewed (e.g., $sd > 1.5, 2.0$) data sets of small sizes (e.g., $n < 15$), Hall's bootstrap method (and also the bootstrap-t method) performs better than the Chebyshev UCL, and for larger samples, the Chebyshev UCL performs better than Hall's and bootstrap-t methods.

Notes: It is noted that 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. In case 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 discussed by Example 2-3.

Example 2-3: Consider the pyrene data set of size 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 consists of several nondetects (NDs) and will be considered in Chapter 4 again. However, here, the data set is considered as an uncensored data set to discuss the various issues associated with skewed data sets consisting of 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 2-7 below confirm that the observation, 2982.45 is an extreme outlier. However, the lognormal distribution accommodated this outlier and the data set consisting of this outlier follows a lognormal distribution (Figure 2-8). Note that the data set with outlier does not follow a gamma distribution.
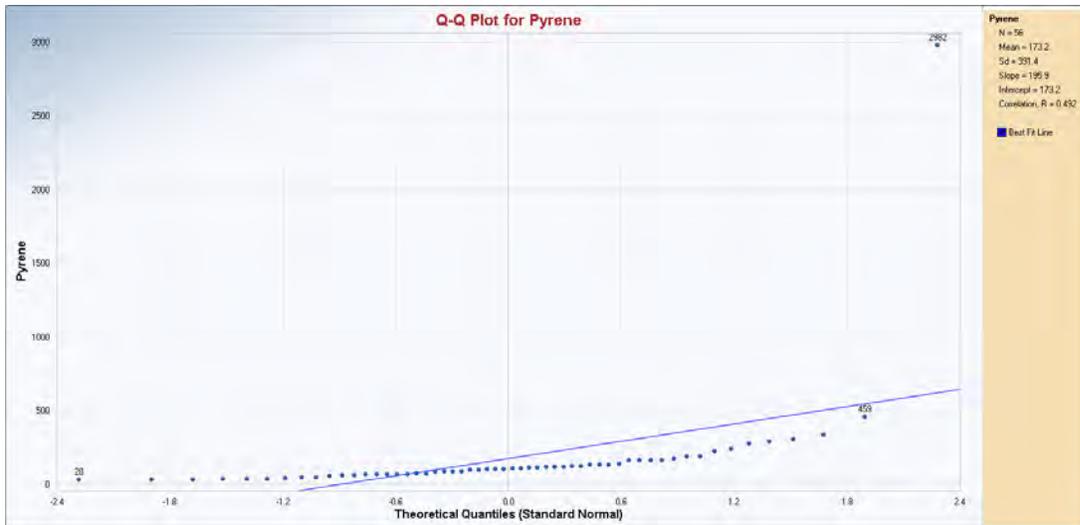
Figure 2-7. Normal Q-Q Plot of She's Pyrene Data Set

Figure 2-8. 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) appears to represent an underestimate of the population mean; a nonparametric UCL such as a 90% Chebyshev or a 95% Chebyshev UCL may be used to estimate the population mean. It should be noted that 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 consisted of 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 2-9. 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 distorts all statistics of interest. The decision statistics should be computed based upon a data set representing the main dominant population.
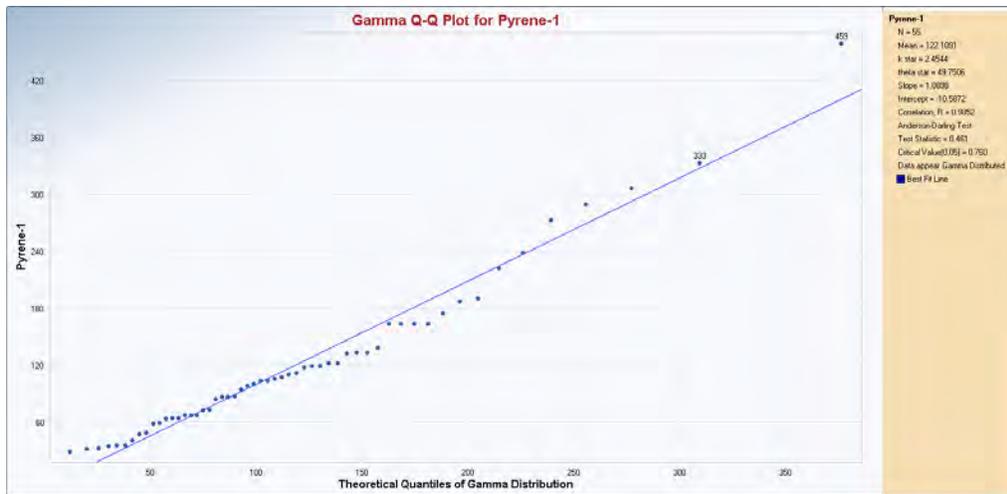
Figure 2-9. Gamma GOF Test on Pyrene Data Set without the Outlier.

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		

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	

2.5   Recommendations and Summary

The recommendations as incorporated in ProUCL about using appropriate UCLs of the mean are summarized in this section. These recommendations are made to help the users in selecting the most appropriate UCL to estimate the EPC term which is routinely used in exposure assessment and risk management studies of the USEPA. The recommendations are based upon the findings of Singh, Singh, and Engelhardt (1997, 1999); Singh, Singh, and Iaci (2002); Singh and Singh (2003); and Singh *et al.* (2006). Recommendations have been summarized for: 1) normally distributed data sets, 2) gamma distributed data sets, 3) lognormally distributed data sets, and 4) nonparametric data sets not following any of the three distributions available in ProUCL.   For a given data set, a correct UCL can be computed by using more than one method. Therefore, depending upon the data size, distribution, and skewness, sometimes ProUCL suggests using more than one UCL. In such situations, the user may use any of the suggested UCLs. If needed, the user may want to consult a statistician for additional insight. When the use of a Chebyshev inequality based UCL (e.g., UCL95) is suggested, the user may want to compare that UCL95 with other UCLs including the Chebyshev UCL90, before deciding upon the use of an appropriate UCL to estimate the population (site) average.

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

For mildly skewed data sets with σ (or $\hat{\sigma}$) < 0.5, most of the parametric and nonparametric methods (excluding Chebyshev inequality which is used on skewed data sets) yield comparable UCL values; any UCL computation method may be used to estimate the EPC term. However, for highly skewed (e.g., $\hat{\sigma}$ >2.0) parametric as well as nonparametric data sets, there is no simple solution to compute a reliable 95% UCL of the population mean, μ_1. Singh, Singh, and Iaci (2002), and Singh and Singh (2003) noted that the UCL95 based upon the skewness adjusted methods, such as the Johnson's modified-t and Chen's adjusted-CLT*,* do not provide the specified coverage (e.g., 95%) to the population mean even for mildly to moderately skewed (e.g., $\hat{\sigma}$ in the interval [0.5, 1.0]) data sets for samples of sizes as large as 100. The coverage of the population mean by these skewness-adjusted UCLs gets poorer (much smaller than the specified coverage of 0.95) for highly skewed data sets, where skewness levels have been defined in Table 2-1 as functions of σ or $\hat{\sigma}$ (standard deviation of log-transformed data). Interested users may want to consult graphs as described in Singh and Singh (2003) for a better understanding of the summary and recommendations described in this section.

2.5.1.1   Normally or Approximately Normally Distributed Data Sets

For normally distributed data sets, several methods including: 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.

- The 95% UCL of the mean given by equation (2-32) based upon the Student's t-statistic (preferably modified-t)   may also be used on non-normal data sets with *sd, s_y* of the log-transformed data less than 0.5, or when the data set follows an approximate normal distribution. A data set is approximately normal when: 1) the normal Q-Q plot displays a linear pattern (without outliers, breaks and jumps) and the resulting correlation coefficient is high (e.g., 0.95 or higher); and/or 2) one of the two GOF tests for a normal distribution incorporated in ProUCL suggests that data are normally distributed.

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

2.5.1.2   Gamma Distributed Data Sets

In practice, many skewed data sets can be modeled both by a lognormal distribution and a gamma distribution. Estimates of the unknown population mean based upon the two distributions can differ significantly (e.g., Example 2- 2 above).For data sets of small size (e.g., <20-50) the 95% H-UCL of the mean based upon a lognormal model often results in unjustifiably large and impractical *95% UCL* values. In such cases, a gamma model, (G *k, θ*), may be used to compute a 95% UCL provided the data set follows a gamma distribution.

- 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), one should 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) with small values of the shape parameter, *k* (e.g., k < 0.5 to 1.0), a gamma UCL95 fails 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 term. The small sample size requirement increases as skewness increases. That is as *k* decreases, the required sample size, *n*, increases.

- For highly skewed gamma distributed data sets of sizes ≥ 15 and small values of the shape parameter, k (e.g., k < 0.5, 1.0), the adjusted gamma UCL95 may be used to estimate the EPC.

- The use of a H-UCL should be avoided for highly skewed (e.g., $\hat{\sigma} > 2.0$) lognormally distributed data sets. For such highly skewed lognormally distributed data sets that cannot be modeled by a gamma or an approximate gamma distribution, the use of nonparametric *UCL* computation methods based upon the Chebyshev inequality (larger samples) or bootstrap-t and Hall's bootstrap methods (smaller samples) is preferred.

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 2-9. Summary Table for the Computation of a 95% *UCL* of the Unknown Mean, μ_1, of a Gamma Distribution

$\hat{k}$	Sample Size, n	Recommendation
$\hat{k} > 1.0$	n>=50	Approximate gamma 95% *UCL*
$\hat{k} > 1.0$	n<50	Adjusted gamma 95% *UCL*
$\hat{k} \leq 1.0$	$n < 15$	95% *UCL* based upon bootstrap-t or Hall's bootstrap method*
$\hat{k} \leq 1.0$	$n \geq 15$	Adjusted gamma 95% *UCL* if available, otherwise use approximate gamma 95% *UCL*

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

2.5.1.3 *Lognormally Distributed Skewed Data Sets*

For lognormally, LN (μ, σ^2), distributed data sets, the H-statistic-based UCL does provide the specified 0.95 coverage for the population mean for all values of σ; however, the H-statistic often results in unjustifiably large UCL values that do not occur in practice. This is especially true when skewness is high (e.g., $\sigma > 1.5\text{-}2.0$) and data set is of small size (e.g., <20-50). For skewed (e.g., with σ (or $\hat{\sigma}$) ≥ 0.5) lognormally distributed data sets, the Student's t-UCL95, modified-t-UCL95, adjusted-CLT UCL95, the standard bootstrap and the percentile bootstrap UCL95 methods fail to provide the specified 0.95 coverage for the population mean for samples of all sizes.

Based upon the results of the research conducted to evaluate the appropriateness of the applicability of a lognormal distribution based estimates of the EPC term (Singh, Singh, and Engelhardt, 1997; Singh, Singh, and Iaci, 2002), the developers of ProUCL suggest avoiding the use of the lognormal distribution to estimate the EPC term. Additionally, the use of the lognormal distribution based Chebyshev (*MVUE*) UCL should also be avoided unless skewness is mild with *sd* of log-transformed data <0.5-0.75.

- ProUCL5.0 computes and outputs H-statistic based UCLs and Chebyshev (*MVUE*) UCLs for historical, research, and comparison purposes as it is noted that some recent guidance documents (e.g., EPA 2009) are recommending the use of lognormal distribution based decision statistics. ProUCL can compute an *H-UCL* of the mean for samples of sizes up to 1000.

- It is suggested that all skewed data sets be first tested for a gamma distribution and for gamma distributed data sets, decisions statistics should be computed using gamma distribution based exact or approximate statistical methods as summarized in Section 2.5.1.2.

- For lognormally distributed data sets that cannot be modeled by a gamma distribution, methods as summarized in Table 2-10 may be used to compute a *UCL* of the mean to estimate the EPC term. The use of nonparametric methods is preferred to compute UCLs for skewed data sets which do not follow a gamma distribution.

- For highly skewed (e.g., with *sd* > 1.5) lognormally distributed data sets, it is re-iterated to avoid the use of lognormal distribution and *UCLs* based upon lognormal distribution. Instead, the use of nonparametric bootstrap methods (Efron and Tibshirani, 1993) and Chebyshev *(Mean, Sd)* UCL is recommended to estimate EPC terms.

2.5.1.4   *Nonparametric Distribution Free Skewed Data Sets without a Discernible Distribution*

For moderately and highly skewed data sets which are neither gamma nor lognormal, one can use a nonparametric Chebyshev UCL, bootstrap-t, or Hall's bootstrap UCL (for small samples) of the mean to estimate the EPC term. For skewed nonparametric data sets with negative and zero values, use a 95% Chebyshev (*Mean, Sd*) UCL for the population mean, μ_1. For all other nonparametric data sets with only positive values, the following procedure may be used to estimate the EPC term. The recommendations described here are based upon the simulation experiments and may not cover all skewed data sets originating from the various real world practical studies and applications.

- As noted earlier, for mildly skewed data sets with σ (or $\hat{\sigma}$) < 0.5, most of the parametric and nonparametric methods (excluding Chebyshev inequality which is meant for skewed data sets) yield comparable UCL95 values; therefore, any of those UCL computation method (as summarized in Table 2-11) may be used to estimate the EPC term. To be more precise, for mildly skewed data sets of smaller sizes (*n* <30) with $\hat{\sigma} \leq 0.5$, one can use the bootstrap BCA method, Student's t-statistic or modified-t- statistic to compute a 95% *UCL* of the mean, μ_1; and for mildly skewed data sets of larger sizes (e.g., *n* ≥30) with $\hat{\sigma} \leq 0.5$ one can use the BCA bootstrap method or the adjusted CLT to compute a 95% UCL of the mean, μ_1.

- For nonparametric moderately skewed data sets (e.g., σ or its estimate, $\hat{\sigma}$ in the interval (0.5, 1]), one may use a 95% Chebyshev (*Mean, Sd*) UCL of the population mean, μ_1. In practice, for values of $\hat{\sigma}$ closer to 0.5, a *95%* Chebyshev (*Mean, Sd*) UCL may represent an over estimate of the EPC term. The user is advised to compare 95% and 90% Chebyshev (*Mean, Sd*) UCLs.

Table 2-10. Summary Table for the Computation of a *UCL* of the Unknown Mean, μ_1, of a Lognormal Population to Estimate the EPC Term

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

*In the case that the Hall's bootstrap or bootstrap-t methods yield an erratic unrealistically large UCL95 value, a UCL of the mean may be computed based upon the Chebyshev inequality: Chebyshev *(Mean, Sd)* UCL

** For highly skewed data sets with $\hat{\sigma}$ exceeding 3.0, 3.5, it is suggested the user pre-processes the data. It is very likely that the data consist of outliers and/or come from multiple populations. The population partitioning methods may be used to identify mixture populations present in the data set.

- For nonparametric moderately and highly skewed data sets (e.g., $\hat{\sigma}$ in the interval [1.0, 2.0]), depending upon the sample size, one may use a 97.5% Chebyshev (*Mean, Sd*) UCL or a 95% Chebyshev (*Mean, Sd*) UCL to estimate the EPC term.

- For highly and extremely highly skewed data sets with $\hat{\sigma}$ in the interval [2.0, 3.0], depending upon the sample size, one may use Hall's UCL95 or the 99% Chebyshev (*Mean, Sd*) UCL or the 97.5% Chebyshev (*Mean, Sd*) UCL or the 95% Chebyshev (*Mean, Sd*) UCL to estimate the EPC.

For skewed data sets with $\hat{\sigma} > 3$, none of the methods considered provide the specified 95% coverage for the population mean, μ_1. The coverages provided by the various methods decrease as σ increases. For such data sets of sizes less than 30, a 95% UCL can be computed based upon Hall's bootstrap method or bootstrap-t method. Hall's bootstrap method provides the highest coverage (but < 0.95) when the sample size is small; and the coverage for the population mean provided by Hall's method (and the bootstrap-t method) does not increase much as the sample size, n, increases. However, as the sample size increases, the coverage provided by the Chebyshev (*Mean, Sd*) UCL increases. Therefore, for larger highly skewed data sets with $\hat{\sigma} > 3$, the EPC term may be estimated by the 99% Chebyshev (*Mean, Sd*) UCL. Large sample size requirement increases as $\hat{\sigma}$ increases. Recommendations are summarized in Table 2-11.

Table 2-11. Summary Table for the Computation of a 95% *UCL* of the Unknown Mean, μ_1, Based Upon a Skewed Data Set (with All Positive Values) without a Discernible Distribution, Where $\hat{\sigma}$ is the *sd* of Log-transformed Data

$\hat{\sigma}$	*Sample Size, n*	*Recommendation*
$\hat{\sigma} < 0.5$	For all n	Student's t, modified-t, or H-UCL Adjusted CLT UCL, BCA Bootstrap UCL
$0.5 \le \hat{\sigma} < 1.0$	For all n	95% Chebyshev *(Mean, Sd)* UCL
$1.0 \le \hat{\sigma} < 1.5$	For all n	95% Chebyshev *(Mean, Sd)* UCL
$1.5 \le \hat{\sigma} < 2.0$	$n < 20$	97.5% Chebyshev *(Mean, Sd)* UCL
	$20 \le n$	95% Chebyshev *(Mean, Sd)* UCL
$2.0 \le \hat{\sigma} < 2.5$	$n < 15$	Hall's bootstrap method
	$15 \le n < 20$	99% Chebyshev *(Mean, Sd)* UCL
	$20 \le n < 50$	97.5% Chebyshev *(Mean, Sd)* UCL
	$50 \le n$	95% Chebyshev *(Mean, Sd)* UCL
$2.5 \le \hat{\sigma} < 3.0$	$n < 15$	Hall's bootstrap method
	$15 \le n < 30$	99% Chebyshev (Mean, Sd)
	$30 \le n < 70$	97.5% Chebyshev *(Mean, Sd)* UCL
	$70 \le n$	95% Chebyshev *(Mean, Sd)* UCL
$3.0 \le \hat{\sigma} \le 3.5^{**}$	$n < 15$	Hall's bootstrap method*
	$15 \le n < 50$	99% Chebyshev(Mean, Sd) UC*L*
	$50 \le n < 100$	97.5% Chebyshev *(Mean, Sd)* UCL
	$100 \le n$	95% Chebyshev *(Mean, Sd)* UCL
$\hat{\sigma} > 3.5^{**}$	For all n	99% Chebyshev *(Mean, Sd)* UCL

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

** For highly skewed data sets with $\hat{\sigma}$ exceeding 3.0 to 3.5, it is suggested that the user pre-processes the data. Data sets with such high skewness are complex and it is very likely that the data consist of outliers and/or come from multiple populations. The population partitioning methods may be used to identify mixture populations present in the data set.

2.5.2 Summary of the Procedure to Compute a 95% UCL of the Unknown Population Mean, $μ_1$, 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.

- Formal GOF tests are performed first so that based the determined data distribution, an appropriate parametric or nonparametric UCL of mean can be computed to estimate the EPC term. ProUCL generates formal GOF Q-Q plots to graphically evaluate the distribution (normal, lognormal, or gamma) of the data set.

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

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

- For 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 2-10.

- For nonparametric data sets, which are not normally, lognormally, or gamma distributed, a nonparametric UCL is used to estimate EPC term. Methods used to estimate EPC terms based upon nonparametric data sets are summarized in Table 2-11. For example for mildly skewed nonparametric data sets of smaller sizes (e.g., $n<30$), one may use a modified-t UCL or BCA bootstrap UCL; and for larger samples one may use a CLT-UCL, adjusted-CLT UCL, or BCA bootstrap UCL. It should be noted that these nonparametric *UCLs* computation methods do not provide desired coverage of the mean for moderately skewed to highly skewed data sets.

- For moderately skewed to highly skewed nonparametric data sets, the use of a Chebyshev (*Mean, Sd*) UCL is suggested. For extremely skewed data sets (e.g., with $\hat{\sigma}$ exceeding 3.0), even a Chebyshev inequality-based *99% UCL* of the mean fails to provide the desired coverage (e.g., 0.95) of the population mean.

- For highly skewed data sets with $\hat{\sigma}$ exceeding 3.0, 3.5, it is suggested the user pre-processes the data. It is very likely that the data consist of outliers and/or come from multiple populations. Population partitioning methods (available in Scout; EPA, 2009) may be used to identify mixture populations present in the data set; and decision statistics, such as EPC terms, may be computed separately for each of the identified sub-population.

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., quantile-quantile [Q-Q] plot) and formal goodness-of-fit (GOF) tests. Depending upon the data distribution, one uses parametric or nonparametric methods to estimate BTVs. An onsite observation in exceedance of a BTV 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 potentially represent unimpacted locations and are considered coming from the background (or comparable to the background) population. This approach, comparing individual site or groundwater 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.

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 (e.g., more than 8 to10 observations) site data are available, one may use hypotheses testing approaches 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 (e.g., 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 nondetect (ND) observations. Statistical limits for data sets consisting of NDs are discussed in Chapter 5.

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 main 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. The BTVs should be estimated by statistics representing the dominant background population represented by the majority of the data set. *Upper limits computed by including a few low probability high outliers (e.g., coming from the far tails of data distribution) tend to represent locations with those elevated concentrations rather than representing the main dominant background population.* It is suggested to compute all relevant statistics using data sets with outliers and without outliers, and compare the results. This extra step often helps the project team to see the potential influence of outlier(s) on the various decision making statistics (e.g., upper confidence limits [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, exposure point concentration [EPC], BTV) under consideration. Since the treatment and handling of outliers in environmental applications is a subjective and controversial topic, it is also suggested that the outliers be treated 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 described in Chapter 7.

Extracting a Site-Specific Background Data Set From a Broader Mixture Data Set: In practice, not many background samples are collected due to resource constraints and difficulties in identifying suitable background areas with anthropogenic activities and natural geological characteristics comparable to onsite areas (e.g., at large Federal Facilities). Under these conditions, due to confounding of site related chemical releases with anthropogenic influences and natural geological variability, it becomes challenging to: 1) identify background/reference areas with comparable anthropogenic activities and geological conditions/formations; and 2) collect adequate amount of data needed to perform meaningful and defensible site versus background comparisons for each geological stratum to determine chemical releases only due to the site related operations and releases. Moreover, a large number of background samples (not impacted by site related chemical releases) may need to be collected representing the various soil types and anthropogenic activities present at the site; which may not be feasible due to several reasons including resource constraints and difficulties in identifying background areas with anthropogenic activities and natural geological characteristics comparable to onsite areas. The lack of sufficient amount of background data makes it difficult to perform defensible background versus site comparisons and computing reliable estimates of BTVs. A small background data set may not adequately represent the background population; and due to uncertainty and larger variability, the use of a small data set tends to yield non-representative estimates of BTVs.

Under these complex conditions present at a site, and using the known fact that that within all environmental site samples (data sets) exist both background level concentrations and concentrations indicative of site-related releases, sometimes it is desirable to extract a site-specific background data set from a mixture data set consisting of all available onsite and offsite concentrations. Several researchers (e.g., Sinclair [1976], Holgresson and Jorner [1978], Fleischhauer and Korte [1990]) have used normal Q-Q plots/probability plots methods to delineate multiple populations potentially present in a mixture data set collected from environmental, geological and mineral exploration studies.

Therefore, when not enough observations are available from reference areas with geological and anthropogenic influences comparable to onsite areas, the project team may want to use population partitioning methods (e.g., Singh, Singh, and Flatman [1994], Fleischhauer and Korte [1990]) on a broader mixture data set to extract a site-specific background data set with geological conditions and anthropogenic influences comparable to those of the various onsite areas. The extraction of a site-specific background data set from a mixture data set is useful when not enough background data are available to properly represent the background of larger sites (e.g., Federal Facilities covering hundreds of acres of

land) consisting of areas with varying geological formations and soil types where it becomes necessary to establish site-specific background.

The topics of population partitioning and the extraction of a site-specific background data set from a mixture data set are beyond the scope of ProUCL software and this technical guidance document. It requires developing a separate chapter describing the iterative population partitioning method including the identification and extraction of a defensible background data set from a mixture data set consisting of all available data collected from background areas (if available), and unimpacted and impacted onsite locations. Currently, work is in progress to develop a background issue paper describing population methods to extract a site-specific background data set from a mixture data set consisting of concentrations from the various onsite areas and offsite areas (if available).

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) – New in ProUCL 5.0

It is noted that 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 various limits with their interpretation is provided in the following sections. Since 0.95 is the 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 (e.g., 0.80, 0.90, 0.95, 0.99) and coverage probabilities (e.g., 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 underlines 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 (e.g., $n > 30$).

Note on Computing Lower Limits: In many environmental applications (e.g., groundwater monitoring), one needs to compute lower limits including: lower prediction limits (LPLs), lower tolerance limits (LTLs), or lower simultaneous limit (LSLs). At present, ProUCL does not directly compute a LPL, LTL, or a LSL. It should be noted that for data sets with and without nondetects, ProUCL outputs the several intermediate results and critical values (e.g., khat, nuhat, K, d2max) needed to compute the interval estimates and lower limits. For data sets with and without nondetects, 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 nondetect* observations, the user may want to use the Scout 2008 software package (EPA 2009c) to 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: 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 a single 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 to represent a single environmental population. Outliers when present in a data set result in inflated values of the various decision statistics including: UPL, UTL, and USL. The use of inflated statistics as BTV estimates tends to result in a higher number of false negatives.

Notes: The user specifies the allowable false positive error rate, α (=1-CC), and 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 single background population and the data set is free of outliers.

Let x_1, x_2, 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.

Upper Prediction Limit (UPL): Based upon an established background data set, a 95% UPL (UPL95) represents that statistic such that an independently collected new/future observation 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. It is noted that 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 10 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:

α_{actual} = 1-(1-α)^k =1 – 0.95^{30} ~1-0.21 = 0.79 (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% 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, it is suggested to use a USL95 to estimate the BTV provided the background data set represents a single population free of outliers.

Upper Simultaneous Limit (USL): Based upon an established background data set free of outliers and representing a single statistical population, 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. 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.

It is noted that by definition, USL95 does not discard any observation. 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 USL95, then that value definitely represents an outlier and may 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 various upper limits used to estimate BTVs.

- Higher statistical limits are associated with higher levels of CCs. For an 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 (e.g., <15-20) samples tend to dismiss contamination as representing background conditions (results in higher number of false negatives, i.e., 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,…) 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 (inflated) estimates of BTVs. Outlying observations which are significantly higher than the majority of the background data may not be used in establishing background data sets and in the computation of BTV estimates. A couple of classical outlier tests cited in environmental literature (EPA, 2006b, Navy, [2002a, 2002b]) are available in the ProUCL software. It is noted that the classical outlier procedures suffer from masking effects as they get distorted by the same outlying observations that they are supposed to find! It is therefore recommended to supplement outlier tests with graphical displays such as box plots, Q-Q plots. On a Q-Q plot, elevated observations which are well-separated from the majority of data represent potential outliers.

