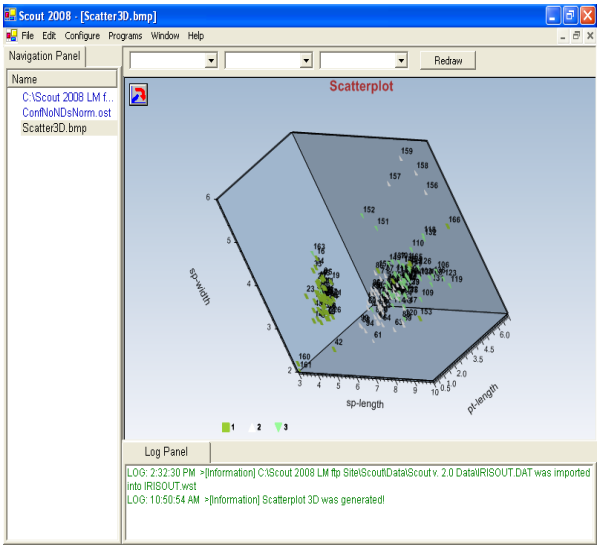
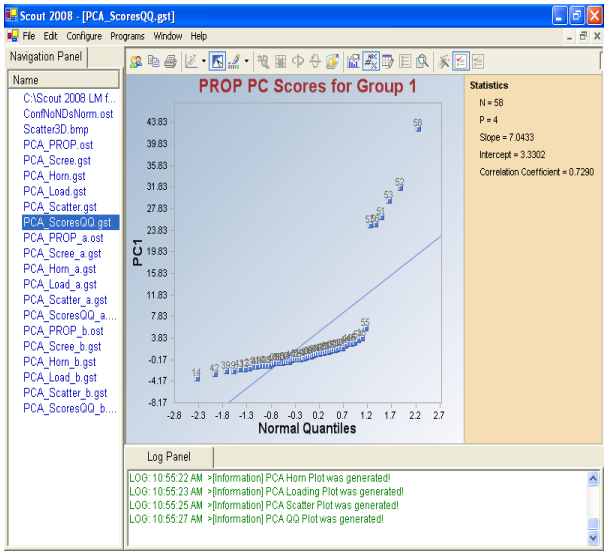
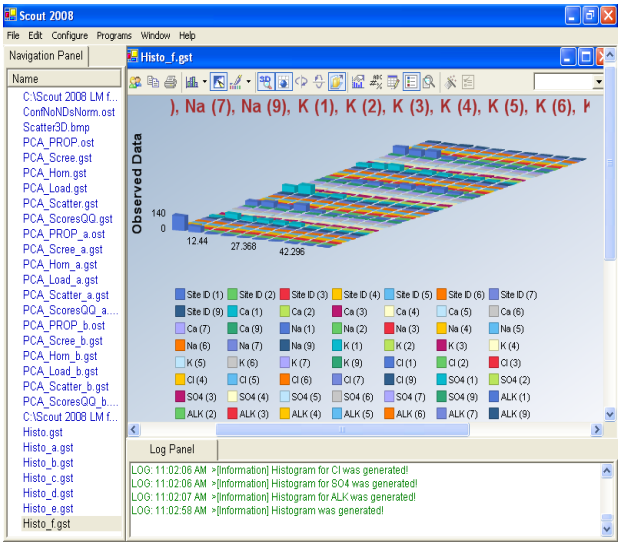
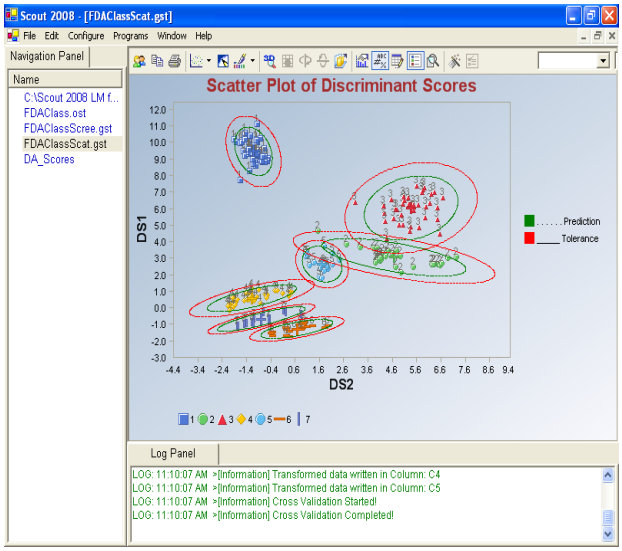
EPA United States Environmental Protection Agency

Scout 2008 Version 1.0 User Guide Part I









RESEARCH AND DEVELOPMENT

EPA/600/R-08/038
February 2009
www.epa.gov

Scout 2008 Version 1.0 User Guide

(Second Edition, December 2008)

John Nocerino

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

Anita Singh, Ph.D.[1]
Robert Maichle[1]
Narain Armbya[1]
Ashok K. Singh, Ph.D.[2]

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

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

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

7663cmb09

Notice

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

The Scout 2008 software was developed by Lockheed-Martin under a contract with the USEPA. Use of any portion of Scout 2008 that does not comply with the Scout 2008 User Guide is not recommended.

Scout 2008 contains embedded licensed software. Any modification of the Scout 2008 source code may violate the embedded licensed software agreements and is expressly forbidden.

The Scout 2008 software provided by the USEPA was scanned with McAfee VirusScan and is certified free of viruses.

With respect to the Scout 2008 distributed software and documentation, neither the USEPA, 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, the Scout 2008 software and documentation are supplied "as-is" without guarantee or warranty, expressed or implied, including without limitation, any warranty of merchantability or fitness for a specific purpose.

Executive Summary

The Scout 2008 version 1.00.01 software package provides a wide variety of classical and robust statistical methods that are not typically available in other commercial software packages. A major part of Scout deals with classical, robust, and resistant univariate and multivariate outlier identification, and robust estimation methods that have been available in the statistical literature over the last three decades. Outliers in a data set represent those observations which do not follow the pattern displayed by the majority (bulk) of the data. It should be pointed out that all of the outlier identification methods are meant to identify outliers in a data set typically representing a single population. Outlier identification methods are not meant to be used on clustered data sets representing mixture data sets, especially when more than two clusters may be present in the data set. On data sets having several clusters, other methods such as cluster analysis and principal component analysis may be used.

Several robust estimation and outlier identification methods that have been incorporated into Scout 2008 include: the iterative classical method, the iterative influence function (e.g., Biweight, Huber, PROP)-based M-estimates method, the multivariate trimming (MVT) method, the least median-of-squared residuals (LMS) regression method, and the minimum covariance determinant (MCD) method. Some initial choices for the iterative estimation of location and scale are also available in Scout 2008, including the orthogonalized Kettenring and Gnanadesikan (OKG) method; the median, median absolute deviation (MAD), or interquartile range (IQR)-based methods; and the MCD method. Scout offers classical and robust methods to estimate: the multivariate location and scale, classical and robust intervals, classical and robust prediction and tolerance ellipsoids, multiple linear regression parameters, principal components (PCs), and discriminant (Fisher, linear, and quadratic) functions (DFs). The discriminant analysis module of Scout can perform cross validation using several methods, including leave-one-out (LOO), split samples, M-fold validation, and bootstrap methods. For both univariate and multivariate data sets, Scout also has a QA/QC module that can be used to compare test (e.g., polluted site, new drug) data set with training (e.g., reference, background, placebo) data set.

Below detection limit (BDL) observations or non-detect (ND) data are inevitable in many environmental and chemometrics applications. Scout has several univariate graphical (e.g., box plots, index plots, multiple quantile-quantile (Q-Q) plots) and inferential methods that can be used on full uncensored data sets and also on left-censored data sets with below detection limit (DL) observations. Specifically, Scout can be used to: compute and graph various interval estimates, perform typical univariate goodness-of-fit (GOF) tests, and perform single and two-sample hypothesis tests on uncensored data sets and left-censored data sets with NDs potentially consisting of multiple detection limits. For univariate data sets with NDs, statistical inference methods (e.g., intervals and hypothesis testing) available in Scout 2008 include simple substitution methods (0, DL/2, and DL), regression on order statistics (ROS) methods, and the Kaplan-Meier (KM)

method. For multivariate data sets with ND observations, Scout can compute mean vector, covariance matrix, prediction and tolerance ellipsoids, and principal components using the Kaplan-Meier method. For multivariate data sets with NDs, Scout can also generate Q-Q plot of Mahalanobis distances (MDs) and prediction and tolerance ellipsoids.

In Scout 2008, emphasis is given to graphical displays of multivariate data sets. Most of the classical and robust methods in Scout are supplemented with formal multivariate classical and robust graphical displays, including the quantile-quantile (Q-Q) plots of the Mahalanobis distances (MDs); control-chart-type index plots of the MDs; distance-distance (D-D) plots; Q-Q plot and index plot of residuals; residual versus leverage distance plots; residual versus residual (R-R) and Y versus Y-hat plots; Q-Q plots of PCs; scatter plots of raw data, PC scores, and DF scores with prediction or tolerance ellipsoids superimposed on the respective scatter plots. Those graphical displays can be formalized by drawing appropriate limits at the critical values of the MDs and Max-MD obtained using the exact scaled beta distribution of the MDs or an approximate chi-square distribution of the MDs. Some graphical methods comparison methods are also available in Scout so that one can graphically compare the performances (e.g., in terms of identifying appropriate outliers and producing best regression fits) of those methods. Specifically, Scout can be used to display multiple D-D plots and R-R plots, multiple linear regression fits, and tolerance ellipsoids or prediction ellipsoids for the various outlier identification methods on the same graph. On these graphs, all observations can be labeled simultaneously or individually by using a mouse. For grouped data, observations can also be labeled by group ID; and group assignment of selected observations can be changed and saved interactively using the computer monitor and mouse.

Scout 2008 also offers GOF test statistics to assess multivariate normality. Several GOF test statistics, including the multivariate kurtosis, the skewness, and the correlation coefficient between the ordered MDs and the scaled beta (or chi-square) distribution quantiles, are displayed on a Q-Q plot of the MDs. The associated critical values of those GOF test statistics (obtained via extensive simulation experiments) are also displayed on the graphical displays of the Q-Q plots of the MDs. Some approximate multinormality GOF test statistics (e.g., standardized kurtosis, omnibus test) and their p-values are also displayed on a Q-Q plot of MDs.

Two standalone software packages, ProUCL 4.00.04 and ParallAX, have also been incorporated into Scout 2008. ProUCL 4.00.04 is a statistical software package developed to address environmental applications, whereas the ParallAX software offers graphical and classification tools to analyze multivariate data using the parallel coordinates.

Acronyms and Abbreviations

% NDs	Percentage of Non-detect observations
ACL	alternative concentration limit
A-D, AD	Anderson-Darling test
AM	arithmetic mean
ANOVA	Analysis of Variance
AOC	area(s) of concern
B*	Between groups matrix
BC	Box-Cox-type transformation
BCA	bias-corrected accelerated bootstrap method
BD	break down point
BDL	below detection limit
BTV	background threshold value
BW	Black and White (for printing)
CERCLA	Comprehensive Environmental Response, Compensation, and Liability Act
CL	compliance limit, confidence limits, control limits
CLT	central limit theorem
CMLE	Cohen's maximum likelihood estimate
COPC	contaminant(s) of potential concern
CV	Coefficient of Variation, cross validation
D-D	distance–distance
DA	discriminant analysis
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 non-detects replaced by half of the respective detection limits
DQO	data quality objective
DS	discriminant scores
EA	exposure area
EDF	empirical distribution function
EM	expectation maximization
EPA	Environmental Protection Agency
EPC	exposure point concentration
FP-ROS (Land)	UCL based upon fully parametric ROS method using Land's H-statistic

Gamma ROS (Approx.)	UCL based upon Gamma ROS method using the bias-corrected accelerated bootstrap method
Gamma ROS (BCA)	UCL based upon Gamma ROS method using the gamma approximate-UCL method
GOF, G.O.F.	goodness-of-fit
H-UCL	UCL based upon Land's H-statistic
HBK	Hawkins Bradu Kaas
HUBER	Huber estimation method
ID	identification code
IQR	interquartile range
K	Next K, Other K, Future K
KG	Kettenring Gnanadesikan
KM (%)	UCL based upon Kaplan-Meier estimates using the percentile bootstrap method
KM (Chebyshev)	UCL based upon Kaplan-Meier estimates using the Chebyshev inequality
KM (t)	UCL based upon Kaplan-Meier estimates using the Student's t-distribution cutoff value
KM (z)	UCL based upon Kaplan-Meier estimates using standard normal distribution cutoff value
K-M, KM	Kaplan-Meier
K-S, KS	Kolmogorov-Smirnov
LMS	least median squares
LN	lognormal distribution
Log-ROS Estimates	estimates based upon data set with extrapolated non-detect values obtained using robust ROS method
LPS	least percentile squares
MAD	Median Absolute Deviation
Maximum	Maximum value
MC	minimization criterion
MCD	minimum covariance determinant
MCL	maximum concentration limit
MD	Mahalanobis distance
Mean	classical average value
Median	Median value
Minimum	Minimum value
MLE	maximum likelihood estimate
MLE (t)	UCL based upon maximum likelihood estimates using Student's t-distribution cutoff value

MLE (Tiku)	UCL based upon maximum likelihood estimates using the Tiku's method
Multi Q-Q	multiple quantile-quantile plot
MVT	multivariate trimming
MVUE	minimum variance unbiased estimate
ND	non-detect or non-detects
NERL	National Exposure Research Laboratory
NumNDs	Number of Non-detects
NumObs	Number of Observations
OKG	Orthogonalized Kettenring Gnanadesikan
OLS	ordinary least squares
ORD	Office of Research and Development
PCA	principal component analysis
PCs	principal components
PCS	principal component scores
PLs	Prediction limits
PRG	preliminary remediation goals
PROP	proposed estimation method
Q-Q	quantile-quantile
RBC	risk-based cleanup
RCRA	Resource Conservation and Recovery Act
ROS	Regression on order statistics
RU	remediation unit
S	substantial difference
SD, *Sd, sd*	standard deviation
SLs	simultaneous limits
SSL	soil screening levels
S-W, SW	Shapiro-Wilk
TLs	tolerance limits
UCL	upper confidence limit
UCL95, 95% UCL	95% upper confidence limit
UPL	upper prediction limit
UPL95, 95% UPL	95% upper prediction limit
USEPA	United States Environmental Protection Agency
UTL	upper tolerance limit
Variance	classical variance
W*	Within groups matrix

WiB matrix	Inverse of W* cross-product B* matrix
WMW	Wilcoxon-Mann-Whitney
WRS	Wilcoxon Rank Sum
WSR	Wilcoxon Signed Rank
Wsum	Sum of weights
Wsum2	Sum of squared weights

Acknowledgements

We wish to express our gratitude and thanks to our colleagues who helped to develop past versions of Scout and to all of the many people who reviewed, tested, and gave helpful suggestions for the development of Scout. We wish to especially acknowledge: Nadine Adkins, Girdhar Agarwal, Anastasia Arteyeva, Chad Cross, Rohan Dalpatadu, Marion Edison, Tim Ehli, Evan Englund, Peter Filzmoser, Kirk Fitzgerald, George Flatman, Forest Garner, Robert Gerlach, Edward Gilroy, Colin Greensill, Anwar Hossain, Kuen Huang-Farmer, Mia Hubert, Alfred Inselberg, Barry Lavine, Maliha Nash, Ramon Olivero, John Palasota, Bruce Rhoads, Brian Schumacher, Cliff Spiegelman, Teruo Sugihara, Martin Stapanian, Valeri Tsarev, Asokan Mulayath Variyath, Suresh Veluchamy, Sabine Verboven, INDUS Corporation, and Computer Sciences Corporation.

Software Used to Develop Scout 2008

Scout 2008 (Scout) has been developed in the Microsoft .NET Framework using the C# programming language to run under the Microsoft Windows XP operating systems. As such, to properly run Scout, the computer using the program must have the .NET Framework pre-installed. The downloadable .NET files can be found at one of the following two Web sites:

- http://msdn2.microsoft.com/en-us/netframework/default.aspx
  Note: *Download .NET version 1.1*

- http://www.microsoft.com/downloads/details.aspx?FamilyId=262D25E3-F589-4842-8157-034D1E7CF3A3&displaylang=en

The Scout source code uses the following embedded licensed software:

Chart FX 6.2 (for graphics), http://www.softwarefx.com

Quinn-Curtis QCChart 3D Charting Tools for .Net (for graphics), http://www.quinn-curtis.com

NMath (for mathematical and statistical libraries), http://www.centerspace.net/

FarPoint (for spreadsheet applications), http://www.fpoint.com/

Table of Contents

Chapter 1

Introduction

This chapter briefly summarizes statistical methods incorporated in Scout, which are not readily available in commercial and freeware software packages. Therefore, only those modules of Scout consisting of such methods are briefly discussed in this chapter. Please note that at the time of writing this Scout 2008 User Guide, resources were not available for producing a Scout 2008 Technical Guide, which would discuss the theory used in the Scout 2008 software in much more detail. A technical guide is planned. However, in the meantime, for theoretical inquiries, please consult the Bibliography given at the end of this user guide.

1.1 Methods to Handle Data Sets with Below Detection Limit Observations

The "Data" module of Scout offers several imputation (e.g., via regression on order statistics) and substitution (e.g., replacing non-detects (NDs) by DLs or DL/2) methods that can be used to estimate or extrapolate non-detect data consisting of multiple detection limits (DLs). Specifically, this module has some univariate imputation (e.g., via regression on order statistics (ROS) – for normal, lognormal, and gamma distributions) and substitution (e.g., replacing NDs by 0, DL, DL/2, or uniform random variables) methods that can be used to estimate and/or extrapolate non-detect observations present in a left-censored data with ND observations. Whenever applicable, transformation and imputation methods in Data module can also be used on data sets consisting of multiple groups (e.g., perform z-transform, log ROS (LROS)). One may use the transformation module on a multivariate data set with NDs before using a multivariate method (e.g., Regression, PCA, and DA) on that data set. It should be noted that for multivariate data sets with NDs, Scout can estimate mean vector and covariance matrix using the Kaplan-Meier (1958) method which does not require the imputation of NDs before using statistical methods such as principal component analysis (PCA). Some basic tools to estimate missing observations and bivariate transformation operations are also available in this Data module. The Stats/GOF module of Scout offers several parametric and nonparametric (including Kaplan-Meier, regression on order statistics (ROS), and bootstrap methods) univariate statistical methods that can be used on left-censored data sets with non-detect observations potentially having multiple detection limits. For both uncensored and left-censored data sets, Scout can compute a variety of parametric and nonparametric interval estimates, including: the confidence interval for the mean, prediction intervals, and tolerance intervals. The Stats/GOF module also has univariate goodness-of-fit (GOF) tests for normal, lognormal, and gamma distributions for uncensored and left-censored data sets. However, it should be noted that it is not easy to verify distributional assumptions for censored data sets consisting of multiple detection limits (DLs). Therefore, use of nonparametric methods is preferable on such left censored data sets. Some single and two-sample hypotheses tests (e.g., Wilcoxon Rank Sum Test,

Gehan Test) for uncensored and left-censored data sets potentially having single or multiple DLs are also available in Scout. The details of methods to compute statistics based upon left-censored data sets can be found in Singh and Nocerino (2001), Helsel (2005), Singh, Maichle, and Lee (2006), and ProUCL 4.00.04 Technical Guide (2007).

1.2 Goodness-of-Fit Test Statistics to Test Multinormality of a Data Set

It is not easy to verify multivariate normality of a data set. Multivariate normality tests such as multivariate kurtosis (MK) and skewness (e.g., Mardia (1970, 1974), Mardia and Kanazawa (1983)) are very sensitive to even small changes in the values of observations of a data set. As a result, it is very hard not to reject the hypothesis of multinormality of a data set. Therefore, it is desirable also to use graphical quantile-quantile (Q-Q) plots (e.g., Singh (1993), Koziol (1993) and Fang and Zhu (1997)) of Mahalanobis distances (MDs) to assess the approximate multinormality of a data set. Singh (1993) proposed to use a correlation-type goodness-of-fit (GOF) tests to assess approximate multivariate normality of a data set. Scout 2008 can compute classical and robust (e.g., based upon iterative M-estimation method, MVT and MCD methods) estimates of multivariate kurtosis and skewness. Scout 2008 can also generate classical and robust Q-Q plots of MDs based upon quantiles of scaled beta distribution and approximate chi-square distribution.

Extensive simulated critical values of the multivariate GOF test statistics including multivariate kurtosis (MK), multivariate skewness (MS), correlation coefficients between order MDs and quantiles of scaled beta (or chi-square) distribution have been generated. The GOF Q-Q plot of MDs is formalized by displaying exact test statistics: MS, MK, and correlation coefficient and their simulated critical values for a specified level of significance, α. Approximate MS (with small sample adjustment), standardized approximate MK, and approximate omnibus multinormality test and their associated p-values are also displayed on these Q-Q graphs. It should be pointed out that there are significant differences between the exact simulated critical values of multivariate kurtosis and skewness, and their approximate critical values as described in the literature. Also, the performance of these approximations (e.g., chi-square distribution for MS and normal distribution for standardized kurtosis) is not well established, especially when the dimension, p becomes larger than 5. These discrepancies can be seen by looking at the various exact and approximate GOF test statistics displayed on the Q-Q plot of MDs. This issue is under further investigation. A linear pattern displayed by data pairs, (theoretical quantiles from the distribution of MDs and ordered observed MDs) on the Q-Q plot of MDs suggests (cautiously) approximate multinormality of the data set. Since, Q-Q plots of MDs are very sensitive to even minor changes in observations and mild outliers, other measures such as Q-Q plot and scatter plot of principal components (also available in Scout) may also be used to assess approximate multinormality (cautiously) of a multivariate data set.

1.3 Robust Methods in Scout

Several options in various modules of Scout (e.g., Robust intervals, Outlier/Estimates, QA/QC, Regression, Method Comparison, PCA, and discriminant analysis) offer robust statistical methods described in the following sections.

1.3.1 Robust Intervals

In addition to classical methods, the Stats/GOF module of Scout has univariate methods to compute robust estimates of location and scale, and robust interval estimates. At present, robust methods are available for uncensored data sets without non-detect observations. The univariate iterative robust estimation methods in Scout 2008 include: Tukey's Bisquare (1975) and Kafadar's version of Tukey's Biweight (1982) influence functions, Huber (1981) and PROP (Singh, 1993) influence functions, and the trimming method. Two choices: (classical mean and *sd*), and (median, 1.48MAD or IQR/1.345) of initial estimates are available for all iterative univariate estimation methods included in Scout. The robust interval module can be used to compute robust confidence intervals of the mean, robust prediction interval for k (≥1) observations, tolerance intervals, and robust simultaneous (with critical value from the distribution of Max (MDs)), and individual (with critical value from the distribution of MDs) intervals. The details of the robust interval estimates can be found in Kafadar (1982), Hoaglin and Mosteller, and Tukey (1983), Singh and Nocerino (1995, 1997), and Horn, Pesce and Copeland (1998).

The robust interval option provides graphical comparison of the various robust and classical interval estimation methods. Depending upon the selected options and methods, some relevant robust statistics such as mean, standard deviation (*sd*), influence function alpha, α, trimming percentage (%), location and scale tuning constants (TCs) are also displayed on these interval method comparison graphs. This option also provides classical and robust control-chart-type interval index plots exhibiting the associated limits for the selected variable. On a single classical or robust (e.g., using Biweight influence function) interval plot (showing all individual data points), one can draw more than one set of intervals including: individual interval, prediction interval, tolerance interval, and simultaneous interval. Specifically, on this control-chart-type interval plot, if Huber option is used, all interval estimates will be computed using the same Huber influence function. These kinds of interval graphs can be quite useful in Quality Assurance/Quality Control (QA/QC) applications including industrial, manufacturing, clinical trials, medical, pharmaceutical, and environmental. Group Analysis option of Robust Interval option can be used to formally compare interval estimates of a characteristic of interest for various groups (e.g., lead concentrations in various areas of a polluted site, arsenic concentrations in monitoring wells, effectiveness of two or more drugs) under study.

Standard terminology, such as coverage (e.g., half samples, h value) and cutoff (influence function α, critical α, trimming percentage) levels used by the robust methods to identify outliers as incorporated in Scout 2008 are described next.

1.3.2 Coverage or Cutoff Levels (Factors) Used by Outlier Identification Methods

Most robust methods available in the literature either use a coverage factor, h (e.g., half samples, h = [(n+p+1)/2] for MCD, best subset of size (p+1), or of size h = [(n+p+1)/2] for LMS), or a critical level, α (e.g., influence function, α for PROP and Huber influence functions, location and scale tuning constants for Biweight function, trimming percentage, α% for multivariate trimming (MVT) method) to identify outliers in a p-dimensional data set of size n. There is a close relationship between the coverage or critical cutoff and the break down (BD) point of an estimate. Specifically, for the MCD and LMS methods, higher values of h may yield MCD and LMS estimates with lower BD points; for influence function-based M-estimation methods (e.g., PROP and Huber), higher values of the influence function, α, may yield estimates with higher BD points; and for MVT method, higher values of trimming percentage tend to yield estimates with higher BD points.

It should be noted that the success of a robust method in identifying outliers depends upon the coverage or cutoff levels used and the behavior of the influence function. In practice, the smooth redescending influence functions, such as the PROP influence, will perform better than nondecreasing influence functions such as the Huber influence function (e.g., Hampel et al. (1986)). In addition to coverage and critical cutoff levels, initial robust starts in iterative process of obtaining robust estimates also play an important role in achieving high break down estimates.

For each of the robust method incorporated in Scout, the user can pick a suitable coverage, h or cutoff level, α. It is suggested that the user uses more than one coverage or cutoff factor for the selected method. For example, for the standard MCD method (also known as very robust MCD) with h = [(n+p+1)/2], the BD is roughly equal to 50%. The use of the very robust MCD method with this coverage, h, tends to find more outliers than actually are present in the data set. Even though it is desirable to use robust methods with high BD points, those robust methods should be efficient enough not to identify inliers (and good leverage points) as outliers (and regression outliers). This issue can be addressed by choosing higher coverage (e.g., 75% coverage) levels. Using Scout 2008, one can perform MCD and LMS methods for user selected coverage levels.

Since the number of outliers present in a data set is not known in advance, it is desirable to use more than one value of the coverage or cutoff level on the same data set. In order to get some idea about the number of outliers present in a data set, the use of graphical displays is recommended before using the outlier identification methods available in Scout (e.g., Huber, MCD, MVT, or PROP) or in any other software package. There is no substitute for graphical displays of multivariate data sets. The graphical displays offer additional information about the patterns and outliers present in a data set. This kind of information cannot be obtained by looking at the statistics computed by the various statistical procedures. Moreover, most computed statistics (e.g., mean vector, covariance matrix, MDs, kurtosis) get distorted by the presence of outliers. The use of graphical displays such as scatter plots of raw data, scatter plots of principal components (PCs), normal quantile-quantile (Q-Q) plot of dependent variable (to identify regression

outliers), and Q-Q plot of Mahalanobis distances (MDs) of explanatory variables (to identify leverage point) is helpful to get some idea about the number (or percentage, k) of outliers that may be present in the data set. The multivariate graphs listed above are also useful to verify if the identified outliers based upon outlier test statistics (e.g., MDs, MS, weights) indeed represent outliers. This step helps the user to pick an appropriate value of h (MCD) or influence function alpha (e.g., PROP), which in turn will help obtain more reliable and accurate estimates of population parameters (e.g., location, scale, regression).

1.3.3 Critical or Cutoff Outlier Alpha Used in Graphical Displays

In Scout 2008, emphasis is given to the graphical displays of multivariate data sets. Graphical methods in Scout 2008 include: 2-dimensional and 3-dimensional scatter plots, Q-Q, Index, and distance-distance (D-D) plots of MDs, prediction and tolerance ellipsoids, Q-Q plots of residuals, and scatter plots of residuals versus unsquared leverage distances, and multiple ellipsoids or regression lines on the same graph. Graphical displays of multiple ellipsoids or regression lines provide useful graphical comparisons of various robust and resistant methods incorporated in Scout 2008. An attempt has been made to formalize these graphical displays by drawing control limits, prediction and tolerance ellipsoids based upon the critical values of the MDs (individual MDs) and Maximum MD (Max-MD) computed using the graphical alpha or regression band alpha. Graphical displays for the MCD and LMS methods use critical values from chi-square distribution at fixed critical level of 0.025 as cited in the literature (e.g., Rousseeuw and van Zomeren (1990)). The LMS method uses fixed cutoff values of -2.5 and +2.5 to identify regression/residual outliers (Rousseeuw and Leroy, 1987).

For other robust (PROP, MVT, Huber), and classical and sequential classical methods, Scout uses critical values of the MDs based upon quantiles of scaled beta (or approximate chi-square) distribution (Singh (1993)). The critical values of MDs and Max-MDs used on these multivariate graphs are computed for user selected outlier critical alpha. Control limits (or prediction and tolerance ellipsoids) drawn at critical values (based upon outlier critical alpha) obtained from the distribution of MDs (prediction ellipsoid) and maximum MD (tolerance ellipsoid) are drawn on the Q-Q plots and index plots of MDs. Critical values of various other statistics displayed on the Q-Q plots of MDs, including MS, MK, and correlation coefficients are also computed for the outlier critical alpha. On scatter plots of raw data, principal component scores, or discriminant score, prediction ellipsoids are drawn at critical value (computed for critical outlier alpha) from the distribution of MDs, and tolerance ellipsoids are drawn at critical value from the distribution of maximum MD (Max-MD). Observations lying outside the outer ellipsoid (tolerance) represent potential outliers, and observations lying between the inner (prediction) and outer (tolerance) ellipsoid may be considered representing borderline outliers.

In regression applications, graphical displays of Q-Q plot or index plot of residuals with control limits drawn at the critical values (associated with selected regression outlier α) of unsquared residual distances (for LMS, these are hard lines drawn at -2.5 and 2.5) are used to determine regression outliers. A semi-formal residual versus unsquared leverage distance plot (Singh and Nocerino (1995)) is also available in Scout to identify regression

outliers (uses regression outlier alpha) and inconsistent (bad) leverage outliers (uses leverage outlier alpha). In most of the graphical displays listed above, Scout 2008 collects and uses user selected critical levels to compute appropriate critical values of the statistics used (e.g., critical values of MDs, critical value of Max MD, critical values for leverage Mahalanobis distances and unsquared regression distances) to generate the graphical displays.

1.3.4 Break Down Point

A brief description of the break down (BP) point (Hampel (1974, 1975), Huber (1981), Maronna, Martin, and Yohai (2006), Hubert, Rousseeuw, and van Aelst (2007)) of an estimate is described as follows.

1.3.4.1 Break Down Point of an Estimation Method

A great deal of emphasis is placed on break down (BD) point of robust outlier identification and estimation methods. The performance of various robust methods (estimates) is evaluated in terms of their BD points (e.g., Hubert, Rousseeuw, and van Aelst (2007)). Robust methods roughly having BD point of about 50% are preferred and often are called "very" robust methods (e.g., Rousseeuw and van Zomeren (1990), Hubert, Rousseeuw, and van Aelst (2007)). It is also noted that the "very" robust estimation methods are inefficient as they often tend to find more outliers than actually are present in a data set (e.g., Maronna, Martin, and Yohai (2006)). The LMS (Rousseeuw (1984), Rousseeuw and Leroy (1987)) and the MCD (Rousseeuw and van Driessen (1999)) methods treat all outliers (e.g., extreme and borderline outliers) equally by assigning the same "zero" weight (hard rejection of outliers). Therefore, it is desirable to use influence function (Hampel (1974, 1985), Huber (1981))-based robust methods possessing soft and smooth rejection of outliers. The PROP influence function (e.g., Singh (1993)) is a redescending smooth influence function. It is noted that iteratively obtained robust M-estimates based upon the PROP influence function (e.g., with initial robust starts) assign reduced-to-negligible weights, respectively, to intermediate and extreme observations; observations coming from the central part of data are assigned full unit weights. Furthermore, the robust estimates based upon the PROP influence function are in close agreement with the classical estimates obtained using the data set without the outliers (Singh and Nocerino (1995)).

The BD point of a method (or of estimates obtained using that method) represents that fraction of observations which can be altered (e.g., can be made very large) arbitrarily without affecting (influencing, distorting, changing drastically) the values of the estimates. That is the BD of a method (e.g., LMS) represents that fraction of outlying observations that can be tolerated by the estimates (e.g., LMS estimates) obtained using that method without distorting (breaking) the estimates. Obviously, the BD point of a classical estimate (e.g., arithmetic mean, OLS regression estimates) is "zero," as even a single arbitrarily selected large value can completely distort (change the estimate without bounds) that classical estimate. It is also noted that the sample median of a data set (and

similarly median of squared residuals) has a BD point of 50% as median of a data set remains unchanged even when about 50% of the data values are altered arbitrarily.

The break down points of LMS and MCD methods are known to be about 50%. Details about LMS and MCD estimates and their break down points are discussed respectively in section 1.3.6 and 1.3.7. Both the LMS regression and the MCD estimation methods are based upon extensive searches of elemental subsets (Hawkins, Bradu, and Kaas (1984), Hawkins (1993)) of size, (p+1). Other variations of the initial subset size such as subsets of size (n+p+1) may also be used. Some of these choices for sizes of the initial subsets searched have been incorporated in the Scout software. In Scout, the MCD method is labeled as the Extended MCD method. It is also known that the theoretical break down point of M-estimates (Maronna, 1976) of p-dimensional multivariate location and scale is no more than 1/ (p+1). However, it should be noted that practical BD of an iteratively obtained robust M-estimate (generalized likelihood estimate) based upon a smooth redescending function such as the PROP (Singh, 1993) influence function can be much higher than 1/(p+1). The break down point of iteratively obtained robust and resistant estimates increases with each iteration (as outlying observations iteratively are assigned reduced weights) until the convergence of M-estimates is achieved. Typically, convergence is achieved in less than 10-15 iterations. More details can be found in Section 1.3.8. Scout generates intermediate results for all intermediate iterations for users to review. It should be noted that higher break down points of iteratively obtained robust estimates (e.g., Huber and PROP) are achieved by using higher values of the influence function alpha, α (or of trimming percentage for MVT method), used to identify outliers. It is observed that a robust method based upon PROP influence function assigns reduced to negligible weights to intermediate and extreme outliers. This is especially true when an initial robust start (e.g., based upon OKG (Devlin, Gnanadesikan, and Kettenring (1975)), Maronna and Zamar (2002) method) is used in the iterative process of obtaining M-estimates.

1.3.5 Initial Estimation Methods Available in Scout 2008

Several initial start robust estimates to compute iteratively obtained M-estimates are available in Scout. It is well known that classical methods have a zero BD point, and they suffer from severe masking effects. This means that the presence of some of the outliers (e.g., extreme outliers) may mask the presence of some other outliers (e.g., intermediate outliers). Even robust outlier identification and estimation methods suffer from masking effects. In order to overcome and reduce the masking effects, robust initial start estimates are used in the iterative process of obtaining robust estimates. Initial start robust estimates as incorporated in Scout can be used with all iterative estimation methods (including sequential classical method) available in Scout.

The initial start estimates as incorporated in Scout include: 1) the classical mean vector and classical scale matrix; 2) the median vector and MAD/0.6745 (or IQR/1.35)-based covariance matrix with off diagonal elements obtained from the classical covariance matrix; 3) the median vector and covariance matrix obtained using the Kettenring and Gnanadesikan (KG) method (1975); and 4) the median vector and orthogonalized KG

(OKG) covariance matrix as proposed by Maronna and Zamar (2002). Here, MAD/0.6745 represents the MAD-based standard deviation of a variable, and the IQR/1.35 represents the IQR standard deviation of a variable. In practice, often the MAD of a variable becomes zero, even when the variance of that variable is not zero (e.g., well known Iris data of size 50). In such cases, an IQR fix is applied, and the IQR/1.35 is used as a robust estimate of the standard deviation for that variable.

It is noted that the OKG estimate as an initial estimate works very well with most iterative estimation methods, including PROP, Huber, and MVT. It is also noted that the use of the OKG method as an initial start estimate also improves the performance (in terms of identification of outliers) of the iterative sequential classical method. However, the computation of the OKG mean vector, as suggested and described in Maronna and Zamar (2002), and Maronna, Martin, and Zamar (2006), does not yield good results, and therefore not included in Scout. The developers of Scout 2008 are currently working on how to compute more reliable estimate of the mean vector based upon OKG method.

1.3.6 Least Median of Squares (LMS) Regression Method

In the LMS regression method, the objective is to find an elemental subset of size (p+1) that minimizes the median of squared residuals (Rousseeuw (1984)). The minimization criterion for the LMS regression is the median of squared residuals. This objective is obtained by searching for elemental subsets of size (p+1), p = number of explanatory variables. The elemental subset that minimizes the median of squared residuals is called the "best" elemental subset. It should be noted that more than one elemental subset can yield the same minimum value of the criterion (median of squared residuals). The use of different LMS subsets (best subsets) may result in different LMS regression estimates.

Depending upon the dimension and size of the data set, the process of searching for the best (global) elemental subset of size (p+1) can be time-consuming. Therefore, in addition to an exhaustive search for all elemental subsets, some quick (1,500 subsets), extensive (3,000 subsets), and user specified search strategies have been incorporated in Scout. As mentioned before, the best subset (minimizing the objective function) of size (p+1) may not be unique, even when the search is exhaustive. Therefore, the LMS regression parameter estimates may not be unique.

Since the median of squared residuals is being minimized, the BD of LMS regression estimates is roughly 50%. The LMS estimates can tolerate about 50% arbitrarily large values (outliers) before the regression estimates break down or get severely distorted by the presence of those outliers. Since the LMS method roughly has 50% BD point, the LMS method tends to identify about 50% the observations as outliers (both regression as well as leverage outliers). It is observed that, in practice, the LMS method identifies some of the inliers (non-outliers for obtaining a regression model) as outliers. That is, the LMS method may find more outliers than actually are present in the data set. This is the reason that the LMS method is known as an inefficient robust method (Maronna, Martin, and Yohai (2006)). To some extent, this problem is overcome by using re-weighted least square regression by assigning zero weights to observations with LMS absolute residuals

greater than 2.5 (Rousseeuw and Leroy (1987), Rousseeuw and van Zomeren (1990)). However, it is noted that even after performing this extra step of re-weighted least square regression, the LMS method tends to find some of the non-outliers as outliers.

It is also noted that, even though, the LMS method identifies most of the leverage points that may be present in a data set, it fails to distinguish between the good and bad leverage points. As a result, the resulting regression model may not be very useful. This issue is illustrated in this user guide by using the LMS method on the Hawkins, Bradu, and Kaas - HBK (1984) data set. This HBK data set has 75 observations and 3 explanatory variables. In the literature, the leverage points are defined as those outliers that are outliers in the space of x-variables (3-dimesional here). The good leverage points enhance the regression model (with higher coefficient of determination, lower scale estimate, and lower standard errors of estimates of regression parameters) and bad leverage points are outliers in both x-space and y-direction of dependent variable. The detailed definition (with graphical displays) of regression outliers, good and bad leverage points can be found in Rousseeuw and Leroy (1987), Rousseeuw and van Zomeren (1990), and Singh and Nocerino (1995). Following the definition of regression outliers, good (consistent) and bad (inconsistent) leverage points, in HBK data set, there are 4 (11, 12, 13, and 14) bad leverage points (and regression outliers) and 10 good leverage points, as the inclusion of 10 good points (1 through 10) enhance the regression model. The LMS regression method identifies observations 1 through 10 as bad leverage points, contradicting the definition of good leverage points as described and graphically illustrated in Rousseeuw and Leroy (1987). Without the first 10 observations, there is no regression model, and the problem reduces to simply an outlier identification problem. Several methods in Scout 2008, such as the PROP method with an OKG start and the MCD method, find the first 14 observations in both 3 (without y-variable) and 4 (with y-variable) dimensional spaces.

Alternatively, instead of minimizing the median of squared residuals, one can minimize some percentile (e.g., 75^{th} percentile, or 90^{th} percentile) of squared residuals. This method is labeled as the least percentile of squares (LPS) regression method in the regression module of Scout software package. The problem of not distinguishing between the good and bad leverage points may be addressed by using the LPS regression (see example in Scout User Guide). Depending upon the number of bad leverage points and regression outliers present in the data set, one may want to use the LMS or the LPS method on the same data set to obtain the appropriate robust fit. Obviously, the LPS regression estimates obtained by minimizing the k^{th} ($k > 50\%$) percentile of squared residuals will have a lower break down point than the LMS estimates. For example, the BD of LPS regression estimates obtained by minimizing the 75^{th} ($k=75\%$) percentile of squared residuals is $(n-[n*0.75]-p+2)/n$, where p is the number of regression variables, and [x] represents the largest integer contained in x.

In order to perform the LPS regression, one needs to have some idea about the value of k, the percentage of outliers (bad leverage points and regression outliers) that may be present in the data set. One may want to perform the LPS regression for a few values of k including $k = 0.5$. As mentioned before, since the number of outliers (both regression

and leverage) are not known in advance, it is suggested to use graphical displays, such as scatter plots of the raw data, scatter plots of the principal components (PCs), a normal quantile-quantile (Q-Q) plot of dependent variable (to identify regression outliers), and a Q-Q plot of Mahalanobis distances (MDs) of explanatory variables (to identify leverage points) to get some idea about the number (or percentage, k) of outliers that may be present in the data set. Based upon the outlier information thus obtained, one may perform an appropriate LMS/LPS regression on the data set. Graphical displays are also useful to perform confirmatory analyses, that is multivariate graphs in Scout can be used to verify if indentified outliers (e.g., based upon MDs and weights) indeed represent outlying and aberrant observations. The BD points for LMS (k~0.5) and the least percentile of squared residuals (LPS, k>0.5) regression methods as incorporated in Scout are summarized in the following table. Note that LMS is labeled as LPS when k>0.5. In the following the fraction, k is given by 0.5≤ k<1. For an example, for the median, the fraction, k = 0.5, for 75th percentile, fraction, and k = 0.75.

Approximate Break Down Point for LMS or LPS Regression Estimates

No. of Explanatory Vars., p = 1		No. of Explanatory Vars., p>1	
Minimizing Squared Residual	**BD**	**Minimizing Squared Residual**	**BD**
Pos = [n/2], k = 0.5	(n-Pos)/n	Pos = [n/2], k = 0.5	(n-Pos-p+2)/n
Pos = [(n+1)/2]	(n-Pos)/n	Pos = [(n+1)/2]	(n-Pos-p+2)/n
Pos = [(n+p+1)/2]	(n-Pos)/n	Pos = [(n+p+1)/2]	(n-Pos-p+2)/n
Pos = [n*k], k>0.5 ~ LPS	(n-Pos)/n	Pos = [n*k], k>0.5	(n-Pos-p+2)/n ~ LPS

Here [x] = greatest integer contained in x, and k represents a fraction: 0.5≤k<1. Pos stands for position/index of an entry in ordered array (of size n) of squared residuals. The squared residual at position, Pos is being minimized. For example, when Pos = [n/2], the median of squared residuals is being minimized.

1.3.7 MCD Method (Extended MCD Method)

For the MCD method, the objective is to find a subset of some specified size, h (n/2≤h≤n), which will minimize the determinant of the covariance matrix based upon that subset of size h. The subset of size h minimizing the determinant of the covariance matrix is termed as the best subset. The positive integer, h is also known as coverage or half sample. The most commonly used and default value of h is [(n+p+1)/2] = largest integer contained in (n+p+1)/2. Just like the LMS method, the search for the best subset of size h, starts with searching through the elemental subsets (subsets of size p+1) or initial subsets of some user specified size. Depending upon the size and dimension of the data set, the search for the best subset of size h can be time-consuming. The fast MCD algorithm as described in Rousseeuw and van Driessen (1999) has been incorporated in Scout. Some variations for the initial subset sizes (e.g., (p+1), (n+p+1), user specified) have been incorporated in Scout. Moreover, the user can choose the number of initial subsets searched (instead of 500) and the number of best subsets (instead of 10) retained to find the final best subset of size h. Just like the LMS method, the MCD estimates are

not unique. It should be noted that different search options may result in different MCD estimates.

The BD point of MCD estimates is given by the fraction (n-h+1)/n. It is noted that there is a direct relation between the coverage value, h, and the BD point of the MCD estimates. Higher values of h yield estimates with a lower BD point. The use of the default value of coverage, h, roughly identifies the optimal (~ about 50%) number of outliers. In practice, the MCD method identifies some of the inliers as outliers. As a result, MCD method is often called to be an inefficient method (Maronna, Martin, and Yohai (2006)). Just like the LMS method, re-weighted estimates of location and scale are obtained by assigning "zero" weights to observations with robust MDs exceeding an approximate chi-square value (0.975) with p degrees of freedom. In practice, it is observed that even after performing this extra step, some of the non-outlying observations are assigned a "zero" weight.

Scout offers some additional options to identify appropriate number of outliers using the MCD method. Instead of finding a "best" subsets of size, h = [(n+p+1)/2], one may find a "best" subset of size h= [n*k], where k represents some percentile >0.5. For example, for k = 0.75, the objective will be to find a subset with minimum determinant of the covariance matrix based upon the best subset consisting of roughly 75% (= [n*.75]) of the observations. The BD of such MCD estimates will be roughly equal 25% (= (n-h+1)/n). The MCD method in Scout is called the Extended MCD method. In order to use this option to appropriately compute the coverage, it is desirable to use graphical displays (or other robust methods) to gain some information about the number of outliers present in the data set. The BD of such MCD estimates will be roughly equal 25% (~ (n-h+1)/n). It should be noted that, the MCD estimates based upon a "best" subset consisting of a higher (> 50%) percentage of data may suffer from masking effects, especially when the data set consists of clustered data. Since all of these options are available in Scout 2008, the user is encouraged to confirm these statements and observations on data sets from their applications.

1.3.8 PROP Influence Function

The PROP influence function (Singh, 1993) represents a smooth redescending influence function assigning full weights to observations coming from the central part of data, and reduced (instead of zero weights) to negligible weights to intermediate and extreme outliers, respectively. The details of this method can be found in Singh (1993, 1996), and Singh and Nocerino (1995, 1997). Even though, theoretical BD of M-estimation methods is not greater than 1/(p+1), it is noted that the practical BD of an iteratively obtained robust M-estimate (generalized likelihood estimate) based upon PROP (Singh, 1993) influence function can be much higher than 1/(p+1). The break down point of robust estimates based upon PROP influence function increases with each iteration. By definition of the PROP influence function, the iterative process identifies multiple outliers smoothly and effectively by reducing the influence of outliers successively in various iterations. This is especially true when an initial robust start based upon OKG

(Devlin, Gnanadesikan, and Kettenring (1975), Maronna and Zamar (2002)) method is used in the iterative process of obtaining M-estimates.

In order to identify potential outliers present in a data set, the PROP function uses an influence function, α, value. Since the number of outliers present in a data set is not known in advance, it is desirable to use more than one value of the influence function, α, on the same data set. As mentioned before, the use of graphical displays is also recommended on methods available in Scout (e.g., Huber, MCD, MVT, or PROP) to get some idea about the number (or % k) of outliers that may be present in the data set; and also to confirm that identified outliers do represent outlying observations. Information gathered from the graphical displays can be used to determine an appropriate critical or influence function alpha, α ($0<\alpha<0.5$), used in Huber and PROP methods, or a trimming percentage value used in the MVT method. Higher values of α or of a trimming percentage are used to identify a larger number of outliers.

The PROP M-estimation method reduces the influence of outliers iteratively. The PROP influence function assigns unit weights to observations coming from the main central part of data (inliers) and reduced to negligible weights to intermediate and extreme outliers. The weights are reduced iteratively till the convergence of estimates if achieved. It is noted that M-estimation based upon PROP influence function performs quite effectively in identify multiple multivariate outliers. Typically, M-estimates based upon the PROP influence function (with initial OKG estimates) roughly assign: 1) full unit weight to observations coming from the central part of data (making the dominant population); 2) reduced weights to intermediate outliers (some of those may represent border line observations coming from overlapping observations); 3) and negligible weights to extreme outliers perhaps representing observations from significantly different population(s). Furthermore, those robust estimates are in close agreement with the classical estimates obtained using the data set without the outliers. The user is encouraged to confirm these observations by using Scout 2008 on his/her own application data sets.

1.4 Outliers/Estimates Module

This module offers both univariate and multivariate outlier identification and estimation methods. For univariate uncensored and left-censored data sets, Scout has some classical outlier tests such as Dixon test, Rosner test, and Grubbs test. For univariate data sets, this module also has Tukey's Biweight (and its variation suggested by Kafadar (1982)) outlier identification and estimation method. Several other univariate robust methods are available as special cases of multivariate robust methods. Multivariate (can also be used on univariate data) outlier identification and estimation methods included in Scout are: sequential classical methods based upon Max-MD and kurtosis; iterative robust and resistant M-estimation methods based upon Huber and PROP influence functions, multivariate trimming (MVT), and re-weighted fast MCD (extended) method. For all iterative robust methods (including Biweight method) in various modules of Scout, several choices (described earlier) for initial estimates of location and scale are available.

1.4.1 Coverage and Influence Function Levels in Robust Outlier Identification Methods

It should be pointed out that the success of a robust method in identifying multiple outliers depends upon the coverage (e.g., h in MCD method) or cutoff levels (e.g., influence function alpha in PROP M-estimation method) and the behavior of the influence function (nondecreasing, redescending, smooth redescending) used to identify those outliers. For an illustration, the MCD method uses the half samples of size h, where the coverage factor, h is typically given by h = [(n+p+1)/2], M-estimation methods based upon PROP and Huber influence functions use a critical or influence function cutoff level, α, and MVT method uses a trimming percentage, α% to identify outliers in p-dimensional data sets of size n. In addition to coverage and critical cutoff levels, initial robust start estimates in the iterative process (e.g., M-estimation) of obtaining robust estimates also play an important role in achieving high break down estimates. It should be noted that there is a direct relationship between the coverage or influence cutoff and the break down (BD) point of an estimate. Specifically, for the MCD (and also LMS regression method) method, higher values of h yield MCD estimates with lower BD points; for influence function based M-estimation methods (e.g., PROP influence function), higher values of influence function, α yield estimates with higher BD points, and for MVT method, higher values of trimming percentage tend to yield estimates with higher BD points.

As a rule of thumb, for appropriate identification of outliers, n should be at least 5p; this is especially true when dimension, p>5. From theoretical point of view, Scout can compute various robust statistics and estimates for values of n > (p+2). However, as well knows, the results (estimate, graphs, and outliers) obtained using such small high dimensional (curse of dimensionality) data sets may not always be reliable and defensible.

1.4.2 Outlier Determination Critical Alpha

In addition to coverage or influence function cutoff levels, all of the outlier methods use a critical level (outlier critical alpha) which is used to determine outliers. Critical values of various test statistics used in all graphical (e.g., Q-Q and index plots, ellipsoids) and outlier identification methods (e.g., MDs, Max-MDs, kurtosis, skewness) are computed using this critical alpha. For an example, MCD method uses a default chi-square (with p degreed of freedom = *df*) cutoff alpha level=0.025 for determination of outliers. Observations with MCD MDs exceeding chi-square (0.975) cutoff with p *df* may represent potential outliers. Similarly, other multivariate outlier methods in Scout including classical, sequential classical, and M-estimation methods (PROP, Huber) use an outlier alpha (user selected) that is used to compute critical values of the test statistics (individual MD, or Max MD) used to determine outliers. Classical and robustified MDs exceeding those critical values may represent potential outliers requiring further investigation.

1.5 QA/QC Module

This module provides univariate and multivariate classical as well as robust methods that can be used in quality assurance and quality control (QA/QC) applications. All classical and robust options and methods available in univariate Interval Module (under Stats/GOF) and Outliers/Estimates module are available in QA/QC module. Specifically, QA/QC module has univariate control-chart-type interval graphs; multivariate control-chart-type index plots; and prediction and tolerance ellipsoids. These graphs can be generated using all observations in a data set or just using observations in a specified training (e.g., background data, placebo) subset data set. These graphs can be used to compare test (site, project, new drug) data with control limits (e.g., prediction, tolerance, simultaneous limits) computed based upon some training (background, reference, controlled) data set. Specifically, this module can be used to compare training (background, reference, upgradient wells) and test (polluted site, groundwater monitoring wells, dredged sediments) data sets. Enough observations from the training data set should be made available to compute defensible control limits and ellipsoids.

The training and test data option is specifically useful to determine if observations from one test group (e.g., polluted site, test group, new treatment) can be considered as coming from the training group (e.g., reference group, background, training group, placebo) perhaps with known well-established acceptable behavior of the contaminant concentrations of potential concern (COPCs). For such graphical displays, relevant statistics and limits are computed using training (controlled, background, reference, placebo) data set, and all points in training and test data sets are plotted on those graphical displays. Test data points (site observations) lying outside the limits (e.g., tolerance and simultaneous limits) may represent out-of-control observations, that is may represent observations not belonging to the controlled population represented by the training data set.

Classical methods included in QA/QC module can handle univariate and multivariate data sets with non-detect observations. For univariate data sets with NDs, the estimates of all relevant statistics (mean, *sd*, standard error of the mean, upper and lower limits) are computed using the Kaplan Meier (1958) method. The individual ND data points displayed on the interval graphs are shown (in red color) based upon the user selected option (e.g., replaced by DL, DL/2, and ROS estimates). KM method is also used to compute relevant multivariate statistics (e.g., mean vector, covariance matrix, prediction and tolerance ellipsoids) based upon training data set. Those KM statistics are used to generate univariate or multivariate control- chart-type graphs. All data (raw or processed) including the imputed data (for NDs) from both training and test data sets are plotted on those control-chart-type graphs. Processed data may represent Mahalanobis distances (used in control-chart-type index plot) or principal component scores (used in prediction or tolerance ellipsoids). It should be noted that for uncensored data sets, classical estimates of location and scale should be in agreement with respective KM estimates.

1.6 Regression Module

Scout can perform multiple linear classical and robust regression using several methods available in the literature. Specifically, Scout can perform least median of squared (LMS) regression as well least percentile of squared (LPS) regression as described earlier in this chapter. Scout can also perform robust regression based upon M-estimation procedure for MVT, and Huber, Biweight, and PROP influence functions. This module generates several formalized graphical displays including Q-Q plot and index plot of residuals with appropriate limits drawn at the critical values of residual unsquared Mahalanobis distances (univariate); scatter plots of residuals versus unsquared leverage distances (Singh and Nocerino (1995)), residual versus residual (R-R) plots, Y versus Y-hat, and Y versus standardized residuals plots. It should be pointed out that residuals are not standardized when the scale estimate (standard deviation of residuals) is very small such as less than 1e-10. The graphical displays included in Scout are useful to identify: regression outliers, inconsistent (bad) leverage points; and distinguish between good (consistent) and bad (inconsistent) leverage points. For most of the graphical displays listed above, Scout 2008 collects and uses user selected critical levels to compute appropriate critical values of statistics plotted (e.g., critical values of MDs, critical value of Max MD) in graphical displays. Scout also generates confidence and prediction bands around fitted regression models including classical linear, quadratic, and cubic; and robust linear models. For the sake of completeness, in addition to robust regression methods, Scout also performs regression diagnostics.

1.6.1 Robust Regression Based Upon M-Estimation and Generalized M-Estimation

Scout can perform robust regression with or without the leverage option. If the leverage option is not used, then iterative M-estimation procedure is used directly on residuals; and when leverage option is used, the generalized M-estimation method is used. In generalized M-estimation method, leverage points (outliers in X-space of explanatory variables) are identified first; and weights thus obtained are used in the first iteration to identify regression outliers (e.g., Singh and Nocerino, 1995). Typically, in practice not all leverage points are regression outliers. It is observed that the generalized M-estimation regression method (e.g., PROP influence function) works quite effectively in identifying regression outliers, and distinguishing between good and bad leverage points. The user may want to use both options (leverage and no leverage) supplemented with graphical displays on a given data set and compare relevant regression statistics (e.g., coefficient of determinations, residual scale estimates, standard errors of estimates of regression coefficients) thus obtained to determine the best multiple linear model fit.

1.7 Principal Component Analysis (PCA) and Discriminant Analysis (DA)

Scout 2008 can perform classical as well as robust principal component and discriminant analyses. The details of robust PCA and DA based upon the MVT method, the PROP and the Huber (Huber, 1981, Gnanadesikan and Kettenring, 1981) influence functions are given in Singh and Nocerino (1995). Additional details about robust PCA and robust discriminant analyses can be found in Campbell (1972), Hubert and Driessen (2002), Hubert and Engelen (2006), Hubert, Rousseeuw and Branden (2005), and Todorov and Pires (2007).

For uncensored data sets without non-detect observations, Scout can perform classical PCA and robust PCA based upon M-estimation methods (e.g., PROP, Huber, MVT), and MCD method. PCA can be performed using covariance as well as correlation matrices. Often for large dimensional data sets, PCA is used as a dimension reduction technique, where future statistical analyses are performed on a much smaller (than p original variables) number, k (<=p) of PCs.

- It is noted that PCA performed using covariance matrix is more informative, especially when PCA is to be used as a dimension reduction technique.

- Q-Q plots and scatter plots of PC scores obtained using the covariance matrix may be used to identify potential outliers. Significant jumps and turns in Q-Q plot of PCs suggest the presence of multiple populations in the data set.

- Q-Q plots and scatter plots of PC scores based upon the correlation matrix may be used to assess approximate multinormality (cautiously).

Based upon the PC statistics and scores thus obtained, this module generates Scree and Horn plots for the eigen values, Scatter plots of PC scores, normal Q-Q plots of PC scores. One can store PC scores in the same or a different worksheet for future analyses. PCA is often used dimension reduction techniques. Typically, first few PCs explain most of the variation that might be present in a data set. The Q-Q plots of the first few PCs and scatter plots of first few PCs can be used to identify variance inflating outliers and/or to identify the presence of mixture data sets. One can draw prediction and tolerance ellipsoids on scatter plot of PC scores.

For multivariate data sets with NDs, not much guidance is available in the statistical literature on how to perform PCA. This topic is still under investigation. Scout 2008 can be used to perform PCA based upon based upon Kaplan-Meier (1958) method (still being investigated). Using the KM covariance (correlation) matrix, one can generate Scree and Horn Plots. For exploratory purposes, one can also impute PC scores based upon KM covariance matrix. However, in order to compute load matrix and PC scores, one needs to replace ND observations with some imputed values. Scout offers several choices for computing such PC scores. These methods include substitution methods (0, DL/2, and DL, uniform random generation of NDs), and regression on order statistics (ROS) methods. It should be noted that for exploratory purposes, one may want to use Data

module of Scout to impute non-detect observations before using PCA module. This step will yield a full data set without any ND observations (NDs replaced by imputed/substituted values). One can then use any of the classical and robust PCA methods available in Scout.