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 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, to include or not include them, in the computation of the various 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. Some examples elaborating on these issues have been discussed in this document.

Notes: It should be pointed out that methods incorporated in ProUCL can be used on any data set with or without nondetects and with or without outliers. It may not be misinterpreted that ProUCL 5.0 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. The intent of the developers of ProUCL software is to inform the users how the inclusion of a few outliers coming from the *low probability tails of the data distribution* can yield distorted values of UCL95, UPLs, UTLs, and various other statistics. The decision limits and test statistics should be computed based upon the majority of data

representing the main dominant population and not by accommodating a few low probability outliers resulting in distorted and inflated values of the decision statistics. The inflated decision statistics tend to represent the locations with those elevated observations rather than representing the main 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 $x_1, x_2, \ldots , x_n$ 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 ordered statistics) is denoted by $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. The ordered statistics are used as nonparametric estimates of upper percentiles, UPLs, UTLs and USLs. Also, let $y_i = \ln (x_i)$; $i = 1, 2, \ldots ,$ 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 pth (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 pth population percentile. It is expected that about $p*100\%$ of the population values will lie below the pth percentile. Specifically, $x_{0.95}$ represents an estimate of the 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. It is noted that 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 $p*n^{th}$ 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:

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

It should be noted that the earlier versions (e.g., ProUCL 4.00.02, 4.00.04) of ProUCL used $p*(n+1)^{th}$ 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, 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.

<u>Notes:</u> 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*100th sample percentile is given by the following statement:

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

Here z_p is the p*100th percentile of a standard normal, N(0, 1), distribution, which means that the area (under the standard normal curve) to the left of z_p is p. If the distributions of the site 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 pth 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*100th percentile is given by the following statement:

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

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:

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

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. If needed, the user may want to use graphical Q-Q plots and perform GOF tests to verify if data sets with smaller values of the *MLE* estimates of k follow normal distributions.

Let $\hat{k}$ and $\hat{\theta}$ represent the *MLE*s 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 * $\theta/2$, 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:

$$x_\alpha = y_\alpha * \theta/2 \tag{3-5}$$

3.4   Upper Tolerance Limits

A UTL $(1-\alpha)$-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,* $(1-\alpha)$. 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,* $(1-\alpha)$. 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 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.

$$UTL = \bar{x} + K * s \tag{3-6}$$

Here, $K = K$ (n, α, 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 $(1-\alpha)$, values of the tolerance factor, K, have been tabulated extensively in the various statistical books (e.g., Hahn and Meeker 1991). Those K values

92

are based upon the non-central t-distribution. Also, some large sample approximations (e.g., Natrella, 1963) are available to compute the *K* values for one-sided tolerance intervals (same for both UTLs and lower tolerance 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* (e.g., 0.90, 0.95, 0.975) and $(1-\alpha)$ values (e.g., 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 $(1-\alpha)$ 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:

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

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

Notes: It is noted that 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 a lognormal UTLs to estimate BTVs should be avoided when skewness is high (e.g., *sd* of logged data > 1 or 1.5) and sample size is small (e.g., < 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. Those approximations are based upon Wilson-Hilferty (WH; 1931) and Hawkins-Wixley (H-W; 1986) approximations. A description of the procedure to compute gamma UTLs is given as follows.

Let $x_1, x_2, ..., x_n$ 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.
- 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:

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

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

$$UTL = \left(\bar{y} + K * s_y\right)^4 \tag{3-9}$$

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

Note: that for mildly skewed to moderately skewed gamma distributed data sets, HW and WH approximations yield fairly comparable UTLs. However for highly skewed data sets (e.g., $k<0.5$) with higher variability, HW method tends to yield higher limits than the WH method. A couple of examples are discussed as follows.

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-$(1 - \alpha)$ providing coverage to $p*100\%$ observations with CC, $(1 - \alpha)$ represents an $(1 - \alpha)*100\%$ UCL for the pth 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 $\leq UTL_{p,(1-\alpha)}$ with CC, $(1 - \alpha)*100$.

Let $x_{(1)} \leq x_{(2)} \leq ...x_{(r)} \leq ... \leq x_{(n)}$ represent n ordered statistics (arranged in ascending order) of a given data set, $x_1, x_2, ..., x_n$. A nonparametric UTL is computed by the 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, $(1 - \alpha)$. It is noted that in comparison with parametric UTLs, nonparametric UTLs require larger data sets to achieve the desired CC; a nonparametric UTL p-$(1 - \alpha)$ computed by order statistics often fails to exactly achieve the specified CC, $(1 - \alpha)$. The formula to compute the order statistic, sample size, and CC achieved by nonparametric UTLs are described as follows. 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).

3.4.4.1   Determining the Order, r, of the Statistic, $x_{(r)}$, to Compute $UTLp,(1-\alpha)$

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 $(1 - \alpha)$. 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 rth (r: $1 \leq r \leq n$) order statistic by using the normal approximation (Conover, 1999) given by the equation (3-10).

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

Depending upon the sample size, p, and $(1 - \alpha)$ 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 rth

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

3.4.4.2 Determining the Achieved Confidence Coefficient, CC_{achieve,} Associated with x_(r)

For a given data set of size, n, once the rth 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 v_1 and v_2 df.

$$CC_{Achieve} = (1 - \alpha_*) = \text{ Probability } (F_{(v_1, v_2)} \le f); \ v_1 = 2(n - r + 1), \text{ and } v_2 = 2r$$

$$f = \frac{r(1 - p)}{(n - r + 1)p}$$

(3-11)

For a given data set of size n, ProUCL 5.0 first computes the order statistic that is used to compute a nonparametric UTL$p,(1-\alpha)$. Once the order statistic has been determined, ProUCL 5.0 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 that can be used 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.

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

(3-12)

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: $1 \le m \le n$; 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 rth order statistic, $x_{(r)}$, is used to compute a UTL. For an 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 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.

For example, when the largest order statistic (with m=1) is used to compute a nonparametric UTL95-95, the approximate minimum sample size needed $0.25*5.99*1.95/0.05 \approx 58.4$ which is rounded upward to 59; and when the second largest value (with m=2) is used to compute a UTL95-95, the approximate minimum sample size needed $[(0.25*9.488*1.95)/0.05]+0.5 \approx 93$. 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. 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.

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 5.0. 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 $p,(1 - \alpha)$. The N bootstrapped nonparametric p^{th} percentiles, $p,(\ i{:=}1,2,...,N)$, are arranged in ascending order: $p_1 \le p_2 \le ....\le p_N$. The $UTL_{p,(1-\alpha)}$ for the target population is given by the value that exceeds the $(1 - \alpha)*100$ 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 and jackknife (deleting one observation at a time) means by the corresponding bootstrap percentiles and jackknife (computed using (n - 1) observations by deleting one observation at a time) 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.

UPL_1 for a Single Future Observation: 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 $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 $(1 - \alpha)*100\%$ prediction limit is given by the following well known equation:

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

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 $(1 - \alpha)*100\%$ lognormal UPL is similarly given by the following equation:

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

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, $x_1, x_2, ..., x_n$ 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), $(1 - \alpha)*100\%$ upper prediction limits for a future observation from the same gamma distributed population are given by:

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

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

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

3.5.4 Nonparametric Upper Prediction Limit

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

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

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 (as they may represent outliers) 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. In order to address this issue, one may want to compute a UPL based upon the Chebyshev inequality. In addition to various parametric and nonparametric upper limits, ProUCL computes Chebyshev inequality based UPL.

3.5.4.1 Upper Prediction Limit Based Upon the Chebyshev Inequality

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

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

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 (e.g., *sd* of log-transformed data is low < 0.75), and the sample size is large (e.g., > 30). The user is advised to use professional judgment before using this method to compute a UPL. Specifically, for larger skewed data sets, instead of using a 95% UPL based upon the Chebyshev inequality, the user may want to compute a Chebyshev UPL with a lower CC (e.g., 85%, 90%) to estimate a BTV. ProUCL can compute a Chebyshev UPL (and all other UPLs) for any user specified CC in the interval [0.5, 1].

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 andnot 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$_k$95 for k ($\geq$1) future observations is designed to compare k future observations; we are 95% sure that "k" future values from the background population will be less than or equal to UPL$_k$95 with CC of 0.95. In addition to UPL$_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$_3$ $\leq$....$\leq$ UPL$_k$....

A normal distribution based UPL$_k$ (1 - α) for k future observations, $x_{n+1}, x_{n+2}, ..., x_{n+k}$ is given by the probability statement:

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

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

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

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, $x_{n+1}, x_{n+2}, ..., x_{n+k}$ is given by the following equation:

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

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,…) 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 $(1 – \alpha) * 100\%$ upper simultaneous limit (USL) based upon an established background data set is meant to provide coverage for <u>all</u> observations, x_i, $i = 1, 2, n$ 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. It is expected that observations coming from the background population will be less than or equal to the USL95 with a 95% CC. 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 recommended, 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 $(1 – \alpha)$ 100% simultaneous interval obtained using the first order Bonferroni inequality (Singh and Nocerino, 1995, 1997) is given as follows:

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

Here, $(d_\alpha^b)^2$ 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, $(d_\alpha^b)^2$ can be found in Singh (1993) and Singh and Nocerino (1997). These values have been programmed in ProUCL 5.0 to compute USLs for any combination of the sample size, n, and CC, $(1 - \alpha)$.

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

$$P\left(x_i \leq \bar{x} + s d_{2\alpha}^b; i := 1, 2, ..., n\right) = 1 - \alpha \ ;$$

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

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

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

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

A gamma distribution based (using WH approximation), one-sided $(1 - \alpha)$ 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 $(1 - \alpha)$ 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.

Notes: 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.

Some examples illustrating the computations of the various upper limits described in this chapter are discussed as follows.

Example 3-1. Consider a real Superfund site data set. The data set has several inorganic constituents of potential concern, including aluminum, arsenic, chromium (Cr), and lead. The computation of background statistics obtained using ProUCL are summarized as follows. The complete data set is given in Appendix 5 of the *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA, 2002a).

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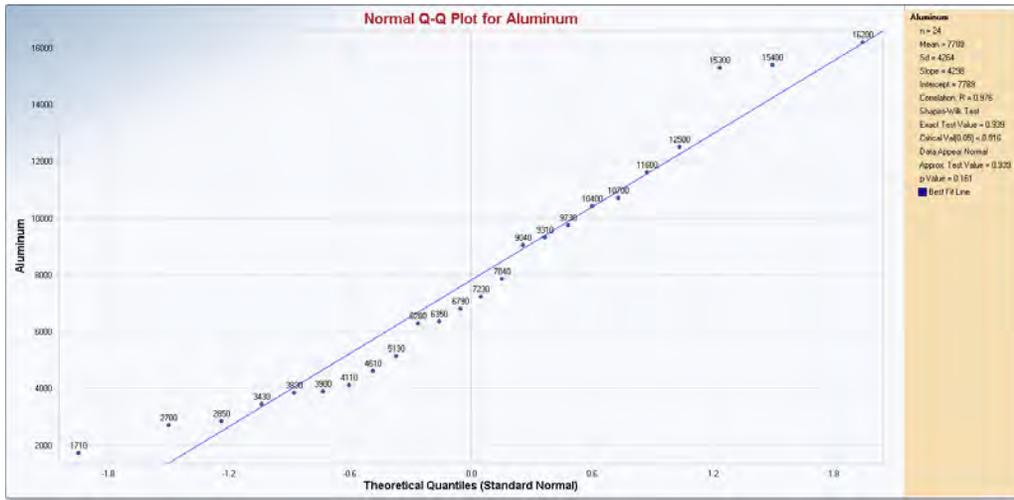
Figure 3-1. Normal Q-Q plot of aluminum with GOF Statistics

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 be coming from the same population as the rest of the 21 observations. The classical outlier tests (e.g., Dixon and Rosner tests) did not identify these 3 data points as outliers. Robust outlier tests (e.g., MCD [Rousseeuw and Leroy, 1987]] and PROP influence function [Singh and Nocerino, 1995] based tests) identified the 3 high values as statistical outliers. The project team should decide whether or not the 3 higher concentrations represent outliers. A brief discussion about robust outlier identification methods is given in Chapter 7. The inclusion of the 3 higher values in the data set resulted in higher upper limits. The various upper limits have been computed with (Table 3-1) and without the 3 high observations (Table 3-2).

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

102

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

The project team should make a determination of which statistics (with outliers or without outliers) should be used to estimate BTVs.

Example 3-2. Chromium concentrations of the superfund site data set used in Example 3-1 follow a lognormal distribution. The computation of background statistics using a lognormal model are shown in -3 3. Figure 3-2 is the lognormal Q-Q plot with GOF statistics.
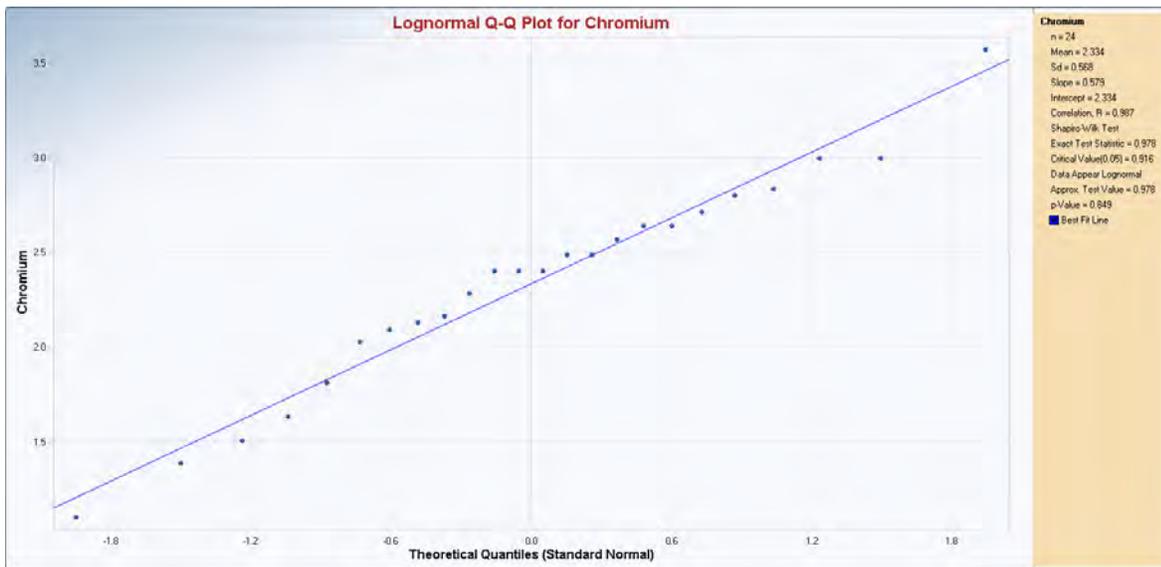
Figure 3-2. Lognormal Q-Q plot of Chromium with GOF Statistics

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 superfund site data set used in Example 3-1 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.
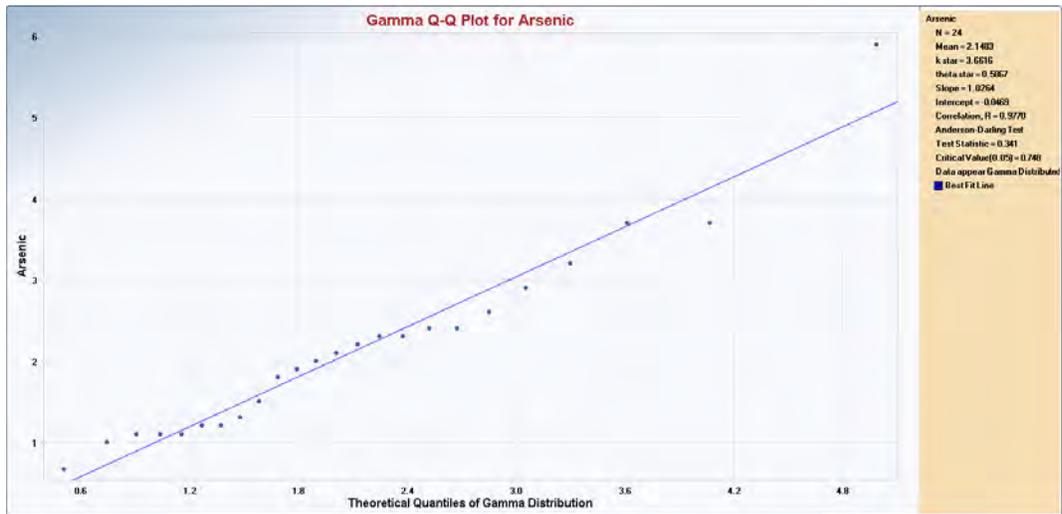
Figure 3-3. Gamma Q-Q plot of Arsenic with GOF Statistics

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 superfund site data set used in Example 3-1 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			

Notes: 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.*

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 UCL95 of the mean. As noted earlier, 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 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. A UPL95 (UTL95-95) based upon a gamma model are 245.3(or 285.6) and 163.5 (or 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 Consisting of Nondetect Observations

4.1   Introduction

Like outliers, nondetects (NDs) are also inevitable in most environmental data sets. First of all, it should be noted that the estimation of the mean and standard deviation (*sd*), and the computation of the various 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 is needed to compute the various decision making statistics such as UCLs, UTLs, and upper prediction limits (UPLs). ProUCL 4.0 (2007) and its higher versions have several methods which can be used:1) to estimate the mean, *sd*, and standard error of the mean; and 2) to compute the various upper limits (e.g., UCLs, UTLs, UPLs) based upon left-censored data sets consisting of multiple detection limits (DLs) or reporting limits (RLs).   Some new methods, based upon the KM estimates, which adjust for data skewness have been incorporated in ProUCL 5.0 to compute upper limits.

For left-censored data sets with NDs, Singh and Nocerino (2002) compared the performances of the various estimation methods to estimate the population mean, μ_1, and *sd,* σ_1 including the: maximum likelihood estimator (MLE; Cohen, [1950, 1959]) method, restricted MLE (RMLE) method (Person and Rootzen, 1977; Gleit, 1985), EPA Delta lognormal method (EPA, 1991; Hinton, 1993), Winsorization (Gilbert, 1987) method, and regression on order statistics (ROS) method (Helsel, 1990). Singh, Maichle, and Lee (EPA 2006) studied the performances of several parametric and nonparametric UCL computation methods for left-censored data sets consisting of ND observations. They also included the Kaplan-Meier (KM) (1958) estimation method in their evaluations. Initially, all evaluated methods including the poor performing methods (e.g., MLE and RMLE, and Winsorization methods) and better performing (in terms of bias in the mean estimate and coverage probabilities associated with UCLs) estimation (e.g., KM method) and UCL computation methods (e.g., BCA bootstrap, bootstrap-t) were incorporated in ProUCL 4.0 (2007). However, due to the poor performances and/or failure to correctly verify probability distributions for data sets with multiple detection limits, the parametric MLE and RMLE methods, the normal ROS and the Winsorization estimation methods to compute upper limits are no longer available in ProUCL version 5.0. The normal ROS method is available only under the Stats/Sample Sizes module of ProUCL 5.0 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 retained in ProUCL 5.0 include the two ROS (lognormal, and gamma) methods and the KM method. These methods can be used on a wide-range of skewed data sets with multiple detection limits and NDs exceeding detected observations. Also, the substitution methods such as replacing NDs by half of the respective detection limits and the H-UCL method (e.g., EPA 2009 recommends its use in Chapter 15) have been retained in ProUCL 5.0 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 DL/2 or H-UCL methods.

The estimation of the mean and *sd* based upon data sets with NDs have been studied extensively by researchers (Cohen, 1991; Schneider 1986, Gilliom and Helsel; 1986, Singh and Nocerino, 2002; and Helsel, 2005). However, until recently (e.g., ProUCL 4.0, 4.1 [EPA, 2007, 2010b]; Singh, Maichle, and Lee, 2006), not much guidance was available on how to accurately compute the various upper limits (e.g., UCLs, UPLs, UTLs) used in the decision making processes of the various environmental applications. For example, in exposure and risk assessment applications, one needs to compute a 95% upper confidence limit (UCL95) of the mean based upon data sets consisting of NDs with multiple DLs; in background evaluation projects, one needs to compute upper prediction limits (UPLs) and upper tolerance limits (UTLs) to estimate and establish background level constituent concentrations. For data sets with multiple detection limits, this chapter describes methods to compute UCL95 of the mean needed to estimate the exposure point concentration (EPC) term. Statistical methods to compute other upper limits (e.g., upper percentiles, UPLs, UTLs, USLs) based upon data sets with NDs are described in Chapter 5.

Several nonparametric estimation methods including the KM (1958)method; and bootstrap and Chebyshev inequality based methods to compute the various upper limits are available in ProUCL 4.0 and its higher versions. In practice, data sets originating from the various environmental studies tend to follow skewed data distributions and UCL95s based upon the percentile and standard bootstrap methods and Student-t type statistic used on the ROS data (Helsel, 2012) or on KM estimates, fail to provide the desired 95% coverage of the population mean. Like uncensored data sets, for left-censored skewed data sets, ProUCL has several UCL computation methods which adjust for data skewness.

Note 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 is useful to decide which UCL should be used to estimate the EPC term for data sets consisting of both detected and nondetected observations. For data sets with NDs, output spreadsheets generated by ProUCL 5.0 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 values coming from population(s) significantly different from the main dominant population whose parameters (mean, upper percentiles) we are trying to estimate based upon the available data set. The main objective of using a statistical procedure is to model the majority of data representing the main dominant population and not to accommodate a few low probability (coming from far and extreme tails) outlying observations potentially representing impacted locations (site related or otherwise). Statistics such as a UCL95 of the mean computed using data sets with a few low probability outliers tend to represent locations exhibiting those elevated low probability outlying observations rather than representing the main dominant population.

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 EPC terms and background threshold values (BTVs). An argument can be made that the outlying observations are inevitable and can be naturally occurring (not impacted by site activities) in some environmental media (and therefore in data sets). For an example, in groundwater applications, a few elevated values (coming from the far tails of the data distribution with low probabilities) may be considered to be naturally occurring and as such may not represent the impacted monitoring well data values. However, the inclusion of a few outliers (impacted or naturally occurring observations) tends to yield distorted and elevated values of the decision statistics of interest (e.g., UCLs, UPLs, and UTLs); and those statistics tend not to represent the main dominant population (MW concentrations). As mentioned earlier, instead of representing the main dominant population, the inflated decision statistics (e.g., UCLs, UTLs) computed using outliers tend to represent those low probability outliers. This is especially true when one is dealing with smaller data sets (e.g., $n<20$-30) and a lognormal distribution is used to model those data sets.

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

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 (e.g., UCL of the mean) in the original space 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 log-transformation on a data set tends to accommodate outliers and hide contaminated locations instead of revealing them. Specifically, an observation that is a potential outlier (representing a contaminated location) in the original raw scale may not appear to be an outlier in the log-scale. This does not imply that the location with elevated concentrations in the original scale does not represent an impacted location.

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 5.0 for data sets consisting of NDs with multiple DLs.

110

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 detection limits (DL/2) is retained in ProUCL 5.0; that is the DL/2 goodness-of-fit (GOF) tests, UCL, UPL, and UTL computation methods have been retained in ProUCL 5.0 for historical reasons, 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 to estimate the EPC term or BTVs unless the data set consists of only a small number of NDs (<5%) and 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 the statements (e.g., EPA, 2006b) made that the DL/2 UCL method can be used for lower ($\leq$ 20%) censoring levels. The coverage provided by the DL/2 (t) UCL method deteriorates fast as the censoring intensity (e.g., percentage of NDs) increases and/or data skewness increases.

4.2.4   Minimum Data Requirement

Whenever possible, it is suggested that sufficient samples be collected to satisfy the requirements for the data quality objectives (DQOs) for the site. Some rule-of-thumb minimum sample size requirements are described in this section. At the minimum, it is suggested to collect a data set consisting of about 10 observations to compute reasonably reliable and accurate estimates of EPC terms (UCLs) and BTVs (e.g., 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 (e.g., 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 for left-Censored Data Sets

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, for data sets with NDs, found in ProUCL include: 1) exclude all NDs; 2) replace NDs by their DL/2s; and 3) ROS methods. 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≤k≤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 5.0, 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 EPC terms 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 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.

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 Kaplan-Meier (KM) (1958) estimation method, also known as the product limit estimation (PLE) method, is 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 (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-Meier (1958) and the report repaired by Bechtel Jacobs Company for the DOE (2000). The properties of the KM method are well researched (e.g., Gillespie, Chen *et al.* 2010, Epidemiology, Vol. 21, Issue 4). Specifically, the KM estimator is a consistent estimator and for large data sets, KM estimator is asymptotically efficient and normally distributed (e.g., Gu, Zhang, 1993). It is noted that the KM method is now also included in a recent EPA RCRA guidance document (2009).

For left-censored data sets, Singh, Maichle, and Lee (EPA 2006) performed extensive simulation experiments to study and evaluate the performance of several UCL computation methods based upon the MLE, ROS, and KM estimation methods. They concluded that KM method based UCLs computed using some bootstrap methods (e.g., BCA bootstrap, bootstrap-t) and Chebyshev inequality perform better (in terms of censoring intensity and coverages) than the various other methods including the MLE and ROS methods. Subsequently, the KM method was incorporated in ProUCL 4.0 (2007) and its higher versions. Currently, the KM method is widely used in environmental applications to compute the upper limits needed to estimate environmental parameters of interest such as the population mean and upper thresholds of a background population.

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 $x'_1 < x'_2, ... <$ 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 $j = 1, ..., n'$, 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

$$\widetilde{F}(x) = 1, \qquad\qquad\qquad x \leq x'_{n'}$$

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

112

$$\widetilde{F}(x) = \widetilde{F}(x_1'), \qquad\qquad\qquad x_{(l)} \le x \le x_1'$$

$$\widetilde{F}(x) = 0 \text{ or undefined}, \qquad\qquad 0 \le x \le x_{(l)}$$

Note that in the last equality statement of $\widetilde{F}(x)$ above, $\widetilde{F}(x) = 0$ when $x_{(l)}$ is a detect, and is undefined when $x_{(l)}$ is a ND. An estimate of the population mean based upon the KM method is given as follows.

$$\hat{\mu}_{KM} = \sum_{i=1}^{n'} x_i'[\widetilde{F}(x_i') - \widetilde{F}(x_{i-1}')], \text{ with } x_0 = 0 \qquad (4-1)$$

Using the PLE (or KM) method, an estimate of the *SE* of the mean is given by the following equation.