Scout 2008 can be used to perform classical and robust (based upon MVT, PROP and Huber influence functions) Fisher linear discriminant analysis (FDA), linear and quadratic discriminant analyses. The classical and robust DA methods are supplemented with graphical displays. The available graphical displays include Scree plots of eigen values and scatter plots of discriminant scores (for Fisher Discriminant Analysis) and original variables used to perform discriminant analysis. On scatter plots of discriminant scores, Scout can draw prediction and/or tolerance ellipsoids. As with all other graphical displays with group assignment options, on scatter plots of discriminant scores, one can reclassify an observation from one group into another group interactively by change group and save changes options. This option can be quite useful for properly classifying border line observations. It should be noted that based upon the discriminant functions (classical or robust), Scout can be used to plot and classify observations with unknown (or new) group memberships into one of the groups used in deriving those discriminant functions.

Several cross validation (CV) methods for DA are also available in Scout 2008. The CV methods in Scout 2008 include: leave-one-out (Lachenbruch and Mickey (1968)), split samples (training and test sets), M-fold CV and bootstrap methods (e.g., Davison and Hall (1992), Bradley and Efron (1997)). In order to use the CV methods properly, the user should make sure that enough data are available in each of the various groups included in the data set.

1.8 Output Generated by Scout 2008

All modules of scout either generate graphical output displays (*.gst file), or Excel-type-spreadsheets (*.ost file), or both graphical displays and excel-type-spreadsheets. The "ost" output file generated by Scout can be saved as an Excel file; and "gst" graphical display can be copied into a Word or WordPerfect file. All of the relevant information, statistics, classical and robust estimates of parameters of interest are displayed on those output sheets. Specifically, all classical estimates, initial robust estimates, final robust estimates, and associated weights are displayed on the output sheet generated by Scout. The user can also save intermediate results in a separate spreadsheet by choosing the Intermediate Iterations option. In addition to graphs, most graphical displays also exhibit relevant estimates, test statistics and associated critical levels and p-values.

1.9 Installing and Using Scout

1.9.1 Minimum Hardware Requirements

- Intel Pentium 1.0 GHz
- 285 MB (396 MB including Scout 2008 resources) of hard drive space
- 512 MB of memory (RAM)
- CD-ROM drive
- Windows 98 or newer. Scout was thoroughly tested on NT-4, Windows 2000, and
- Windows XP operating systems. Limited testing has been conducted on Windows ME.

1.9.2 Software Requirements

Scout has been developed in the Microsoft .NET Framework using the C# programming language. As such, to properly run Scout, the computer using the program must have the .NET Framework pre-installed. The downloadable .NET files can be found at one of the following two Web sites:

- http://msdn2.microsoft.com/en-us/netframework/default.aspx
  Note: *Download .NET version 1.1*

- http://www.microsoft.com/downloads/details.aspx?FamilyId=262D25E3-F589-4842-8157-034D1E7CF3A3&displaylang=en

The first Web site lists all of the downloadable .NET Framework files, while the second Web site provides information about the specific file(s) needed to run Scout. Download times are estimated at 57 minutes for a dial-up connection (56K), and 13 minutes on a DSL/Cable connection (256K).

1.9.3 Installation Instructions

Scout 2008 v. 1.00.01 Installation Instructions from the CD

Open Windows Explorer and create a new directory called Scout 2008 v. 1.00.01.

Download (save) the Scout 2008 v. 1.00.01 files from the CD to the Scout 2008 v. 1.00.01 directory.

Using Windows Explorer, right click on the Scout 2008 v. 1.00.01 main directory and make sure that the read-only attribute is off.

Using Windows Explorer, create a shortcut (optional) by right-clicking on the file, Scout.exe (application), in the Scout directory; left click on "Send To" and left click on "Desktop (create shortcut)" to create a shortcut icon the desktop (optional: rename to Scout 2008 v. 1.00.01).

Using Windows Explorer, start Scout 2008 v. 1.00.01 by left double-clicking on the file, Scout.exe (application), in the Scout directory, or by left double-clicking on the Scout shortcut icon on the desktop, or by using the RUN command from the Start Menu to locate and run Scout.exe.

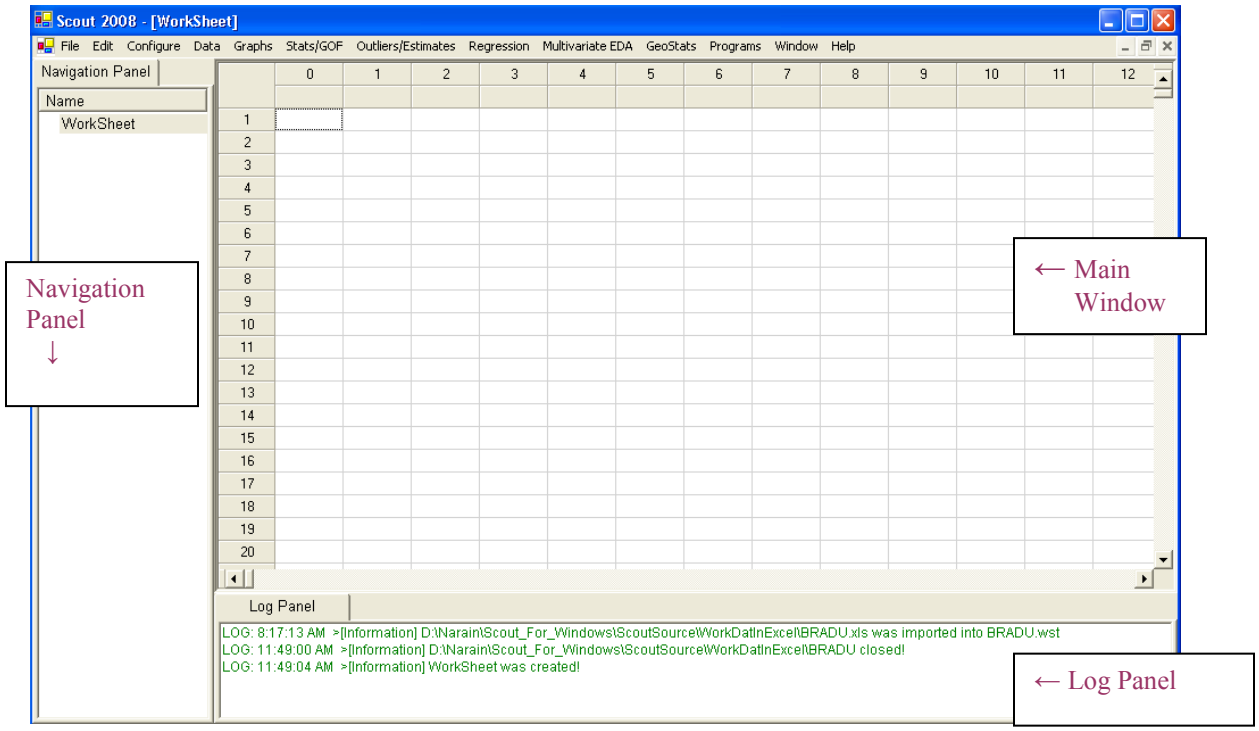
Try to open an example file in the Scout sub-directory, Data. If the file does not open, be sure that the read-only attribute is off (right-click on the Data sub-directory).

If the computer does not have .NET Framework 1.1 installed (either a pre-2002 Windows operating system or a late version of Windows XP), then it will be necessary for the end user to download it from Microsoft. A Google search for "NET Framework 1.1" will yield several download locations.

1.9.4 Getting Started

The functionality and the use of the methods and options available in Scout have been illustrated using "Screen Shots" of output screens generated by Scout. Scout uses a pull-down menu structure, similar to a typical Windows program.

The screen below appears when the program is executed.



The screen consists of three main window panels:

- The **MAIN WINDOW** displays data sheets and outputs from the procedure used.

- The **NAVIGATION PANEL** displays the name of data sets and all generated outputs.
    - At present, the navigation panel can hold at most 20 outputs. In order to see more files (data files or generated output files), one can click on Widow Option.

- The **LOG PANEL** displays transactions in green, warnings in orange, and errors in red. For an example, when one attempts to run a procedure meant for censored data sets on a full-uncensored data set, Scout will print out a warning message in orange in this panel.

    - Should both panels be unnecessary, you can click **Configure ► Panel ON/OFF**.

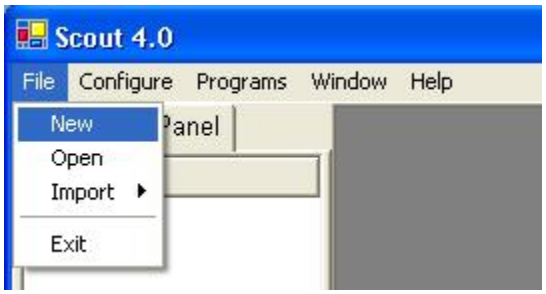
The use of this option will give extra space to see and print out the statistics of interest. For an example, one may want to turn off those panels when multiple variables (e.g., multiple Q-Q plots) are analyzed and GOF statistics and other statistics may need to be captured for all of the variables.

Chapter 2

Working with Data, Graphical Output, and Non-Graphical Output

2.1 Creating a New Spreadsheet (Data Set)

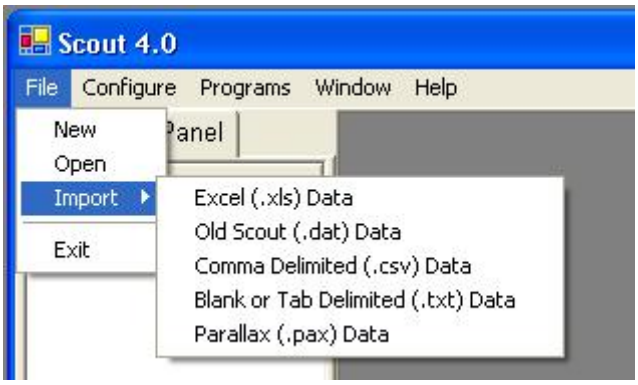
To create a new worksheet: click **File ► New**



2.2 Open an Existing Spreadsheet (Data Set)

If your data sets are stored in the Scout data format (*.wst), Scout output format (*.ost), Scout graphical format (*.gst) or an Excel spreadsheet (*.xls), then click **File ► Open**.

- If your data sets are stored in the Microsoft Excel format (*.xls), or in the DOS-Scout format (*.dat) or ParallAX format (*.pax), then choose **File ► Import ►Excel** or **Old Scout** or **ParallAX.**



- Make sure that the file that you are trying to import is not currently open. Otherwise, there will be the following warning message in the Log panel:

*"[Information] Unable to open C:***.xls." Check the validity of this file.*

Note: *.csv files and *.txt files will be available in later versions of Scout.

2.3 Input File Format

- The program can read Excel files (*.xls files), data files (*.dat files for DOS versions of GeoEas and Scout software packages), ParallAX files (*.pax files), comma delimited data files (*.csv files), and tab or space delimited files (*.txt files).

- The user can perform typical Cut, Paste, and Copy operations, as in Microsoft Excel.

- The first row in all input data files should consist of alphanumeric (strings of numbers and characters) variable names representing the header row. Those header names may represent meaningful variable names such as Arsenic, Chromium, Lead, Temperature, Weight, Group-ID, and so on.

    - The Group-ID column has the labels for the groups (e.g., Background, AOC1, AOC2, 1, 2, 3, a, b, c, Site1, Site2, and so on) that might be present in the data set. The alphanumeric strings (e.g., Surface, Sub-surface) can be used to label the various groups.

    - The data file can have multiple variables (columns) with unequal number of observations. NOTE: Some of the robust methods require all of the variables to have an equal number of observations.

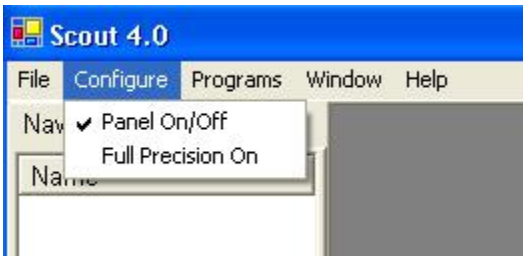
    - Except for the header row and columns representing the group labels, only numerical values should appear in all of the other columns.

    - All of the alphanumeric strings and characters (e.g., blank, other characters, and strings), and all of the other values (that do not meet the requirements above) in the data file are treated as missing values.

    - Also, a large value denoted by 1E31 (= $1x10^{31}$) can be used to represent missing data values. All of the entries with this value are ignored from the computations. Those values are counted when missing data values are tracked.

2.4 Number Precision

- You may turn Full Precision on or off by choosing: **Configure ► Full Precision On/OFF.**

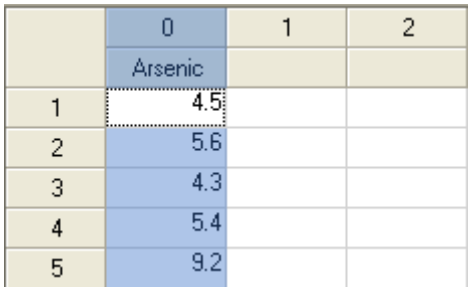


- By leaving the Full Precision turned on, Scout will display numerical values using an appropriate (the default) decimal digit option. However, by turning the Full Precision off, all of the decimal values will be rounded to the nearest thousandths place.

- Full Precision On option is specifically useful when one is dealing with data sets consisting of small numerical values (e.g., <1) resulting in small values of the various estimates and test statistics. Those values may become very small with several leading zeros (e.g., 0.00007332) after the decimal. In such situations, one may want to use the Full Precision option to see nonzero values after the decimal.

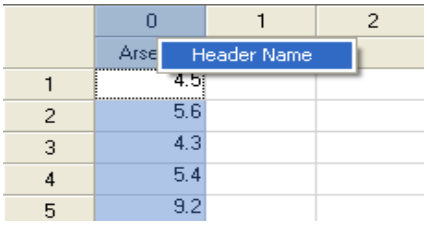
2.5 Entering and Changing a Header Name

1. Highlight the column whose header name (variable name) you want to change by clicking either the column number or the header as shown below.



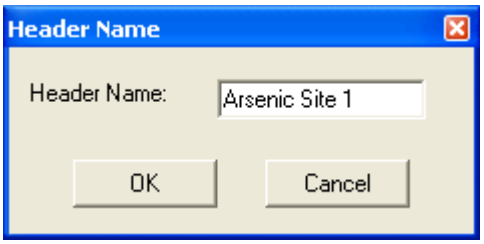
	0	1	2
	Arsenic		
1	4.5		
2	5.6		
3	4.3		
4	5.4		
5	9.2		

2. Right-Click and then click "**Header Name**"



	0	1	2
	Arse	Header Name	
1	4.5		
2	5.6		
3	4.3		
4	5.4		
5	9.2		

3. Change the Header Name.

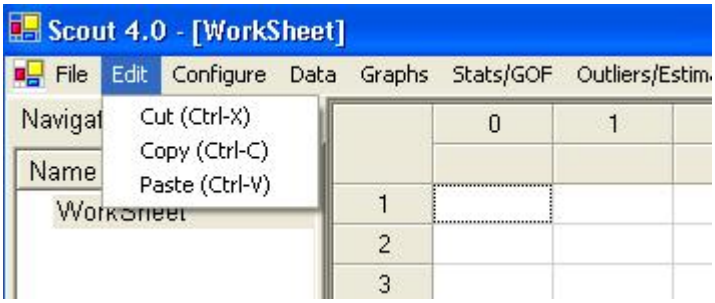


4. Click the "**OK**" button to get the following output with the changed variable name.

	0	1	2
	Arsenic Site 1		
1	4.5		
2	5.6		
3	4.3		
4	5.4		
5	9.2		

2.6 Editing

Click on the Edit menu item to reveal the following drop-down options.



The following Edit drop-down menu options are available:

- Cut option: similar to a standard Windows Edit option, such as in Excel. It performs standard edit functions on selected highlighted data (similar to a buffer).
- Copy option: similar to a standard Windows Edit option, such as in Excel. It performs typical edit functions on selected highlighted data (similar to a buffer).
- Paste option: similar to a standard Windows Edit option, such as in Excel. It performs typical edit functions of pasting the selected (highlighted) data to the designated spreadsheet cells or area.
- Note that the Edit option could also be used to Copy Graphs.

2.7 Handling Non-detect Observations

Scout can handle data sets with single and multiple detection limits.

For a variable with non-detect observations (e.g., arsenic), the detected values, and the numerical values of the associated detection limits (for less than values) are entered in the appropriate column associated with that variable.

Specifically, the data for variables with non-detect values are provided in two columns. One column consists of the detected numerical values with less than ($< DL_i$) values entered as the corresponding detection limits (or reporting limits), and the second column represents their detection status consisting of only 0 (for less than values) and 1 (for detected values) values. The name of the corresponding variable representing the detection status should start with d_, or D_ (not case sensitive) and the variable name. The detection status column with variable name starting with a D_ (or a d_) should have only two values: 0 for non-detect values, and 1 for detected observations.

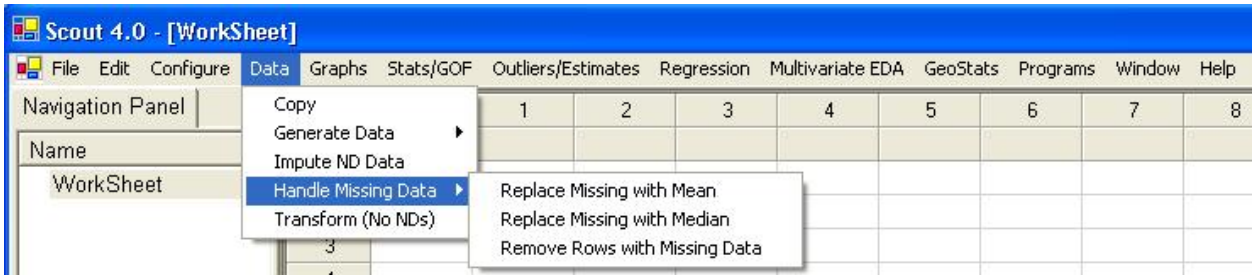
For an example, the header name, D_Arsenic, is used for the variable, Arsenic having non-detect observations. The variable D_Arsenic contains a 1 if the corresponding Arsenic value represents a detected entry, and contains a 0 if the corresponding entry for variable, Arsenic, represents a non-detect.

There should not be any missing value in the non-detects column. If there exists an observation with no indication of "0" or "1" in the non-detects column, then that observation should be deleted if the various methods for non-detects are to be used. Otherwise the methods for detected data (i.e., methods which do not require a non-detects column) can be used.

D:\example.wst

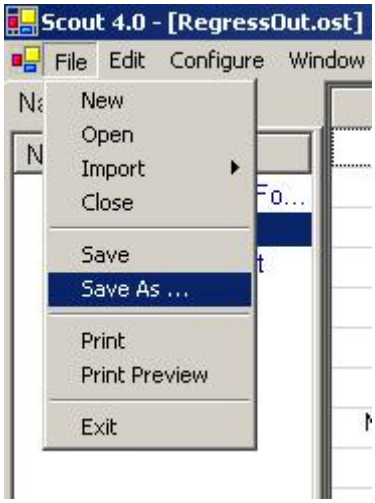
	0	1	2	3	4	5	6
	Arsenic	D_Arsenic	Mercury	D_Mercury	Vanadium	Zinc	Group
1	4.5	0	0.07	1	16.4	89.3	Surface
2	5.6	1	0.07	1	16.8	90.7	Surface
3	4.3	0	0.11	0	17.2	95.5	Surface
4	5.4	1	0.2	0	19.4	113	Surface
5	9.2	1	0.61	1	15.3	266	Surface
6	6.2	1	0.12	1	30.8	80.9	Surface
7	6.7	1	0.04	1	29.4	80.4	Surface
8	5.8	1	0.06	1	13.8	89.2	Surface
9	8.5	1	0.99	1	18.9	182	Surface
10	5.65	1	0.125	1	17.25	80.4	Surface
11	5.4	1	0.18	1	17.2	91.9	Subsurface
12	5.5	1	0.21	1	16.3	112	Subsurface
13	5.9	1	0.29	1	16.8	172	Subsurface
14	5.1	1	0.44	1	17.1	99	Subsurface
15	5.2	1	0.12	1	10.3	90.7	Subsurface
16	4.5	0	0.055	1	15.1	66.3	Subsurface
17	6.1	1	0.055	1	24.3	75	Subsurface
18	6.1	1	0.21	1	18	185	Subsurface
19	6.8	1	0.67	1	16.9	184	Subsurface
20	5	1	0.1	1	12	68.4	Subsurface
21			0.8	1			
22			0.26	1			
23			0.97	1			
24			0.05	1			
25			0.26	1			

2.8 Handling Missing Values



Section 4.4 details how missing values are treated in Scout.

2.9 Saving Files

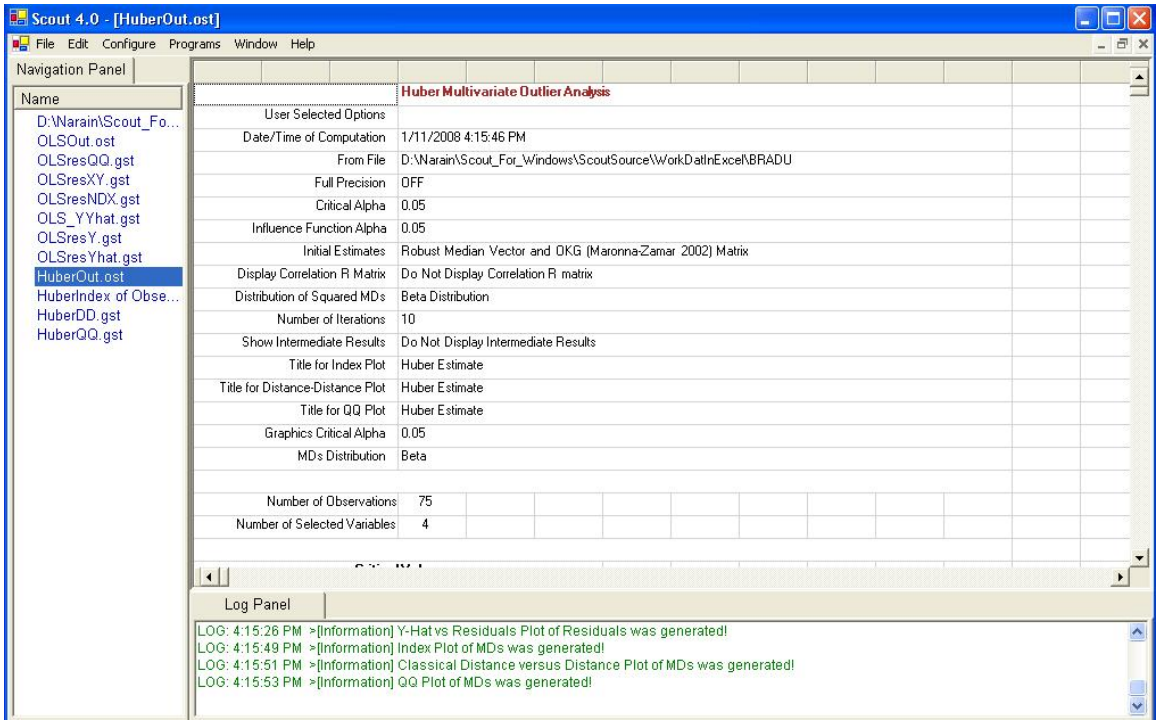


- The Save option allows the user to save the active window.

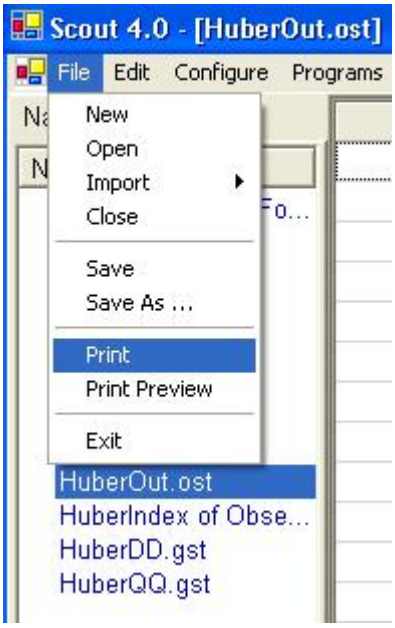
- The Save As option allows the user to save the active window. This option follows typical Windows standards, and saves the active window to a file in Excel (*.xls) format or an output sheet (*.ost) format.

2.10 Printing Non-Graphical Outputs

1. Click the output you want to copy or print in the **Navigation Panel**.



2. Click **File ► Print**.

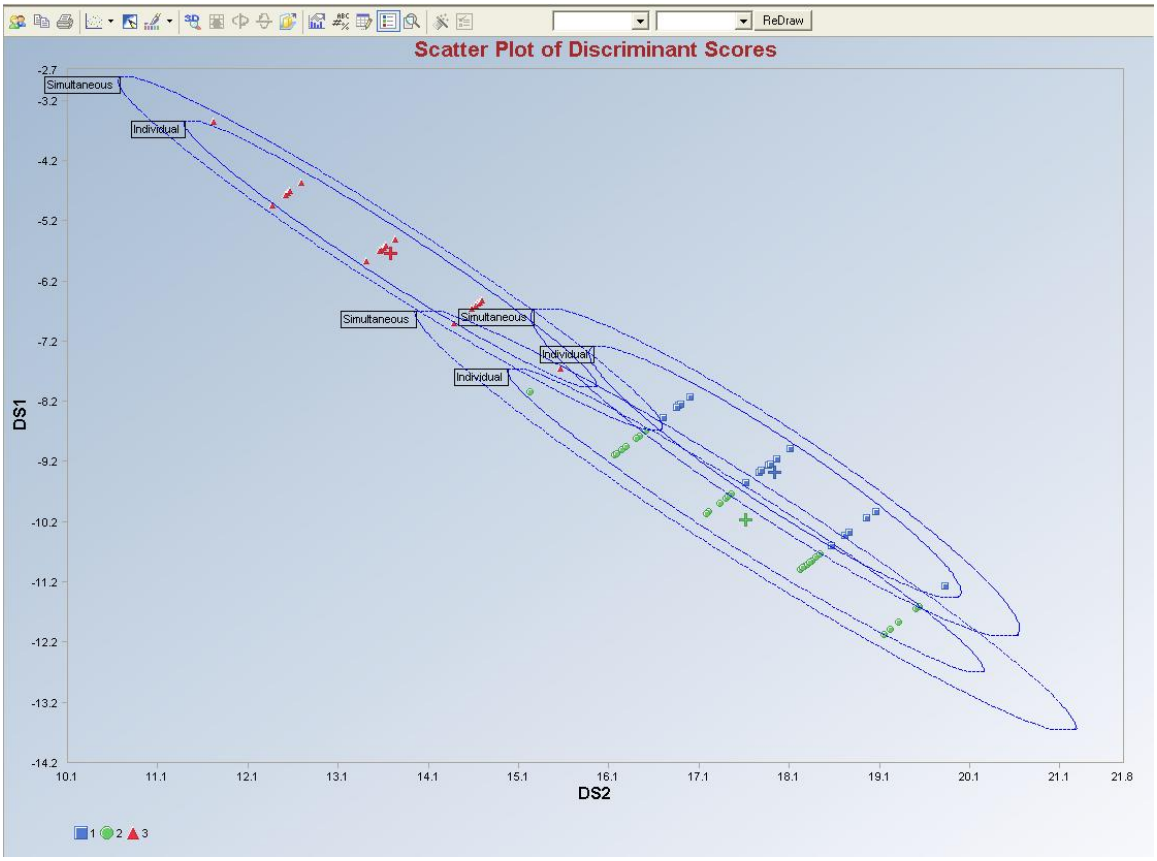


2.11 Working with Graphs

Advanced users are provided with two sets of tools to modify graphics displays. A graphics tool bar is available above the graphics display, and as the user right clicks on the desired object within the graphics display, a drop-down menu will appear. The user can select an item from the drop-down menu list by clicking on that item. This will allow

the user to make desired modifications as available for the selected menu item. An illustration is given below.

2.11.1 Graphics Toolbar



The user can change fonts, font sizes, vertical and horizontal axes, and select new colors for the various features and text. All of those actions are generally used to modify the appearance of the graphic display. The user is cautioned that those tools can be unforgiving and may put the user in a situation where the user cannot go back to the original display. Users may want to explore the robustness of those tools and become more experienced in their use before actually trying to use those graphic tools on real data sets.

Another feature in this graphics tool bar is the presence of one, two, or three drop-down variable selection boxes, depending upon the type of graph.

- The XY Plot in Regression has only one drop-down variable selection box for different X variables.

- The Scatter Plots in 2D Graphs, Principal Component Analysis, and Discriminant Analysis have two drop-down variable selection boxes for

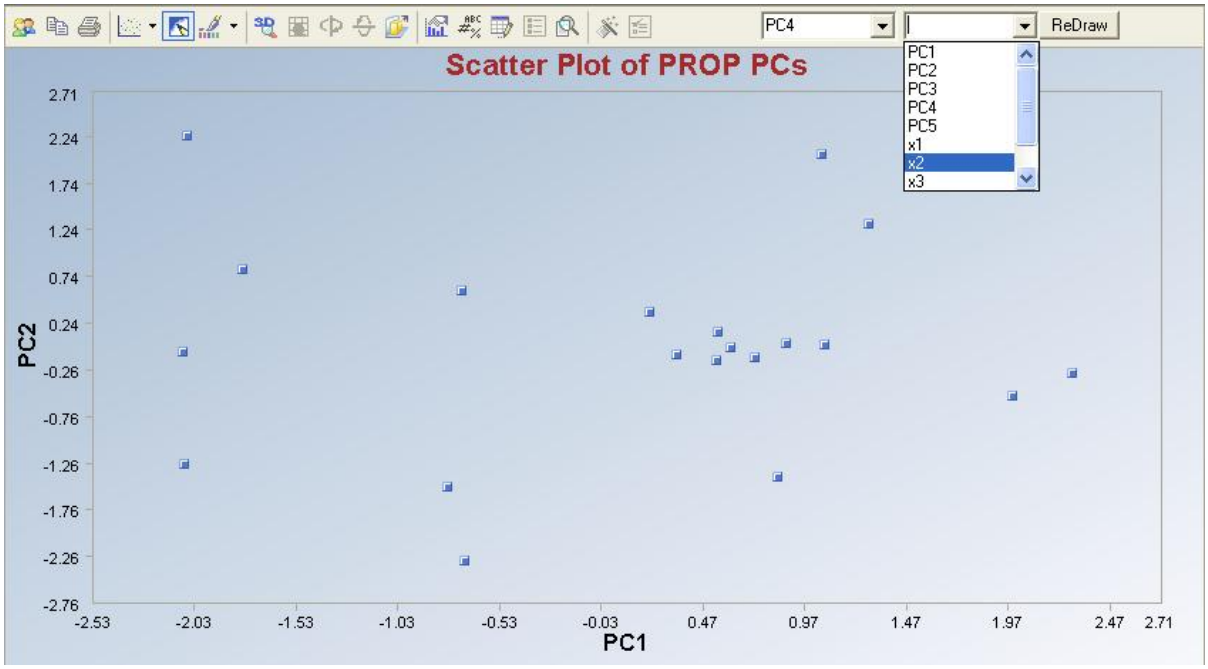
selecting different X and Y variables. The first box is for the X variable and the second box is for the Y variable.

- Scatter Plots in 3D Graphs have three drop-down variable selection boxes for selecting different X, Y and Z variables.

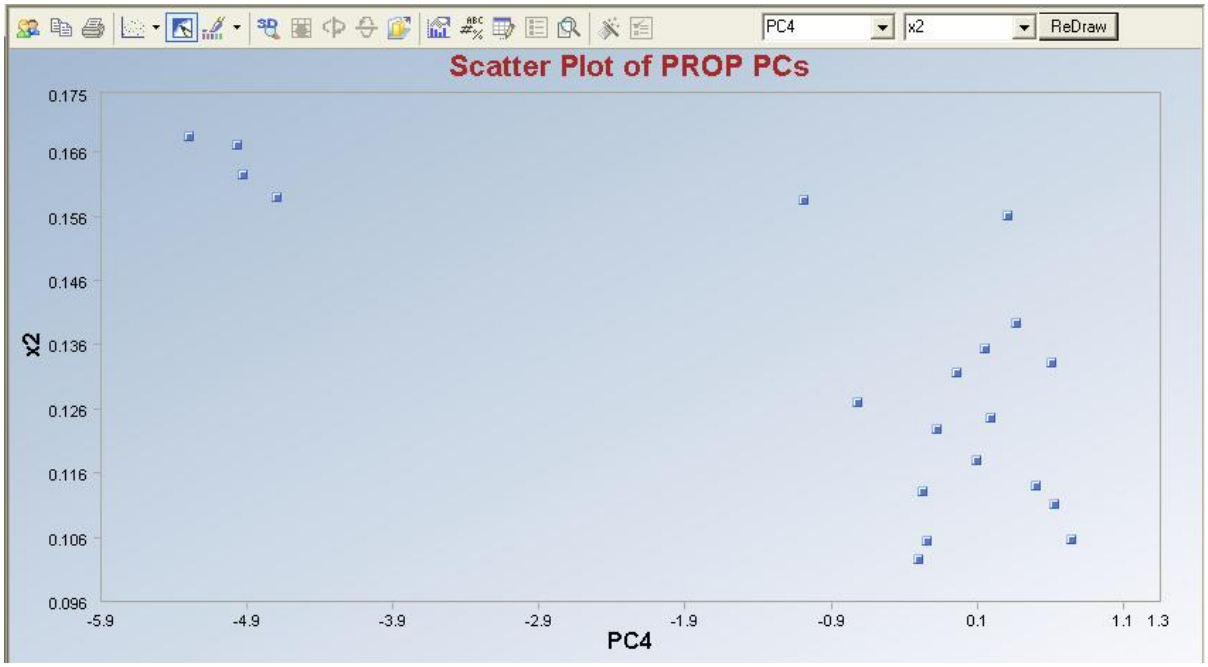
- The user can select the required variables and the new graph is obtained by clicking the "Redraw" button. An example is given below.

***Note**: One can select variables from the graph itself, as shown in the following figure.*

Graph: PROP principal components scatter plot.
Data Set used: Well-known Wood data set. All five of the X-variables were selected to derive the PCs.
Default Graph Obtained: PC1 is drawn along the X-axis and PC2 is drawn along the Y-axis.
Changing X-axis variable to PC4 and Y-axis variable to variable X2.

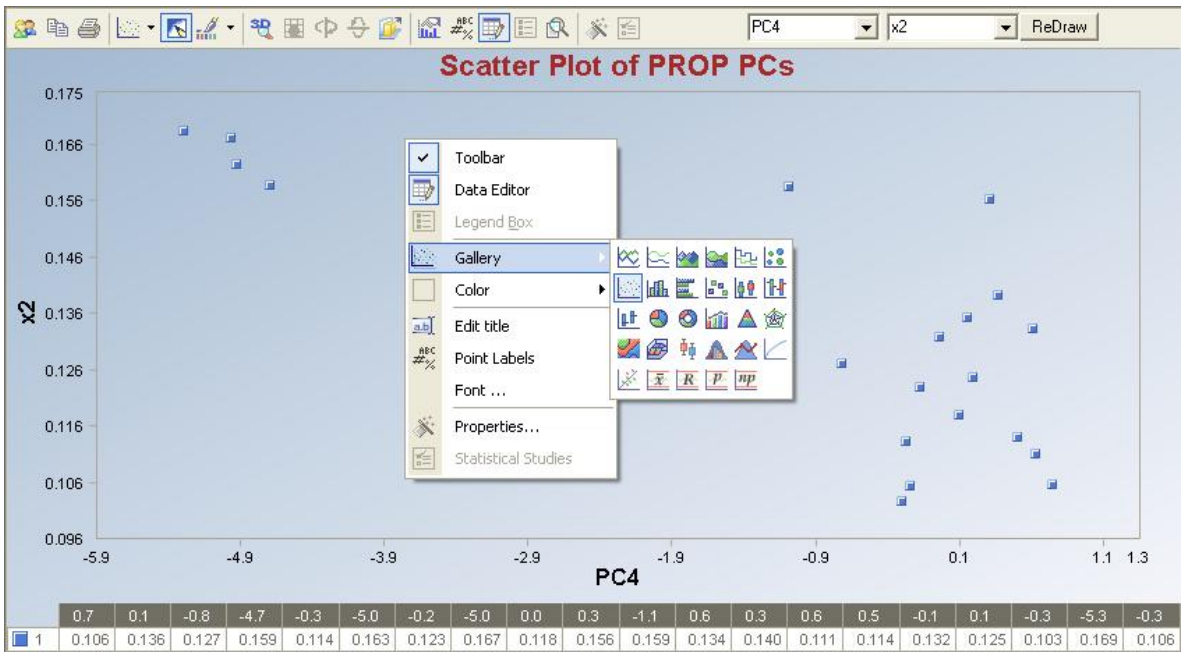


The X-axis variable is PC4 and the Y-axis variable is variable X2.



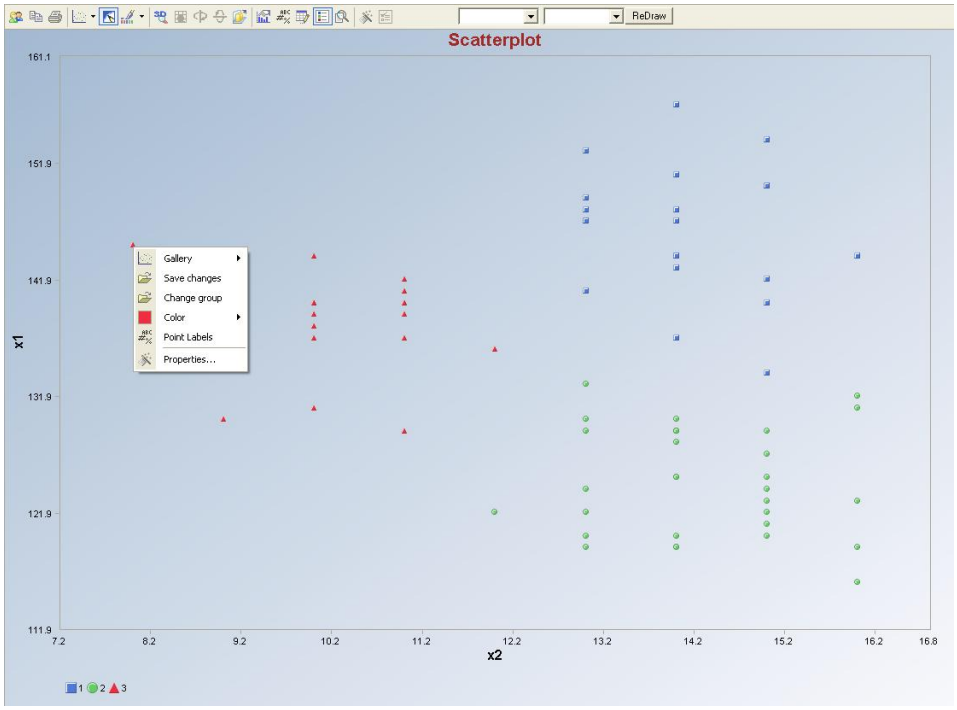
2.11.2 Drop-Down Menu Graphics Tools

Those tools allow the user to move the mouse icon to a specific graphic item such as an axis label or a display feature. The user then right clicks the mouse button and a drop-down menu appears. This menu presents the user with available options for that particular control or graphic object. If one is not careful and experienced, then there is a small risk of making an unrecoverable error when using those drop-down menu graphics tools. As a cautionary note, the user can always delete the graphics window and redraw the graphical displays by repeating their operations from the datasheet and menu options available in Scout. An example of a drop-down menu obtained by right clicking the mouse button on the background area of the graphics display is given as follows. Some of the options are: changing the color of the observations, changing the type of graph, viewing the observation numbers (**Point Labels**), and editing the title of the graph.

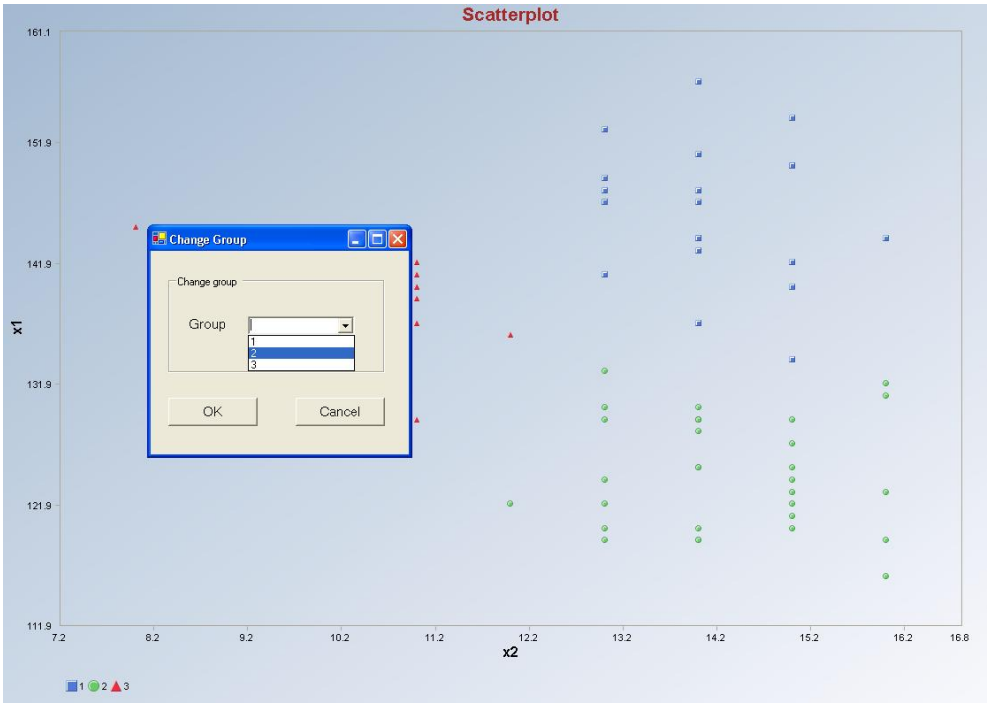


Scout provides a different Drop-Down Menu Graphic Tool in the presence of observations of various groups. This can be used to change the grouping of the observations on the graph. To perform this feature, move the mouse icon to the particular observation and click the right click button on the mouse. A menu comes up. Click the "**Change Group**" option. A window comes up with "**Change Group Drop-Down Box**." Select the new group of the observation and click "**OK**" to continue or "**Cancel**" to cancel the option. Once a selection has been made, move the mouse icon to that particular observation and click on the left mouse button. This will change the observation group assignment and the observation will belong to the new group shown on the graph.

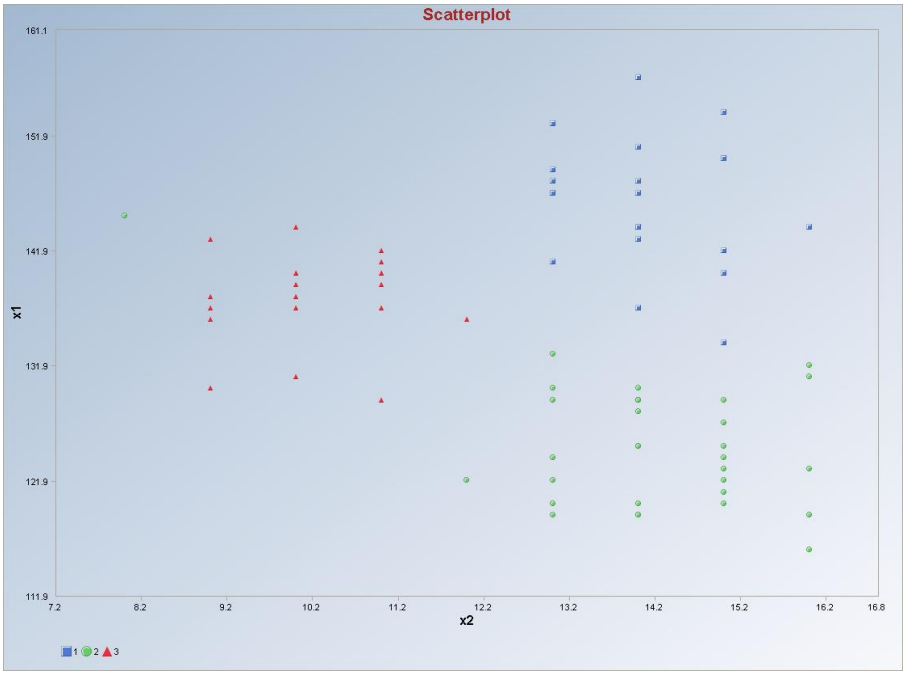
Graph of 2D scatter plot with groups from graphs.
Data Set used: Beetles.
Changing the left-most observation from Group 3 (red triangle) to Group 2 (green circle).



Change group option brings up a Change Group window, as shown below.



The left-most observation from Group 3 (red triangle) now belongs to Group 2 (green circle) on the graph.



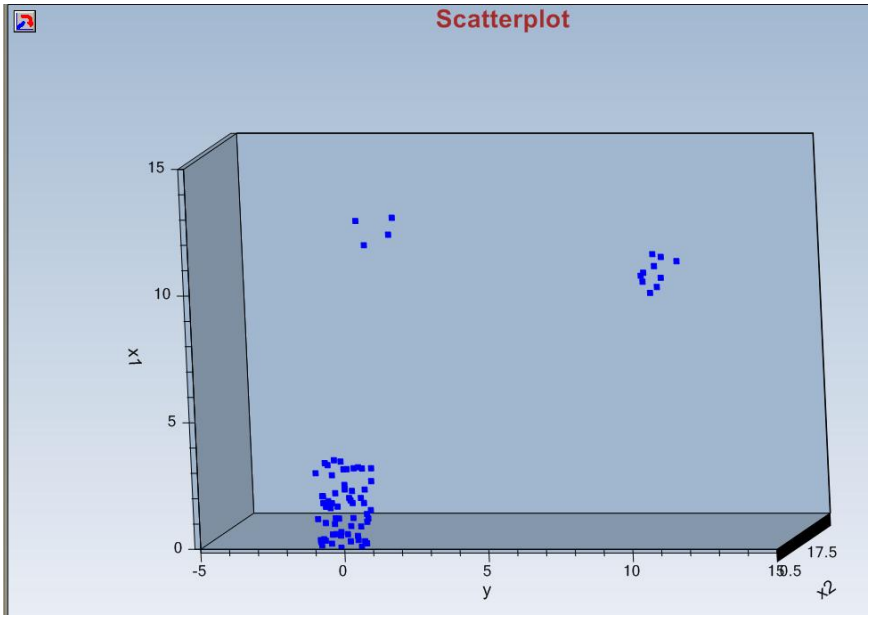
To incorporate the changes in the graph to the worksheet, click the "Save Changes" option after using the right-click button on the mouse. This saves the new grouping to the first available column on the worksheet as "**newGrp**."

Observation 53 changed from Group 3 to Group 2.

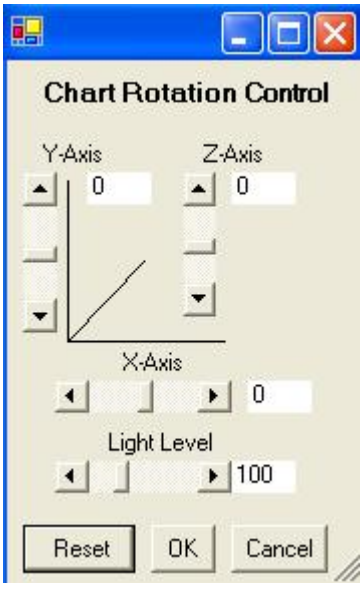
	0	1	2	3	4
	Group	x1	x2	newGrp	
43	2	124	15	2	
44	2	120	13	2	
45	2	119	16	2	
46	2	119	14	2	
47	2	133	13	2	
48	2	121	15	2	
49	2	128	14	2	
50	2	129	14	2	
51	2	124	13	2	
52	2	129	14	2	
53	3	145	8	2	
54	3	140	11	3	
55	3	140	11	3	
56	3	131	10	3	
57	3	139	11	3	
58	3	139	10	3	
59	3	136	12	3	
60	3	129	11	3	
61	3	140	10	3	
62	3	137	9	3	

2.11.3 3D Graphics Chart Rotation Control Button

The axes in a 3D scatter plot can be rotated using the Chart Rotation Control button present on the top-left corner of the 3D scatter plot.



When this Chart Rotation Control button is clicked, the Chart Rotation Control tool box appears. This tool box has three scroll bars for the three axes and a fourth scroll bar for adjusting the brightness of the graph. The scroll bars can be used to rotate any or all of the three axes. When the "**Reset**" button is clicked, the graph is reset to the standard front view. The "**Cancel**" button brings the graph to its default view.



The angle of rotation for the three axes ranges from -120 to +111 degrees. The positive sign is for rotation in clockwise direction and the negative sign counter-clockwise direction. The Light Level scroll bar ranges from 0 for black to 391 for the white (brightest) level.

References

ProUCL 4.00.04. (2009). "ProUCL Version 4.00.04 User Guide." The software ProUCL 4.00.04 can be downloaded from the web site at:

http://www.epa.gov/esd/tsc/software.htm.

Chapter 3

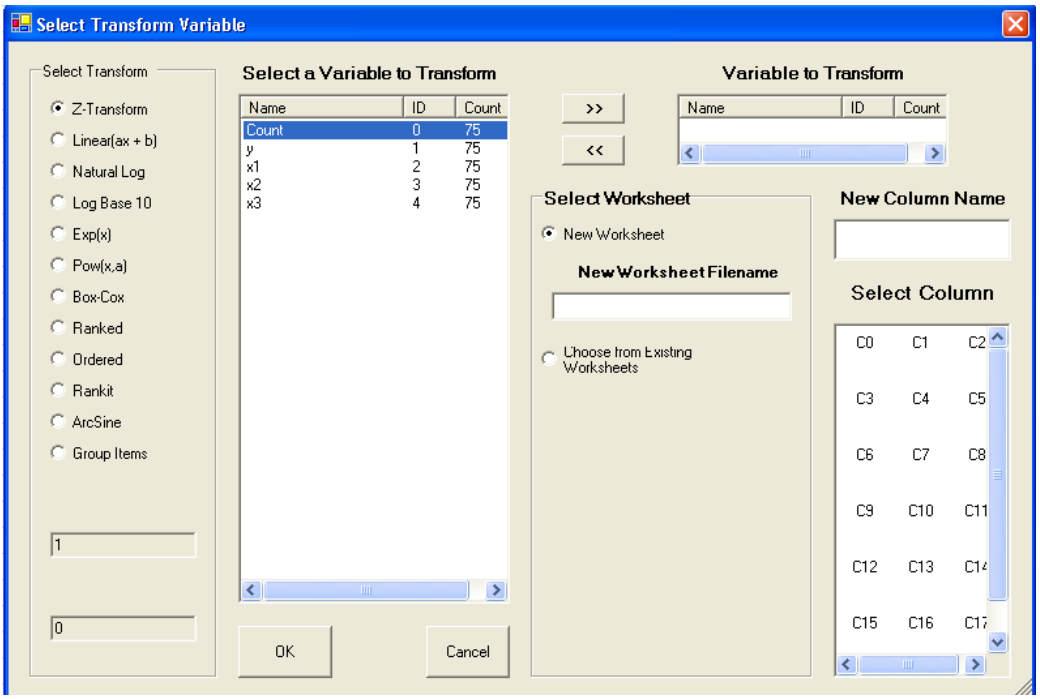
Select Variables Screens

Scout provides a number of variable selection screens for different types of statistical analysis. Most of them are illustrated here.

3.1 Data Drop-Down Menu

3.1.1 Transform (No NDs)

- When the user clicks **Data ► Transform (No NDs)**, the following window will appear:

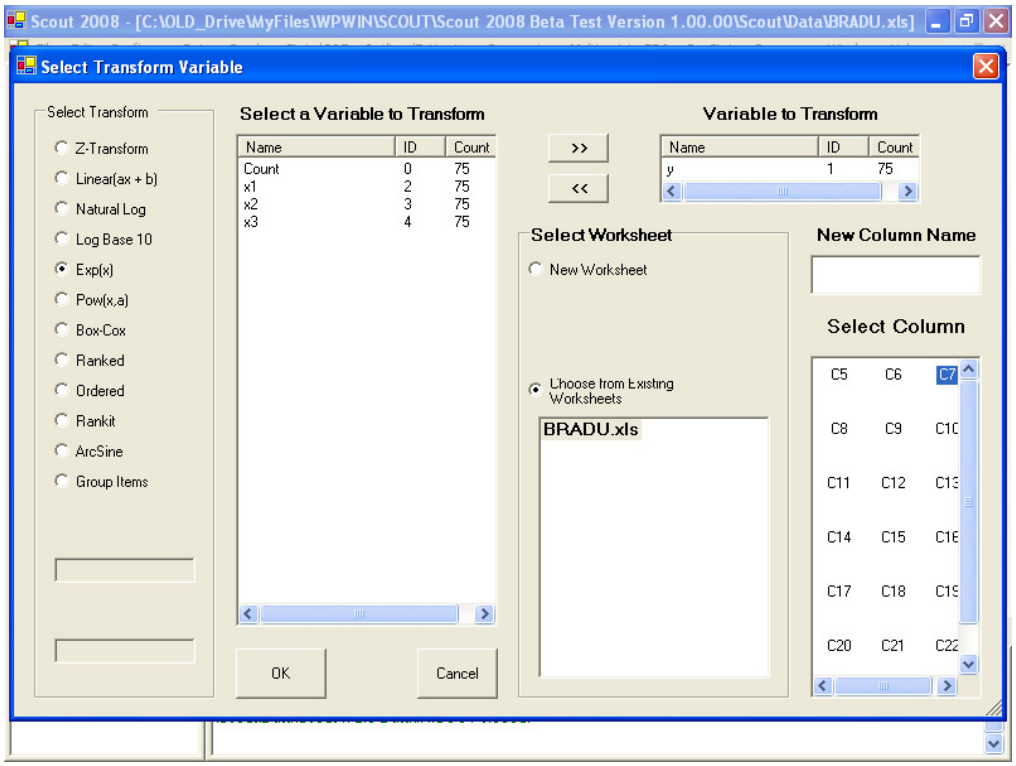


- This screen allows the user to transform a single variable. The transformations available are in the "**Select Transform**" box.

- A single variable is selected and that variable appears in the "**Variable to Transform**" box.

- The user can select the worksheet to store the transform using the "**New Worksheet**" or the "**Other Worksheets**" and a set of available columns appear in the "**Select Column**" box. The user has to specify a name for the new column.

- An example of the selections made is shown below.



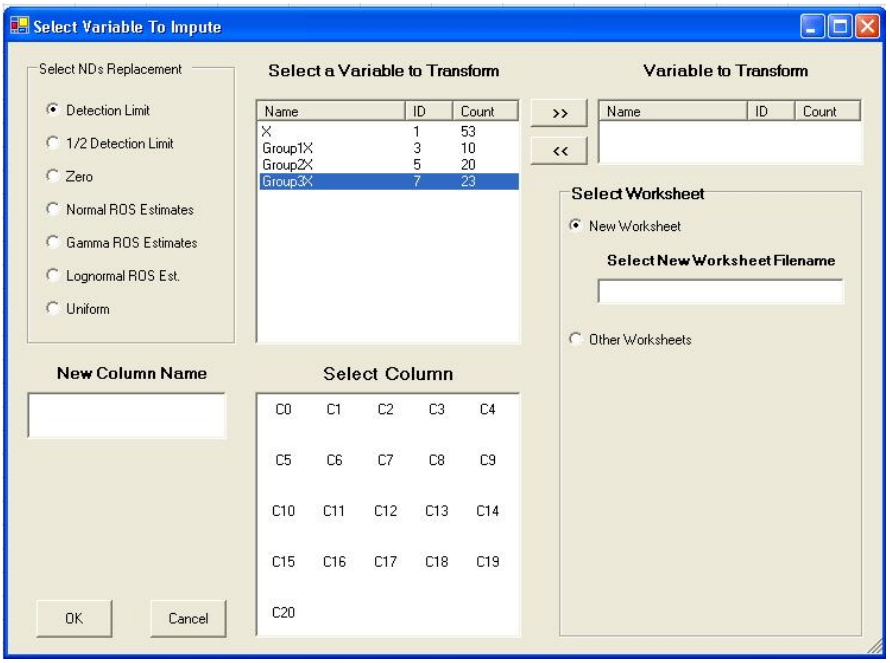
3.1.2 Impute: Transform Two Columns to a Column (NDs)

- When the user clicks Data ► Impute (NDs) the window given below will appear.

- This selection screen comes up only for data sets having non-detects. If the file does not have columns for indicating non-detects, then an error message is displayed in the Log Panel.

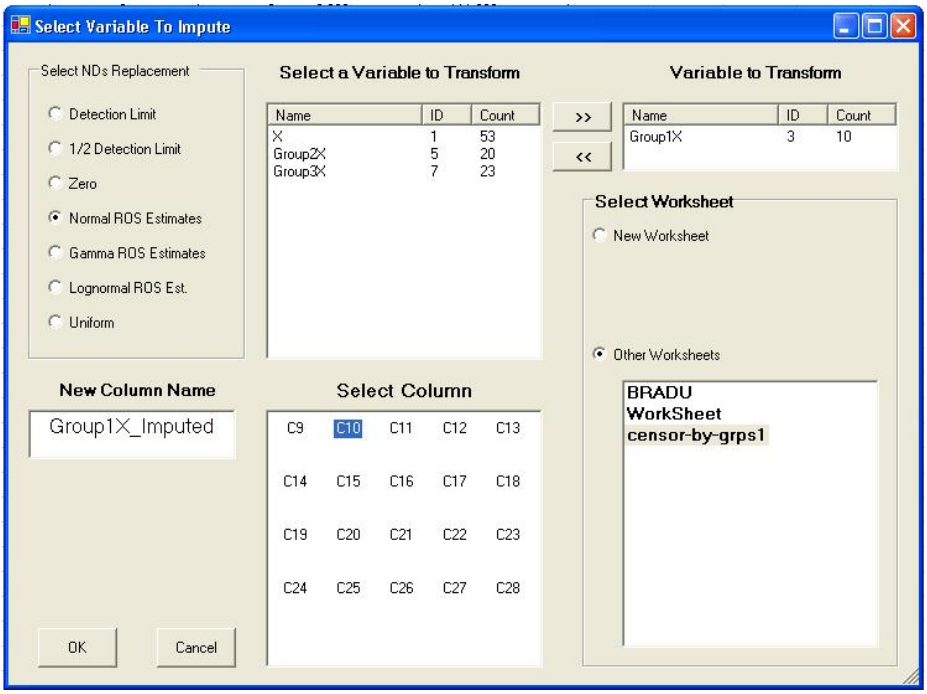
- This screen allows the user to transform a single variable. The transformations available are in the "**Select Transform**" box.