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

Where k = number of ND observations, and

$$a_i = \sum_{j=1}^{i} (x_{j+1}' - x_j')\widetilde{F}(x_j'), \text{ } i: =1, 2, \ldots, n'-1.$$

The KM variance is computed as follows:

$$\hat{\sigma}_{KM}^2 = \hat{\mu}_{(x^2)-KM} - \left(\hat{\mu}_{(x)-KM}\right)^2 \qquad (4-3)$$
$$\hat{\mu}_{(x)-KM} = \text{KM mean of the data, } x$$
$$\hat{\mu}_{(x^2)-KM} = \text{KM mean of the square of the data, } x \text{ (second raw moment)}$$

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 EPC terms whereas the variance is used to compute BTV estimates. 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 not be a trivial exercise 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 estimate of *sd* (e.g., Helsel, 2012) 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, 2012). 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, 2012, 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.

It is well known that the KM method does yield a good (e.g., 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 does not guarantee that the t-UCL 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) 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 are available in ProUCL 5.0 (were not available in ProUCL 4.1) which adjust for data skewness.

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 are used to impute NDs based upon a hypothesized distribution such as a gamma or a lognormal distribution. The "Stats/Sample Sizes" menu option of ProUCL 5.0 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 principal component analysis. 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 5.0 also performs GOF tests on data sets obtained using the log ROS (LROS) and gamma ROS (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 EPC terms 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 to compute the various upper limits. The 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 version 5.0. Instead, when detected data follow a normal distribution, the use of KM estimates in normal equations is suggested to compute 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 LROS method are the same as for the normal distribution; and when the data sets consists of multiple DLs, the plotting positions for the lognormal and gamma distributions are also similar. For a full data set of size n, the most commonly used plotting

position for the ith observation (ordered) is given by $(i − ⅜) / (n + ¼)$ or $(i − ½)/n; i:=1,2,...,n$. These plotting positions are routinely used to generate Q-Q plots based upon full uncensored data sets (e.g., 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, $(i − ⅜) / (n + ¼)$ for normal and lognormal distributions, and uses empirical percentiles given by $(i − ½)/n$ 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: $P(Z \leq q_{(i)}) = (i − ⅜) / (n + ¼)$, $i := 1, 2, …, n$, where Z represents a standard normal variable (SNV). The gamma quantiles are computed using the probability statement: $P(X \leq q_{(i)}) = (i − ½)/n$, $i := 1, 2, …, n$, where X represents a gamma (~constant *chi-square) random variable.

In case multiple detection limits are present with NDs potentially exceeding the detected observations, the plotting positions (percentiles) are computed using methods that adjust for multiple detection limits. The details of the computation of such plotting positions (percentiles), p_i; $i: =1, 2, ... ,n$, for data sets with multiple detection limits or with ND observations exceeding the detection limits 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:

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

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

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, 2012) of size n=56, the plotting positions and LROS and GROS quantiles (denoted by Q) generated by ProUCL are summarized in Table 4-1 as follows. 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 by using the $(n - k)$ pairs, $(q_{(i)}, x_{(i)})$; $i:= k + 1, k + 2, …, n$, where $x_{(i)}$ are the *(n-k)* detected values arranged in ascending order. The OLS regression line fitted to the $(n - k)$ pairs $(q_{(i)}, x_{(i)})$; $i:= k + 1, k + 2, …, n$ corresponding to the detected values is given by:

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

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, $x_{(i)}$; $i := k + 1, k + 2,$ …, 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 ith ND, the imputed ND is given by $x_{(ndi)} = a + bq_{(ndi)}$; $i:=1,2,...k.$ When there is only a single detection limit (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 detection limits, and when some of those detection limits 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.

Notes: 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 needed to compute defensible upper limits, enough (e.g., > 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 log ROS (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, $q_{(\text{ndi})}$ 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.

$$\hat{\mu}_{Org} = \exp(\hat{\mu}_{Ln} + \hat{\sigma}_{Ln}^2 / 2), \text{ and } \hat{\sigma}_{Org}^2 = \hat{\mu}_{Org}^2 (\exp(\hat{\sigma}_{Ln}^2) - 1) \quad (4-5)$$

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 EPC terms and BTVs. 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.

Notes: The use of the robust ROS method has become quite popular. Helsel (2012) 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. 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 (Helsel, 2012) of size 24 with 13 NDs. The detected data set of size 11 follows a lognormal distribution as shown in Figure 4-1.The LROS data set consisting of 11 detects and 13 imputed NDs also follows a lognormal distribution as shown in Figure 4-2. Summary statistics and LROS UCLs are summarized in Table 4-2.
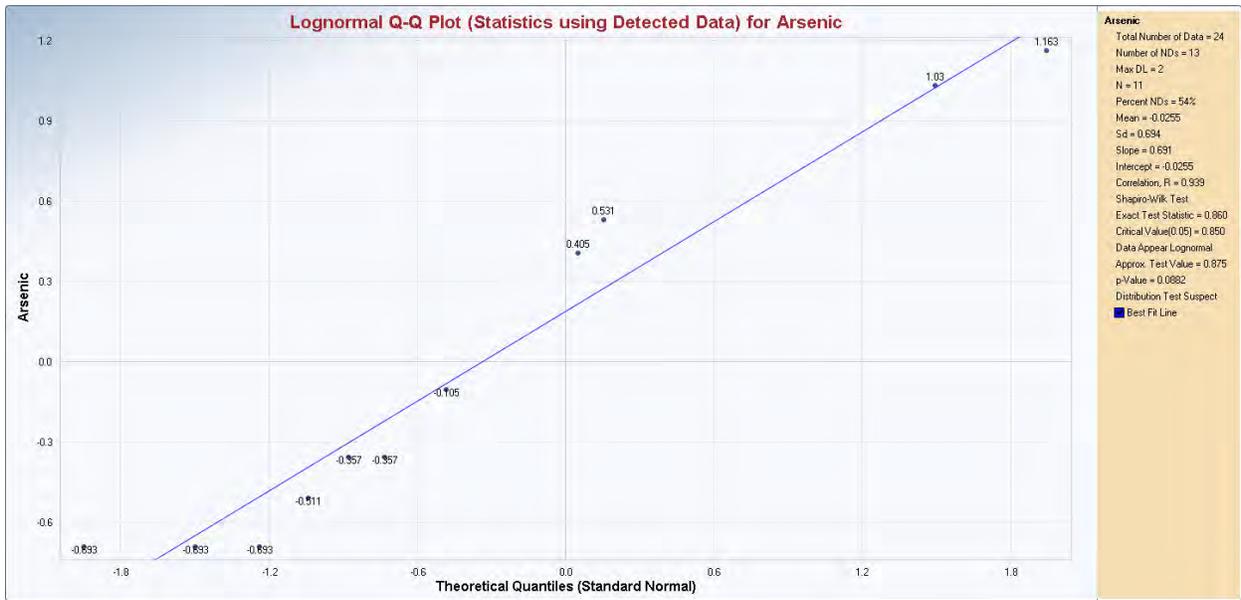
Figure 4-1. Lognormal GOF Test on Detected Concentrations of the Oahu Data Set.
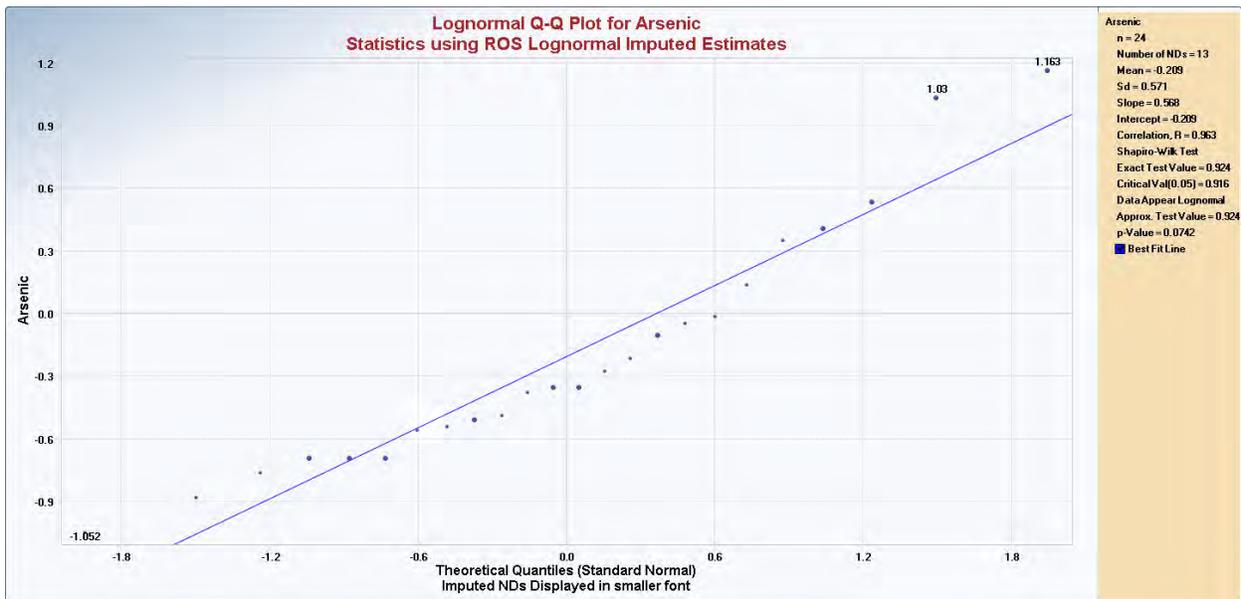
Figure 4-2. Lognormal GOF Test on LROS Data Obtained Using the Oahu Data Set.

118

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 term. However, the data set also follows 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 EPC terms and BTVs (e.g., 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 $x_{(k+1)} \leq x_{(k+2)} \leq ... \leq x_{(n)}$ 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 $(n-k)$ detects) of size n where the detected $(n-k)$ 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; $i := 1, 2, …, n$ 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 $(n-k)$ detected values. This process may have some effect on the accuracy of the estimated gamma quantiles (which use estimated value of the shape parameter, k), and consequently on the accuracy of the imputed NDs. The availability of enough (e.g., at least 8-10) detected gamma distributed observations is suggested to compute 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, $x_{0i} = z_{0i}\hat{\theta}/2$; $i := 1, 2, …, n$, where quantiles z_{0i} (already ordered) are obtained by using the inverse chi-square distribution given as follows:

$$\int_0^{z_{0i}} f(\chi^2_{2\hat{k}})d\chi^2_{2\hat{k}} = (i - 1/2)/n; \quad i := 1, 2, …, n \quad \text{(Single DL Case)} \tag{4-6}$$

$$\int_0^{z_{0i}} f(\chi^2_{2\hat{k}})d\chi^2_{2\hat{k}} = p_i; \quad i := 1, 2, …, n \quad \text{(Multiple DL Case)} \tag{4-7}$$

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, $i := 1, 2, …, n$, for left-censored data sets with multiple detection limits 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: The GROS method <u>may not</u> 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 value of k <0.5). When the estimated value (MLE) of the shape parameter, k based upon detected data is small (e.g., < 0.5), or when the data set consists of many tied NDs at multiple DLs with a high percentage of NDs (e.g., >50%), the GROS may not perform well in imputing NDs, as the GROS method tends to yield negative imputed NDs (e.g., due to outliers distorting 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).*

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 used in normal UCL equations. This hybrid gamma-KM method to compute upper limits is available in ProUCL 5.0. It is noted that the hybrid KM-gamma method yields reasonable UCLs by accounting for data skewness as shown in Example 4-5 to follow.

The use of the hybrid gamma-KM upper limits is preferred when GROS method yields negative imputed values; that is when estimated value of the shape parameter, k based upon detected data is small(e.g., < 0.5).

Example 4-4 (Oahu Data Set Continued): The detected data set of size 11 follows a gamma distribution as shown in Figure 4-3.The GROS data consisting of 11 detects and 13 imputed NDs also follows a gamma distribution as shown in Figure 4. Summary statistics and GROS UCLs are summarized in Table 4-3 following Figure 4. Since the data set is only mildly skewed all methods (GROS and Hybrid KM-Gamma) yield comparable results.
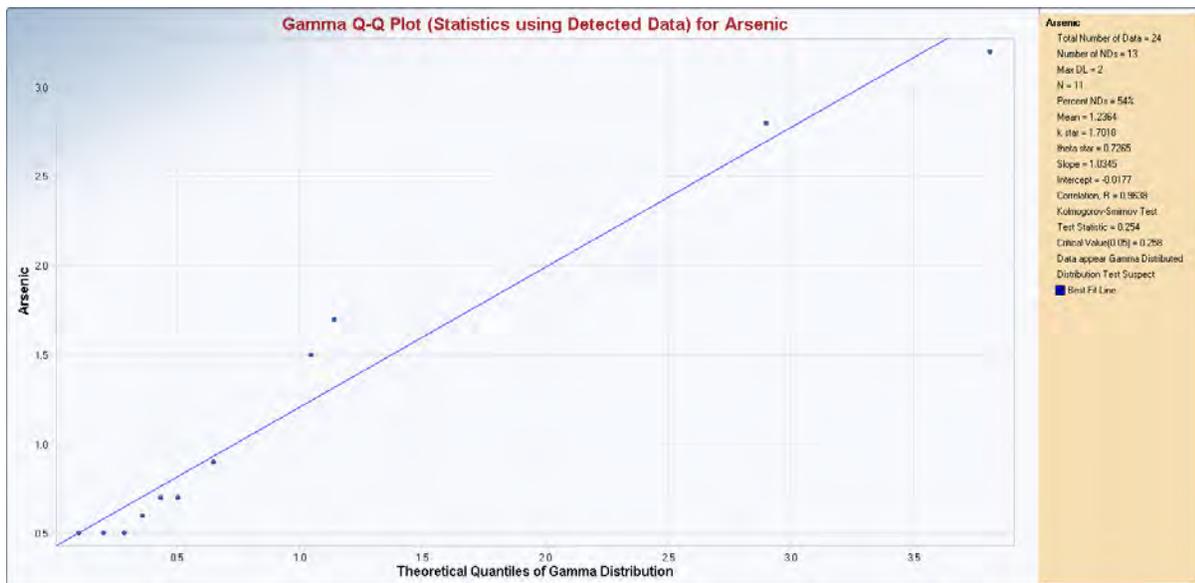
Figure 4-3. Gamma GOF Test on Detected Concentrations of the Oahu Data Set.
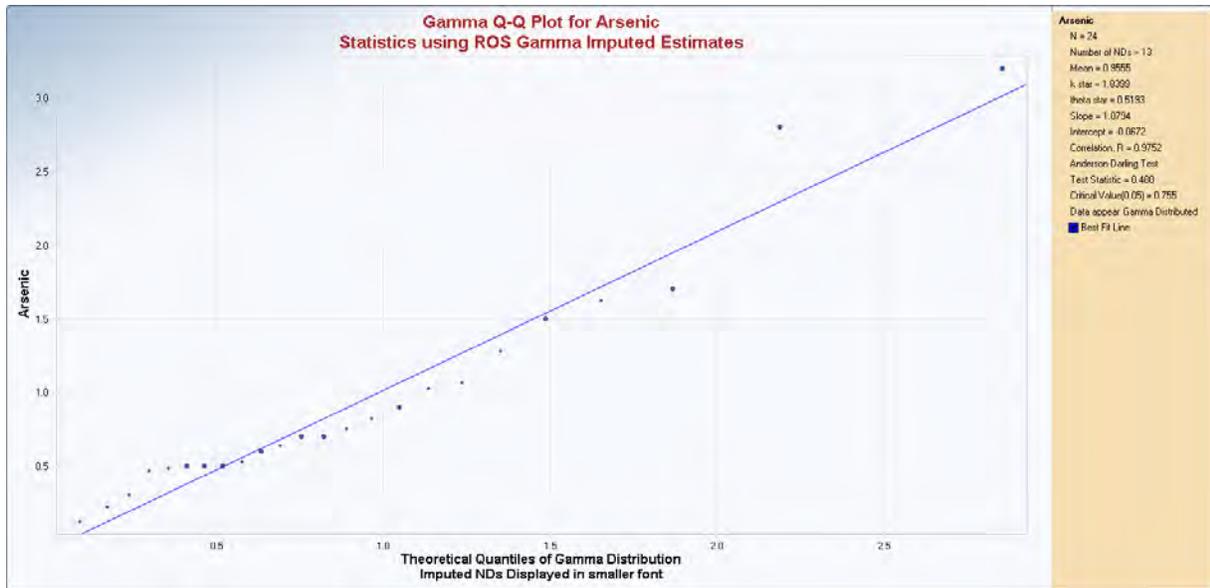
Figure 4-4. Gamma GOF Test on GROS Data Obtained Using the Oahu Data Set.

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 recommends using GROS UCL of 1.27.

122

4.6 A Hybrid KM Estimates Based Approach to Compute Upper Limits for Skewed Data Sets – New in ProUCL 5.0

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 to use the KM method 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 the various 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, ProUCL5.0 computes upper limits using KM estimates in parametric (normal, lognormal, and gamma) equations to compute the various upper limits. It is noted that the use of this hybrid approach has also been suggested in Chapter 15 of EPA (2009) to compute upper limits using KM estimates in the lognormal distribution based equations to compute the various upper limits.

ProUCL 5.0 and its earlier versions compute 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 5.0 (new in ProUCL 5.0) also computes KM UCLs and other upper limits using a lognormal or a gamma distribution. These methods account for skewness present in the data set. For 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<0.5), or consists 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 5.0 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:

$$KM\ UCL95\ (t) = \hat{\mu} + t_{0.95,(n-1)} \sqrt{\hat{\sigma}_{SE}^2} \qquad (4\text{-}8)$$

The above KM UCL(t) represents a good estimate of the EPC term 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, GROS method tends to yield infeasible (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**θ, and
Variance = *k**θ^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:

$$\hat{k} = \hat{\mu}^2_{KM} / \hat{\sigma}^2_{KM}$$

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 5.0 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 detection limits (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 5.0 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 (2009) is 1.155 which appears to be a little higher than the other LROS BCA bootstrap UCL of 1.308 (Table 4-2).

Example 4-6. A real TRS 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 5 and 6. 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).
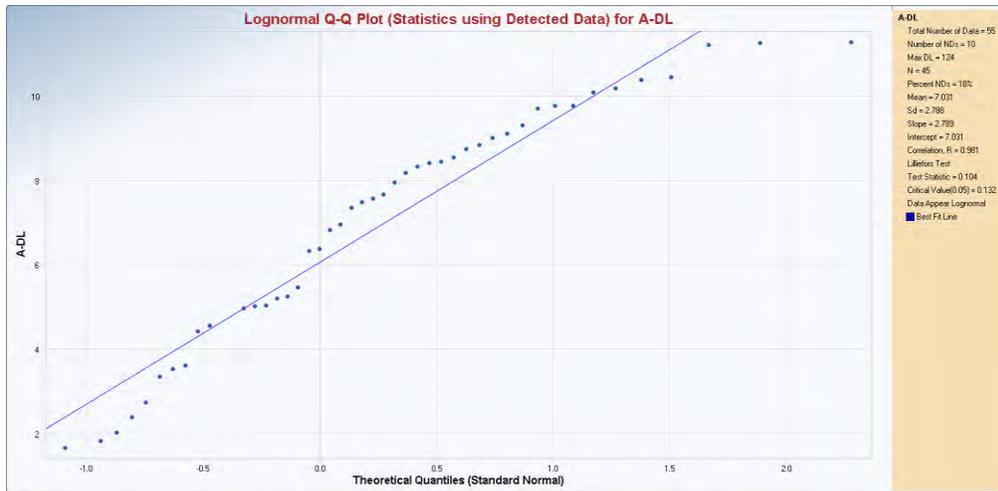
Figure 4-5. Lognormal GOF test on Detected TRS Data Set.

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		
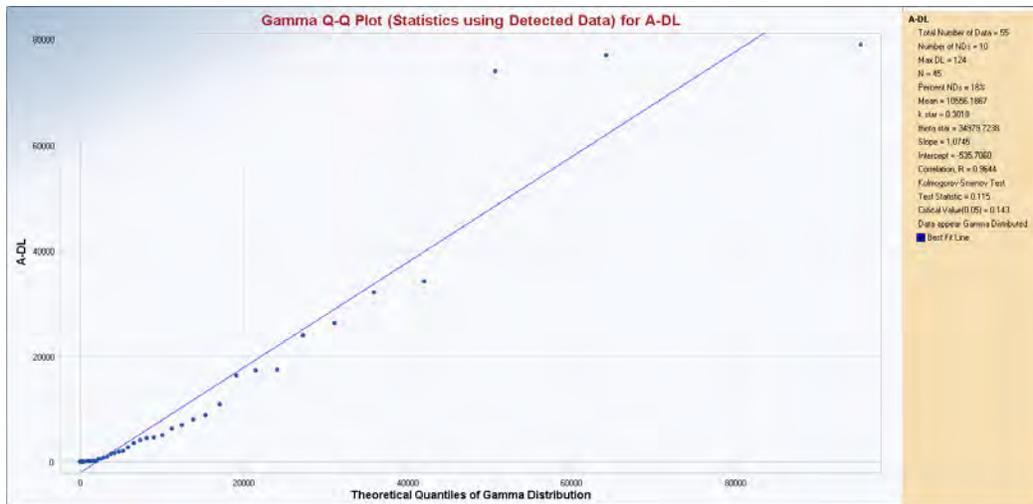
Figure 4-6. Gamma GOF Test on Detected TRS Data Set

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

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 the above table, it is noted that the percentile bootstrap method on LROS method as described in Helsel (2012) 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 5.0.

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. The estimates thus obtained 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 to compute UCL95 for data sets with ND observations. The four bootstrap methods include 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 to compute upper limits based upon 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}, ..., x_{ndk}, x_{k+1}, x_{k+2}, ..., 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 pth upper percentile (used to compute bootstrap UTLs), x_p, that needs to be estimated based upon 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 pth upper percentile, then the estimate, $\hat{\theta}$, may represent the pth 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 an 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 (see Efron, 1981; Barber and Jennison, 1999). The indicator variables, $I_{j :} j:=1,2,...,n$ represent the detection status of the sampled observations, x_j ; $j: = 1, 2,..., n$. A large number, N (1000, 2000) of two-dimensional bootstrap resamples, (x_{iJ}, I_{iJ}), $j:= j: = 1, 2,..., N$, and $i: = 1, 2,..., n$, of size n are drawn with replacement. The indicator variable, I, takes on a value $= 1$ when a detected value is selected and $I = 0$ if a nondetected value is selected. The two-dimensional bootstrap process keeps track of the detection status of each observation in a bootstrap re-sample. In this setting, the detection limits are fixed 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 Nth 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 should be noted there is a positive chance 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 least15to 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 Kaplan-Meier (1958) 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.

<u>Step 1.</u> Let $(x_{i1}, x_{i2}, \dots , x_{in})$ represent the ith bootstrap resample of size n with replacement from the original left-censored data set $(x_1, x_2, \dots, x_n)$. 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 ith bootstrap resample, $i:= 1, 2, \dots, N$.

<u>Step 2.</u> 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 $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N$. 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 mean or N KM means). The bootstrap estimate of the standard error is given by:

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (\bar{x}_i - \bar{x}_B)^2} \ . \tag{4-9}$$

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 $\hat{\theta}_i$, $i:=1,2,\dots N$. 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 ith 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.

<u>Notes:</u> Just like for small uncensored data sets, for small left-censored data sets (e.g., <8-10) with only a few distinct values (e.g., 2,3), 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:

$$t = \frac{\hat{\theta} - \theta}{\hat{\sigma}_B} \ . \tag{4-10}$$

A $(1 - \alpha)*100\%$ standard bootstrap UCL for θ is given as follows:

$$UCL = \hat{\theta} + z_\alpha \hat{\sigma}_B \tag{4-11}$$

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 $(1 - \alpha)*100\%$ 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 $(1 - \alpha)*100\%$ UCL of the mean based upon the bootstrap-t method is given as follows.

$$UCL = \bar{x} - t_{(\alpha N)} \frac{s_x}{\sqrt{n}} \tag{4-12}$$

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, the use of an appropriate Chebyshev inequality-based UCL is recommended. 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:

$$95\% \, Percentile - UCL = 95^{\text{th}}\% \, \bar{x}_i \, ; \, i: = 1, \, 2, \, ..., \, N \tag{4-13}$$

For example, when $N = 1000$, a simple 95% percentile-UCL is given by the 950th ordered mean value given by $\bar{x}_{(950)}$. It is observed that for skewed (lognormal and gamma) data sets, the BCA bootstrap method performs (described below) slightly better (in terms of coverage probability) than the simple percentile method.

4.7.1.4 Bias-Corrected Accelerated (BCA) Percentile Bootstrap Procedure

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:

$$\textit{(1- }\alpha\textit{)*100\% UCL}_{PROC} = BCA - UCL = \bar{x}_{PROC}^{\alpha_2} \tag{4-14}$$

Here $\bar{x}_{PROC}^{\alpha_2}$ is the $\alpha_2 100^{\text{th}}$ 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:

130

$$\alpha_2 = \Phi\left[\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{\alpha}(\hat{z}_0 + z^{(1-\alpha)})}\right] \qquad (4\text{-}15)$$

$\Phi(Z)$ 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:

$$\hat{z}_0 = \Phi^{-1}\left[\frac{\#(\bar{x}_{PROC,i} < \bar{x}_{PROC})}{N}\right], \, i: = 1, \, 2, \, ..., \, N \qquad (4\text{-}16)$$

$\Phi^{-1}(x)$ 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:

$$\hat{\alpha} = \frac{\sum(\bar{x}_{PROC} - \bar{x}_{-i,PROC})^3}{6[\sum(\bar{x}_{PROC} - \bar{x}_{-i,PROC})^2]^{1.5}} \qquad (4\text{-}17)$$

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 $(n\text{-}1)$ observations without the ith 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:

$$P(-k\sigma_1 \leq X - \mu_1 \leq k\sigma_1) \geq 1 - 1/k^2.$$

A $(1-\alpha)*100$ UCL of population mean, μ_1, can be obtained by:

$$UCL = \bar{x} + \sqrt{((1/\alpha) - 1)} s_x / \sqrt{n}. \qquad (4\text{-}20)$$

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 but reasonable UCL of the mean.

4.9 Saving Imputed NDs Using Stats/Sample Sizes Module of ProUCL 5.0

Using this option, NDs are imputed based upon the selected distribution (e.g., normal, lognormal, or gamma) of the detected observations. Using the menu option, "Imputed NDs using ROS Methods" ProUCL 5.0 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 5.0 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).

Example 4-7. Pyrene Data Set continued: Even though this data set has been used extensively in the literature (Helsel, 2012; Helsel, 2013 [NADA Package for R]), not much guidance is provided on how to compute upper limits such as a UCL of the mean based upon 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 left-censored data sets. This data set also illustrates the impact of a few outliers (not wrong values) on the various estimates and statistics. Table 4-6a has statistics computed using the outlier and Table 4-6b has statistics computed without the outlier, 2982. It is noted that the detected data with the outlier does not follow a gamma or a lognormal distribution; and detected data set without the outlier follows a lognormal distribution.

Table 4-6a. 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; and 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 4-6b 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 term.

Table 4-6b. 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 4-6b).

4.10   Summary and Recommendations

Most of the parametric MLE methods (e.g., MLE, RMLE, and expectation maximization [EM]) assume that there is only one detection limit. It is not easy to verify the data distribution of a left-censored data set consisting of detects and NDs with multiple detection limits, therefore some poor performing estimation methods including the parametric MLE methods and the Winsorization method are not retained in ProUCL 5.0. In ProUCL 5.0, 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 the 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 main 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 to avoid the use of the DL/2 (t) UCL method, 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 (e.g., EPA, 2006a) made that the DL/2 method can be used for lower (e.g., ≤ 20%) censoring levels. The coverage provided by the DL/2 (t) method deteriorates fast as the censoring intensity increases. The DL/2 (t) method is not recommended by the authors or developers of this text and ProUCL 5.0.

The use of the KM estimation method is a preferred method as it can handle multiple detection limits. Therefore, the use of KM estimates is suggested to compute decision statistics based upon methods which adjust for data skewness. Depending upon the data set size, distribution of the detected data, and data skewness, the various nonparametric and hybrid KM UCL95 methods including KM (BCA), bootstrap-t KM UCL, Chebyshev KM UCL, and Gamma-KM UCL based upon the KM estimates provide good coverages for the population mean. All of these methods are available in ProUCL 5.0. Suggestions regarding the selection of a 95% UCL of the mean are provided to help the user to select the most appropriate 95% UCL. These recommendations are based upon the results of the simulation studies summarized in Singh, Singh, and Iaci (2002) and Singh, Maichle, and Lee (2006). For additional insight, the user may want to consult a statistician.

134

CHAPTER 5

Computing Upper Limits to Estimate Background Threshold Values Based Upon Data Sets Consisting of Nondetect (ND) Observations

5.1   Introduction

Background evaluation and comparison studies require the computation of upper prediction limits (UPLs) and upper tolerance limits (UTLs) based upon left-censored data sets consisting of nondetect (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 upper confidence limits (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 Kaplan-Meier (KM), Chebyshev inequality, and other computer intensive bootstrap methods to compute upper limits needed to estimate background threshold values (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 regression on order statistics (ROS) methods (e.g., gamma ROS [GROS] and log ROS [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, ordinary least squares (OLS) estimates) of interest. OLS regression estimates (e.g., 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 to manually identify potential outliers present in a data set before proceeding with the computation of the various BTV estimates as described in this chapter. As mentioned in earlier chapters, upper limits computed by including a few low probability high outliers tend to represent locations with those elevated concentrations rather than representing the main dominant population. Compute all relevant statistics using data sets with outliers and without outliers, and compare the results. 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). This in turn helps the project team 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, exposure point concentration [EPC], BTV) under consideration. Since the treatment and handling of outliers in environmental applications is a subjective and controversial topic, it is also suggested that the outliers be treated on a site-specific basis using all existing knowledge about the site and reference areas under investigation.

A couple of classical outlier tests (e.g., Dixon and Rosner tests) are available in the ProUCL software. These tests can be used on data sets with or without ND observations. Additionally, one can use graphical displays such as Q-Q plots and box plots to visually identify high outliers in a left-censored data set. It should be pointed out, that 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. For the purpose of the identification of high outliers, one may replace ND values by their respective detection limits or half of the detection limits or may just ignore them (e.g., especially when high reporting limits are associated with NDs) from the outlier tests. A similar approach can be used to generate graphical displays: quantile-quantile (Q-Q) plots and histograms. Except for the identification of high outlying observations, the outlier test statistics (computed with NDs or without NDs) are not used in any of the estimation and decision making processes. Therefore, for the purpose of testing for high outliers, it does not matter how the ND observations are treated.

5.3   Estimating BTVs Based Upon Left-Censored Data Sets

This section describes parametric and nonparametric methods to compute upper limits (e.g., UPLs, UTLs, USLs, upper percentiles) that may be used to estimate BTVs and other not-to-exceed levels based upon 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, 2009, 2011]) to compute UPLs and UTLs based upon statistics (e.g., mean, *sd*) obtained using MLE, KM, or ROS methods. The methods 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 data sets with the *sd* of the log-transformed data less than 0.75 to 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 UCL95 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 would also be less than acceptable. For skewed data sets, parametric (e.g., gamma distribution on KM estimates) or nonparametric methods (e.g., Chebyshev inequality) which account for data skewness should be used to compute BTV estimates. A brief description of those methods is described in the following sections.

136

5.3.1 Computing Upper Prediction Limits (UPLs) for Left-Censored Data Sets

This section describes some parametric and nonparametric methods to compute UPLs based upon left-censored data sets.

5.3.1.1 UPLs Based Upon Normal Distribution of Detected Observations and KM Estimates

When detected observations in a data set consisting of NDs follow a normal distribution (which can be verified by using GOF module of ProUCL), one may use the normal distribution on KM estimates to compute the various upper limits needed to estimate BTVs (already available in ProUCL 4.1). A $(1 - \alpha)*100$ 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:

$$UPL_1 = \hat{\mu}_{KM} + t_{((1-\alpha),(n-1))} \sqrt{\hat{\sigma}_{KM}^2 (1+1/n)} \qquad (5\text{-}1)$$

Here $t_{((1-\alpha),(n-1))}$ is the critical value of Student's t-distribution with $(n-1)$ degrees of freedom (df). 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 $UPL_1 95$ 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 (incorporated in ProUCL 5.0).

5.3.1.2 UPL Based Upon the Chebyshev Inequality

The Chebyshev inequality can be used to compute a reasonably conservative but stable UPL and is given as follows:

$$UPL = \bar{x} + [\sqrt{((1/\alpha)-1)*(1+1/n)}]s_x \qquad (5\text{-}2)$$

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, UPL based upon the Chebyshev inequality tends to yield higher estimate of BTVs than the other methods. This is especially true when skewness is mild (e.g., sd of log-transformed data is low < 0.75), and the sample size is large (e.g., > 30). The user is advised to use professional judgment before using this method to compute a UPL. Specifically, for larger skewed data sets, instead of using a 95% UPL based upon the Chebyshev inequality, the user may want to compute a Chebyshev UPL with a lower CC (e.g., 85%, 90%) to estimate a BTV. ProUCL can compute a Chebyshev UPL (and all other UPLs) for any user specified CC in the interval [0.5, 1].

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 $k \geq 1$ observations. These limits are described below when $k=1$.

$$\text{\textit{Wilson-Hilferty (WH) UPL}} = \max\left(0, \left(\overline{y}_{KM} + t_{((1-\alpha),(n-1))} * s_{yKM} * \sqrt{1+\frac{1}{n}}\right)^3\right)$$

$$\text{\textit{Hawkins-Wixley (HW) UPL}} = \left(\overline{y}_{KM} + t_{((1-\alpha),(n-1))} * s_{yKM} * \sqrt{1+\frac{1}{n}}\right)^4$$

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 $k \geq 1$ observations. These limits are described below when $k=1$. An upper $(1 - \alpha)*100\%$ lognormal UPL is given by the following equation:

$$UPL = \exp(\overline{y} + t_{((1-\alpha),(n-1))} * s_y * \sqrt{(1+1/n)})$$

Here $t_{((1-\alpha),(n-1))}$ is a critical value from Student's t-distribution with $(n-1)$ df, and $\overline{y}$ 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 pth percentile given as given as follows:

$$\hat{x}_p = \hat{\mu}_{KM} + z_p \sqrt{\hat{\sigma}^2_{KM}} \tag{5-3}$$

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 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:

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

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 HW approximation, the transformed detected data $Y = X^{1/4}$ follow an approximate normal distribution. Let $\bar{y}$ and s_y represent 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:

$$\hat{x}_p = \max\left(0, \left(\bar{y} + z_p * s_y\right)^3\right)$$

$$\hat{x}_p = \left(\bar{y} + z_p * s_y\right)^4$$

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 5.0 generates additional columns consisting of the *(n-k)* detected values, and k imputed ND values. Once, the ND observations have been estimated, an experienced user may use any of the parametric or nonparametric percentile computation methods (e.g., gamma percentiles) for full data sets as described in Chapter 3. 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 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:

$$UTL = \hat{\mu}_{KM} + K_{n,\alpha,p} * \sqrt{\hat{\sigma}^2_{KM}} \tag{5-4}$$

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 (e.g., 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:

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

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

$$UTL = \left(\bar{y} + K * s_y\right)^4$$

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 5.0 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.

5.3.4   Computing Upper Simultaneous Limits (USLs) for Left-Censored Data Sets

Parametric and nonparametric Upper simultaneous limit (USL) computation methods for left-censored data sets 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 $(1 - \alpha)$ 100% USL providing $(1 - \alpha)$ 100% coverage for all sample observations is given by:

$$USL = \hat{\mu}_{KM} + d_{2\alpha}^b * \sqrt{\hat{\sigma}_{KM}^2}$$

Here $(d_{2\alpha}^b)^2$ 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 one-sided *(1 – α) 100%* USL is given by the following equation:

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

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

*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 HW approximation, the transformed detected data $Y = X^{1/4}$ follow an approximate normal distribution. Let $\bar{y}$ and s_y represent KM mean and *sd* of the transformed data (detects and NDs), y. 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 $(1-\alpha)$ 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. Several examples illustrating the computations of the various limits described in this section are discussed next.

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 5.0 are described in Tables 5-1 through 5-3 as follows. 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. 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		

Table 5-1 mostly has normal distribution based upper limits computed using KM estimates as described in Helsel (2012) 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, it is noted that 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 (TRS data): 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 gamma distribution. This data set is considered here to illustrate the merits of the gamma distribution based upper limits. It was noted 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 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.

A-DL
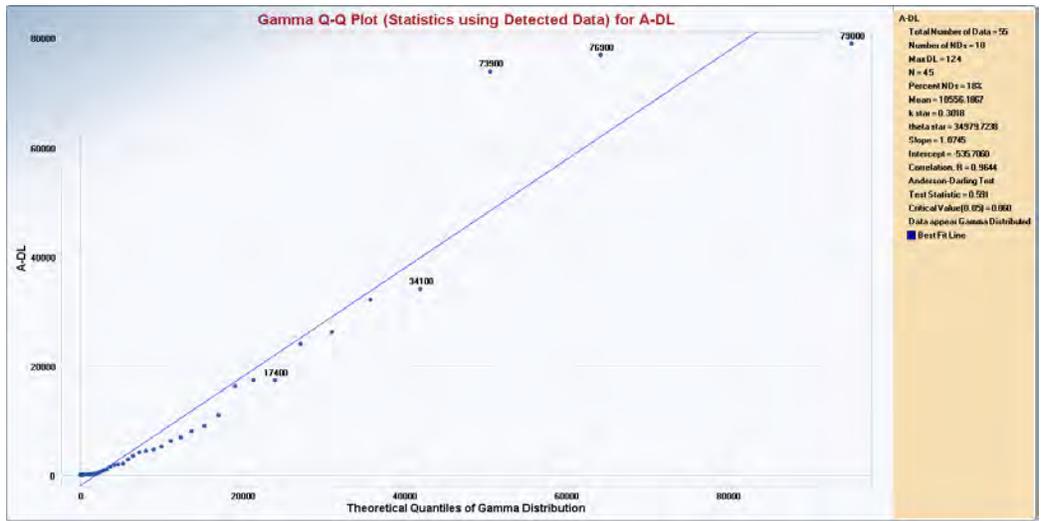
General Statistics			
Total Number of Observations	55	Number of Missing Observations	0
Number of Distinct Observations	53		
Number of Detects	45	Number of Non-Detects	10
Number of Distinct Detects	45	Number of Distinct Non-Detects	8
Minimum Detect	5.2	Minimum Non-Detect	3.8
Maximum Detect	79000	Maximum 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

KM method based estimates of the mean, SE of the mean, and *sd* are:

Mean	8638
SD	18246
Standard Error of Mean	2488

Notes: Direct estimate of KM SD = 18246
Indirect Estimate of KM SD (Helsel, 2012) = 18451.5

The gamma goodness-of-fit (GOF) test on detected data and various upper limits including UCLs obtained using the GROS method and gamma distribution on KM estimates are described in Table 5-4; and lognormal GOF test on detected data and the various upper limits obtained using the LROS method and lognormal distribution on KM estimates are described in Table 5-5 . Table 5-6 has the summary of the main upper limits computed using the lognormal and gamma distribution of the detected data.

Table 5-4. 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

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

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

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

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 5-5. Upper Limits Using LROS, 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		

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 5-6.

146

Table 5-6. 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- ProUCL5.0	121,584	Helsel (2012), EPA 2009
UTL95-95 (ROS)	47,429	WH- ProUCL 5.0	394,791	Helsel (2012), EPA 2009
UPL95 (KM)	32,961	WH-ProUCL 5.0	106,741	EPA (2009)
UTL95-95 (KM)	46,978	WH-ProUCL 5.0	334,181	EPA(2009)
UCL95 (ROS)	14,445	ProUCL5.0	14,659	bootstrap-t, ProUCL 5.0
			12,676	percentile bootstrap, Helsel (2012)
UCL (KM)	14,844	ProUCL 5.0	1,173,988	H-UCL, KM mean and *sd* on logged data - EPA (2009)

The statistics described in Tables 5-4 and 5-5, and summarized in Table 5-6 demonstrate the merits of using the gamma distribution to compute practical and meaningful estimates (upper limits) of the decision parameters (e.g., mean, upper percentile) of interest.

5.4   Computing Nonparametric Upper Limits Based Upon Higher Order Statistics

It is noted that 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 detection limits, the USLs, UTLs, UPLs, and upper percentiles can be estimated 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 main 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 to compare constituent concentrations of two (e.g., site versus background) or more (several monitoring wells [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 nondetect (ND) observations. For data sets consisting of multiple nondetects, a new two-sample hypothesis test, the Tarone-Ware (T-W; 1978) test has been incorporated in the current ProUCL, version 5.0. 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 extract, 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. It is noted that some recent EPA guidance documents (e.g., EPA 2009) are also recommending 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 5.0 and its earlier versions.

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 an upper confidence level [UCL] of the mean) is used (assuming normality of site data, or when the sample size is large such as 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 use 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 (e.g., associated detection limits) 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, then a test for proportion or a test for percentile (or equivalently a 95% upper tolerance limit with 95% [UTL95-95] coverage, UTL95-90) can be used to compare the site proportion, p of exceedances of the action level, A_0 by site observations to some pre-specified allowable proportion, P_0 of exceedances of A_0 by site observations (perhaps after some remediation activity). It is noted that the proportion test can also handle NDs provided all NDs are below the action level, A_0.

It is desirable to use DQOs 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 single-sample hypotheses tests listed above. As mentioned before, if only a small number (e.g., < 6) 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 DQOs 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 inherent greater heterogeneity/variability potentially 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 5.0. Some details of those methods are described in this chapter for interested users. It is recommended to supplement statistical results and test statistics 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.

149

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 EPA (2003) and the ITRC Technical Regulatory guidance document (2012).

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 that a hypothesis test statistic (e.g., t-statistic, WMW test statistic) will evaluate and determine. 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 many 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≥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 area of concern (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.

150

6.3.2 Test Form 2

The null hypothesis (H₀): 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	
	H₀ is True	**H₀ is not true**
H₀ is not rejected	Correct Decision: $(1 - \alpha)$	Type II Error: False Negative (β)
H₀ 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 (e.g., 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% ($\alpha = 0.10$) and the power should be at least 80% ($\beta = 0.20$).

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.

152

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; there are no 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 goodness-of-fit (GOF) tests to determine data distribution (such as the Shapiro-Wilk [S-W] test for normality) often fail if there are not enough observations (e.g., <20 observations), or 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, and 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 the 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. It should be noted that typically, nonparametric tests and statistics require data sets of larger sizes to derive correct conclusions. Several two-sample nonparametric hypotheses tests: the WMW test, Gehan test, and Tarone-Ware test are available in ProUCL. Like the Gehan test, the Tarone-Ware test (T-W) is used on data sets consisting of NDs with multiple RLs. The T-W test is new in ProUCL 5.0.

The relative performances of different testing procedures can be assessed by comparing, *p-values* associated with those tests (e.g., GOF 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 consisting 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, Mann-Kendall 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. It should be noted that 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 the parametric test may work well provided that the data sets are not heavily skewed. For heavily skewed data sets, the sample size requirement associated with the central limit theorem (CLT) can become quite large such as larger than 100. A brief simulation study elaborating

on the sample size requirements to apply 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 (for example, less than 20 data points), then the p-value may be inaccurate because the central limit theory does not apply in this case.

- If a nonparametric test is applied to 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 desired DQOs. The collected sample should represent a random sample representing the site area of concern 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 to use this test when dealing with data sets consisting of 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 (e.g., *sd* of log-transformed data exceeding 1, 1.5), even when the data set is of a large size such as 100.

6.6.1.2   Directions for the One-Sample t-Test

Let $x_1, x_2, \ldots, x_n$ represent a random sample (data set of analytical results) of size, n, collected from a population (AOC). The user should test for the normality of this data set before using t-test (ProUCL verifies normality automatically).

STEP 1: Specify an average cleanup goal or action level, μ_0 (C_s), and choose one of the following null hypotheses:

Form 1: *H_0: site $\mu \leq \mu_0$ vs. alternative H_A: site $\mu > \mu_0$*
Form 2: *H_0: site $\mu \geq \mu_0$ vs. alternative H_A: site $\mu < \mu_0$*
Two-Sided: *H_0: site $\mu = \mu_0$ vs. alternative H_A: site $\mu \neq \mu_0$.*
Form 2 with substantial difference, S: *H_0: site $\mu \geq \mu_0 + S$ vs. alternative H_A: site $\mu < \mu_0 + S$, here $S \geq 0$.*

STEP 2: Calculate the test statistic:

$$t_0 = \frac{\bar{x} - \mu_0 - S}{\dfrac{sd}{\sqrt{n}}} \tag{6-1}$$

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 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 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 (LCL) 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 lower confidence limits (LCLs) and upper confidence limits (UCLs) for normally distributed data sets. The use of a 95% upper confidence limit (UCL95) is very common to determine if a polluted site has attained the cleanup standard, μ_0 after remediation. If a UCL95 is less than 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), 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: *H₀: site $\mu \leq \mu_0$ vs. alternative H$_A$: site $\mu > \mu_0$*
Form 2: *H₀: site $\mu \geq \mu_0$ vs. alternative H$_A$: site $\mu < \mu_0$*
Two-Sided: *H₀: site $\mu = \mu_0$ vs. alternative H$_A$: site $\mu \neq \mu_0$.*

Based upon 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

$$\bar{x} - t_{n-1,0.05} * sd / \sqrt{n} > \mu_0; \text{ or when } LCL95 > \text{cleanup goal}, \mu_0.$$

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

$$\bar{x} < \mu_0 - t_{n-1,0.05} * sd / \sqrt{n}$$

Equivalently reject the null hypothesis and conclude that site meets the cleanup standard when

$$\bar{x} + t_{n-1,0.05} * sd / \sqrt{n} < \mu_0; \text{ or when } UCL95 < \mu_0.$$

The site is concluded clean when UCL95 < μ_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 (detection limits, 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:

$$z_0 = \frac{p + c - P_0}{\sqrt{P_0(1-P_0)\big/ n}} \tag{6-2}$$

$$\text{Where } c = \begin{cases} \dfrac{-0.5}{n} \, if, p > P_0 \\ \dfrac{0.5}{n} \, if, p < P_0 \end{cases} \quad \text{and} \quad p = \frac{x(\# \text{ of site values} > A_0)}{n}$$

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 5.0 also performs the proportion test based upon the exact binomial distribution when the sample size is small and one may not be able to use the normal approximation as described above. ProUCL 5.0 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 5.0 performs this test based upon the BD, when the above conditions are not satisfied. In such cases, ProUCL 5.0 computes the critical values and p-values based upon the BD and its cumulative distribution function (CDF). The probability statements concerning the computation of p-values can be found in Conover (1999).

STEP 2: Conclusion Based Upon 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 α /2; b_1 and b_2 are given by the probability statements: P $(B \leq b_1)$ ~ $\alpha/2$, and $P(B > b_2)$ ~ $\alpha/2$. 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 5.0 is described below.

Let b be the calculated value of the binomial random variable, B under the null hypothesis. ProUCL 5.0 computes the *p*-values using the following 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 $b \leq (n - b)$: *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,...,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,

$$z_0 = \frac{B - \frac{n}{2} - S}{\sqrt{\frac{n}{4}}}.$$

(6-3)

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, 2α)*, then reject the null hypothesis that the population median is less than the cleanup level, C.

Form 2: If $B \leq B_{UPPER}$ *(n, 2α)*, then reject the null hypothesis that the population median is more than the cleanup level.

Two-Sided: If $B \geq B_{UPPER}$ *(n, α)* or $B \leq B_{UPPER}$ *(n, α) - 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, 2α),* 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 detection limits with NDs exceeding the detects are present in the data set.

For data sets with NDs with a single detection limit, DL, a surrogate value of DL/2 is used for all ND values (EPA, 2006). The presence of multiple detection limits makes this test less powerful. It is suggested not to use this test when multiple detection limits are present with NDs exceeding the detected values. Per EPA (2006) guidance, when multiple detection limits are present, then all detects and NDs less than the largest detection limit may be censored which tends to reduce the power of the test. In ProUCL 5.0, 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 detection limit, 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 detection limits. 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* $\leq C$ *vs.* H_A: *Site location* $> C$
Form 2: H_0: *Site location* $\geq C$ *vs.* H_A: *Site location* $< C$
Two-Sided: H_0: *Site location* $= C$ *vs.* H_A: *Site location* $\neq C$
Form 2 with substantial difference, S: H_0: *Site location* $\geq C + S$ *vs.* H_a: *Site location* $< C + S$, here $S \geq 0$.

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 $T^+ = \sum\limits_{\{i:R_i>0\}} R_i$, where T^+ is the sum of the positive signed ranks.

For $n > 20$, use a normal approximation and compute the test statistic given by

$$z_0 = \frac{T^+ - n(n+1)/4}{\sqrt{\operatorname{var}(T^+)}} \tag{6-4}$$

Here $\operatorname{var}(T^+)$ is the variance of T^+ and is given by

$\operatorname{var}(T^+) = \dfrac{n(n+1)(2n+1)}{24} - \dfrac{1}{48} \sum\limits_{j=1}^{g} t_j\left(t_j^2 - 1\right)$; g = number of tied groups.

163

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 $T^+ \geq \dfrac{n(n+1)}{2} - w_\alpha = w_{(1-\alpha)}$, 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 $T^+ \geq \dfrac{n(n+1)}{2} - w_{\alpha/2}$ 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.

<u>Notes:</u> In the above, w_α represents the αth 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:

$P(T^+ \leq w_{1-\alpha}) = 1 - \alpha,$ 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 real data set used in Chapter 3, which was collected from a Superfund site to illustrate the usage of single-sample hypotheses tests described above. 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 results (Form 2), Proportion test (2-sided) and WRS test (Form 1) are shown below.

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		H0: Sample Mean >= 10000 (Form 2)	
Raw Statistics			
Number of Valid Observations	24	Test Value	-2.54
Number of Distinct Observations	24	Degrees of Freedom	23
Minimum	1710	Critical Value (0.05)	-1.714
Maximum	16200	P-Value	0.00915
Mean	7789		
Median	7010	**Conclusion with Alpha = 0.05**	
SD	4264	Reject H0, Conclude Mean < 10000	
SE of Mean	870.4	P-Value < Alpha (0.05)	

Conclusion: Reject the null hypothesis and conclude that mean aluminum concentration <10000.

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.

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	**H0: Sample Mean/Median <= 0.2 (Form 1)**
Maximum	0.456	
Mean	0.147	
Median	0.07	Large Sample z-Test Statistic
SD	0.133	Critical Value (0.05)
SE of Mean	0.0271	P-Value
Number Above Action Level	6	
Number Equal Action Level	0	**Conclusion with Alpha = 0.05**
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
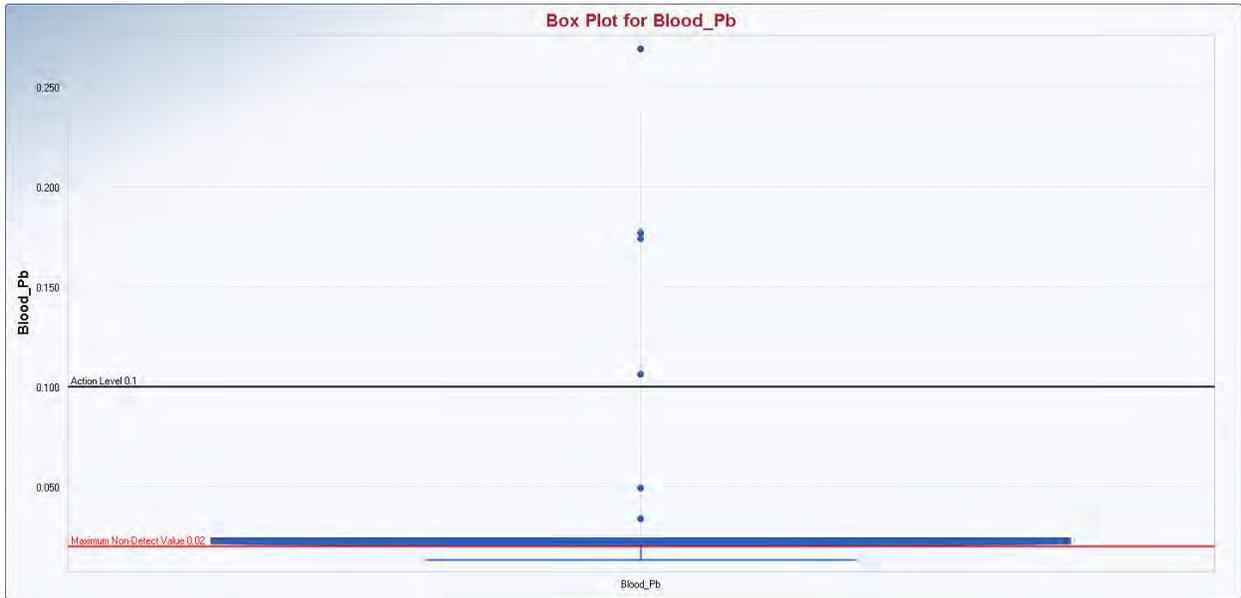
Figure 6-1. Box Plot of Lead in Blood Data Comparing Pb Concentrations with the Action Level of 0.1.