- A single variable is selected and that variable appears in the "**Variable to Transform**" box.



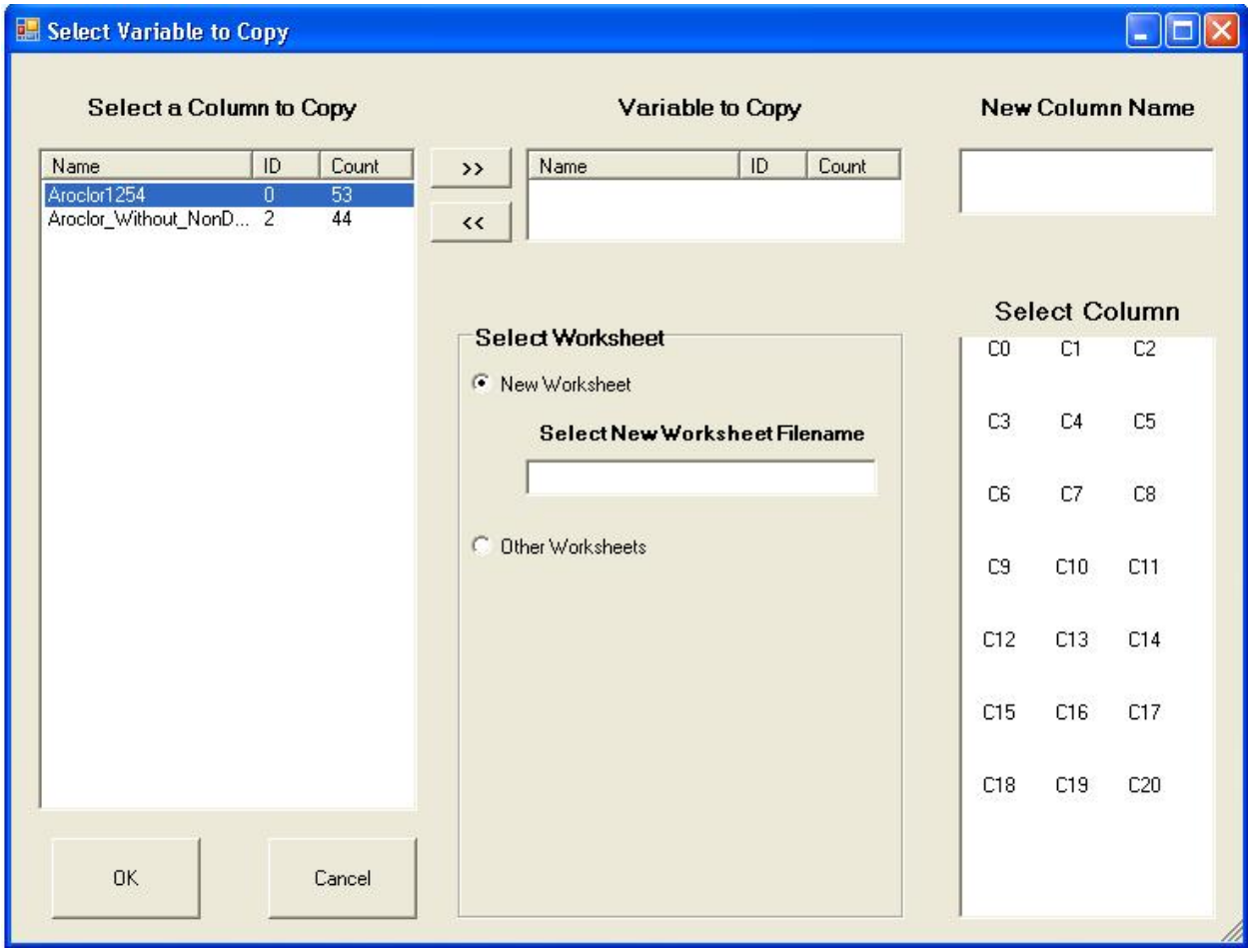
- The user can select the worksheet to store the transform using the "**New Worksheet**" or the "**Other Worksheets**" and a set of available columns appear in the "**Select Column**" box. The user has to specify a name for the new column.

- An example of the selections made is shown below:



3.1.3 Copy

- When the user clicks Data ► Copy, the following window will appear:

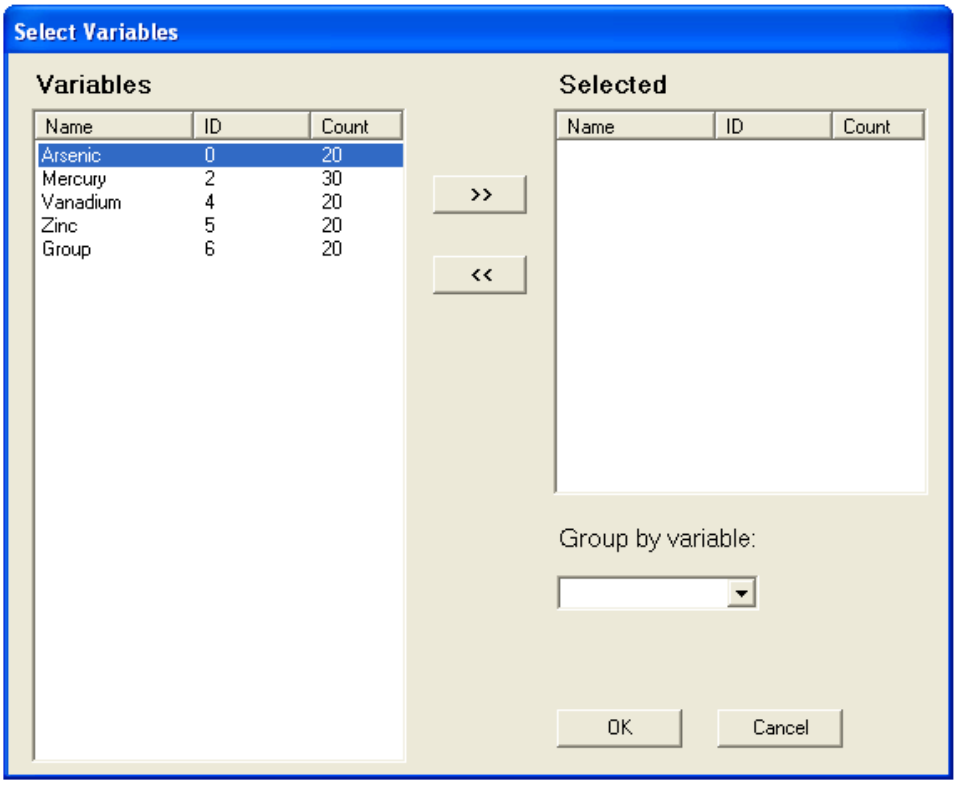


- This screen allows the user to copy a single variable to a new column.

3.2 Graphing and Statistical Analysis of Univariate Data

- Variables need to be selected to perform statistical analyses.

- When the user clicks on any drop-down menu (Except Background vs. Site Comparison option), the following window will appear.



- The Options button is available in certain menus. The use of this option leads to a different pop-up window.

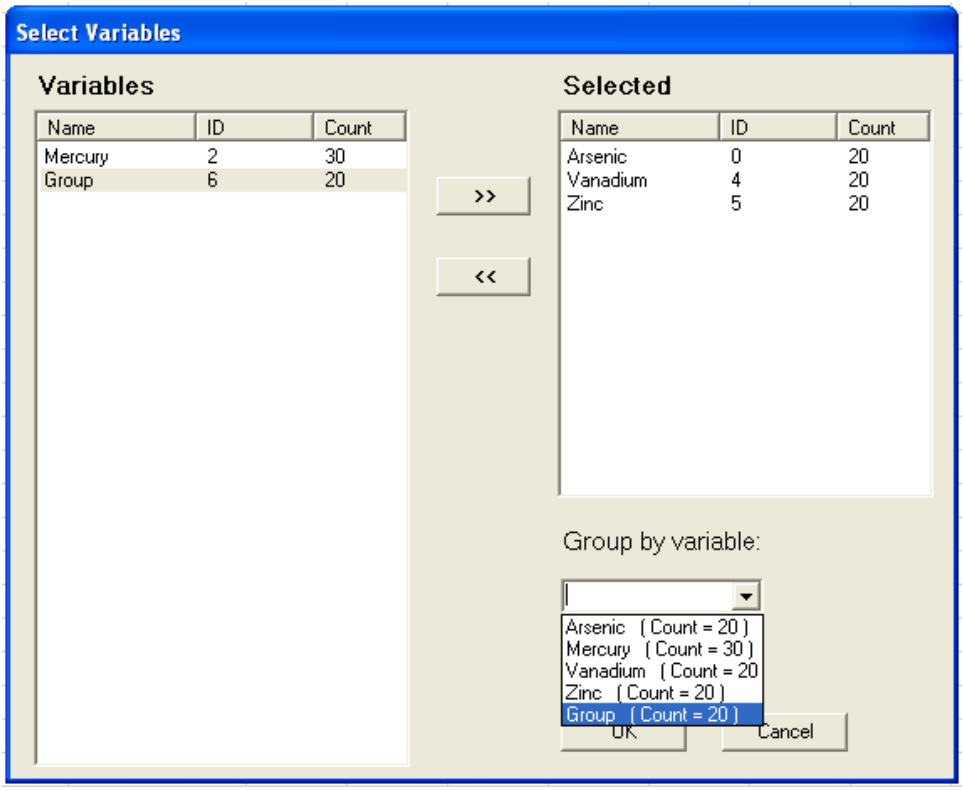
- Multiple variables can be processed simultaneously in Scout.

- Moreover, if the user wants to perform a statistical analysis on a variable (e.g., contaminant) by a Group variable, click on the arrow below the "**Group by Variable**" to get a drop-down list of the available variables to select an appropriate group variable. For an example, a group variable (e.g., Site Area) can have alphanumeric values, such as AOC1, AOC2, AOC3, and Background. Thus, in this example, the group variable name, Site Area, takes 4 values, such as AOC1, AOC2, AOC3, and Background.

- The Group variable is particularly useful when data from two or more samples need to be compared.

- Any variable can be a group variable. However, for meaningful results, only a variable that really represents a group variable (categories) should be selected as a group variable.

- The number of observations in the group variable and the number of observations in the selected variables (to be used in a statistical procedure) should be the same. In the example below, the variable, "Mercury," is not selected because the number of observations for Mercury is 30; in other words, Mercury values have not been grouped. The group variable, and each of the selected variables, has 20 data values.



Caution: Care should be taken to avoid misrepresentation and improper use of group variables. It is recommended not to assign any missing values for the group variable.

More on Group Option

- The group option provides a powerful tool to perform various statistical tests and methods (including graphical displays) separately for each of the groups (samples from different populations) that may be present in a data set. For an example, the same data set may consist of samples from the various groups (populations). The graphical displays (e.g., box plots, Q-Q plots) and statistics of interest can be computed separately for each group by using this option.

- In order to use this option, at least one variable representing the group ID (alphanumeric characters) should be included in the data set. The various values of that group variable represent different group categories.

- Note that the number of values (representing group membership) in a group variable should equal the number of values in the variable (e.g., Arsenic) of interest that needs to be partitioned into various groups (e.g., monitoring wells).

- The group column can be any qualitative group ID representing different species, laboratories, shifts, regions, and so on. For an example, in environmental applications, data for the various groups represent data from the various site areas (e.g., background, AOC1, AOC2, …), or from monitoring wells (e.g., MW1, MW2, …).

3.2.1 Graphs by Groups

- Individual or multiple graphs (Q-Q plots, box plots, and histograms) can be displayed on a graph by selecting the "**Graphs by Groups**" option.

- Individual graphs for each group (specified by the selected group variable) are produced by selecting the "**Individual Graph**" option.

- Multiple graphs (e.g., side-by-side box plots, multiple Q-Q plots on the same graph) are produced by selecting the "Group Graph" option for a variable categorized by a group variable. Using this "Group Graph" option, multiple graphs can be displayed for all of the sub-groups included in the Group variable. This option is useful when data to be compared are given in the same column and are classified by the group variable.

- Multiple graphs (e.g., side-by-side box plots, multiple Q-Q plots) for selected variables are produced by selecting the "Group Graph" option. Using the "**Group Graph**" option, multiple graphs can be displayed for all selected variables. This option is useful when data (e.g., lead) to be compared are given in different columns, perhaps representing different populations.

***Note**: It is the users' responsibility to provide an adequate amount of detected data to perform the group operations. For an example, if the user desires to produce a graphical Q-Q plot (using only detected data) with regression lines displayed, then there should be at least two detected points (to compute slope, intercept, sd) in the data set. Similarly if the graphs are desired for each of the group specified by the group ID variable, there should be at least 2 detected observations in each group specified by the group variable. Scout generates a warning message (in orange color) in the lower panel of the Scout screen. Specifically, the user should make sure that a variable with non-detects and categorized by a group variable should have enough detected data in each group to perform the various methods (e.g., GOF tests, Q-Q plots with regression lines) as incorporated in Scout.*

The analyses of data categorized by a group ID variable such as:

1) Surface vs. Subsurface,

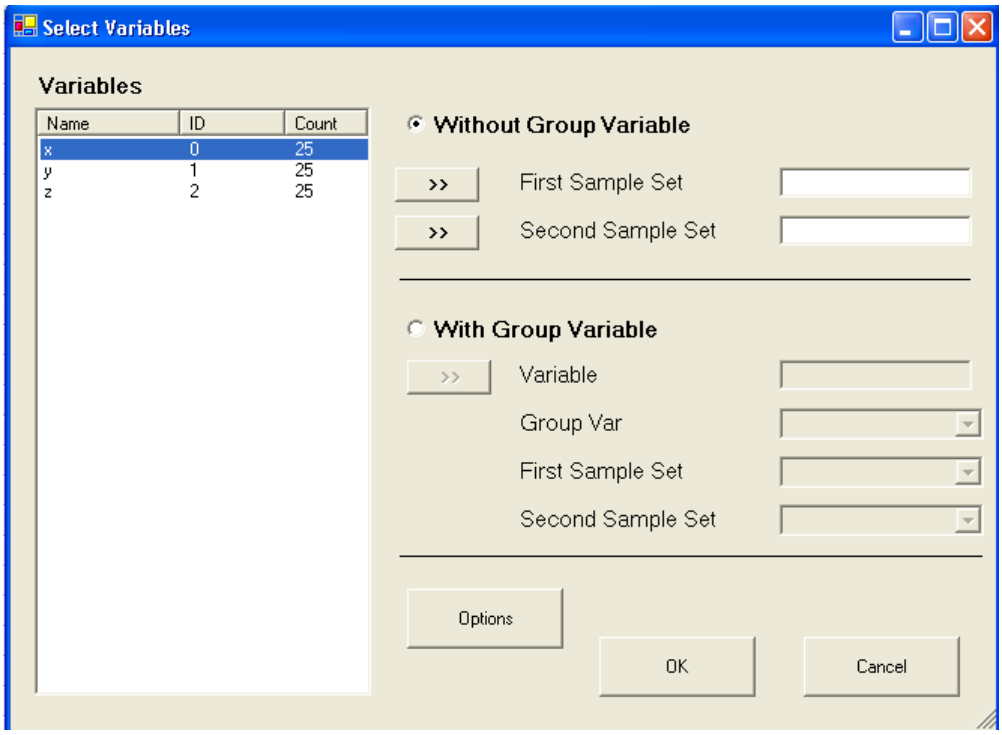
2) AOC 1 vs. AOC 2,

3) Site vs. Background, and

4) Upgradient vs. Downgradient monitoring wells, are quite common in many environmental applications.

3.2.2 Select Variables Screen for Two-Sample Hypothesis Testing

The variables selection screen is different for two-sample hypothesis testing when compared to single sample hypothesis testing. The "**Select Variables**" screen is as shown.



3.2.2.1 Without Group Variable

- The first sample set (e.g., background concentration) and the second sample set (e.g., site concentration) of variables (e.g., COPC) are selected.

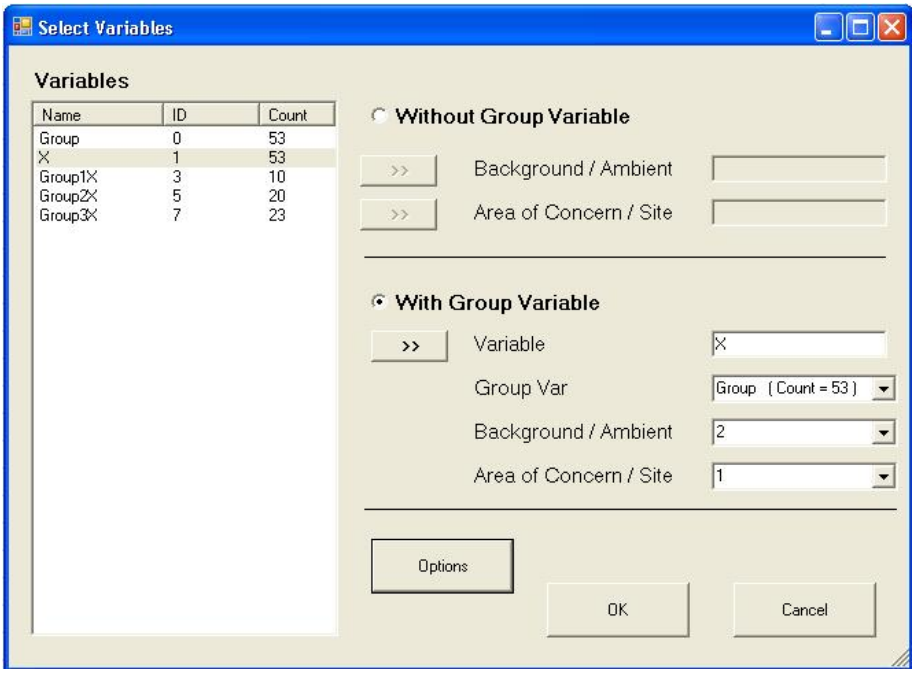
- The "**Options**" button provides the various options available with the selected test.

3.2.2.2 With Group Variable

- This option is used when data values of the variable (e.g., COPC) for the first sample set (e.g., site) and the second sample set (e.g., background) are given in the same column. The values are separated into different populations (groups) by the values of an associated group variable. The group variable may represent

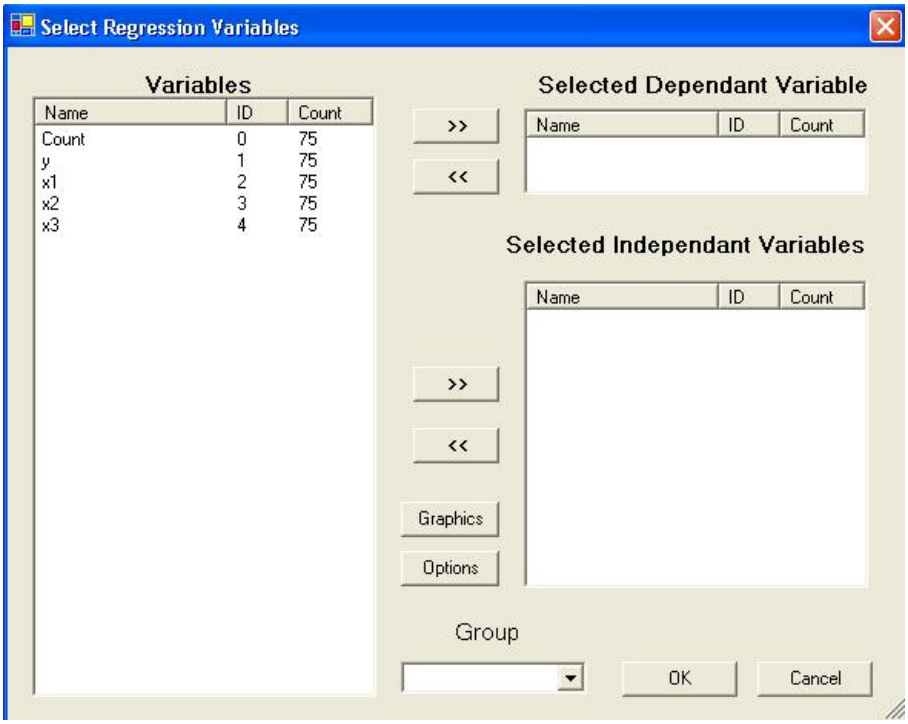
several populations (e.g., several AOCs, MWs). The user can compare two groups at a time by using this option.

- When using this option, the user should select a group variable by clicking the arrow next to the **Group Var** option for a drop-down list of available variables. The user selects an appropriate (meaningful) variable representing groups, such as Background and AOC. The user is allowed to use letters, numbers, or alphanumeric labels for the group names. A sample variables selection screen is shown below.



3.3 Regression Menu

- When the Regression Menu is clicked on, the following window pops up.

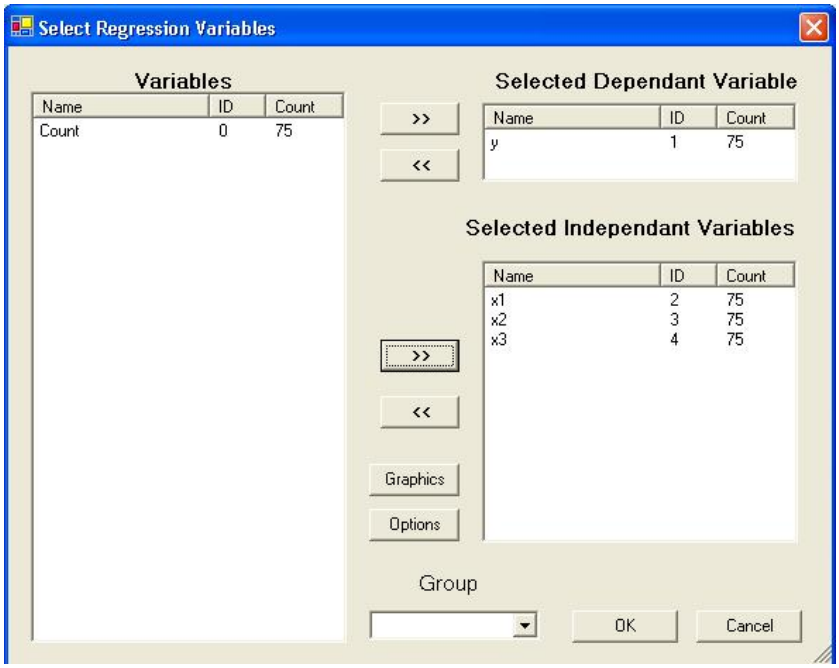


- Both dependent and independent variables need to be selected.

- The use of the "Options" button leads to a new options window. The methods on regression drop-down menu have different "**Options**" and "**Graphics**" screens. They are discussed in Chapter 8.

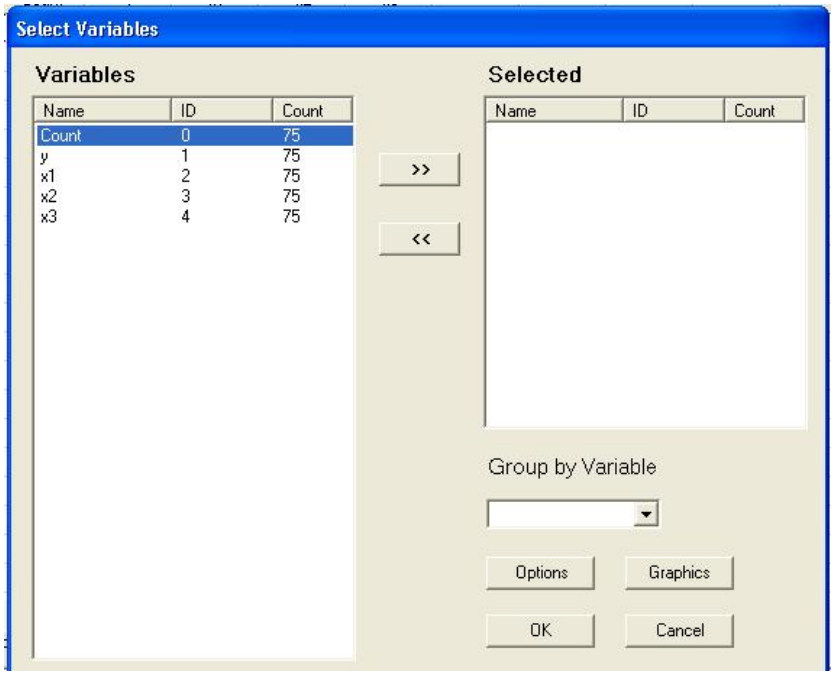
- Grouping works in the same way as for univariate data.

- An example of the selected screen is shown below.



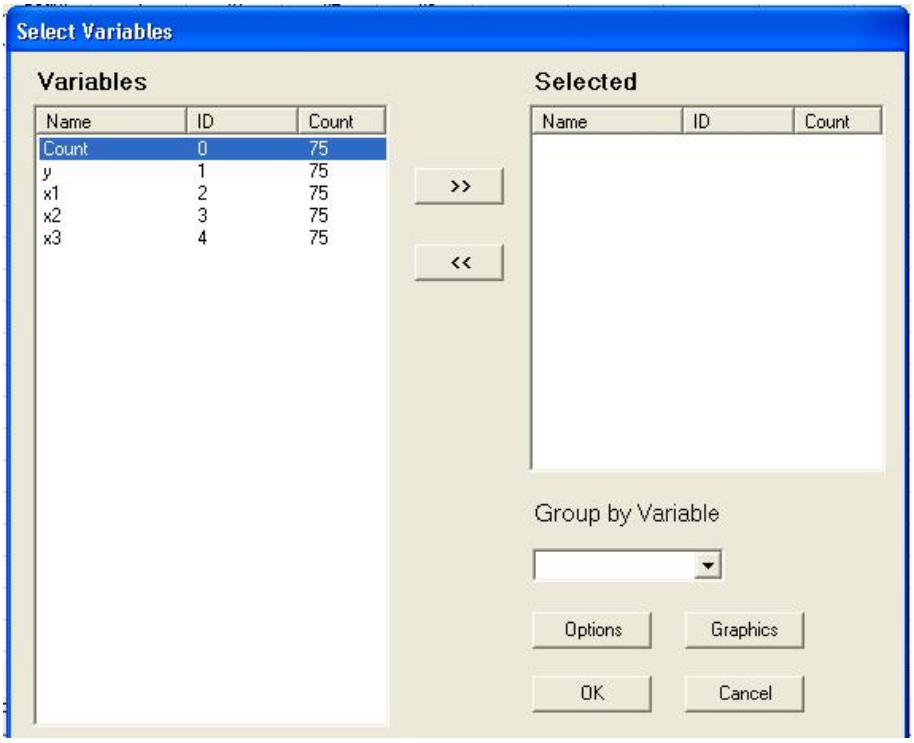
3.4 Multivariate Outliers and PCA Menu

- For multivariate outliers or multivariate PCA, the following "**Select Variables**" screen appears:



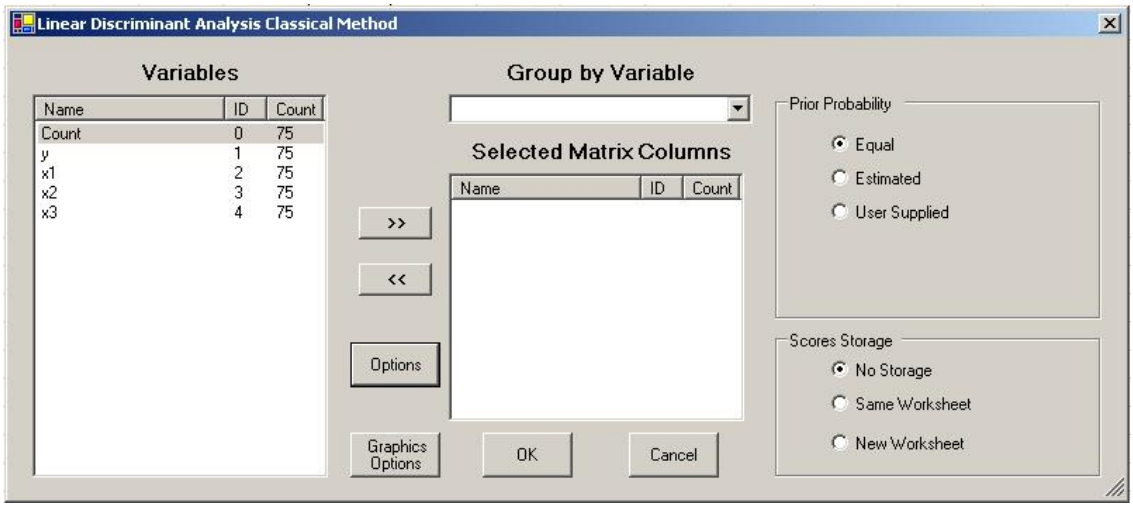
- The variables that are to be considered for the analyses are selected and the "**Options**" button may be clicked to select from the various options available. Those options are discussed in Chapters 7 and 9.