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	**H0: Sample Median >= 0.1 (Form 2)**
Mean of Detects	0.107	
Median of Detects	0.0776	
SD of Detects	0.0911	Large Sample z-Test Statistic -3.667
Median of Processed Data used in WSR	0.01	Critical Value (0.05) -1.645
Number Above Action Level	4	P-Value 1.2291E-4
Number Equal Action Level	0	
Number Below Action Level	23	**Conclusion with Alpha = 0.05**
T-plus	39	Reject H0, Conclude Mean/Median < 0.1
T-minus	339	P-Value < Alpha (0.05)

167

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

$X_1, X_2, ..., X_n$ represent site samples and $Y_1, Y_2, ..., Y_m$ 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 5.0 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, then nonparametric procedures such as the WMW may be applied. This test is not robust to outliers because sample means and standard deviations are sensitive to outliers. 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 (new in ProUCL 5.0) 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.*

168

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 $\overline{X}$ and the sample variance S_X^2 for the site (e.g., Population 1, Sample 1) data and compute the sample mean $\overline{Y}$ 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:

$$ S_p = \sqrt{\frac{(n-1)s_x^2 + (m-1)s_y^2}{(m-1)+(n-1)}} \; ; \tag{6-5} $$

$$t_0 = \frac{(\bar{X} - \bar{Y}) - S}{S_p\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}}$$

(6-6)

Here S = 0, except when used in Form 2 hypothesis with substantial difference, $S \geq 0$.

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 $\overline{X}$ and the sample variance S_X^2 for the site data and compute the sample mean $\overline{Y}$ and the sample variance S_Y^2 for the background data.

STEP 3: Use the F-test described below (in ProUCL) to verify if the variances of the two populations are comparable. Compute the t-statistic:

$$t_0 = \frac{\overline{X} - \overline{Y} - S}{\sqrt{\dfrac{s_X^2}{n} + \dfrac{s_Y^2}{m}}}$$ (6-7)

Here S = 0, except when used in Form 2 hypothesis with substantial difference, $S \geq 0$.

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:

$$df = \frac{\left[\dfrac{s_X^2}{n} + \dfrac{s_Y^2}{m} \right]^2}{\dfrac{s_X^4}{n^2(n-1)} + \dfrac{s_Y^4}{m^2(m-1)}}$$ (6-8)

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.

<u>P-Values for Two-sample t-Test</u>

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 <u>required</u> 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 5.0 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 6-2.
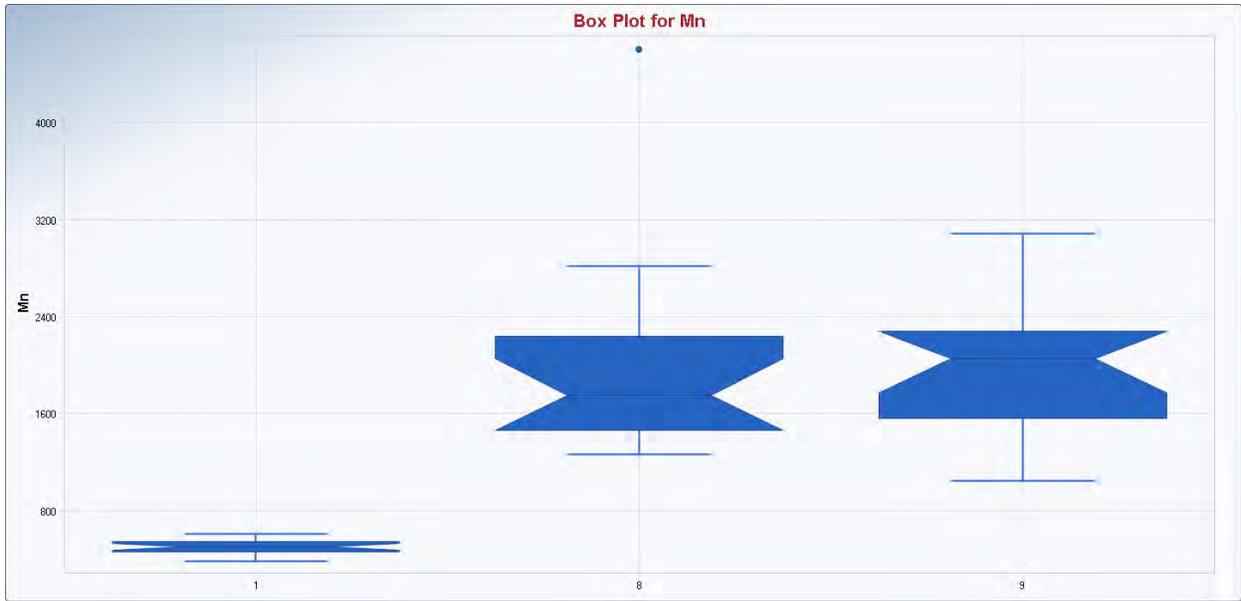
Figure 6-2. Box Plots Comparing Concentrations of Three Wells: One Upgradient and Two Downgradient.

Table 6-2. T-Test Comparing Mn in MW8 vs. MW9

H$_0$: Mean Mn concentrations of MW 8 and MW9 are comparable

	Sample 1 vs Sample 2 Two-Sample t-Test					
	H0: Mean of Sample 1 = Mean of Sample 2					
Selected Null Hypothesis	Sample 1 Mean = Sample 2 Mean (Two Sided Alternative)			t-Test	Lower C.Val	Upper C.Val
Alternative Hypothesis	Sample 1 Mean <> Sample 2 Mean	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 Varianc	24.5	0.123	-2.064	2.064	0.903
Sample 1 Data: Mn-89(8)		Pooled SD: 690.548				
Sample 2 Data: Mn-89(9)		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				
Raw Statistics			**Test of Equality of Variances**			
	Sample 1	Sample 2				
Number of Valid Observations	16	16				
Number of Distinct Observations	16	15	Variance of Sample 1	703523		
Minimum	1270	1050	Variance of Sample 2	250190		
Maximum	4600	3080				
Mean	1998	1968	Numerator DF	Denominator DF	F-Test Value	P-Value
Median	1750	2055	15	15	2.812	0.054
SD	838.8	500.2	Conclusion with Alpha = 0.05			
SE of Mean	209.7	125	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 detection limits.

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 detection limit (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 the Tarone-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 to supplement WMW test results with graphical displays.

6.9.1.2   WMW Test in the Presence of Nondetects

If there are *t* ND values with a single detection limit, then they are considered as "ties" and are assigned the average rank for this group. If more than one detection limit is present in the data set, then WMW test censors all of the observations below the largest detection limit, and are treated as NDs at the largest detection limit. 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 ($n \leq 20$ or $m \leq 20$)

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 detection limit or at an arbitrary (when no detection limit 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:

$$E(U) = nm/2$$
$$Var(U) = nm(n+m+1)/12$$

<u>Notes:</u> 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. $\tilde{\mu}_X - \tilde{\mu}_Y < S$

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 5.0.

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 detection limits and /or NDs exceeding the detected values are present in the data sets. Other tests such as the Tarone-Ware and Gehan tests also available in ProUCL 5.0 are better suited for data sets consisting of many NDs with multiple detection limits 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 5.0 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:

$$Z_0 = \frac{W_s - \frac{n(n+m+1)}{2}}{\sqrt{\frac{nm(n+m+1)}{12}}} \tag{6-10}$$

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:

$$Z_0 = \frac{W_s - n(n+m+1)/2}{\sqrt{\frac{nm}{12}\{(n+m+1) - \frac{\sum_{j=1}^{g} t_j(t_j^2 - 1)}{(n+m)(n+m-1)}\}}} \tag{6-11}$$

Here g represents the number of tied groups and t_j is the number of tied values in the jth group.

STEP 5: For large data sets with both n and $m \geq 20$, ProUCL computes approximate test statistic given by equations (6-10) and (6-11) and computes 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.

178

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 dataset used here can be downloaded from the ProUCL website. The data set consists of several tied observations. The test results are summarized in Table 6-3.

Table 6-3. 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.
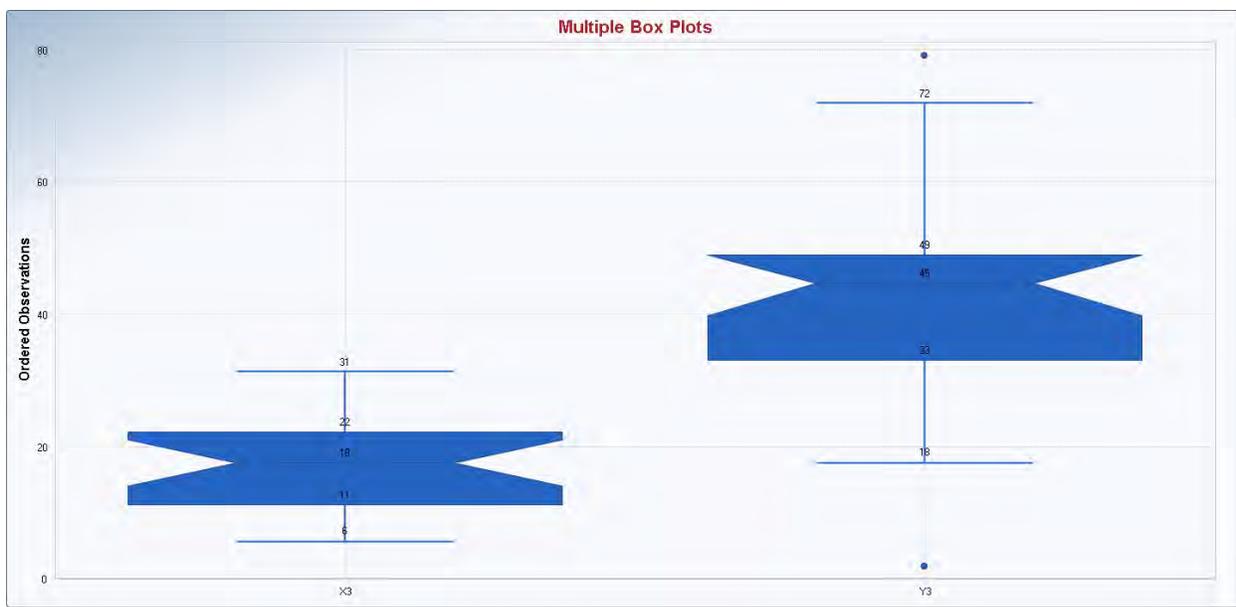
Box Plots Comparing Values of Two Groups used in Example 6-4.

6.9.2   The Gehan Test

The Gehan test 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 detection limits. 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 detection limits. 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 $\geq$ 10, the normal distribution based approximate Gehan's test statistic is described as follows.

180

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 5.0. 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 detection limits 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, ..., N*.

STEP 5: Compute the Gehan statistic, *G*, as follows:

$$G = \frac{\sum_{i=1}^{N} h_i a(R_i)}{\left[mn \sum_{i=1}^{N} \frac{[a(R_i)]^2}{N(N-1)} \right]^{\frac{1}{2}}} \tag{6-12}$$

Where $\begin{cases} h_i = 1 \\ h_i = 0 \end{cases}$

$h_i = $ *1* if the ith datum is from the site population
$h_i = $ *0* if the ith datum is from the background population
$N = n + m$

$a(R_i) = 2 R_i - N - 1$, 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 detection limits. The T-W test as described below has been incorporated in ProUCL 5.0. 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 detection limits 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 detection limits 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 (e.g., <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.

182

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 using ProUCL 5.0:

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 <u>distinct detected </u>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, $i=1,2,....N$ of detects and NDs (= reported as detection limits 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 ; $i=1,2,....N$

STEP 5: Repeat Step 4 on the site data set. That is count the number, $m_i, i=1,2,....N$ of detects and NDs (reported as detection limits 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 detection limits 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:

$$E_{Site}(Detection) = c_i * m_i / n_i \tag{6-13}$$

$$V_{Site}(Detection) = c_i * (n_i - c_i)m_i l_i /(n_i^2(n_i - 1)) \tag{6-14}$$

STEP 7: Compute the normal approximation of the TW test statistic using the following equation:

$$T - W = \frac{\sum_{i=1}^{N} \sqrt{n_i} (d_i - E_{Site}(Detection))}{\sqrt{\sum_{i=1}^{N} n_i (V_{Site}(Detection))}}$$ (6-15)

STEP 7: 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 $T\text{-}W \leq\text{-} z_{1-\alpha}$, 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 6-3 and box plots comparing Zn concentrations in the two zones are shown in Figure 6-4.
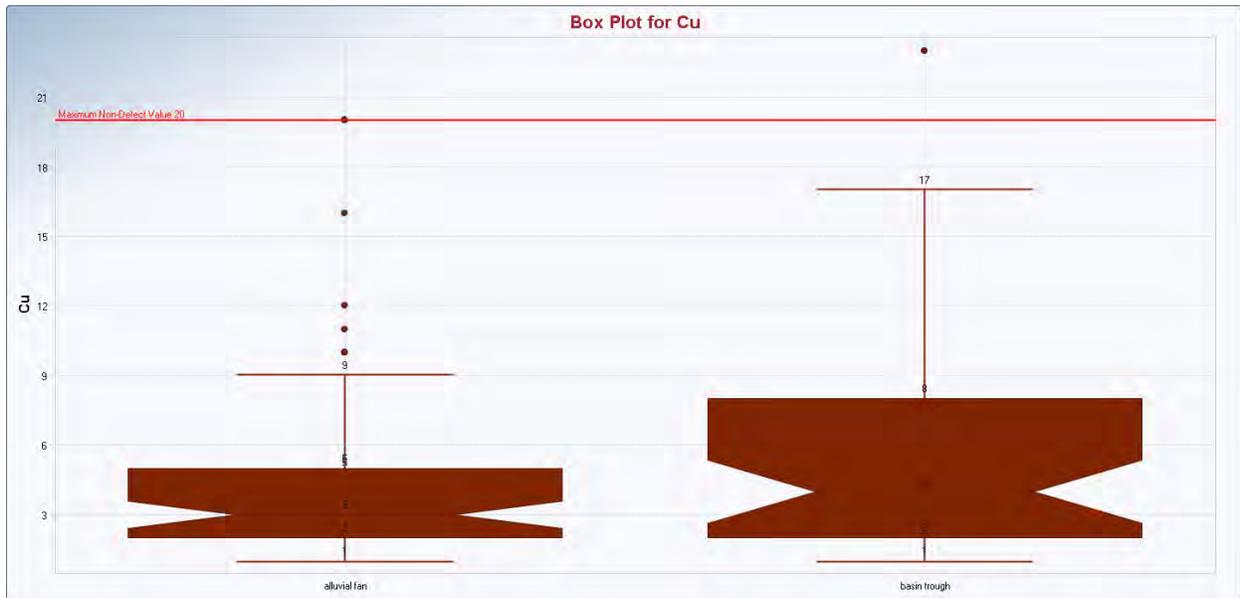
Figure 6-3. Box plots Comparing Cu in Two Zones: Alluvial Fan versus Basin Trough

184
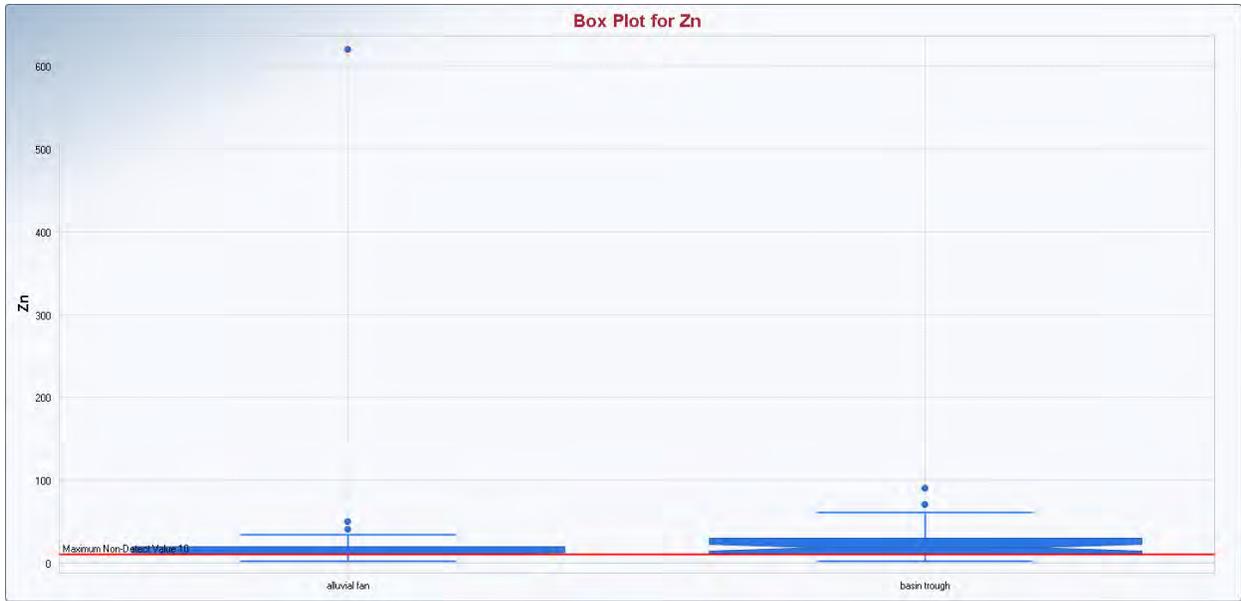
Figure 6-4. Box Plots Comparing Zn in Two Zones: Alluvial Fan versus Basin Trough

Table 6-4. 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 6-4, 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 6-5. 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 6-4 and the T-W test results summarized in Table 6-5, the null hypothesis is rejected, and it is concluded that the Zn concentrations in groundwaters of two zones are not comparable (p-value = 0.0346).

186

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 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 <u>wrong/incorrect</u> values, which often result from transcription errors, data-coding errors, or instrument breakdown errors. Such wrong values could be outlying (well-separated, coming from 'low' probability far tails), with respect to the rest of the data set; these outliers need to be fixed and corrected (or removed) before performing a statistical method. However, a sampled data set can also consist of some <u>correct</u> measurements that are extremely large or small relative to the majority of the data, and therefore those low probability extreme values are suspected of misrepresenting the main dominant background population from which they were collected. Typically, correct extreme values represent observations coming from population(s) other than the main dominant population; and such observations are called outliers with respect to the main dominant population.

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 main 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 main 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, goodness-of-fit (GOF) statistics, ordinary least squares (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. The use of such distorted statistics (e.g., two-sample tests, 95% upper confidence limits [UCL95], 95% upper tolerance limits with 95% coverage [UTL95-95]) may lead to incorrect cleanup decisions which may not be cost-effective or protective of human health and the environment.

A distorted estimate (e.g., UCL95) computed by accommodating a few low probability outliers (coming from far tails) tends to represent the population area represented by those outliers and not the main dominant population of interest.

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. In order to compute robust and resistant estimates of the population parameters of interest (e.g., exposure point concentrations [EPC] terms, background threshold values [BTV]s), EPA NERL-Las Vegas, NV developed a multivariate statistical software package, Scout 2008, Version 1.0 (EPA 2009) consisting of

several univariate and multivariate robust outlier identification and estimation methods. Scout software can be downloaded from the following EPA website: http://www.epa.gov/nerlesd1/databases/datahome.htm.

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 quantile-quantile (Q-Q) plots and box plots (e.g., 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, Shapiro-Wilk (S-W) test). Graphical displays help identify observations that are much larger or smaller than the bulk (majority) of the data. The statistical tests (e.g., Rosner test) alone cannot determine whether a statistical outlier should be investigated further. Based upon historical and current site and regional information, graphical displays, and outlier test results, 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).

Notes: In practice, extreme outliers represent: 1) low probability observations potentially coming from the extreme far tails (with low to negligible probability) of the distribution of the main population under consideration, or 2) observations coming from population(s) different from the main dominant population of interest. On a normal Q-Q plot, observations well-separated (sticking out, significantly higher than the majority of the data) from the majority of observations represent extreme physical outliers; and the presence of a few high outlying observations distorts the normality of a data set. That is, many data sets follow a normal distribution after the removal of identified outliers.

It should be pointed out that methods incorporated in ProUCL can be used on any data set (with or without nondetects [NDs]) with or without the outliers. Specifically, it should not be misrepresented that ProUCL 5.0 is restricted and can be used only on data sets without the 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 how the inclusion of a *few low probability* 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., 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.

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 consisting of outliers will not follow a normal distribution unless 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 data set with outliers) needs to follow a normal distribution. There appears to be some confusion among practitioners (e.g., Helsel and Gilroy, 2012) and they mistakenly assume that one can perform Dixon and Rosner tests only when the data set, including outliers, follows a normal distribution which is true only rarely.

As noted earlier, a lognormal model tends to accommodate outliers (Singh, Singh, and Engelhardt, 1997), and a data set with outliers can follow a lognormal distribution. This does not imply that the outlier potentially representing the impacted location does not exist! Those impacted locations may need further investigations. Outlier tests should be performed on raw data, as the cleanup decision needs to be made based upon concentration values in the raw scale and not in log-scale or some other transformed scale (e.g., cube root). 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. Over the last 25 years, several modern iterative robust outlier identification methods have been developed (Rousseeuw and Leroy, 1987; Singh and Nocerino, 1995) which are beyond the scope of ProUCL. Some of those methods are available in the Scout 2008 version 1.0 software (EPA, 2009)

7.3   Outlier Tests for Data Sets without Nondetect Observations

A couple of classical outlier tests described 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 areas of concern [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 $X_{(1)}$, $X_{(2)}$, . . . , $X_{(n)}$ represent the data ordered from smallest to largest. Check that the data without the suspect outlier are normally distributed. If normality fails, then apply a different outlier identification method such as a robust outlier identification procedure. *It is suggested to avoid the use of a transformation such as a log-transformation to achieve normality to be able to use the Dixon test.* All cleanup and remediation decisions are made based upon the data set in raw scale. Therefore, outliers (perhaps representing isolated contaminated locations) should be identified in the original scale. As mentioned before, the use of a log-transformation tends to hide and accommodate outliers (instead of identifying them).

STEP 2: $X_{(1)}$ is a potential outlier (Case 1): Compute the test statistic, C, where

$$C = \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}} \text{ for } 3 \leq n \leq 7, \quad C = \frac{X_{(3)} - X_{(1)}}{X_{(n-1)} - X_{(1)}} \text{ for } 11 \leq n \leq 13,$$

$$C = \frac{X_{(2)} - X_{(1)}}{X_{(n-1)} - X_{(1)}} \text{ for } 8 \leq n \leq 10, \quad C = \frac{X_{(3)} - X_{(1)}}{X_{(n-2)} - X_{(1)}} \text{ for } 14 \leq n \leq 25,$$

STEP 3: If C exceeds the critical value for the specified significance level α, then $X_{(1)}$ is an outlier and should be further investigated.

STEP 4: $X_{(n)}$ is a potential outlier (Case 2): Compute the test statistic, C, where

$$C = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}} \text{ for } 3 \leq n \leq 7, \quad C = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(2)}} \text{ for } 11 \leq n \leq 13,$$

$$C = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(2)}} \text{ for } 8 \leq n \leq 10, \quad C = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(3)}} \text{ for } 14 \leq n \leq 25,$$

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 *without* the suspected outliers follows a normal distribution as a data set with outliers tends not to follow a normal distribution. A graphical

display (Q-Q plot) can be used to identify suspected outliers needed to perform the Rosner test. Like the Dixon test, 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 ($r_0 \leq 10$), 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:

$$[\bar{x}^{(0)}, s^{(0)}, y^{(0)}]; [\bar{x}^{(1)}, s^{(1)}, y^{(1)}]; \ldots, [\bar{x}^{(r_0-1)}, s^{(r_0-1)}, y^{(r_0-1)}] \text{ where}$$

$$\bar{x}^{(i)} = \frac{1}{n-i}\sum_{j=1}^{n-i} x_j \,, \ s^{(i)} = \sqrt{\frac{1}{n-i}\sum_{j=1}^{n-i}(x_j - \bar{x}^{(i)})^2} \,, \text{ and } y^{(i)} \text{ is the farthest value } \bar{x}^{(i)}.$$

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: $R_r = \dfrac{|y^{(r-1)} - \bar{x}^{(r-1)}|}{s^{(r-1)}}$ 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 high outliers (coming from the right tail of data distribution) 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 detection limits (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, or replacing them by DL/2) to identify outliers are available in ProUCL for data sets consisting of 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 procedures as described in Chapters 4 through 6.