- A "**Graphics**" button is provided for Robust/Iterative methods and Principal Component Analysis methods as shown below. Those options are discussed in Chapters 7 and 9.



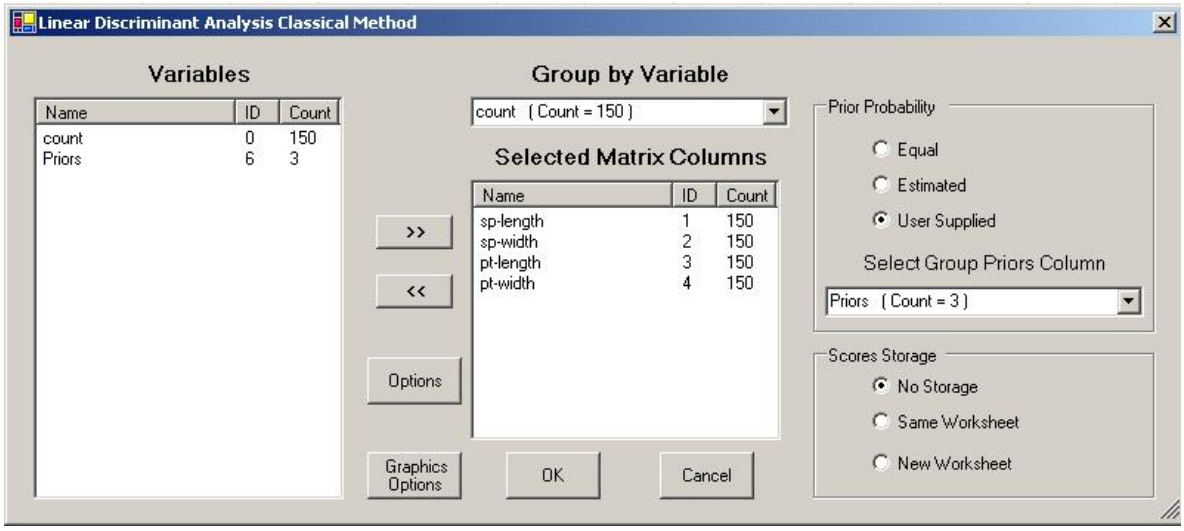
3.5 Multivariate Discriminant Analysis Menu

- When the Multivariate EDA ► Discriminant Analysis is clicked on, the following window appears.



- There should be a group column specifying the various groups present.
- The group variable is selected from the "**Group by Variable**" drop-down bar.
- The various variables required for the analysis are then selected.
- If the prior probabilities are supplied by the user, then a column should exist in the work sheet for the prior probabilities and the probabilities can be selected from the "**Select Group Priors Column**" drop-down bar.

- An example is illustrated below.



Note: *The Prior Probability box is not available for the Fisher Discriminant Analysis since equal priors are assumed.*

Chapter 4

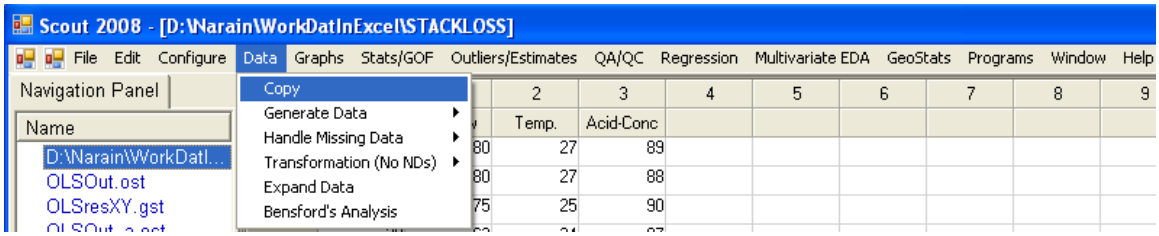
Data

Scout provides the user with an array of options to modify the given data, both without non-detects and with non-detects. The various options include:

- Copy: copies data from one column to another.
- Generate: generates univariate and multivariate data.
- Impute: generates estimated data for non-detect observations.
- Missing: handles missing observations.
- Transform: transforms data without non-detects using mathematical functions.

4.1 Copy

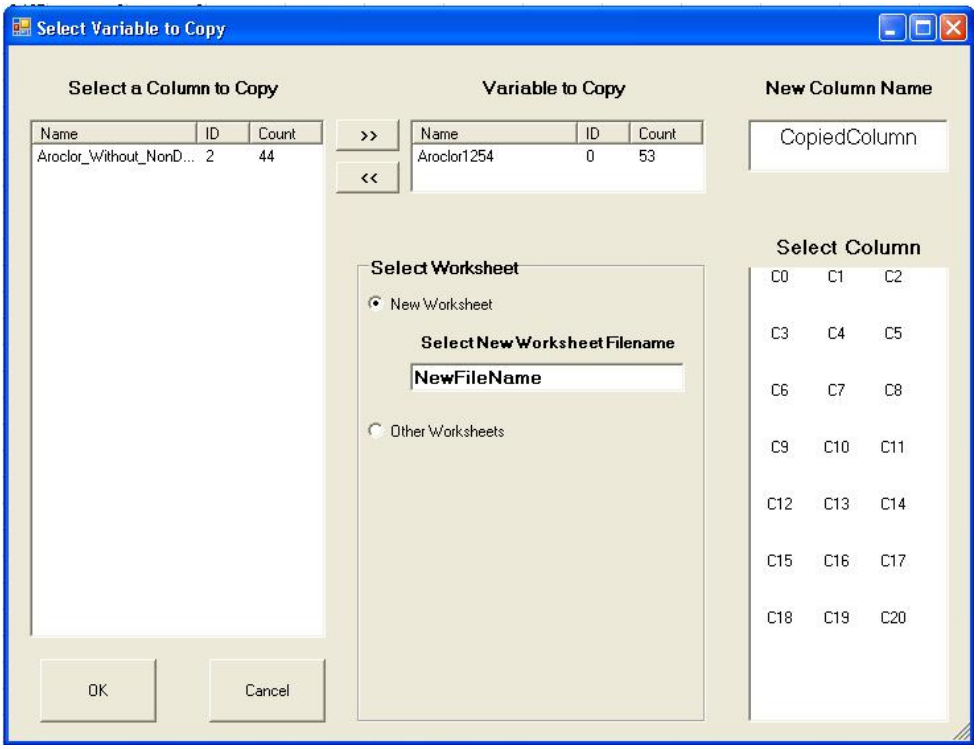
1. Click **Data ► Copy.**

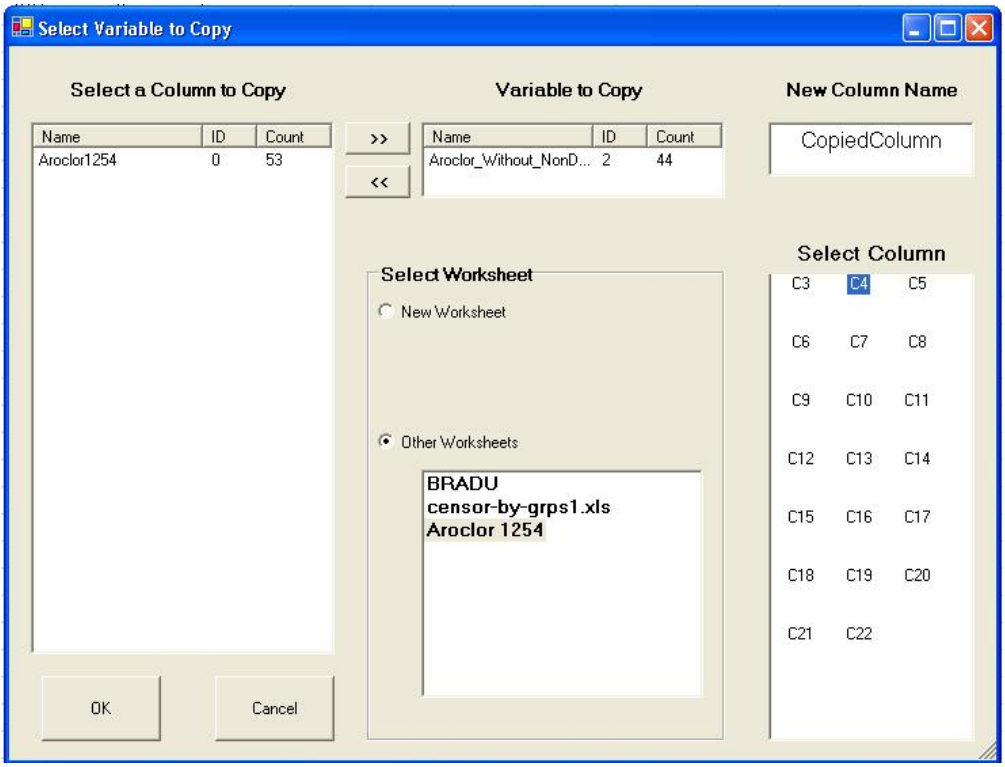


2. The "**Select Variable to Copy**" screen (Section 3.1.3) will appear. Also, see example screens shown below.

    - A single variable is selected and that variable appears in the "**Variable to Copy**" box.
    - The user can select the preferred worksheet in storing the transformed data using the "**New Worksheet**" or the "**Other Worksheets**" and a set of available columns appear in the "**Select Column**" box. If the "**New Worksheet**" option is selected, then the data is copied onto the new worksheet. If the "**Other Worksheets**" option is selected, a set of available worksheets are displayed and the columns available for the selected "**Other Worksheet**" are also displayed. The user has to specify a name for the new column.

- Examples for the selections using "**New Worksheet**" and "**Other Worksheet**" are shown below.



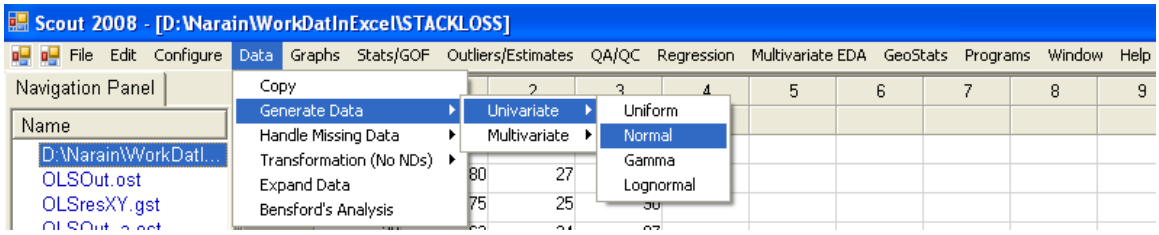


4.2 Generate

The Generate option generates univariate uniform, normal, gamma and lognormal distributed random numbers, and also multivariate normal data.

4.2.1 Univariate

1. Click **Data ► Generate ►Univariate.**

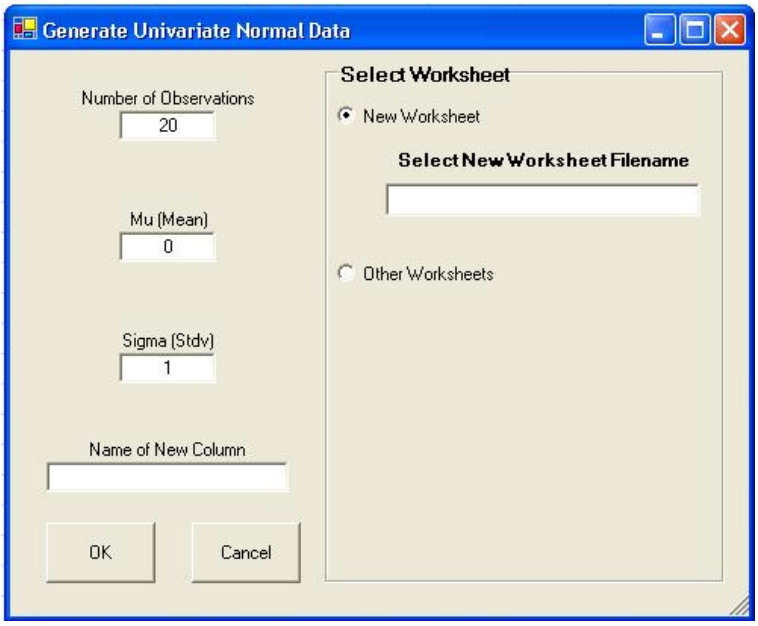


2. Random numbers from the four different distributions are generated:

   - Uniform distribution: input parameters are "**a**" (lower limit) and "**b**" (upper limit).
   - Normal distribution: input parameters are "**Mu**" (mean) and "**Sigma**" (standard deviation) of raw data.
   - Gamma distribution: input parameters are "**Alpha**" (scale parameter) and "**Beta**" (shape parameter).
   - Lognormal distribution: input parameters are "**Mu**" (mean) and "**Sigma**" (standard deviation) of data is log-transformed space (logged data).

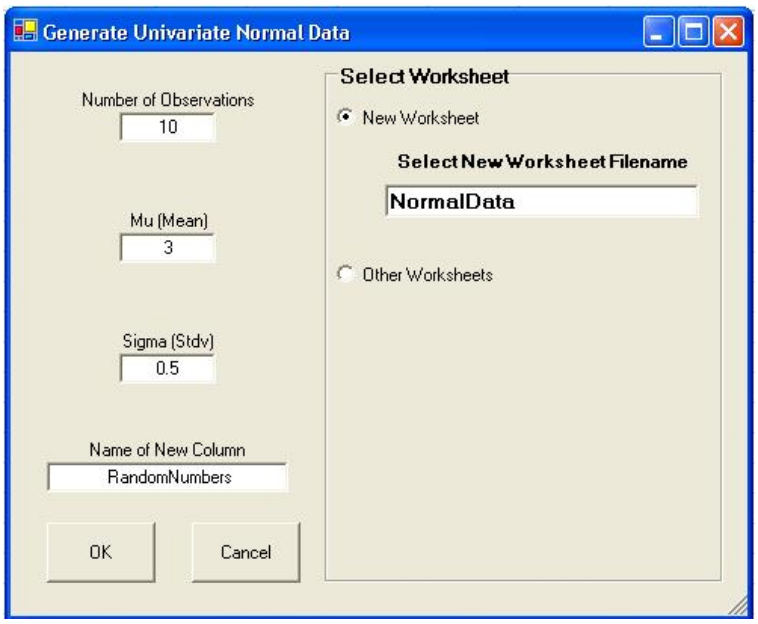
3. An example for the normal distribution is illustrated.

- Click **Data ► Generate ► Univariate ► Normal.**

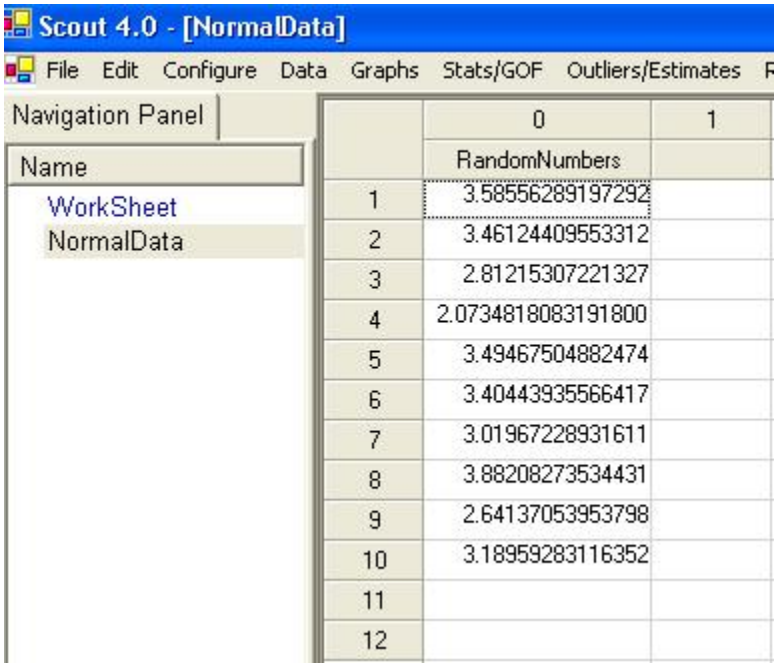


  - Specify the number of observations required. The default is "**20**."
  - Specify "**Mu**" (mean) and "**Sigma**" (standard deviation). The defaults are "**0**" and "**1**," respectively.
  - Specify the name of the new column.
  - Select the worksheet into which the new data is to be generated.

- Click "**OK**" to continue or "**Cancel**" to cancel the Generate option.



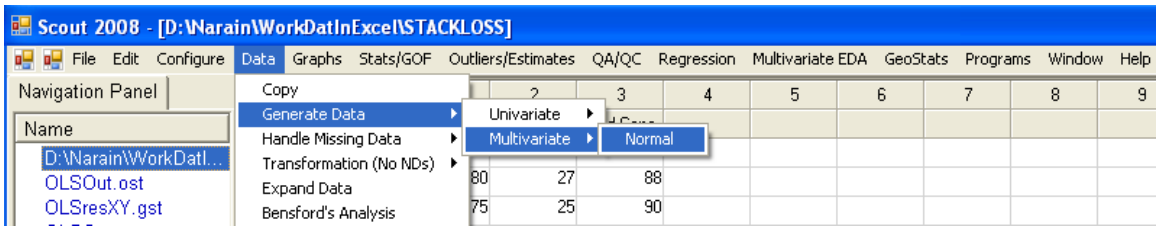
Output Screen for Univariate Normal Data.

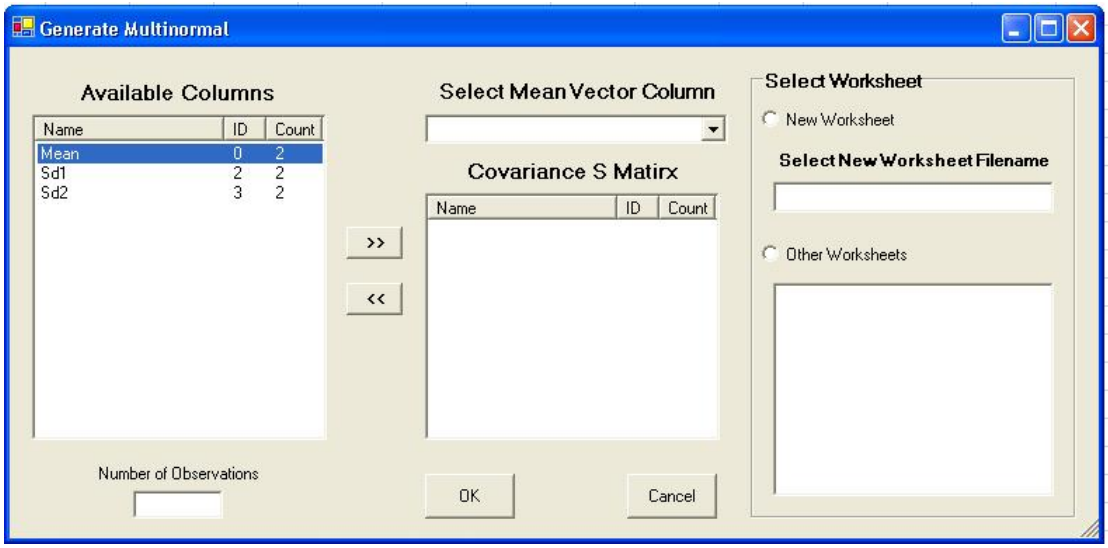


The new worksheet has been named "Normal Data," as seen in the **Navigation Panel**.

4.2.2 Multivariate

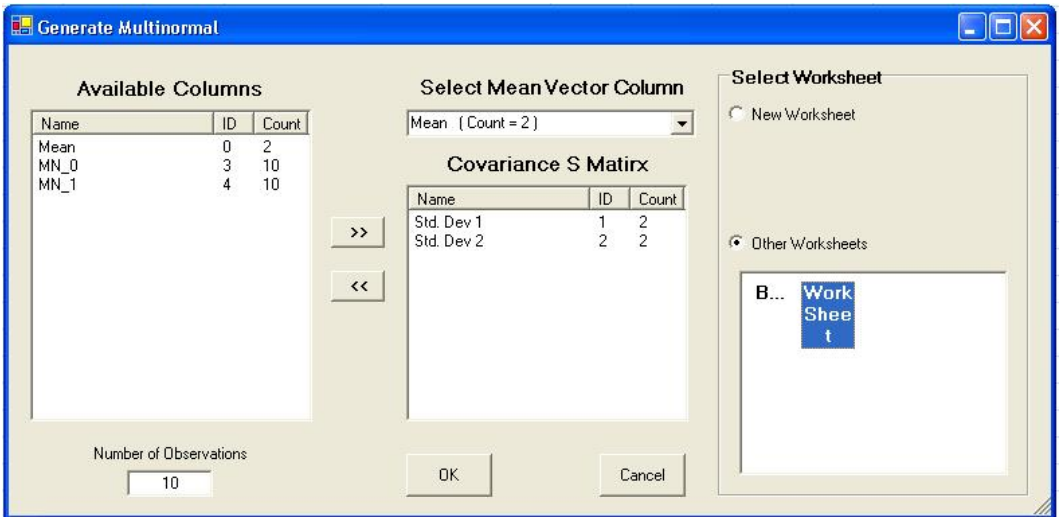
1. Click **Data ► Generate ► Multivariate ► Normal.**



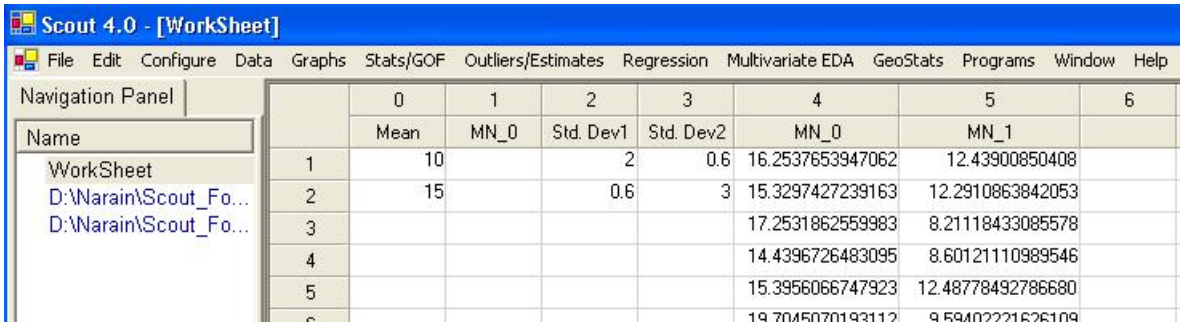


***Note**: In order to use this option, the user should make sure that there is a column for the mean vector and p columns for the variance covariance matrix, where p is the number of variables in the matrix.*

- The mean vector is chosen from the "**Select Mean Vector Column**" drop-down bar and the columns representing the columns of variance-covariance matrix are chosen for the "**Covariance S Matrix**."
- The selected worksheet represents the worksheet where the new generated data would be stored. The generated data then can be used in various other modules of Scout or some other software.
- If the "**New Worksheet**" is selected, then a name for the worksheet has to be specified.
- Click "**OK**" to continue or "**Cancel**" to cancel the Generate option.



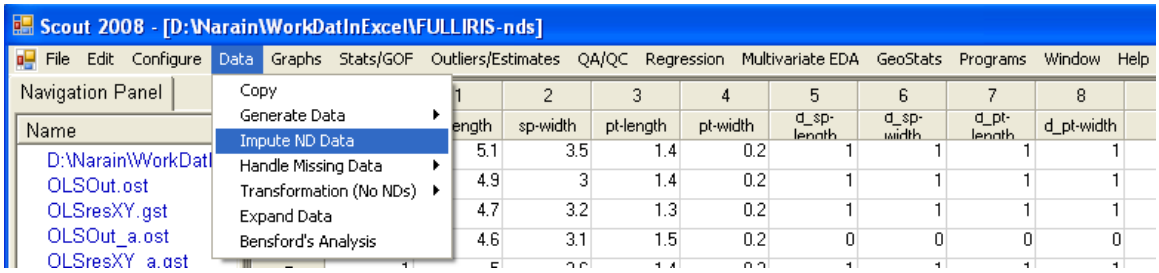
Output Screen for Multivariate Normal Data.



4.3 Impute (NDs)

Data sets with non-detect observations are transformed using the impute option. Various options are available to impute (estimate or extrapolate) the non-detect observations. The use of this option generates additional columns consisting of all of the extrapolated non-detects and detected observations. Those columns can be appended to the any of the existing open spreadsheets or in a new worksheet.

1. Click **Data ► Impute (NDs).**



2. The "**Select Variable to Impute**" screen (see Section 3.1.2 and the screen below) appears. The various options available are:

   - **Detection Limit**: the non-detect observations are given the values of the detection limit.
   - **½ Detection Limit**: the non-detect observations are given the values of the one-half of the detection limit.
   - **Zero**: the non-detect observations are given zero values.
   - **Normal ROS**: Regression on Order Statistics (ROS) is used to extrapolate the non-detect observations using a normal model.
   - **Gamma ROS**: Regression on Order Statistics (ROS) is used to extrapolate the non-detect observations using a gamma model.

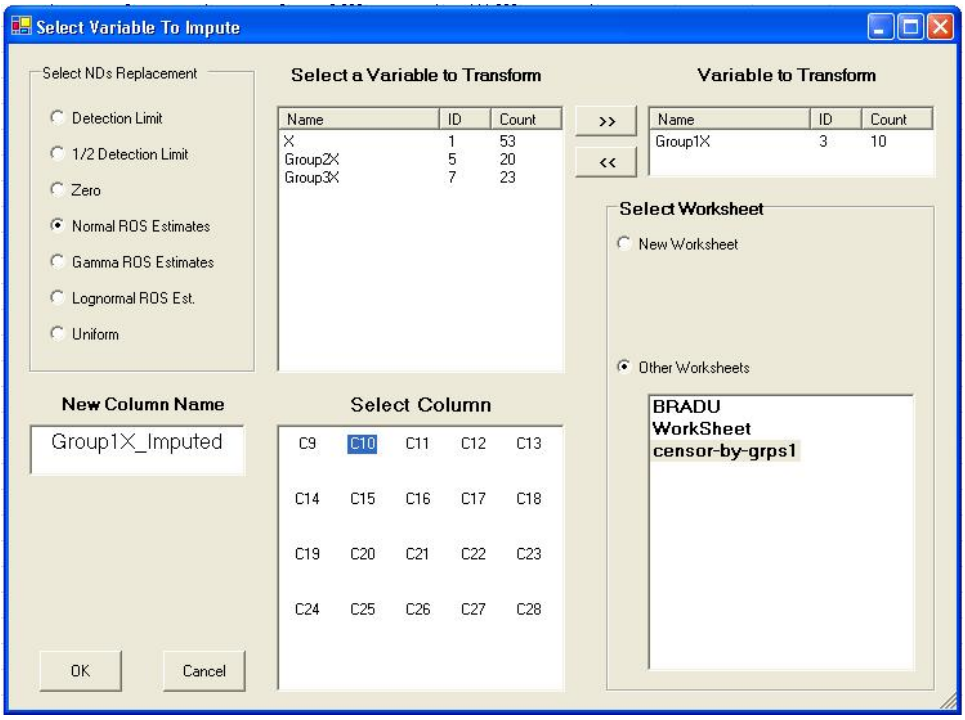
- **Lognormal ROS**: Regression on Order Statistics (ROS) is used to extrapolate non-detect observations using a lognormal model.

- **Uniform:** the non-detect observations are given a value of a uniform distribution random number with the lower limit as zero and upper limit as the detection limit.

3. An example for the Normal ROS is illustrated.

   - Click **Data ► Impute (NDs).**

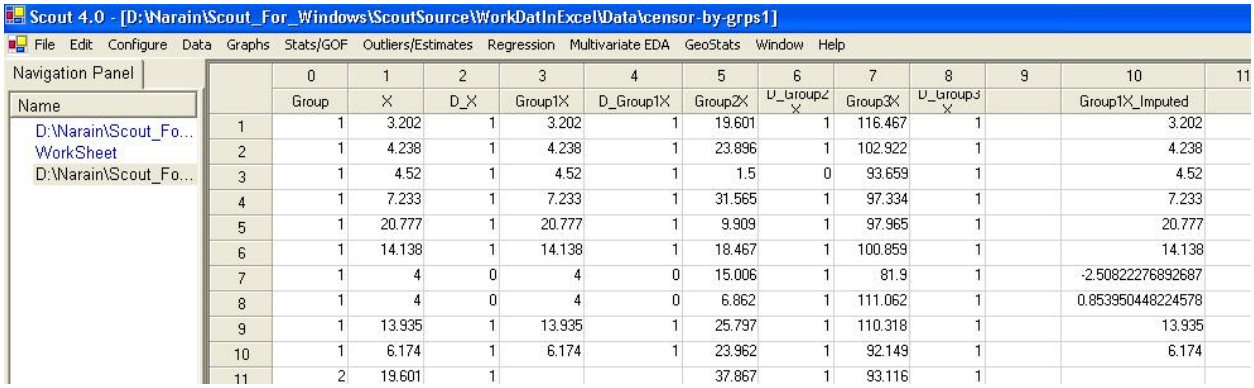
     - In the "**Select Variable To Impute**" screen, the following options are selected.