Example 7-1. Consider a real 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 lead data is shown in Figure 7-1 below. Figure 7-1 immediately suggests that the data set has some outliers. Dixon test cannot directly identify all outliers present in a data set (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.
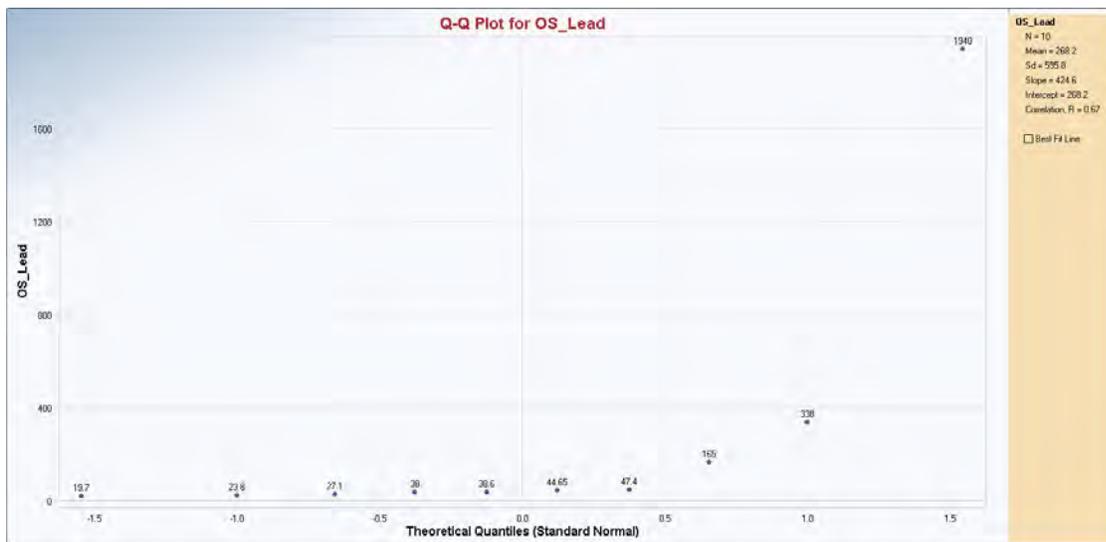
Figure 7-1. Normal Q-Q Plot Identifying Outliers

Table 7-1. 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 7-2, and the normal Q-Q plot without NDs is shown in Figure 7-2 below.
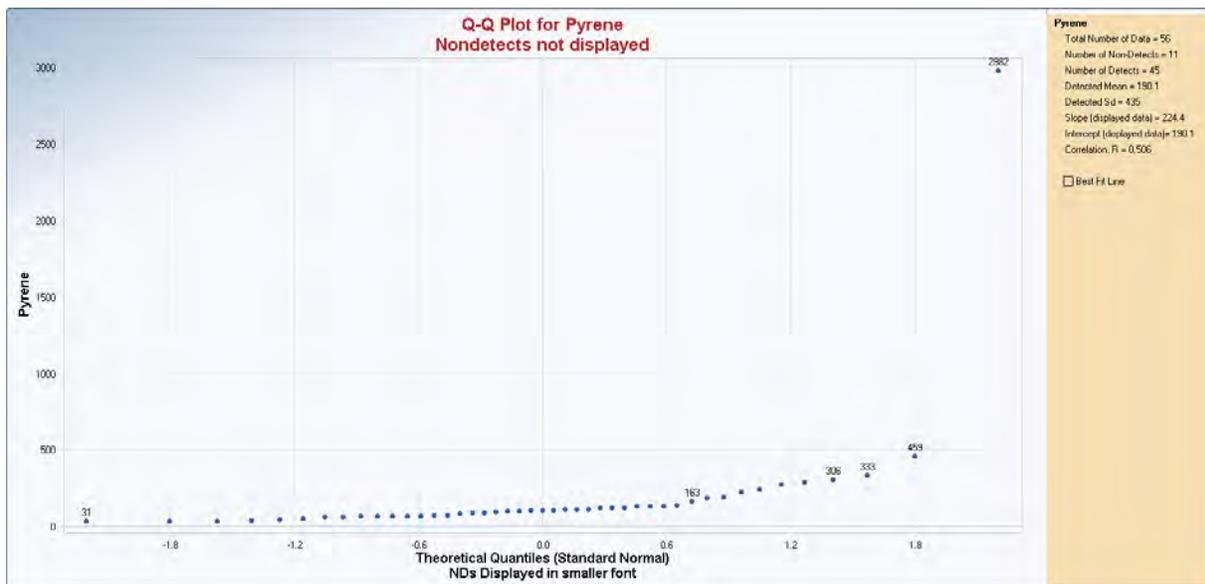
Figure 7-2. Normal Q-Q Plot of Pyrene Data Set Excluding NDs

193

Table 7-2. 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 7-3 below. Figure 7-3 suggests that there are 4 outliers (at least the observation=30,000) present in the data set. Rosner test results are shown in Table 7-3. Due to masking, the Rosner test could not even identify the outlying observation of 30,000.
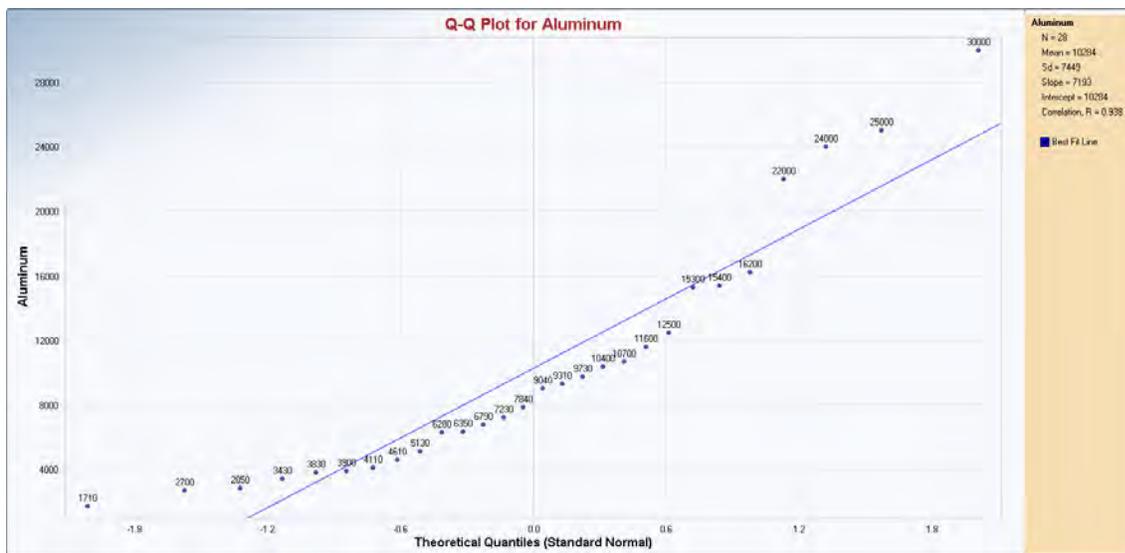
Figure 7-3. Normal Q-Q Plot of Aluminum Concentrations

Table 7-3. 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				
			Potential	Obs.	Test	Critical	Critical
#	Mean	sd	outlier	Number	value	value (5%)	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 potentially present in a data set. Several of those methods are incorporated in Scout 2008 (EPA 2009). 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 7-4 and 7-5. 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 30000. All observations exceeding the horizontal lines displayed at critical values of the Largest Mahalanobis Distance (MD; Wilks, 1963; Barnett and Lewis, 1994) represent outliers.
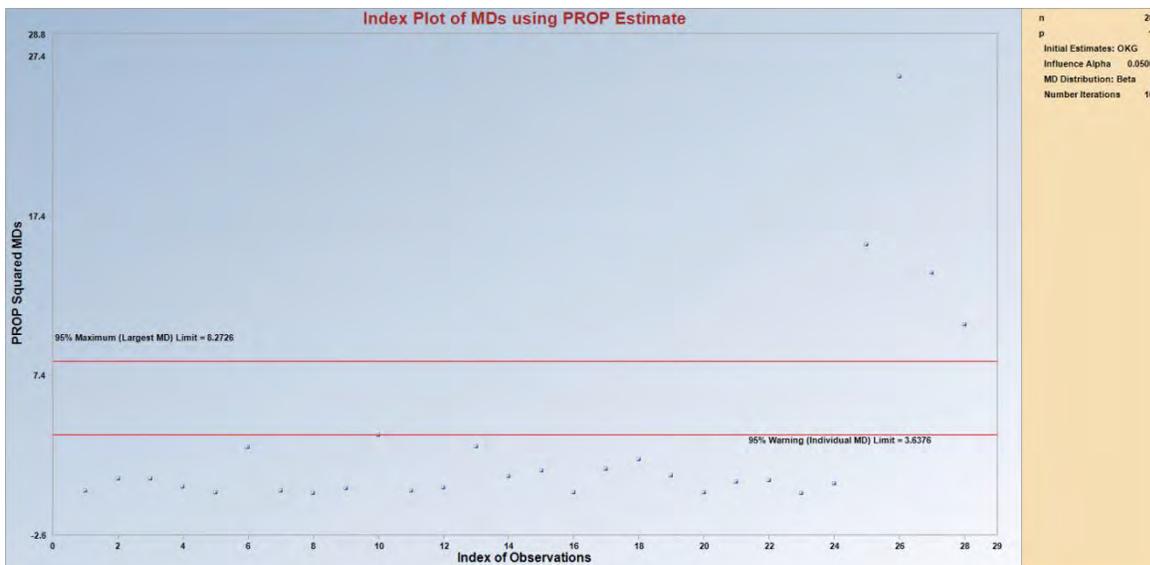
Figure 7-4. Robust Index Plot of MDs Based Upon the PROP Influence Function

195
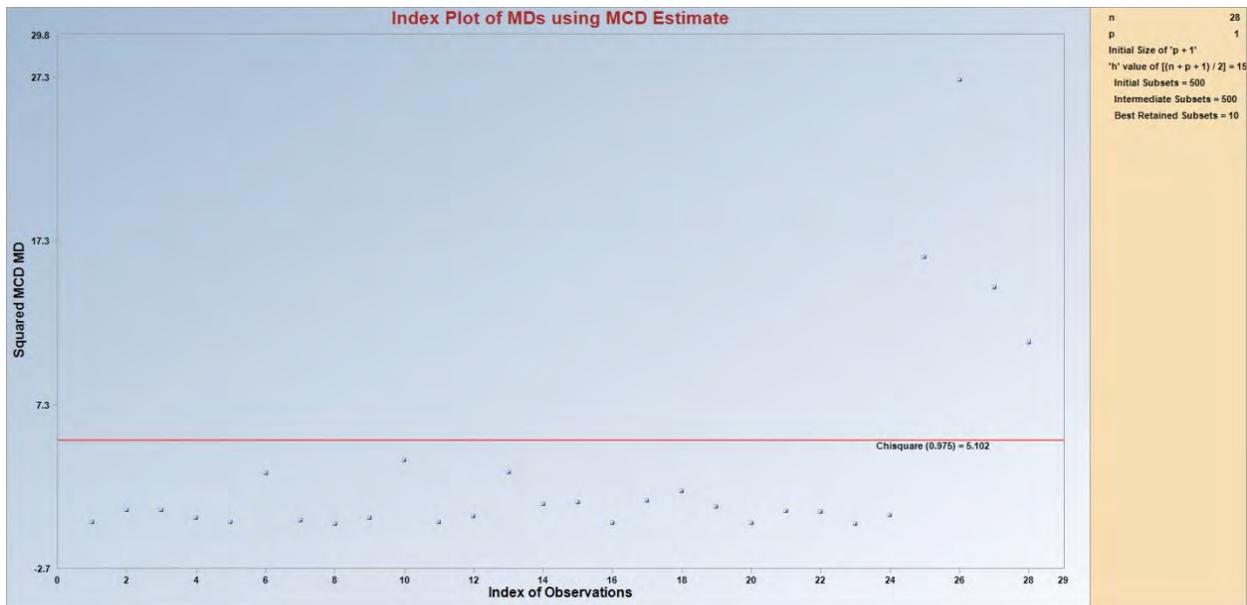
Figure 7-5. Robust Index Plot of MDs Based Upon the MCD Method

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 of the various 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: confidence coefficient, *(1-α)* and the allowable error margin, *Δ* in the estimate (e.g., maximum allowable difference between the estimate[e.g., sample mean] and the unknown mean, *μ*) from the unknown true population parameter (e.g., mean, *μ*). 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 Type I (false positive error rate, *α*) and Type II (false negative error rate, *β*=1-power) error rates; 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 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 error margin and width of the gray region. A brief description of these terminology is described 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 (e.g., 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 has the DQOs based parametric minimum sample size formula to estimate the population mean assuming that the sample mean follows a normal distribution (e.g., can also be assumed due to central limit theorem [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., area of concern [AOC] or monitoring well [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 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 any). 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: The Sample size module in ProUCL 5.0 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 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 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 available 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 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 user's 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 95% upper confidence limit of the population mean is used to estimate the exposure point concentration (EPC) term. Several variations listed below are available in the literature to compute the minimum sample size, n, needed to estimate the population mean with specified confidence coefficient (CC), $(1 - \alpha)$, 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),

$$n = \sigma^2 z^2_{1-(\alpha/2)} / \Delta^2,$$
(8-1)

or by using the following approximate standard normal distribution based equation (when population variance is not known),

$$n = s^2 z^2_{1-(\alpha/2)} / \Delta^2$$
(8-2)

or, alternatively, by using the t- distribution based equation (when population variance is not known):

$$n = s^2 t^2_{(n-1),(1-\alpha/2)} / \Delta^2$$
(8-3)

Here Δ represents the allowable error margin ($\pm$) in the mean estimate. The computed sample size assures that the sample mean will be within $\pm \Delta$ units of the true population mean with probability $(1-\alpha)$.

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 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 $\Delta = 10$, 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 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 only by 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, Wilcoxon-Rank-Sum [WRS] test). This adjustment is also used in the various sample size formulae described in the EPA guidance documents (e.g., 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.

Example 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 the distribution 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 formula to determine 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 (e.g., EPA 1989). 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 versus (vs.) H_A: site mean, $\mu > AL$ or a C_s

Gray Region: Range of the mean concentrations where the consequences of deciding that the site mean is less than the action level 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, Wilcoxon Signed Rank [WSR] test) considered later in this chapter.
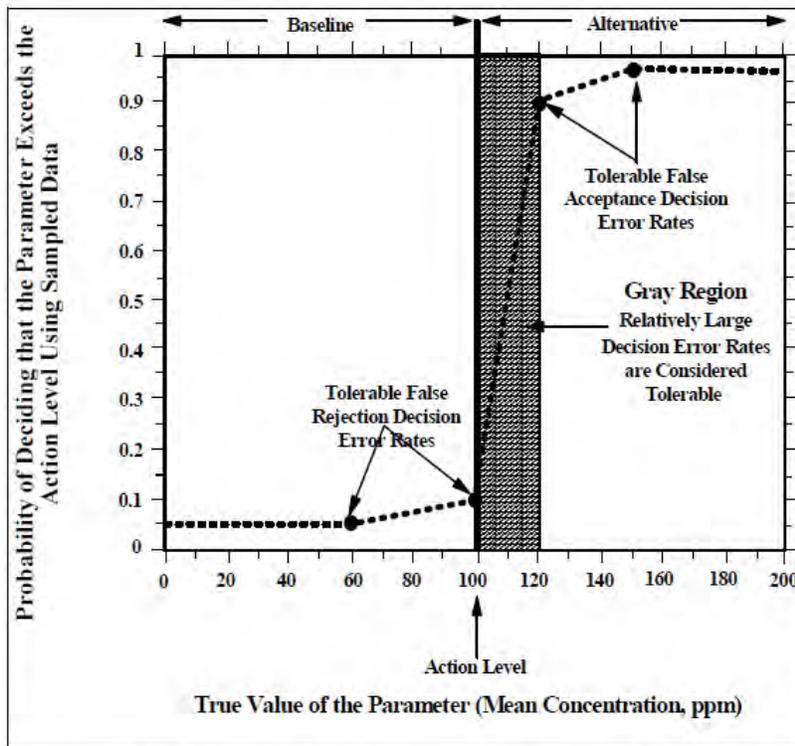
Diagram Where the Alternative Condition <u>Exceeds</u> the Action Level

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,* $\mu < AL$ *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.
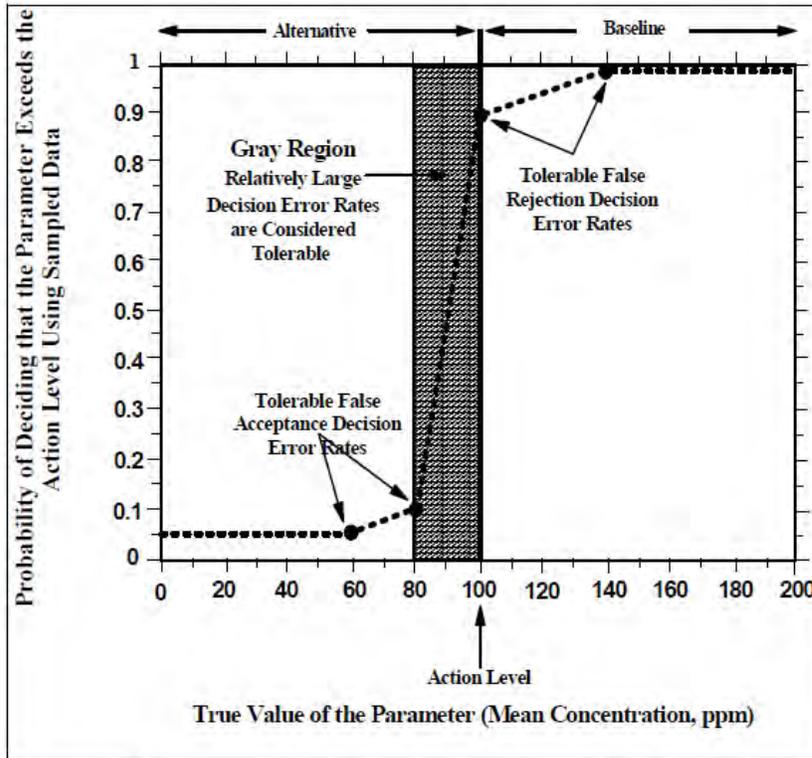
Diagram Where the Alternative Condition <u>Falls Below</u> the Action Level

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

$$n = \left(z_{1-\alpha} + z_{1-\beta}\right)^2 \left(\frac{s}{\Delta}\right)^2 + \frac{z_{1-\alpha}^2}{2} \tag{8-5}$$

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:

$$n = \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(\frac{s}{\Delta}\right)^2 + \frac{z_{1-\alpha/2}^2}{2} \tag{8-6}$$

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

204

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; and

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 5.0 sample size GUI draws user's 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 hypotheses, ProUCL computes the minimum sample size to be 9 (rounding up), and sample size of 11 is computed for a two-sided alternative hypothesis.

Example 8-2. Sample size for single-sample t-test sample sizes ($\alpha = 0.05$, $\beta = 0.2$, $s = 10.41$, $\Delta = 10$)

	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 $\leq$ 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 $\geq$ 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 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

$$n = \left(\frac{z_{1-\alpha}\sqrt{P_0(1-P_0)} + z_{1-\beta}\sqrt{P_1(1-P_1)}}{P_1 - P_0} \right)^2 \qquad (8\text{-}7)$$

8.2.2.2   Case III (Two-Sided Alternative Hypothesis)

H_0: population proportion $=$ specified value (P_0) vs. H_A: population proportion $\neq$ specified value (P_0)

The following procedure is used to determine the minimum sample size needed to conduct a two-sided proportion test.

$$a = \left(\frac{z_{1-\alpha/2}\sqrt{P_0(1-P_0)} + z_{1-\beta}\sqrt{P_1(1-P_1)}}{P_1 - P_0} \right)^2 \text{ for Right-Sided alternative;}$$

when $P_1 = P_0 + \Delta$; and

$$b = \left(\frac{z_{1-\alpha/2}\sqrt{P_0(1-P_0)} + z_{1-\beta}\sqrt{P_1(1-P_1)}}{P_1 - P_0} \right)^2 \text{ for Left-Sided alternative;}$$

when $P_1 = P_0 - \Delta$

P_0 = specified proportion
P_1 = outer bound of the gray region.

$$\Delta = \text{width of the gray region} = |P_0 - P_1| = \text{abs}(P_0 - P_1)$$

The sample size, n, for two-sided proportion test (Case III) is given by

$$n = \max(a, b) \tag{8-8}$$

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.

Example 8-3. Output for Single-Sample proportion test sample size ($\alpha = 0.05$, $\beta = 0.2$, $P_0 = 0.2$, $\Delta = 0.05$)

	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.

Example 8-4. Output for single-sample proportion test sample sizes ($\alpha = 0.05$, $\beta = 0.2$, $P_0 = 0.7$, $\Delta = 0.8$)

	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* $\leq$ *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* $\geq$ *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:

$$n = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{4\left(Sign\ P - 0.5\right)^2}\ , \text{where} \tag{8-9}$$

$$Sign\ P = \Phi\left(\frac{\Delta}{sd}\right) \tag{8-10}$$

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:

$$n = \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{4\left(Sign\ P - 0.5\right)^2}$$

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 5.0 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.

Example 8-5. Output for single-sample sign test sample sizes ($\alpha = 0.05$, $\beta = 0.1$, $sd = 3$, $\Delta = 2$)

	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 Action Level or Cleanup Standard 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 $\leq$ 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 $\geq$ 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.

$$n = 1.16 \left(\frac{sd^2 \left(z_{1-\alpha} + z_{1-\beta} \right)^2}{\Delta^2} + \frac{z_{1-\alpha}^2}{2} \right) \tag{8-11}$$

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 5.0 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 $\neq$ 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:

$$n = 1.16 \left(\frac{sd^2 \left(z_{1-\alpha/2} + z_{1-\beta} \right)^2}{\Delta^2} + \frac{z_{1-\alpha/2}^2}{2} \right) \tag{8-12}$$

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 5.0 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 one-sided alternative hypotheses, and 14 for a two-sided alternative hypothesis.

Example 8-6. Output for single-sample WSR test sample sizes ($\alpha = 0.1$, $\beta = 0.2$, $sd = 130$, $\Delta = 100$)

	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 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:

$$m = n = 2\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \left(\frac{S_p}{\Delta}\right)^2 + \frac{z_{1-\alpha}^2}{4} \tag{8-13}$$

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:

$$m = n = 2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(\frac{S_p}{\Delta}\right)^2 + \frac{z_{1-\alpha/2}^2}{4} \tag{8-14}$$

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 single sided alternative hypotheses and 7 for two-sided alternative hypothesis.

212

Note: S_p represents the pooled estimate of the populations under comparison. During the planning stage, one will input a preliminary estimate of variance while computing the minimum sample sizes; and while assessing the power associated with the t-test, one will input 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 two sample t-test output sheet; ProUCL 5.0 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.

Example 8-7. Output for two-sample t-test sample sizes ($\alpha = 0.05$, $\beta = 0.2$, $s_p = 1.467$, $\Delta = 2.5$)

	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 $\leq$ 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 $\geq$ 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

$$m = n = 1.16 \left(2\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \left(\frac{sd}{\Delta}\right)^2 + \frac{z_{1-\alpha}^2}{4} \right) \qquad (8\text{-}15)$$

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 planning stage; and

sd^2 = pooled standard deviation 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 5.0 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 $\neq$ 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:

$$m = n = 1.16 \left(2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \left(\frac{sd}{\Delta}\right)^2 + \frac{z_{1-\alpha/2}^2}{4} \right) \qquad (8\text{-}16)$$

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 planning stage; and

sd^2 = pooled standard deviation computed using the collected data to be used when assessing the power of the test in retrospect.

Note: ProUCL 5.0 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 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.

Example 8-8. 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	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

8.4   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 (e.g., 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 associated UTL will not exceed (e.g., '0' or no exceedance rule) 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:

$$n = \frac{m-1}{2} + \left(\frac{(1+p)}{4(1-p)} \right) \chi_\alpha^2 (2m) \tag{8-17}$$

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
$\alpha = 1 -$ confidence coefficient, and
$\chi^2_{\alpha,2m}$ = the cumulative percentage point of a chi-square distribution with *2m*   degrees of freedom; 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 $x_1 < x_2 < ... < x_n$ be an ordered sample size *n* from this distribution.

For a pre-assigned proportion, *p* and confidence coefficient $1 - \alpha$, let the following probability statement given by equation (8-10) be true.

$$P\left\{ \int_{x_r}^{x_{n+1-s}} f(x)dx > p \right\} = 1 - \alpha \tag{8-18}$$

The statement given by (8-18) implies that the interval $\left(x_r, x_{n+1-s} \right)$ contains at least a proportion, *p*, of the distribution with the probability, $1 - \alpha$. The interval, $\left(x_r, x_{n+1-s} \right)$, whose endpoints are the rth smallest and s[th] largest observations in a sample size of *n*, is a nonparametric *100p%* tolerance interval with a confidence coefficient of $1 - \alpha$, 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:

$$g(z) = \frac{1}{B(n-m+1,m)} z^{n-m} (1-z)^{m-1}, \quad 0 < z < 1$$
$$= 0, \qquad\qquad\qquad\qquad \text{otherwise} \tag{8-19}$$

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:

$$P(z \geq p) = \sum_{t=0}^{n-m} \binom{n}{t} p^t (1-p)^{n-t} \tag{8-20}$$

For given values of *m*, *p* and α, the minimum sample size, *n*, for acceptance sampling is obtained by solving the inequality:

$$P(z \geq p) \geq 1 - \alpha \tag{8-21}$$

216

$$\sum_{t=0}^{n-m} \binom{n}{t} p^t \left(1-p\right)^{n-t} \geq 1-\alpha \tag{8-22}$$

Where:

m = number of non-conforming items (always greater than 1)
$p = 1 - proportion$
$proportion$ = pre-specified proportion of non-conforming items; and
$\alpha = 1 - confidence\ coefficient$.

An example output generated by ProUCL is given as follows.

Example 8-9. 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 5.0 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. It should be noted that classical Oneway ANOVA represents a generalization of the two-sample t-test (Chapter 6). ProUCL has goodness-of-fit (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 Wilcoxon-Mann-Whitney (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 ith group, $i = 1, 2, ..., g$
n_i: number of observations in the ith group
j: an index used for the jth observation in a group; for the ith, $j = 1, 2, ..., n_i$
x_{ij}: the jth observation of the response variable in the ith group
n: total number of observations $= n_1 + n_2 + ... + n_g$

$\sum_{j=1}^{n_i} x_{j,i}$ = sum of all observations in the ith group

$\overline{x}_i$ = mean of the observations collected from the ith group

$\overline{x}$ = mean of all, n_t (the observations)

μ_i = true (unknown) mean of the ith 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 ith group, and errors, $e_{i,j}$ are assumed to be independently and normally distributed with mean = 0 and 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:

$$SS_{Between\ Groups} = \sum_{i=1}^{g} n_i \left(\bar{x}_i - \bar{x} \right)^2 \tag{9-2}$$

- 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:

$$SS_{Total} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} \left(x_{j,i} - \bar{x} \right)^2 \tag{9-4}$$

- Between Groups Degrees of Freedom (*df*): *g-1*

- Within Groups *df*: *n-g*

- Total *df*: *n-1*

- Mean Squares Between Groups is given by:

$$MS_{Between\ Groups} = \frac{SS_{Between\ Groups}}{g-1}$$

(9-5)

- Mean Squares Within Groups:

$$MS_{Within\ Groups} = \frac{SS_{Within\ Groups}}{n-g}$$

(9-6)

- Scale estimate is given by:

$$S = \sqrt{MS_{Within\ Groups}}$$

(9-7)

- R^2 is given by:

$$R^2 = 1 - \frac{SS_{Within\ Groups}}{SS_{Total}}$$

(9-8)

- Decision statistic, F, is given by:

$$F\ \boldsymbol{Statistic} = \frac{MS_{Between\ Groups}}{MS_{Within\ Groups}}$$

(9-9)

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:

$$H_0 : m_1 = m_2 = \cdots = m_i = \cdots = m_g$$
$$H_A : At\ least\ one\ of\ the\ \textbf{g}\ medians\ is\ not\ equal\ to\ others$$

(9-10)

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.

<u>K-W Test on Data Sets with NDs:</u> 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 nondetects (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 ith Group, $\bar{R}_i$: Average of the ranks (in the combined data set of size, n) of the n_i observations in the ith 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):

$$Z_i = \frac{\bar{R}_i - \bar{R}}{\sqrt{\frac{\left((n+1)\left(\frac{n}{n_i} - 1 \right) \right)}{12}}}$$

(9-11)

n = total number of observations = $n_1 + n_2 + \ldots + n_g$
n_i = observation in the ith 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 ith group and the overall average rank, R, of the combined data set of sized n.

- Kruskal-Wallis H-Statistic (without ties) is given by:

$$H = \frac{12 \sum_{i=1}^{g} n_i \left(\bar{R}_i - \bar{R} \right)^2}{n(n+1)}$$

(9-12)

- K-W H-Statistic adjusted for ties is given by:

$$H_{adj-ties} = \frac{H}{1 - \left[\dfrac{\sum\limits_{i=1}^{g} t_i^3 - t_i}{n^3 - n} \right]}$$ (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.

Example 9-1. Consider Fisher's famous Iris data set (Fisher, 1936) with 3 iris species. The classical Oneway ANOVA results comparing petal widths of 3 iris species are summarized as follows.

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.

Example 9-2 (Iris Data). The K-W Oneway ANOVA results comparing petal widths of 3 iris species are summarized as follows.

		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

The 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 5.0. Methods to perform trend analysis and OLS Regression with graphical displays are available under the Statistical Tests module of ProUCL 5.0. 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 certain period of time. Specifically, the OLS regression with time or a simple index variable as the predictor variable can be used to determine 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:

$$Y = b_0 + b_1 t + e;$$
$$E[Y] = b_0 + b_1 t = \textit{mean response at t}$$

(10-1)

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, $N(0, \sigma 2)$, with mean 0 and unknown variance, σ^2. Let $(t_i, y_i); i: = 1, 2,....n$ 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:

$$y_i = b_0 + b_1 t_i + e_i;$$
$$E[y_i] = b_0 + b_1 t_i = \textit{mean response when } t = t_i$$

(10-2)

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

exactly lie 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 monitoring well [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, $i:=1,2,...,n$

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:

$$S_{ty} = \sum_{i=1}^{n} t_i y_i - (\sum_{i=1}^{n} t_i \sum_{i=1}^{n} y_i)/n; and$$

$$S_{tt} = \sum_{i=1}^{n} t_i^2 - (\sum_{i=1}^{n} t_i)^2/n; and \ S_{yy} = \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n$$

(10-3)

The OLS estimates of slope and intercept are given as follows:

$$\hat{b}_1 = S_{ty}/S_{tt}; \ and$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{t}$$

$$\bar{t} = \sum_{i=1}^{n} t_i/n$$

(10-4)

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 $\hat{e}_i = y_i - \hat{y}_i$; $i = 1, 2,....,n$. 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 $i:=1,2,...,n$.

The residual sum of squares is given by:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

(10-5)

Estimate of the error variance, σ^2, and variances of the OLS estimates, $\hat{b}_0$ *and* $\hat{b}_1$ are given as follows:

$$\hat{\sigma}^2 = MSE = SSE/(n-2)$$

$$Var(\hat{b}_0) = \sigma^2(\frac{1}{n} + \frac{\overline{t}^2}{S_{tt}}) \tag{10-6}$$

$$Var(\hat{b}_1) = \sigma^2/S_{tt}$$

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:

$$t = \hat{b}_1 / SE(\hat{b}_1) \tag{10-7}$$

Under normality of the responses, y_i (and the random errors, e_i), 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: Mann-Kendall and Theil-Sen 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:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad\qquad (10\text{-}8)$$

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:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = SSR + SSE \qquad\qquad (10\text{-}9)$$

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)_.

$$F = \frac{MSR}{MSE} \qquad\qquad (10\text{-}10)$$

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:

$$R^2 = 1 - \frac{SSE}{SST}$$ (10-11)

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 also 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:

$$\hat{y}_i \pm t_{((1-\alpha/2),(n-2))} sd[\hat{y}_i]$$ (10-12)

Where the estimated standard deviation, $sd(\hat{y}_i)$ of the mean response, $\hat{y}_i$ is given by:

$$sd[\hat{y}_i] = \sqrt{MSE(\frac{1}{n} + \frac{(t_i - \bar{t})^2}{S_{tt}})}; i = 1, 2, ..., n$$

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 lower and upper prediction limits (LPL and UPL) making the prediction band are given by:

$$\hat{y}_0 \pm t_{((1-\alpha/2),(n-2))} sd(\hat{y}_0); with \ \hat{y}_0 = \hat{b}_0 + \hat{b}_1 x_0$$ (10-13)

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 2009). 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

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, 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: $t_1 < t_2 < t_3, ... < t_n$.

The Trend Analysis module of ProUCL has two trend tests: Mann-Kendall (M-K) test and Theil-Sen (T-S) test. The trend tests in ProUCL are performed on time series data sets. Both M-K and T-S tests 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,...) 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. These trend tests in ProUCL software also 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.

10.2.1 Mann–Kendall Test

The Mann-Kendall (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.

It is noted that 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; or
- H_A: Data set exhibits a downward trend; or
- H_A: Data set exhibits a trend (two-sided alternative).

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 Central Limit Theorem (CLT). The exact p-values for the M-K test are available for sample sizes up to 22 which 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:

$$sd(S) = \sqrt{\frac{1}{18}\left[n(n-1)(2n+5) - \sum_{j=1}^{g} t_j(t_j-1)(2t_j+5) \right]}$$

(10-14)

Where n is the sample size, g represents the number of groups of ties (if any) in the data set, and tj is the number of tied observations in the jth group of ties. If no ties or nondetects (NDs) are present, the equation reduces to the simpler form:

$$sd(S) = \sqrt{\frac{1}{18}\left[n(n-1)(2n+5) \right]}$$

(10-15)

The standardized S statistic denoted by Z for an increasing (or decreasing) trend is given as follows:

$$Z = \frac{(S-1)}{sd(S)} \quad if\ S > 0;$$
$$Z = 0 \quad if\ S = 0; and$$
$$Z = \frac{(S+1)}{sd(S)} \quad if\ S < 0$$

(10-16)

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.

232

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., Theil-Sen 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, ..., 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:

$$\text{sgn}\left(y_j - y_i\right) = \begin{cases} 1 & if\left(y_j - y_i\right) > 0 \\ 0 & if\left(y_j - y_i\right) = 0 \\ -1 & if\left(y_j - y_i\right) < 0 \end{cases}$$

(10-17)

Step 2. Compute the M-K test statistic, *S*, given by the following equation:

$$S = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{sgn}\left(y_j - y_i\right)$$

(10-18)

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 $\alpha \le \alpha_{cp}$, 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:
H₀: no trend, vs. *Hₐ: upward trend*
or
H₀: no trend vs. *Hₐ: 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 above, Mann-Kendall 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 real 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. The M-K trend statistics are summarized as follows.

234

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.
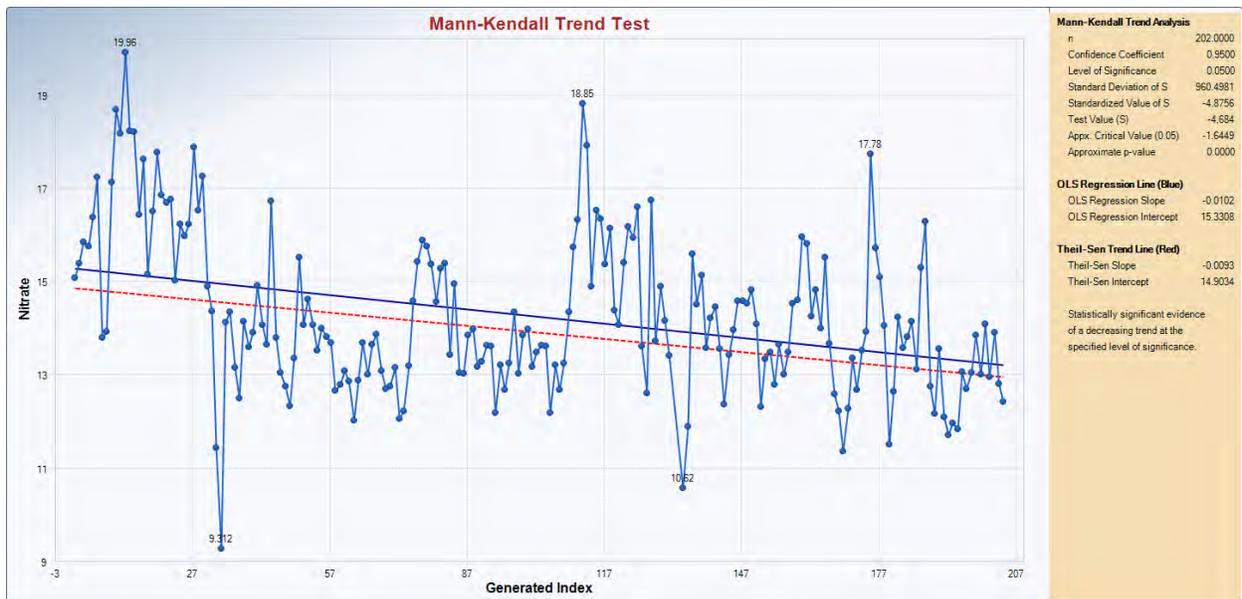
Figure 10-2. Trend graph with M-K Test Results and OLS Line and Nonparametric Theil-Sen Line

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 (*i.e.*, 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.

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 Theil-Sen 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 Theil-Sen test statistic. In such cases, the user may want to pre-process the data before using the Theil-Sen 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 Theil-Sen test in ProUCL 5.0 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.

236

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 $y_1, y_2, ..., y_n$ 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 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 detection limit (DL) or the reporting limit (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 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 $m(1), m(2), ..., m(N)$. 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:

$$Q = \begin{cases} m_{([N+1]/2)} & if \ N = odd \\ \left(m_{([N/2])} + m_{([N+2]/2)} \right) \Big/ 2 & if \ N = even \end{cases} \tag{10-19}$$

Step 3. Arrange the n measurements in ascending order from smallest to the largest value: $y(1), y(2), ..., y(n)$. Determine the median measurement using the following algorithm:

$$\tilde{y} = \begin{cases} y_{(n+1/2)} & if \ n = odd \\ \left(y_{(n/2)} + y_{((n+2)/2)} \right) \Big/ 2 & if \ n = even \end{cases} \tag{10-20}$$

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: $m(1), m(2),\ldots, m(N)$. Compute S given in (10-18) and its *sd* given below:

$$sd(S) = \sqrt{\frac{1}{18}\left[n(n-1)(2n+5) - \sum_{j=1}^{g} t_j (t_j - 1)(2t_j + 5)\right]} \qquad (10\text{-}21)$$

ProUCL can be used to test the following hypotheses:

H_0: Data set does not exhibit sufficient evidence of any trends (stationary measurements) vs.

    I.   *H_A: Data set exhibits a trend (two-sided alternative)*
    II. *H_A: Data set exhibits an upward trend*; or
    III. *H_A: Data set exhibits a downward trend*.

<u>Case I.</u> 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:

$$M_1 = \left[\frac{N - C_\alpha}{2}\right]; \qquad \text{and} \qquad M_2 = \left[\frac{N + C_\alpha}{2}\right]$$

- Obtain the M_1^{th} largest and M_2^{th} largest slopes, $\left(m_{(M_1)}\right)$ and $\left(m_{(M_2)}\right)$, 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:

$$P\left(m_{(M_1)} < Q < m_{(M_2)}\right) = 1 - \alpha$$

On ProUCL output, $\left(m_{(M_1)}\right)$ is labeled as LCL and $\left(m_{(M_2)}\right)$ 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.

<u>Cases II and III:</u> 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:

$$C_\alpha = Z_\alpha * sd(S)$$

$$M_1 = \left[\frac{N - C_\alpha}{2}\right] \quad \text{and} \quad M_2 = \left[\frac{N + C_\alpha}{2}\right]$$

- Obtain the M_1^{th} largest and M_2^{th} largest slopes, $\left(m_{(M_1)}\right)$ and $\left(m_{(M_2)}\right)$ from the set consisting of all $n(n-1)/2$ slopes.
- Conclusion:

If $\left(m_{(M_1)}\right) > 0$, then the data set exhibits a significant upward trend.

If $\left(m_{(M_2)}\right) < 0$, then the data set exhibits a significant downward trend.

Example 10-3. Time series data (time event, concentration) were collected from several groundwater monitoring wells (MW) 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.
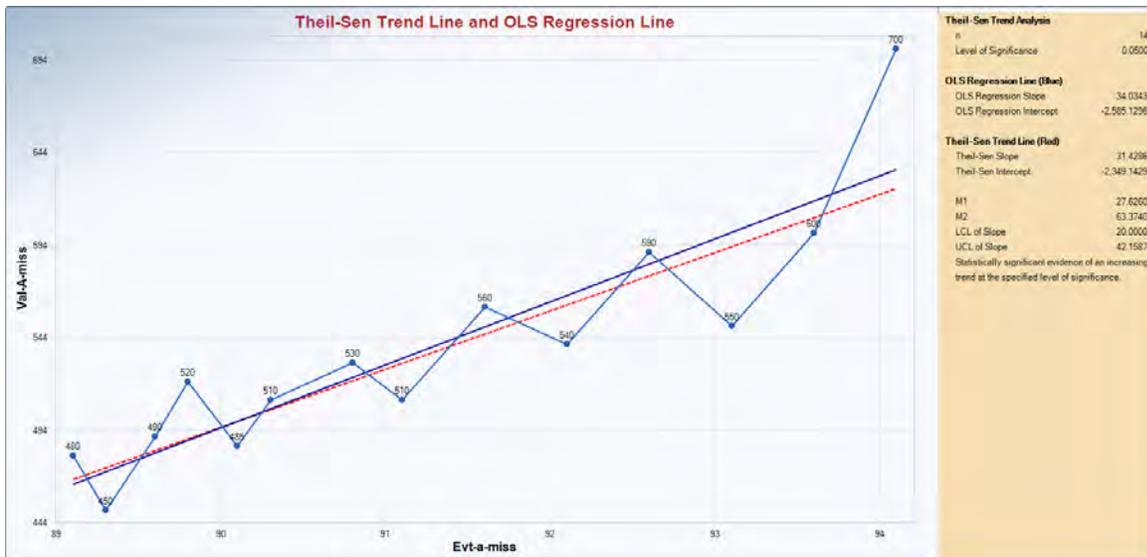
Figure 10-3. Time Series Plot and OLS and Theil-Sen Results with Missing Values

The Excel output sheet, generated by ProUCL and showing all relevant results, is shown as follows:

Val-A-miss			Approximate inference for Theil-Sen Trend Test	
			Mann-Kendall Statistic (S)	72
General Statistics			Standard Deviation of S	18.24
			Standardized Value of S	3.893
Number of Events	14		Approximate p-value	4.9562E-5
Number Values Observations	16		Number of Slopes	91
Number Values Missing	2		Theil-Sen Slope	31.43
Number Values Reported (n)	14		Theil-Sen Intercept	-2349
Minimum	450		M1'	30.5
Maximum	700		One-sided 95% lower limit of Slope	21.36
Mean	536.8		95% LCL of Slope (0.025)	20
Geometric Mean	533.6		95% UCL of Slope (0.975)	42.16
Median	525			
Standard Deviation	62.99		**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). 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 too influenced by such irregularities in the data collection and sampling events. In such circumstances, the user should make a conclusion based upon 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 to compare 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.
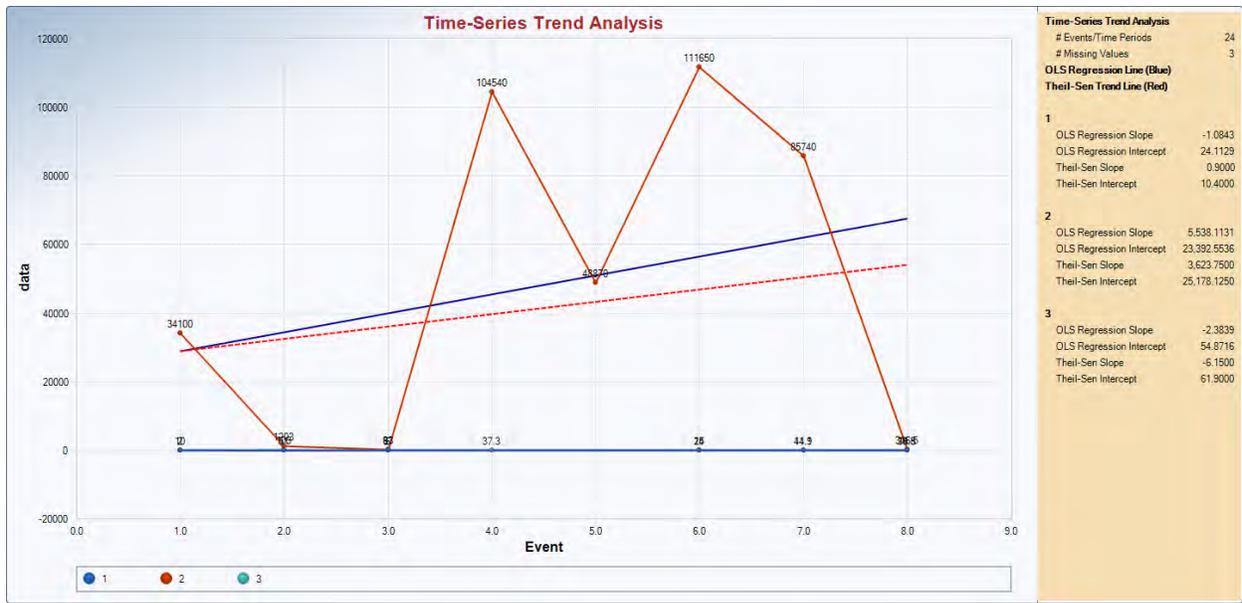
Figure 10-4. Time Series Plot Comparing concentrations of Multiple Wells over a Period of Time

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 be 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 year, ...). 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,..) used in generating 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.

241

CHAPTER 11

Background Incremental Sample Simulator (BISS)
Simulating BISS Data from a Large Discrete Background Data

The Background Incremental Sample Simulator (BISS) module has been incorporated in ProUCL5.0 at the request of the Office of Superfund Remediation and Technology Innovation (OSRTI). However, this module is currently under further investigation and research, and therefore it is not available for general public use. This module may be released in a future version of the ProUCL software, along with strict conditions and guidance for how it is applied. The main text for this chapter is not included in this document for release to general public. Only a brief placeholder write-up is provided here.

The following scenario describes the Site or project conditions under which the BISS module could be useful: Suppose there is a long history of soil sample collection at a Site. In addition to having a large amount of Site data, a robust background data set (at least 30 samples from verified background locations) has also been collected. Comparison of background data to on-Site data has been, and will continue to be, an important part of this project's decision-making strategy. All historical data is from discrete samples, including the background data. There is now a desire to switch to incremental sampling for the Site. However, guidance for incremental sampling makes it clear that it is inappropriate to compare discrete sample results to incremental sample results. That includes comparing a Site's incremental results directly to discrete background results.

One option is to recollect all background data in the form of incremental samples from background decision units (DUs) that are designed to match Site DUs in geology, area, depth, target soil particle size, number of increments, increment sample support, etc. If project decision-making uses a background threshold value (BTV) strategy to compare Site DU results one at a time against background, then an appropriate number (the default is no less than 10) of background DU incremental samples would need to be collected to determine the BTV for the population of background DUs. However, if the existing discrete background data show background concentrations to be low (in comparison to Site concentrations) and fairly consistent (relative standard deviation, RSD <1), there is a second option described as follows.

When a robust discrete background data set that meets the above conditions already exists, the following is an alternative to automatically recollecting ALL background data as incremental samples.

Step 1. Identify 3 background DUs and collect at least 1 incremental sample from each for a minimum of 3 background incremental samples.

Step 2. Enter the discrete background data set (n ≥ 30) and the ≥3 background incremental samples into the BISS module (the BISS module will not run unless both data sets are entered).

- The BISS module will generate a specified (default is 7) simulated incremental samples from the discrete data set.

- The module will then run a t-test to compare the simulated background incremental data set (e.g., with n = 7) to the actual background incremental data set (n ≥ 3).

o   If the t-test finds no difference between the 2 data sets, the BISS module will combine the 2 data sets and determine the statistical distribution, mean, standard deviation, potential UCLs and potential BTVs for the combined data set. Only this information will be supplied to the general user. The individual values of the simulated incremental samples will not be provided.

o   If the t-test finds a difference between the actual and simulated data sets, the BISS module will not combine the data sets nor provide a BTV.

o   In both cases, the BISS module will report summary statistics for the actual and simulated data sets.

Step 3. If the BISS module reported out statistical analyses from the combined data set, select the BTV to use with Site DU incremental sample results. Document the procedure used to generate the BTV in project reports. If the BISS module reported that the simulated and actual data sets were different, the historical discrete data set cannot be used to simulate incremental results. Additional background DU incremental samples will need to be collected to obtain a background DU incremental data set with the number of results appropriate for the intended use of the background data set.

The objective of the BISS module is to take advantage of the information provided by the existing background discrete samples. The availability of a large discrete data set collected from the background areas with geological formations and conditions comparable to the Site DU(s) of interest is a requirement for successful application of this module. There are fundamental differences between incremental and discrete samples. For example, the sample supports of discrete and incremental samples are very different. Sample support has a profound effect on sample results so samples with different sample supports should not be compared directly, or compared with great caution.

Since incremental sampling is a relatively new approach, the performance of the BISS module requires further investigation. If you would like to try this strategy for your project, or if you have questions, contact Deana Crumbling, crumbling.deana@epa.gov.

244

Appendix A

Simulated Critical Values for Gamma GOF Tests, the Anderson-Darling Test and the Kolmogorov-Smirnov Test, and Summary Tables for Recommendations of UCL95

Updated Critical Values of Gamma GOF Test Statistics (New in ProUCL 5.0): 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 (e.g., $\leq$ *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'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). 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 ≤ 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* (≤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 (e.g., D'Agostino and Stephens [1986], Schneider and Clickner [1976] and 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*.

246

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

DECISION SUMMARY TABLES

Table A-7. 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
$0.5 \leq \sigma < 1.0$	Mild skewness to moderate skewness
$1.0 \leq \sigma < 1.5$	Moderate skewness to high skewness
$1.5 \leq \sigma < 2.0$	High skewness
$2.0 \leq \sigma < 3.0$	Very high skewness (moderate probability of outliers and/or multiple populations)
$\sigma \geq 3.0$	Extremely high skewness (high probability of outliers and/or multiple populations)

Table A-8. Summary Table for the Computation of a 95% UCL of the Unknown Mean, μ_1, of a Gamma Distribution

$\hat{k}$	Sample Size, n	Recommendation
$\hat{k} > 1.0$	n>=50	Approximate gamma 95% UCL
$\hat{k} > 1.0$	n<50	Adjusted gamma 95% UCL
$\hat{k} \leq 1.0$	$n < 15$	95% UCL based upon bootstrap-t or Hall's bootstrap method*
$\hat{k} \leq 1.0$	$n \geq 15$	Adjusted gamma 95% UCL if available, otherwise use approximate gamma 95% UC*L*

*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.

Table A-9. Summary Table for the Computation of a 95% *UCL* of the Unknown Mean, μ_1, of a Lognormal Population

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

*In the case that Hall's bootstrap or bootstrap-t methods yield an erratic unrealistically large UCL value, UCL of the mean may be computed based upon the Chebyshev inequality: Chebyshev *(Mean, Sd) UCL*

** For highly skewed data sets with $\hat{\sigma}$ exceeding 3.0, 3.5, it is suggested the user pre-processes the data. It is very likely that the data consist of outliers and/or come from multiple populations. The population partitioning methods may be used to identify mixture populations present in the data set.

Table A-10. Summary Table for the Computation of a 95% *UCL* of the Unknown Mean, μ_1, Based Upon a Skewed Data Set (with all Positive Values) without a Discernible Distribution, Where $\hat{\sigma}$ is the *sd* of Log-transformed Data

$\hat{\sigma}$	*Sample Size, n*	*Recommendation*
$\hat{\sigma} < 0.5$	For all *n*	Student's t, modified-t, or H-UCL Adjusted CLT UCL, BCA Bootstrap UCL
$0.5 \le \hat{\sigma} < 1.0$	For all *n*	95% Chebyshev *(Mean, Sd)* UCL
$1.0 \le \hat{\sigma} < 1.5$	For all *n*	95% Chebyshev *(Mean, Sd)* UCL
$1.5 \le \hat{\sigma} < 2.0$	*n* < 20	97.5% Chebyshev *(Mean, Sd)* UCL
	$20 \le n$	95% Chebyshev *(Mean, Sd)* UCL
$2.0 \le \hat{\sigma} < 2.5$	*n* < 15	Hall's bootstrap method
	$15 \le n < 20$	99% Chebyshev *(Mean, Sd)* UCL
	$20 \le n < 50$	97.5% Chebyshev *(Mean, Sd)* UCL
	$50 \le n$	95% Chebyshev *(Mean, Sd)* UCL
$2.5 \le \hat{\sigma} < 3.0$	*n* < 15	Hall's bootstrap method
	$15 \le n < 30$	99% Chebyshev (*Mean, Sd*)
	$30 \le n < 70$	97.5% Chebyshev *(Mean, Sd)* UCL
	$70 \le n$	95% Chebyshev *(Mean, Sd)* UCL
$3.0 \le \hat{\sigma} \le 3.5^{**}$	*n* < 15	Hall's bootstrap method*
	$15 \le n < 50$	99% Chebyshev(*Mean, Sd*) UCL
	$50 \le n < 100$	97.5% Chebyshev *(Mean, Sd)* UCL
	$100 \le n$	95% Chebyshev *(Mean, Sd)* UCL
$\hat{\sigma} > 3.5^{**}$	For all *n*	99% Chebyshev (*Mean, Sd*) UCL

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

** For highly skewed data sets with $\hat{\sigma}$ exceeding 3.0 to 3.5, it is suggested that the user pre-processes the data. Data sets with such high skewness are complex and it is very likely that the data consist of outliers and/or come from multiple populations. The population partitioning methods may be used to identify mixture populations present in the data set.