     - Select the method to replace NDs ("**Select NDs Replacement**"), the variable to transform, the New Column Name, and the worksheet.

   - Click "**OK**" to continue or "**Cancel**" to cancel the impute option.

Output Screen for Impute using Normal ROS.

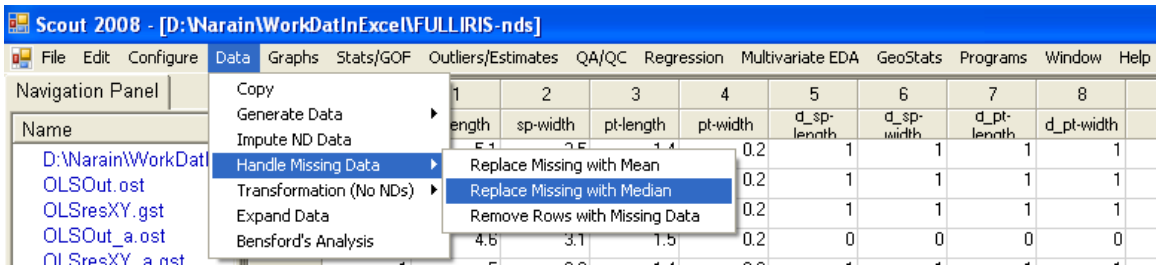


	0	1	2	3	4	5	6	7	8	9	10	11
	Group	X	D_X	Group1X	D_Group1X	Group2X	D_Group2X	Group3X	D_Group3X		Group1X_Imputed	
1	1	3.202	1	3.202	1	19.601	1	116.467	1		3.202	
2	1	4.238	1	4.238	1	23.896	1	102.922	1		4.238	
3	1	4.52	1	4.52	1	1.5	0	93.659	1		4.52	
4	1	7.233	1	7.233	1	31.565	1	97.334	1		7.233	
5	1	20.777	1	20.777	1	9.909	1	97.965	1		20.777	
6	1	14.138	1	14.138	1	18.467	1	100.859	1		14.138	
7	1	4	0	4	0	15.006	1	81.9	1		-2.50822276892687	
8	1	4	0	4	0	6.862	1	111.062	1		0.853950448224578	
9	1	13.935	1	13.935	1	25.797	1	110.318	1		13.935	
10	1	6.174	1	6.174	1	23.962	1	92.149	1		6.174	
11	2	19.601	1			37.867	1	93.116	1			

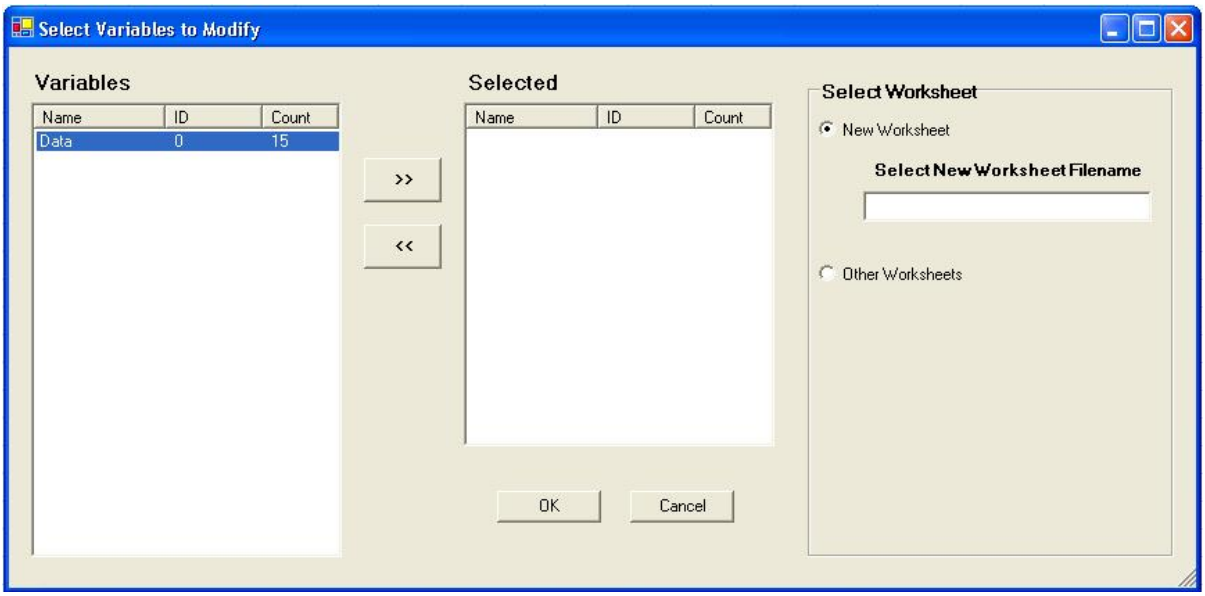
4.4 Missing

Scout has three methods to handle missing observations. The first method replaces the missing observations by the mean of the data, the second method replaces the missing observations by the median of the data and the third method removes the rows with missing observations. A new column is created for the selected variable using the selected option. This new column can be added to a new worksheet or an existing worksheet. Note that observations are given values 1E-31 or 1E+31 (considered to be missing).

1. Click **Data ► Missing ► Replace Missing with Median.**



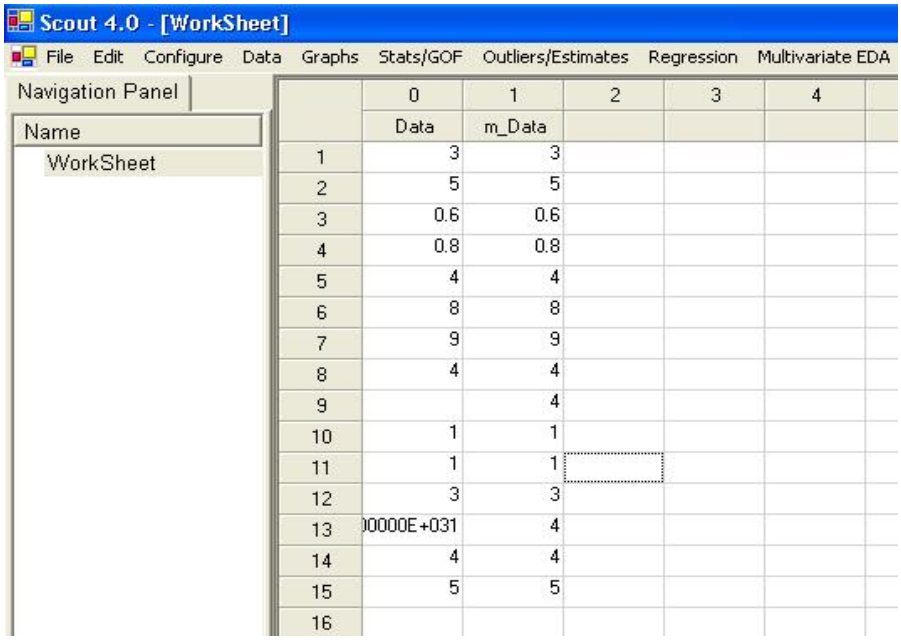
2. The following screen appears:



- Select the variable to modify ("**Variables**").
- Specify whether the new column should be added to a "**New Worksheet**" or to existing "**Other Worksheets**" (under "**Select Worksheet**").

- Click "**OK**" to continue "**Cancel**" to cancel the missing option.

Output Screen for Missing (Replace rows with the median).

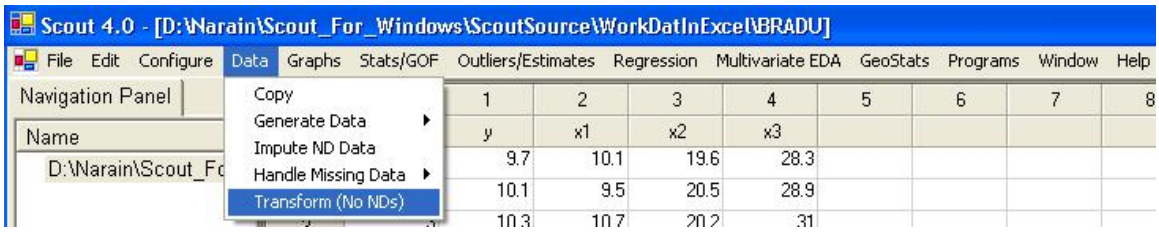


4.5 Transform (No NDs)

Scout offers a number of options to transform the variables without non-detects:

- **z – transform**: standardizes the variable; i.e., the mean of the observations is subtracted and the result is divided by the standard deviation.
- **Linear (ax + b)**: gives a linear transformation of x. The values of "a" and "b" are entered by the user.
- **Natural Log**: gives the natural logarithm transform of the variable.
- **Log Base 10**: gives the logarithm to the base 10 transform of the variable.
- **Exp(x)**: gives the exponential transformation of the variable.
- **Pow(x, a)**: gives the value of the variable "x" raised to power "a."
- **Box-Cox**: gives the Box-Cox transformation of the variable; i.e., $\left(\frac{x^a - 1}{a}\right)$; the value of "a" is entered by the user.
- **Ranked**: gives the order number of the observations in the variable after sorting.
- **Ordered**: sorts the data in ascending order.
- **Rankit**: gives the expected values of ordered statistics of the standard normal distribution corresponding to the data points in a manner determined by the order in which the data points appear.
- **Arcsine**: gives the arc-sine value of the observations in the selected variable.
- **Group Items**: this option is used in conjunction with the Discriminant Analysis for data sets with groups. This option outputs the group names in a sorted order in the selected column. This option is useful when the user wants to input the values of prior probabilities for the groups.

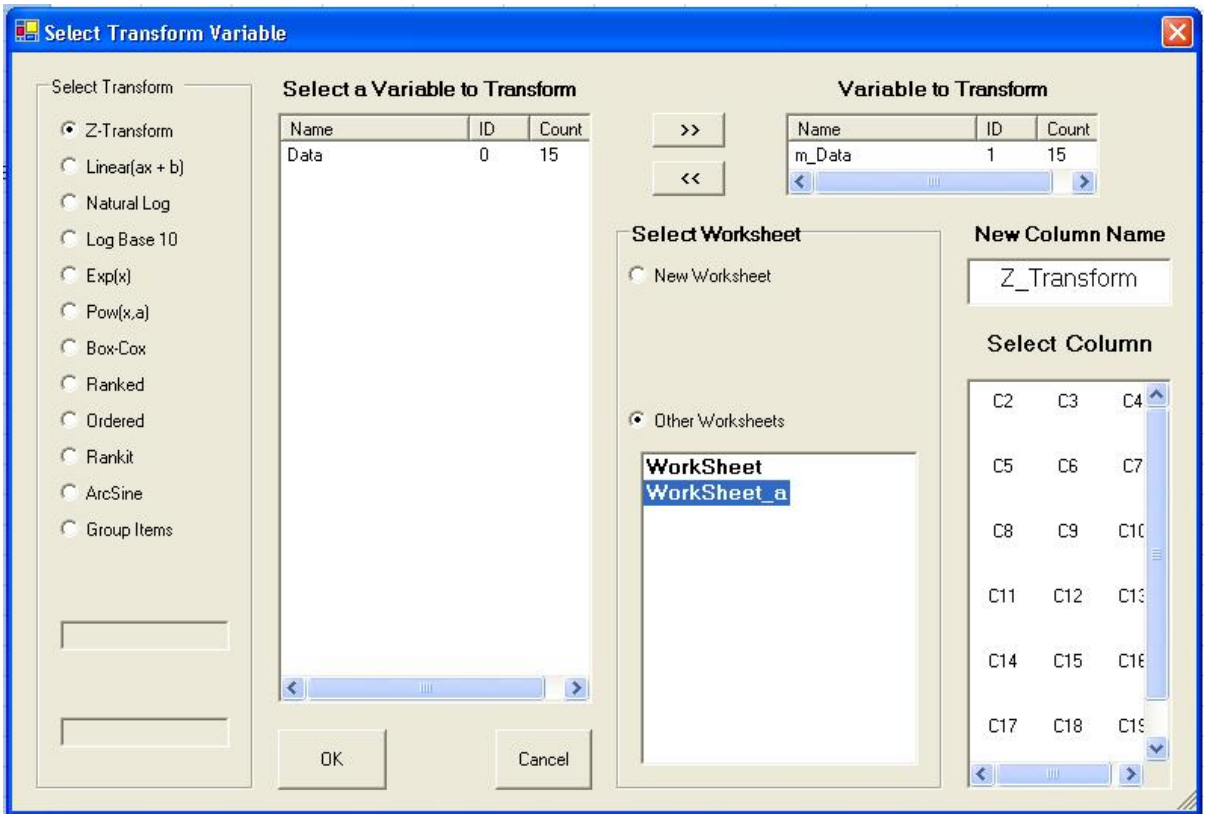
1. Click **Data ► Transform (No NDs).**



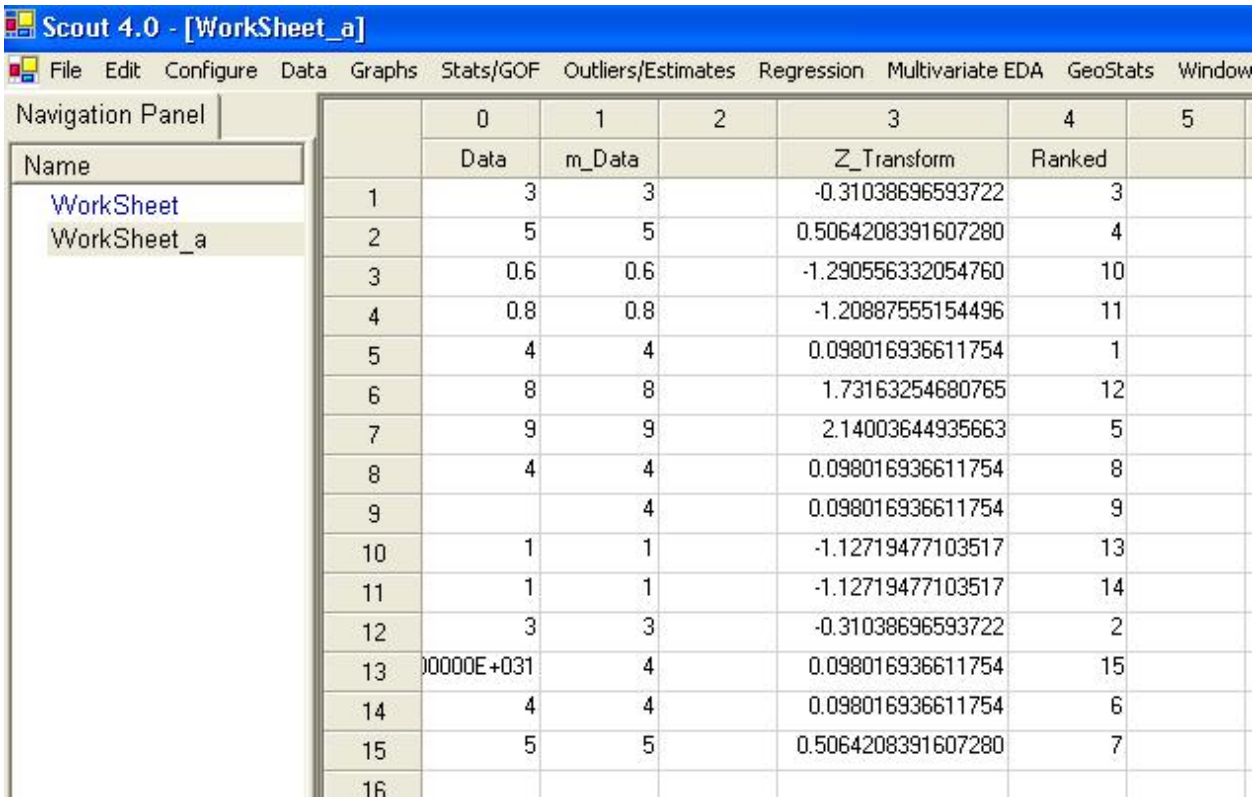
2. The "**Select Transform Variable**" screen (See also Section 3.1.1) appears.

   - Specify the transform to apply ("**Select Transform**").
   - Specify a variable to transform ("**Select a Variable to Transform**").
   - Specify whether the new column should be added to a "**New Worksheet**" or existing, "**Other Worksheets**" (under "**Select Worksheet**"; then, enter a name for the transformed variable (under "**New Column Name**").

- Click "**OK**" to continue or "**Cancel**" to cancel the Transform option.



Output Screen for Transform (No NDs).
Selected options: z – transform and Ranked.

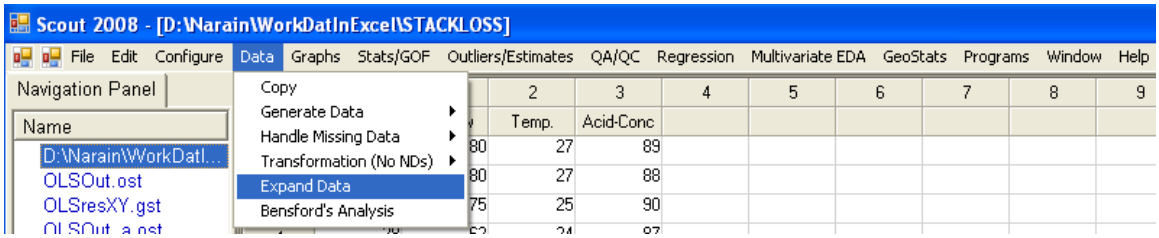


	0	1	2	3	4	5
	Data	m_Data		Z_Transform	Ranked	
1	3	3		-0.31038696593722	3	
2	5	5		0.5064208391607280	4	
3	0.6	0.6		-1.290556332054760	10	
4	0.8	0.8		-1.20887555154496	11	
5	4	4		0.098016936611754	1	
6	8	8		1.73163254680765	12	
7	9	9		2.14003644935663	5	
8	4	4		0.098016936611754	8	
9		4		0.098016936611754	9	
10	1	1		-1.12719477103517	13	
11	1	1		-1.12719477103517	14	
12	3	3		-0.31038696593722	2	
13	)0000E+031	4		0.098016936611754	15	
14	4	4		0.098016936611754	6	
15	5	5		0.5064208391607280	7	
16						

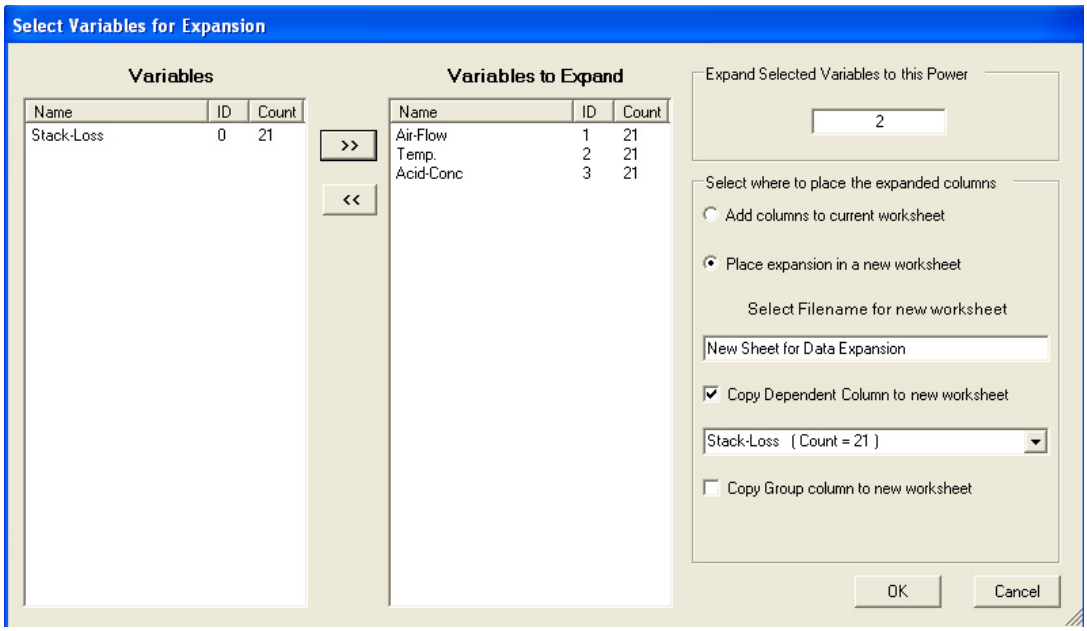
4.6 Expand Data

Scout allows the user to generate the interaction terms using the available variables. This part of the Scout program was developed so that the user can generate interaction terms for regression analysis. The highest power supported by Scout is 10. But the user is cautioned that the maximum number of interaction terms supported by Scout is 256. If more than 256 terms are generated, then those terms will not be displayed on the worksheet. The user is also cautioned that generating interaction terms with high degrees takes up considerable computer resources and computing time.

1. Click **Data ► Expand Data.**



2. The following "**Select Transform Variable**" screen appears.



- Specify the variable to expand ("**Variables to Expand**").
- Specify the power /degree ("**Expand Selected Variables to this Power**").
- Specify whether the new columns should be added to a "**New Worksheet**" or existing, "**Other Worksheets**" (under "**Select Worksheet**"; then, enter a name for the transformed variable (under "**New Column Name**").
- If new worksheet option is selected specify if the dependent variable used in regression should be copied to the new worksheet.
- If new worksheet option is selected specify if the group column should be copied to the new worksheet.

- Click "**OK**" to continue or "**Cancel**" to cancel this option.

Scout 2008 - [New Sheet for Data Expansion]

File Edit Configure Data Graphs Stats/GOF Outliers/Estimates QA/QC Regression Multivariate EDA GeoStats Progr

Navigation Panel

Name

D:\Narain\WorkDatI...

New Sheet for Data...

Expansion.ost

	0	1	2	3	4	5	6	7
	Stack-Loss	AA	AB	AC	BB	BC	CC	
1	42	6,400	2,160	7,120	729	2,403	7,921	
2	37	6,400	2,160	7,040	729	2,376	7,744	
3	37	5,625	1,875	6,750	625	2,250	8,100	
4	28	3,844	1,488	5,394	576	2,088	7,569	
5	18	3,844	1,364	5,394	484	1,914	7,569	
6	18	3,844	1,426	5,394	529	2,001	7,569	
7	19	3,844	1,488	5,766	576	2,232	8,649	
8	20	3,844	1,488	5,766	576	2,232	8,649	
9	15	3,364	1,334	5,046	529	2,001	7,569	
10	14	3,364	1,044	4,640	324	1,440	6,400	
11	14	3,364	1,044	5,162	324	1,602	7,921	
12	13	3,364	986	5,104	289	1,496	7,744	
13	11	3,364	1,044	4,756	324	1,476	6,724	
14	12	3,364	1,102	5,394	361	1,767	8,649	
15	8	2,500	900	4,450	324	1,602	7,921	
16	7	2,500	900	4,300	324	1,548	7,396	
17	8	2,500	950	3,600	361	1,368	5,184	
18	8	2,500	950	3,950	361	1,501	6,241	
19	9	2,500	1,000	4,000	400	1,600	6,400	
20	15	3,136	1,120	4,592	400	1,640	6,724	
21	15	4,900	1,400	6,370	400	1,820	8,281	

***Note**: A second output sheet called "**Expansion.ost**" will be generated. This output sheet will indicate what the variables in the column header stand for in the interaction terms.*

Scout 2008 - [Expansion.ost]

File Edit Configure Programs Window Help

Navigation Panel

Name

D:\Narain\WorkDatI...

New Sheet for Data...

Expansion.ost

	Expansion Legend
Date/Time of Computation	10/29/2008 12:49:41 PM
From File	D:\Narain\WorkDatInExcel\STACKLOSS
To New Worksheet	New Sheet for Data Expansion
Expanded to the	2nd Power
Representation	Actual Variable Name
"A"	Air-Flow
"B"	Temp.
"C"	Acid-Conc

4.7 Benford's Analysis

Benford's law (see separate pdf file of Appendix C for details), less commonly known as Newcomb's law, the first digit law, the first digit phenomenon, and the leading digit phenomenon, was independently discovered first by Simon Newcomb (1881), and then by Frank Benford (1938). Each noticed that the beginning tables of books of logarithms were "dirtier" at the beginning (due to use) rather than at the end, noting that some particular first digits should occur with a greater "natural" frequency.

Newcomb's form of the law is given as

$$p\left(d_1\left(i\right)=i\right)=\log_{10}\left[1+\frac{1}{d_1(i)}\right]\;;\quad \text{i} = 1, 2, 3, ..., 9$$

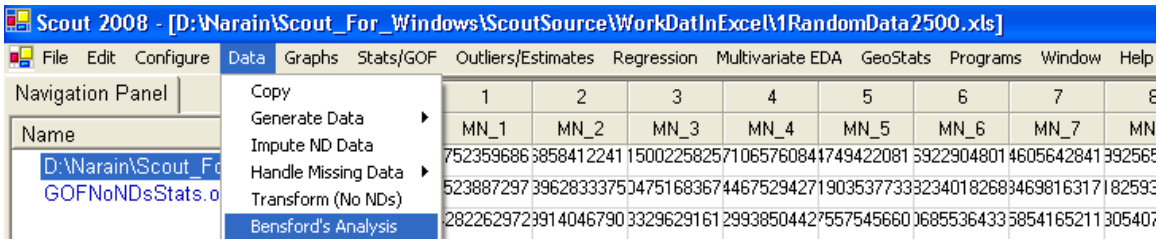
And the equivalent Benford's form of the law is given as

$$p\left(d_1\left(i\right)=i\right)=\log_{10}\left[\frac{d_1(i)+1}{d_1(i)}\right]\;;\quad \text{i} = 1, 2, 3, ..., 9$$

where $p(\mathrm{d}_1(i) = i)$ is the probability that the first place, $j = 1$ ($j = 1, 2, 3, \ldots, n$), significant non-zero integer digit, $\mathrm{d}_j(i) = d_1(i)$, of a number, N, has a particular integer value, i. Those logarithmically distributed significant digits can be calculated and summarized as

First Place Digit Integer, $d_1(i)$ $i = 1, 2, 3, \ldots, 9$	Probability of Occurrence $p(d_1(i) = i)$ $i = 1, 2, 3, \ldots, 9$
1	0.30103
2	0.17609
3	0.12494
4	0.09691
5	0.07918
6	0.06695
7	0.05799
8	0.05115
9	0.04578

1. Click **Data ► Benford's Analysis.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If graphs have to be produced by using a group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select an appropriate variable representing a group variable.
- Click "**OK**" to continue or "**Cancel**" to cancel Benford's analysis.

Output example: The data set "**RandomData2500.xls**" was used. The results of the first digit analysis and the second digit analysis were computed.

Output for Benford's Analysis.

	Benford Analysis
User Selected Options	
Date/Time of Computation	1/30/2008 5:53:14 PM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\1RandomData2500.xls
Full Precision	OFF

MN_0

Number of Valid Observations	2500
Number of Distinct Observations	2500

Benford's First Digit Analysis

	0	1	2	3	4	5	6	7	8	9
Expected	0.00000	0.30103	0.17609	0.12494	0.09691	0.07918	0.06695	0.05799	0.05115	0.04576
Actual	0.00000	0.40280	0.20040	0.07920	0.05080	0.05480	0.05600	0.05360	0.05040	0.05200

Benford's Second Digit Analysis

	0	1	2	3	4	5	6	7	8	9
Expected	0.11968	0.11389	0.10882	0.10433	0.10031	0.09668	0.09337	0.09035	0.08757	0.08500
Actual	0.11760	0.11400	0.12520	0.10520	0.10640	0.09640	0.09200	0.08080	0.08920	0.07320

References

F. Benford, "The Law of Anomalous Numbers." Proceedings of the American Philosophical Society, 78, 551-572 (1938).

ProUCL 4.00.04. (2009). "ProUCL Version 4.00.04 Technical Guide." The software ProUCL 4.00.04 can be downloaded from web site at: http://www.epa.gov/esd/tsc/software.htm.

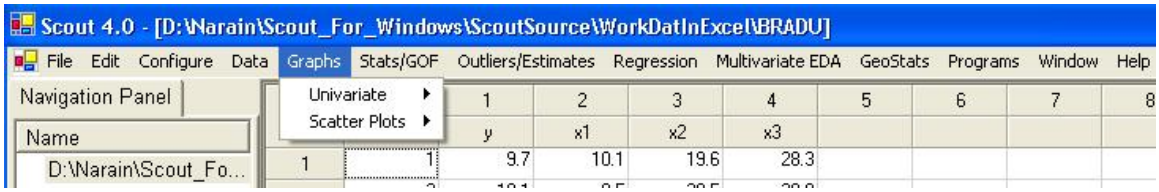
ProUCL 4.00.04. (2009). "ProUCL Version 4.00.04 User Guide." The software ProUCL 4.00.04 can be downloaded from the web site at: http://www.epa.gov/esd/tsc/software.htm.

S. Newcomb, "Note on the Frequency of Use of the Different Digits in Natural Numbers," American Journal of Mathematics, 4, 39-40 (1881).

Chapter 5

Graphs

The Graphs option provides graphical displays for both univariate and multivariate data.



5.1 Univariate Graphs

Three commonly used graphical displays are available under the Univariate Graph Option:

- Box Plots
- Histogram
- Multi-Q-Q

- The box plots and multiple Q-Q plots can be used for full data sets without non-detects and also for data sets with non-detect values.