Notes: Suggestions regarding the selection of a 95% UCL are provided to help the user to select the most appropriate 95% UCL. These recommendations are based upon the results of the simulation studies summarized in Singh, Singh, and Iaci (2002) and Singh and Singh (2003). For additional insight, the user may want to consult a statistician.

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 correct 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 correct 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 nondetects (NDs) with varying detection limits 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 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 cannot 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 (e.g., Singh, Singh, and Engelhardt, 1997; Singh, Singh, and Iaci, 2002; Singh and Singh, 2003). 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 correct decisions at the various 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 based upon 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 (e.g., 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 central limit theorem (CLT) for positively skewed distributions (for details see Chapter 2 of this Technical Guide). Based upon statistical theory, it is noted that the CLT yields UCL results slightly smaller than the Student's t-UCL; and the adjusted CLT and 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. Taking these facts into consideration, only UCLs based upon the modified t-statistic were used in the simulation experiments summarized in Singh and Singh (2003). The simulation results summarized in Singh and Singh (2003) 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 (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.

To elaborate on the behavior and properties of the various UCL computation methods, some graphical displays comparing the performances of the various UCL95 based upon skewed data sets (from Singh and Singh, 2003) are presented in Figures B-1 through B-7. 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 (e.g., G(5,50), LN(5,0.05)); moderately skewed (e.g., G(2,50), LN(5,1)); and highly skewed (e.g., 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/exposure point concentration (EPC) term for most skewed data sets.
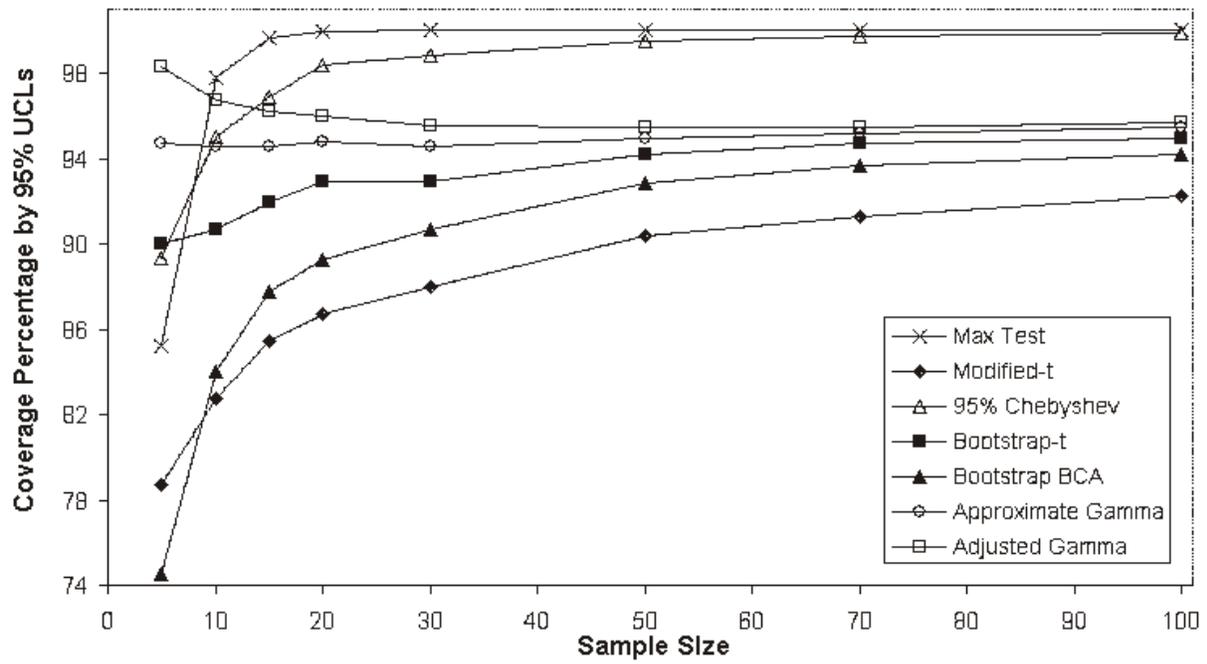
258

Figure B-1. Graphs of Coverage Probabilities by 95% UCLs of the mean of G (k=0.50, Θ=50)
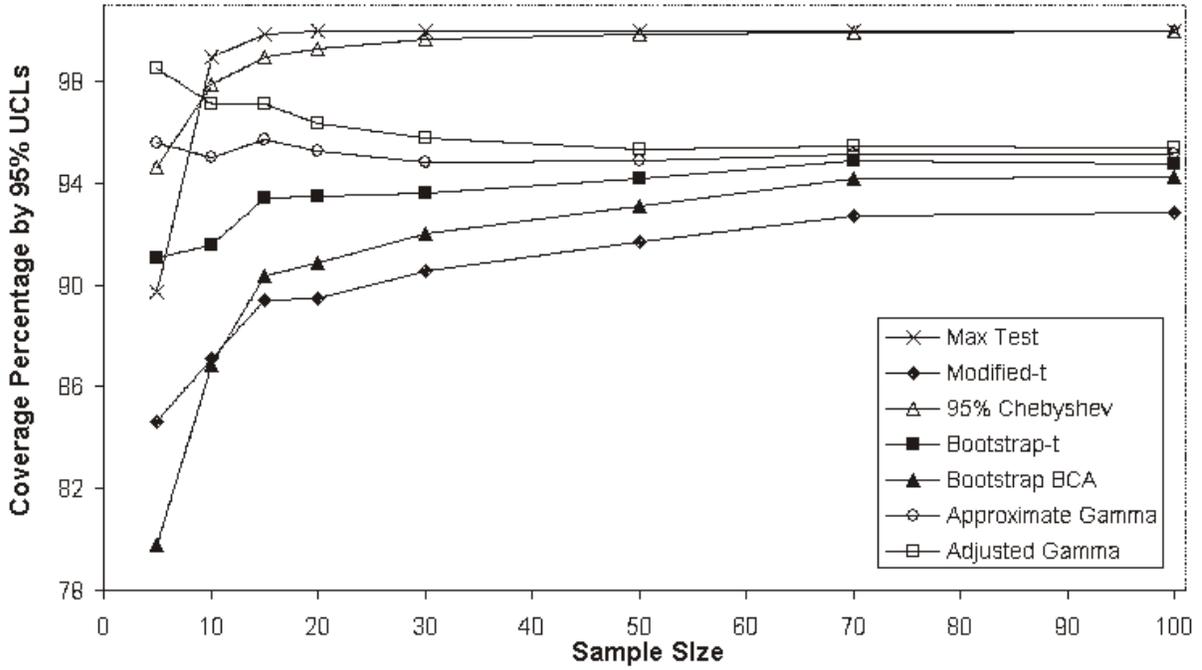
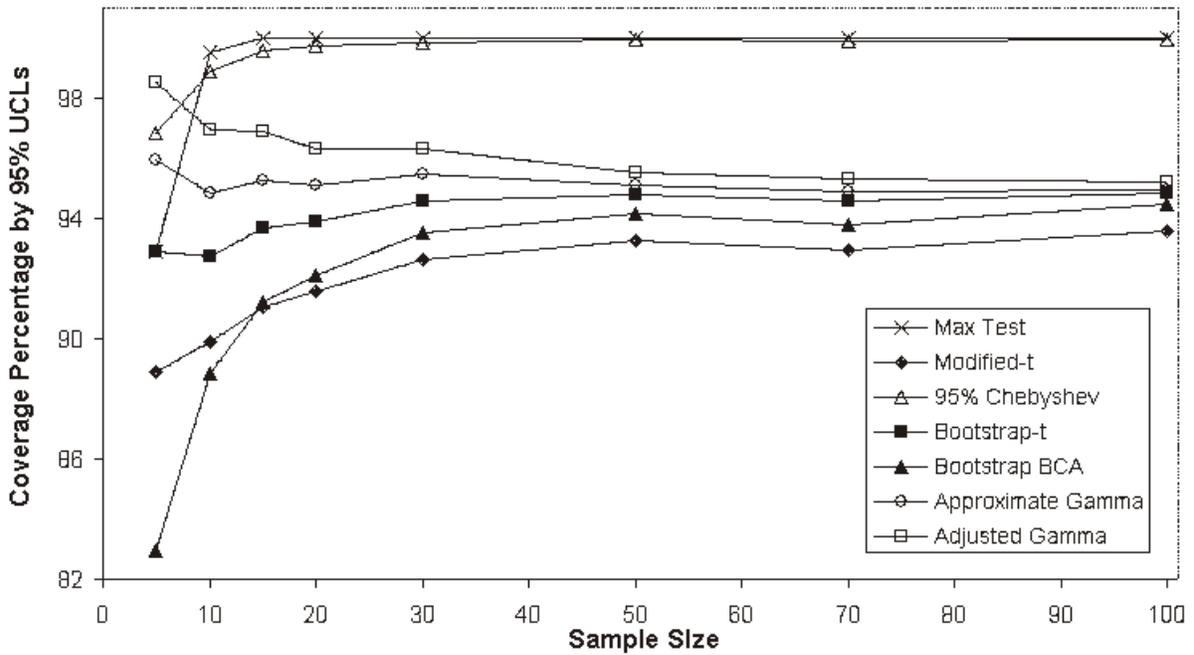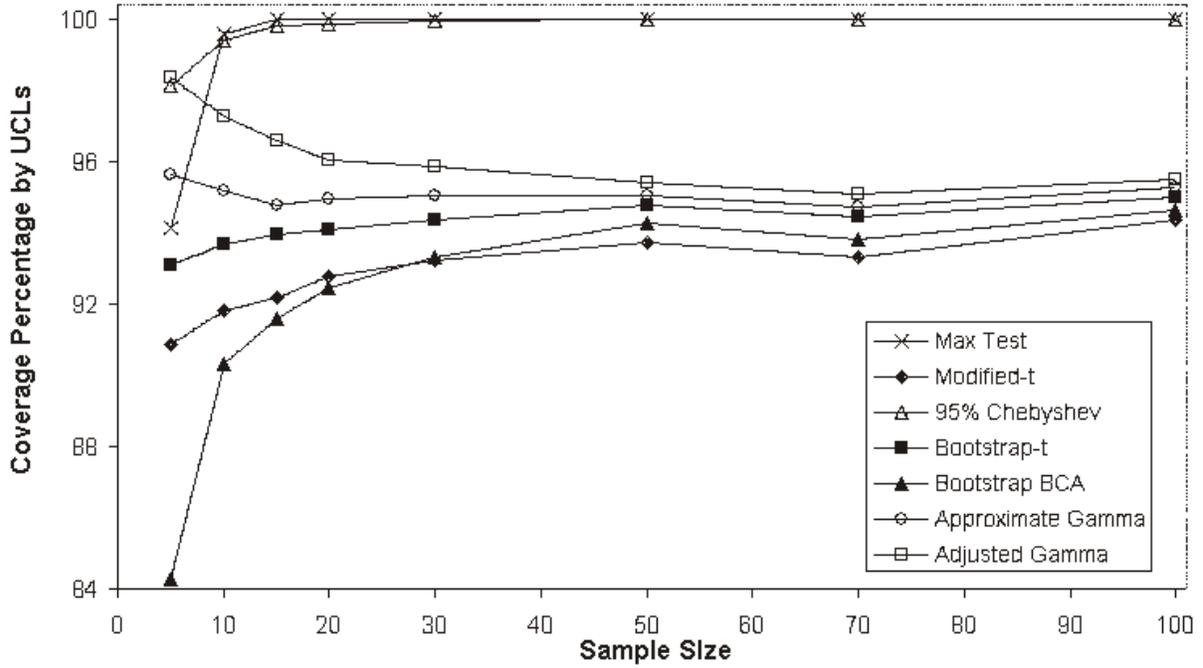Figure B-2. Graphs of Coverage Probabilities by 95% UCLs of Mean of G(k=1.00,Θ=50)

Figure B-3. Graphs of Coverage Probabilities by 95% UCLs of Mean of G$(k$=2.00, Θ=50)

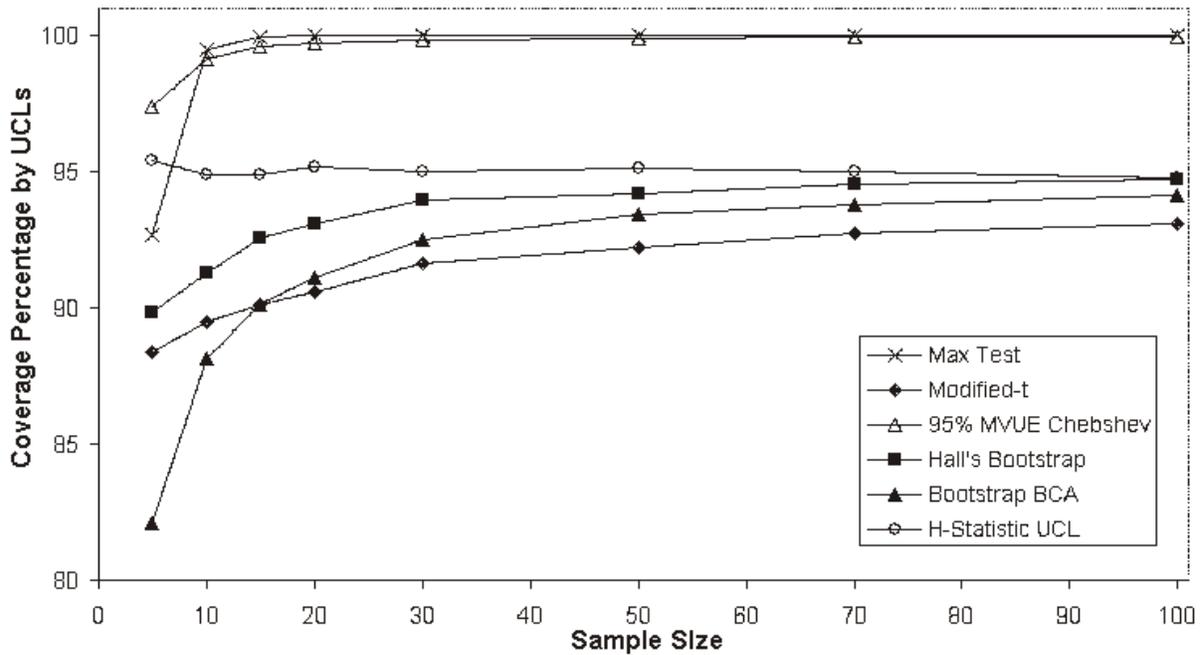Figure B4. Graphs of Coverage Probabilities by 95% UCLs of Mean of G(k=5.00,Θ=50)

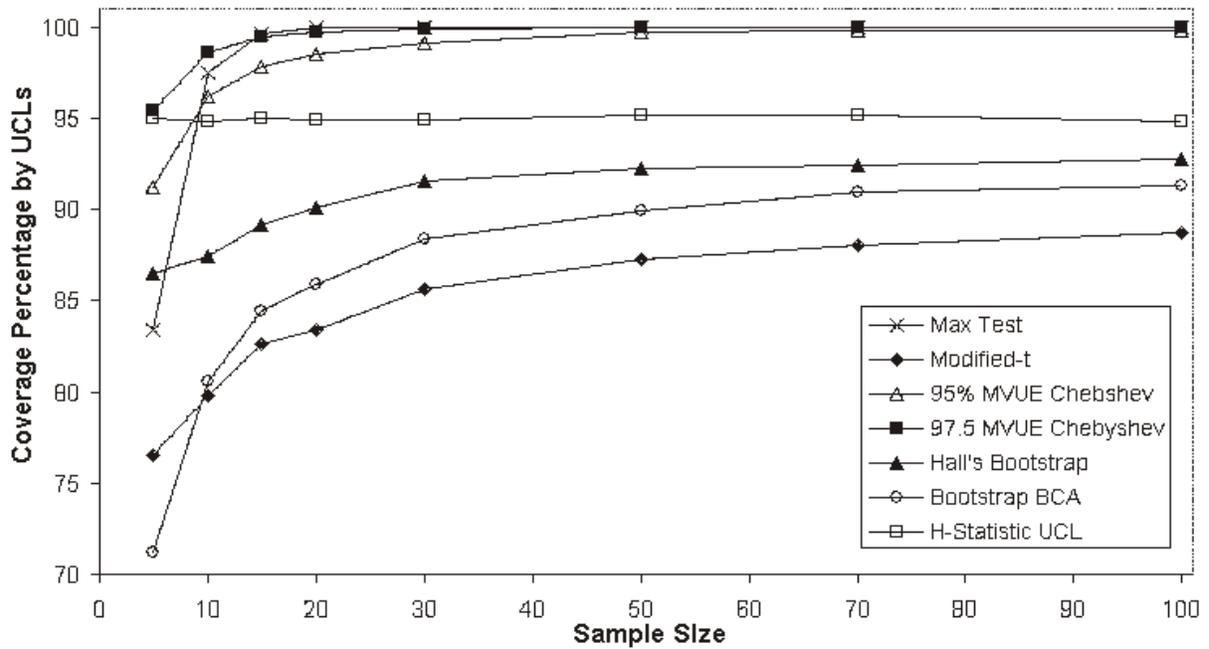Figure B-5. Graphs of Coverage Probabilities by UCLs of Mean of LN(μ=5,σ=0.5)

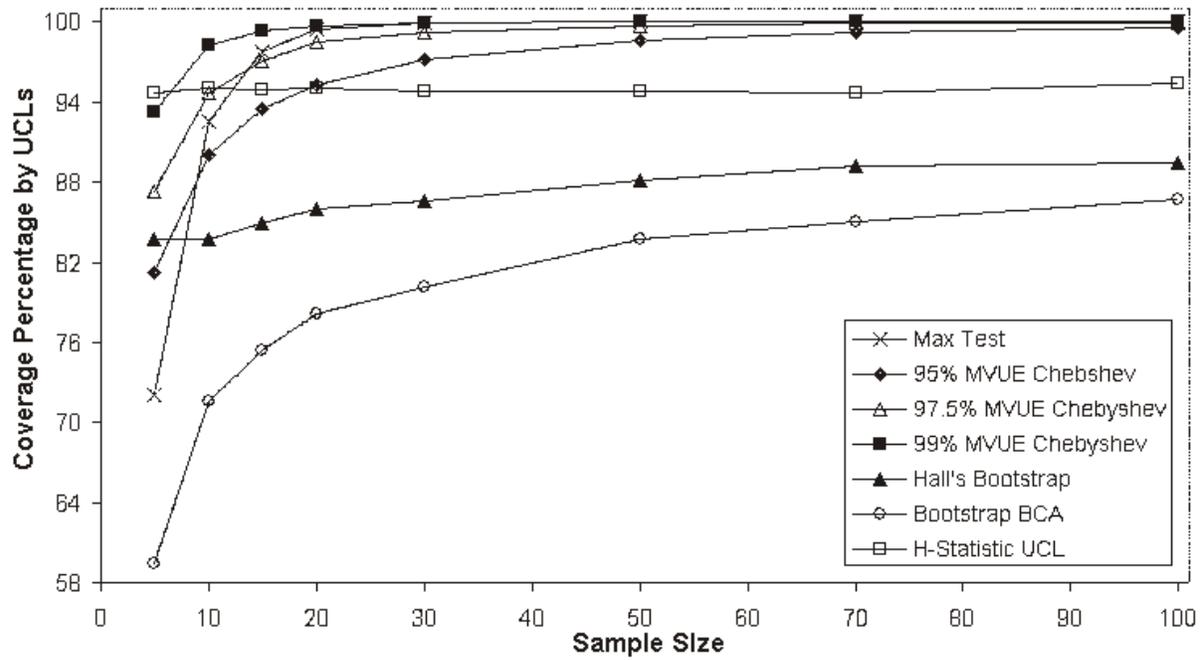Figure B-6. Graphs of Coverage Probabilities by UCLs of Mean of LN(μ=5,σ=1.0)

Figure B-7. Graphs of Coverage Probabilities by UCLs of Mean of LN(μ=5,σ=1.5)

262

REFERENCES

Aitchison, J. and Brown, J.A.C. 1969. *The Lognormal Distribution*, Cambridge: Cambridge University Press.

Anderson, T.W. and Darling, D. A. 1954. *Test of goodness-of-fit*. Journal of American Statistical Association, Vol. 49, 765-769.

Bain, L.J., and M. Engelhardt. 1991. *Statistical Analysis of Reliability and Life Testing Models*, Theory and Methods. 2nd Edition. Dekker, New York.

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 (Navy). 1998. Procedural Guidance for Statistically Analyzing Environmental Background Data, Naval Facilities Engineering Command.

Department of Navy. 1999. *Handbook for Statistical Analysis of Environmental Background Data*. Prepared by SWDIV and EFA West of Naval Facilities Engineering Command. July 1999.

Department of Navy. 2002a. *Guidance for Environmental Background Analysis*. Volume 1 Soil. Naval Facilities Engineering Command. April 2002.

Department of Navy. 2002b. *Guidance for Environmental Background Analysis*. Volume 2 Sediment. Naval Facilities Engineering Command. May 2002.

Department of Navy (Navy). 2003. Guidance for Environmental Background Analysis, Naval Facilities Engineering Command, Volume II: Sediment, UG-2054-ENV.

Dixon, W.J. 1953. *Processing Data for Outliers. Biometrics 9: 74-89*.

Draper NR & 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.

Fleischhauer, H. and Korte, N. 1990. *Formation of Cleanup Standards Trace Elements with Probability Plot.* Environmental Management, Vol. 14, No. 1. 95-105.

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.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York.

Gillian, Mimmack, Manas, G. and Meyer, D. 2001. *Introductory Statistics for Business.* Pearsons South Africa, p. 21.

Gilliespie, B.W., Chen, Q., *et al*. 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. 2012. *Statistics for Censored Environmental Data Using Minitab and R.* Second Edition. John Wiley and Sons, NY.

Helsel, D. 2013. Nondetects and Data Analysis for Environmental Data, NADA in R

Helsel, D.R. 2102a. *Practical Stats Webinar on ProUCL v4*. The Unofficial User Guide; October 15, 2012.

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.

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

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 Ingram Olkin, Harold Hotelling, et alia. Stanford University Press. pp. 278–292.

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.*

Natrella, M.G. 1963. *Experimental Statistics*. National Bureau of Standards, Hand Book No. 91, U.S. Government Printing Office, Washington, DC.

Kunter, M. J., C. J. Nachtsheim, J. Neter, and Li W. 2004. *Applied Linear Statistical Methods.* Fifth Edition. McGraw-Hill/Irwin.

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. 1982. *An extension of Shapiro and Wilk's W test for normality to large samples*. Applied Statistics, 31, 115–124.

Royston, P. 1982a. *Algorithm AS 181: The W test for Normality*. Applied Statistics, 31, 176–180.

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.

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. 1978. *Kolmogorov-Smirnov Test Statistic for the Gamma Distribution with Unknown Parameters*, Dissertation, Department of Statistics, Temple University, Philadelphia, PA.

Schneider, B.E. and Clickner, R.P. 1976. *On the Distribution of the Kolmogorov-Smirnov Statistic for the Gamma Distribution with Unknown Parameters*. Mimeo Series Number 36, Department of Statistics, School of Business Administration, Temple University, Philadelphia, PA.

Schneider, H. 1986. *Truncated and Censored Samples from Normal Populations*. Vol. 70, Marcel Dekker Inc., New York, 1986.

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.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. 1993. *Omnibus Robust Procedures for Assessment of Multivariate Normality and Detection of Multivariate Outliers*. In Multivariate Environmental Statistics, Patil G.P. and Rao, C.R., Editors, pp. 445-488. Elsevier Science Publishers.

Singh, A., Singh, A.K., and G. Flatman 1994. *Estimation of Background Levels of Contaminants*. Math Geology, Vol. 26, No, 3, 361-388.

Singh, A. and Nocerino, J.M. 1995. *Robust Procedures for the Identification of Multiple Outliers*. Handbook of Environmental Chemistry, Statistical Methods, Vol. 2.G, pp. 229-277. Springer Verlag, Germany.

Singh, A. 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.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, Anita, 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. 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., Singh, A.K., and Iaci, R.J. 2002. *Estimation of the Exposure Point Concentration Term Using a Gamma Distribution,* EPA/600/R-02/084, October 2002.

Singh, A. and Singh, A.K. 2003. *Estimation of the Exposure Point Concentration Term (95% UCL) Using Bias-Corrected Accelerated (BCA) Bootstrap Method and Several other methods for Normal, Lognormal, and Gamma Distributions*. Draft EPA Internal Report.

Singh, A. 2004. *Computation of an Upper Confidence Limit (UCL) of the Unknown Population Mean Using ProUCL Version 3.0*. Part I. Download from: www.epa.gov/nerlesd1/tsc/issue.htm

Singh, A., Maichle, R., 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 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

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). 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, 2007.

U.S. Environmental Protection Agency (EPA). 2009. *Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities – Unified Guidance*. EPA 530-R-09-007, 2009.

U.S. Environmental Protection Agency (EPA). 2009a. *ProUCL Version 4.00.04 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. *ProUCL4.00.04 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). 2009c. *Scout 2008 – A Robust Statistical Package*, Office of Research and Development, February 2009.
http://www.epa.gov/esd/databases/scout/abstract.htm#Scout2008v101
http://www.epa.gov/nerlesd1/databases/datahome.htm

U.S. Environmental Protection Agency (EPA). 2010. *Scout 2008 User Guide (Draft)* EPA/600/R-08/038, Office of Research and Development, April 2010.
http://www.epa.gov/nerlesd1/databases/datahome.htm

U.S. Environmental Protection Agency (EPA). 2010a. *A Quick Guide to the Procedures in Scout (Draft),* Office of Research and Development, April 2010.
http://www.epa.gov/nerlesd1/databases/datahome.htm
http://www.epa.gov/esd/databases/scout/abstract.htm

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). 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

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.