- Three options are available to draw Q-Q plots with non-detect (ND) observations. Specifically, Q-Q plots are displayed only for detected values, with NDs replaced by ½ detection limit (DL) values, or with NDs replaced by the respective detection limits. The statistics displayed on a Q-Q plot (mean, sd, slope, and intercept) are computed according to the method used. The NDs are displayed with a smaller font and in red color.

- Scout can display box plots for data sets with NDs. This kind of graph may not be very useful if many NDs are present in the data set.

    - A few choices are available to construct box plots for data sets with NDs. For an example, all non-detects below the largest detection limit (DL) and portion of the box plot below the largest DL are not shown on the box plot. A horizontal line is displayed at the largest detection limit level.

    - Scout constructs a box plot using all of the detected and non-detect (using DL values) values. Scout shows the full box plot; however, a horizontal line is displayed at the largest detection limit.

- When multiple variables are selected, one can choose to: 1) produce multiple graphs on the same display by choosing the "**Group Graphs**" variable option, or 2) produce "**Individual Graphs**" for each selected variable.

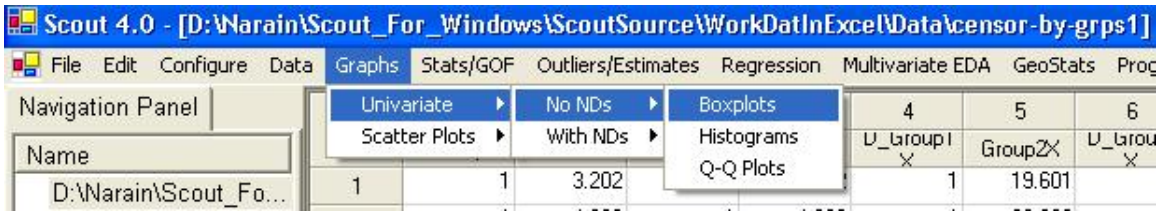
- The "**Graph by Group**" variable option produces side-by-side box plots, multiple Q-Q plots, or histograms for the groups of the selected variables representing samples obtained from multiple populations (groups). Those multiple graphs are particularly useful to perform two (background vs. site) or more sample visual comparisons.

    - Additionally, the box plot has an optional feature which can be used to draw lines at statistical limits (e.g., upper limits of background data set) computed from one population on the box plot obtained using the data from another population (e.g., a site area of concern). This type of box plot represents a useful visual comparison of site data with background threshold values (background upper limits).

    - Up to four (4) statistics can be added to a box plot. If the user inputs a value in the value column, then the check box in that row will get activated. For example, the user may want to draw horizontal lines at 80th percentile, 90th percentile, 95th percentile, or a 95% UPL on a box plot.

5.1.1 Box Plots

1. Click **Graphs ► Univariate ► No NDs or With NDs ► Box Plot.**

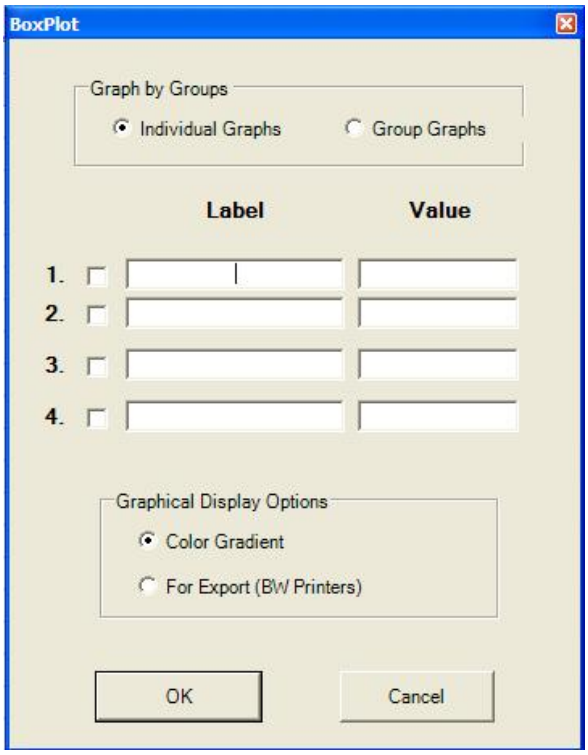


2. The "**Select Variables**" screen (Section 3.2) will appear.

    - Select one or more variables from the "**Select Variables**" screen.

    - If graphs have to be produced by using a group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select an appropriate variable representing a group variable.

- When the "**Options**" button is clicked, the following window appears.



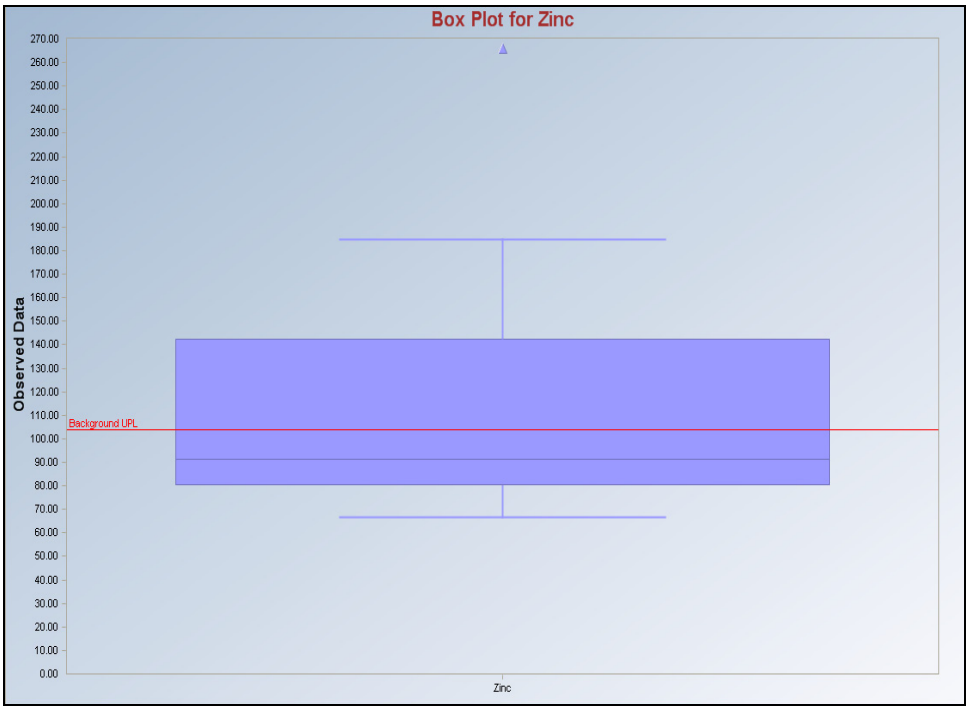
  - The default option for "**Graph by Groups**" is "**Individual Graphs**." This option will produce one graph for each selected variable. If you want to put all the selected variables into a single graph, then select the "**Group Graphs**" option. This group graphs option is used when multiple graphs categorized by a group variable have to be produced on the same graph.

  - The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to use and import graphs in black and white into a document or report, then check the radio button next to "**For Export** (**BW Printers**)."

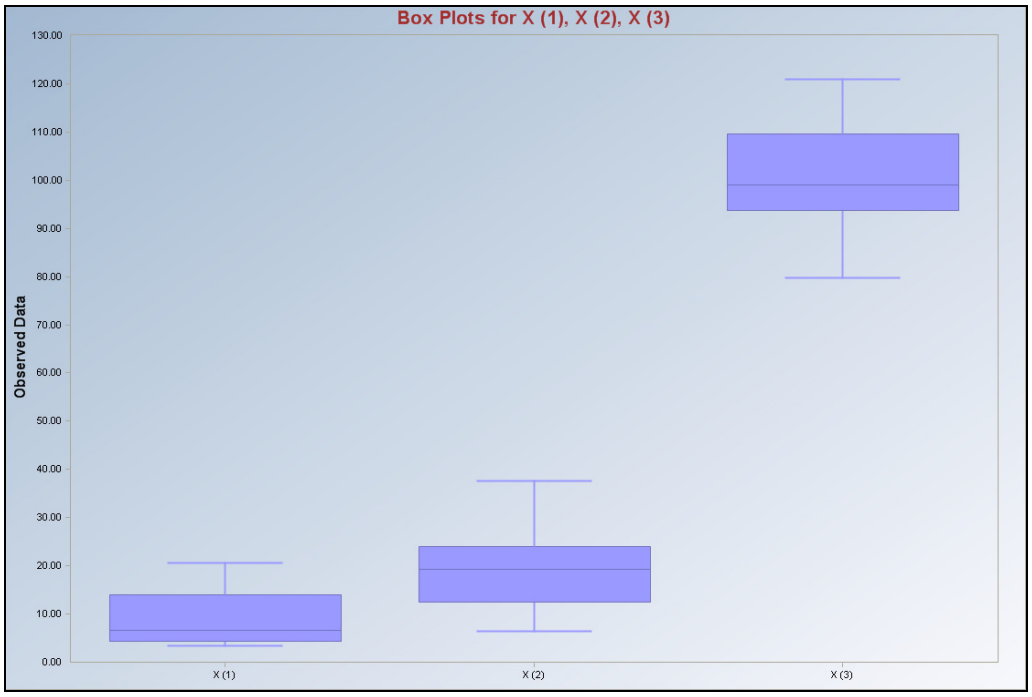
  - Click on the "**OK**" to continue or "**Cancel**" to cancel the options.

- Click on the "**OK**" to continue or "**Cancel**" to cancel the Box Plot.

Box Plot Output Screen (Single Graph).
Selected options: Label (Background UPL), Value (103.85), Individual Graphs, and Color Gradient.



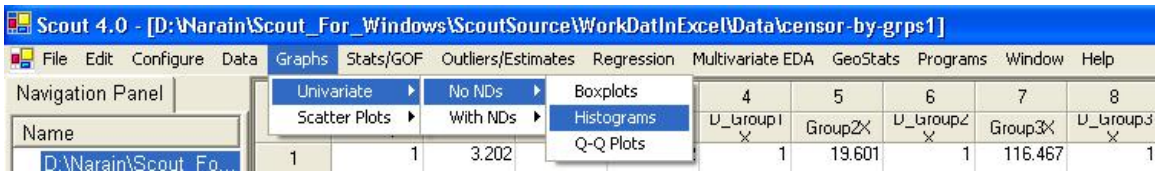
Box Plot Output Screen (Group Graphs).
Selected options: Group Graphs and Color Gradient.



5.1.2 Histograms

5.1.2.1 No NDs

1. Click **Graphs ► Univariate ► No NDs ► Histograms.**

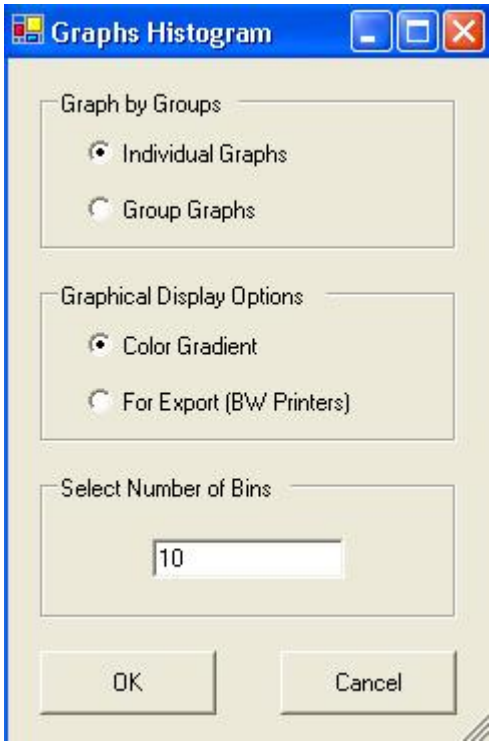


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

   - If graphs have to be produced by using a group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select an appropriate variable representing a group variable.

   - When the "**Options**" button is clicked, the following window appears.



     - The default selection for "**Graph by Groups**" is "**Individual Graphs**." This option produces a histogram (or other graphs), separately for each selected variable. If multiple graphs or graphs by

groups are desired, then check the radio button next to "**Group Graphs**."

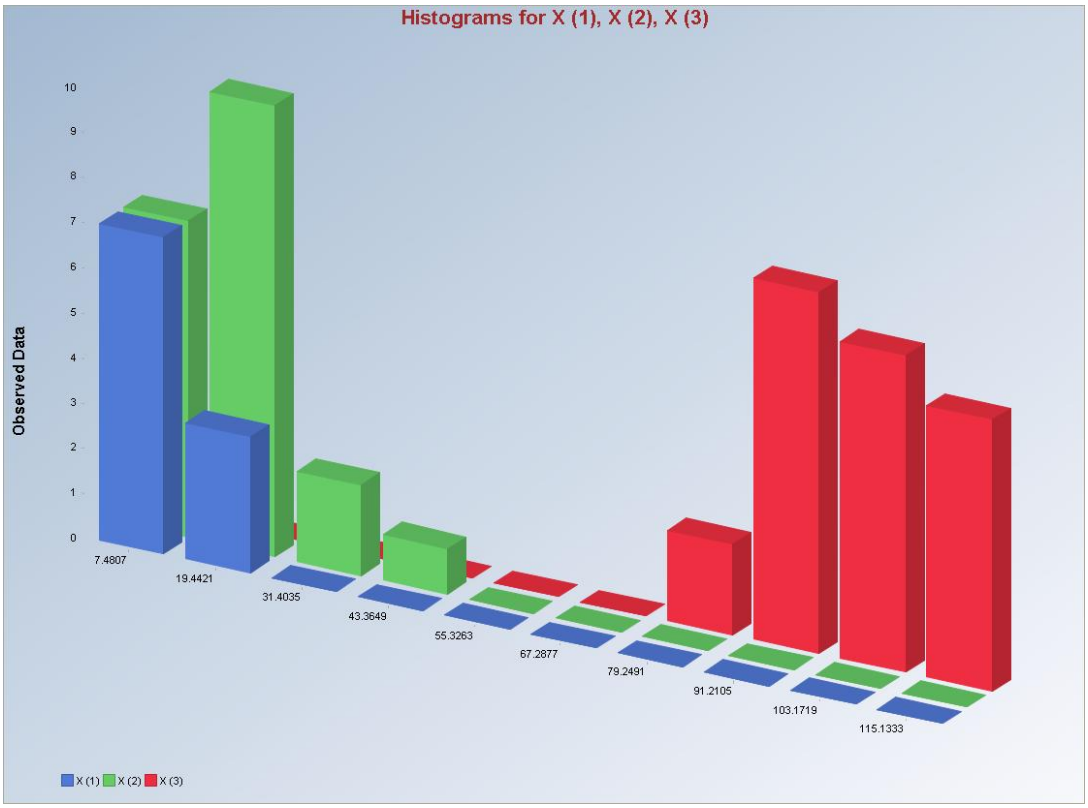
- The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to use and import graphs in black and white into a document or report, then check the radio button next to "**For Export (BW Printers)**."

- Specify the number of bins for the selected variable in "**Select Number of Bins**" text box. The default is "**10**."

- Click "**OK**" to continue or "**Cancel**" to cancel the option.

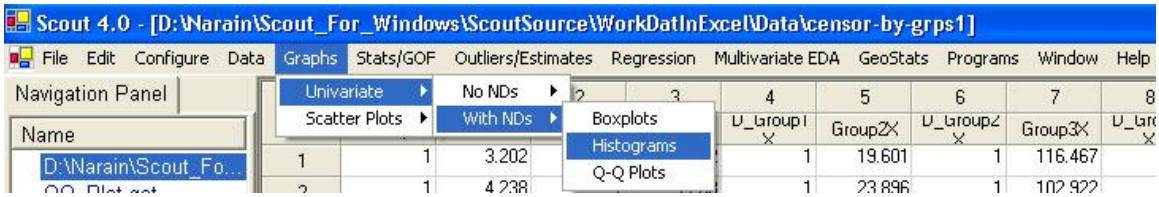
- Click "**OK**" to continue or "**Cancel**" to cancel the Histogram.

Histogram Output Screen.
Selected options: Group Graphs and Color Gradient.



5.1.2.2 With NDs

1. Click **Graphs ► Univariate ► With NDs ► Histograms.**

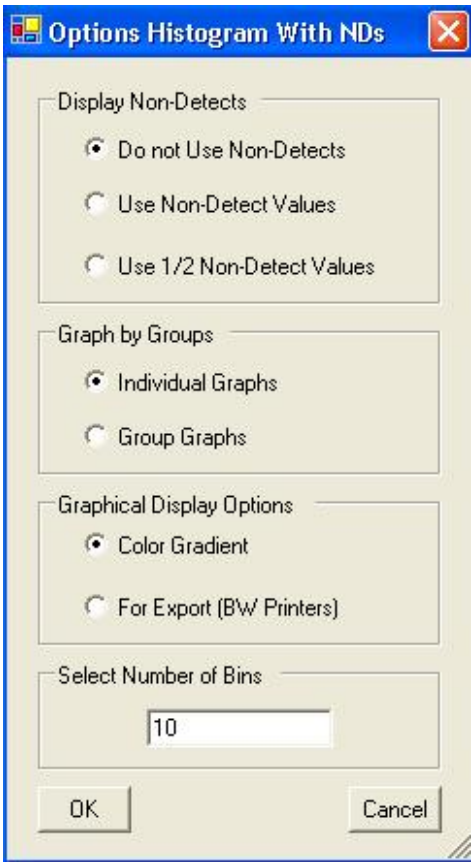


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

   - If graphs have to be produced by using a group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select an appropriate variable representing a group variable.

   - When the "**Options**" button is clicked, the following window appears.



- Specify the "**Use Non-detects**" option. The default is "**Do not Use Non-detects**."

  Do not Use Non-detects*: Selection of this option excludes the NDs detects and uses only detected values on the associated histogram.*

  Use Non-detect Values*: Selection of this option treats detection limits as detected values and uses those detection limits and detected values on the histogram.*

  Use ½ Non-detect Values*: Selection of this option replaces the detection limits with their half values, and uses half detection limits and detected values on the histogram.*

- The default selection for "**Graph by Groups**" is "**Individual Graphs**." This option produces a histogram (or other graphs) separately for each selected variable. If multiple graphs or graphs by groups are desired, then check the radio button next to "**Group Graphs**."

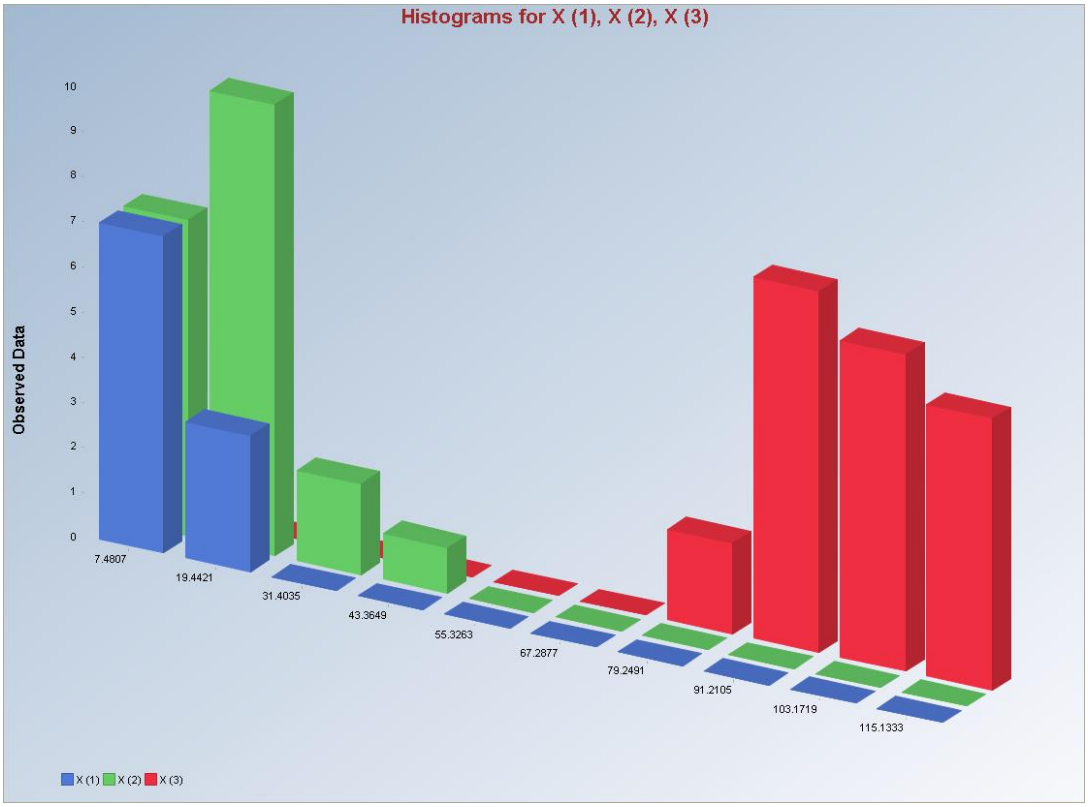
- The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to use and import graphs in black and white into a document or report, then check the radio button next to "**For Export (BW Printers)**."

- Specify the number of bins for the selected variable in "**Select Number of Bins**" text box. The default is "**10**."

- Click "**OK**" to continue or "**Cancel**" to cancel the option.

- Click "**OK**" to continue or "**Cancel**" to cancel the Histogram.

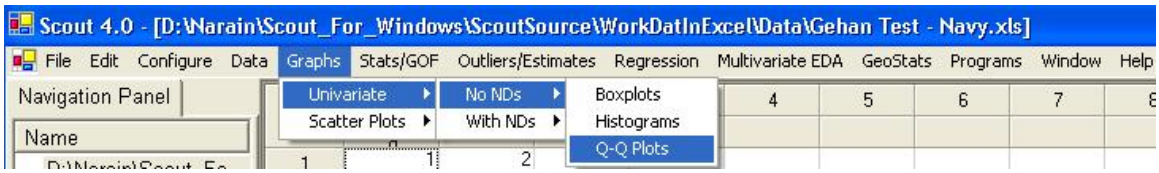
Histogram Output Screen.
Selected options: Group Graphs and Color Gradient.



5.1.3 Q-Q Plots

5.1.3.1 No NDs

1. Click **Graphs ► Univariate ► No NDs ► Q-Q Plots.**

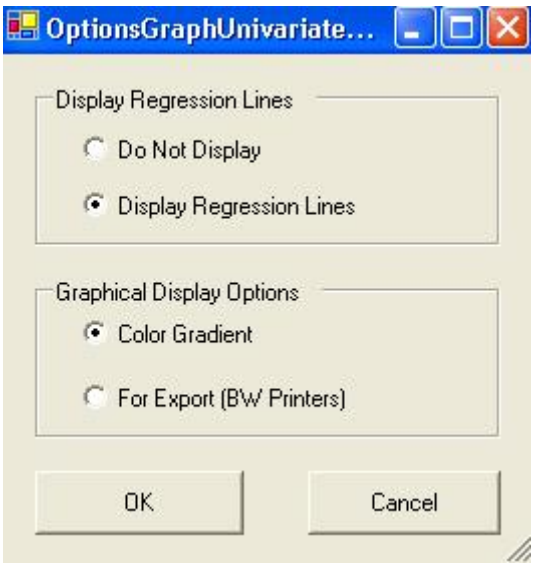


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

   - If graphs have to be produced by using a group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select an appropriate variable representing a group variable.

- When the "**Options**" button is clicked, the following window appears.



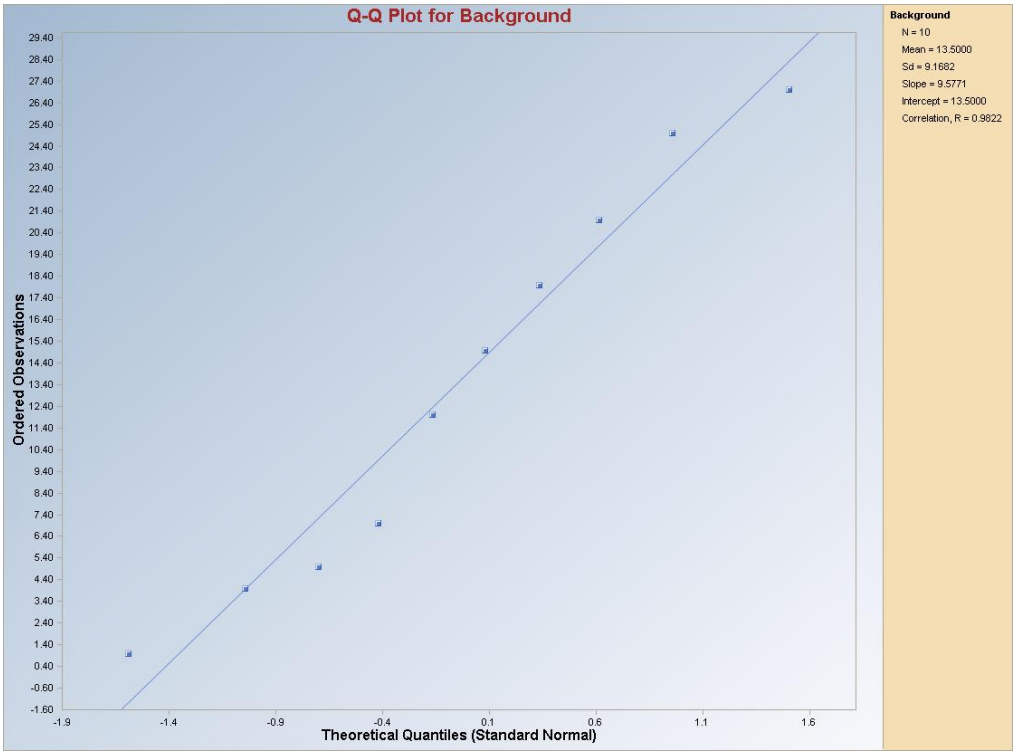
  - The default option for "**Display Regression Lines**" is "**Do Not Display**." If you want to see regression lines, then check the radio button next to "**Display Regression Lines**."

  - The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to use and import graphs in black and white into a document or report, then check the radio button next to "**For Export (BW Printers)**."

  - Click "**OK**" to continue or "**Cancel**" to cancel the option.

- Click "**OK**" to continue or "**Cancel**" to cancel the Q-Q Plot.

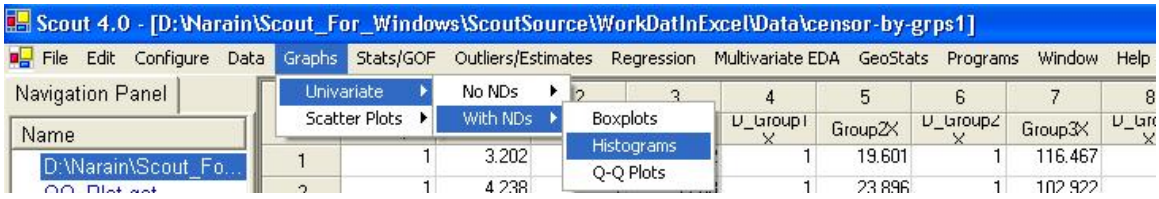
Q-Q Plot for No NDs Output Screen.



Note: *For Multi-Q-Q plot option, for both "Full" as well as for data sets "With NDs," the values along the horizontal axis represent quantiles of a standardized normal distribution (Normal distribution with mean 0 and standard deviation 1). Quantiles for other distributions (e.g., Gamma distribution) are used when using Goodness-of-Fit (GOF) test option.*

5.1.3.2 With NDs

1. Click **Graphs ► Univariate ► With NDs ► Q-Q Plots.**

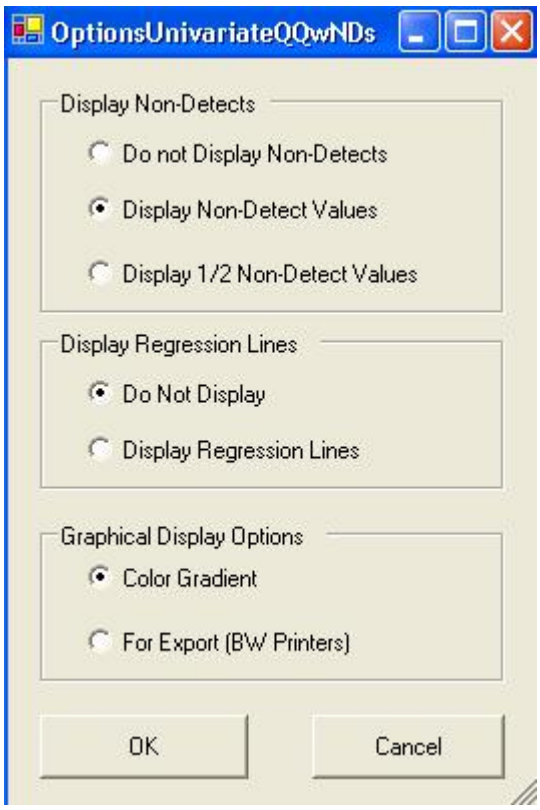


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

   - If graphs have to be produced by using a group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select an appropriate variable representing a group variable.

- When the "**Options**" button is clicked, the following window appears.



  - Specify the "**Display Non-detects**" option. The default is "**Do not Display Non-detects**."

***Do not Display Non-detects**: Selection of this option excludes the NDs detects and displays only detected values on the associated Q-Q Plot.*

***Display Non-detect Values**: Selection of this option treats detection limits as detected values and displays those detection limits and detected values on the Q-Q Plot.*

***Display ½ Non-detect Values**: Selection of this option replaces the detection limits with their half values, and it displays half detection limits and detected values on the Q-Q Plot.*

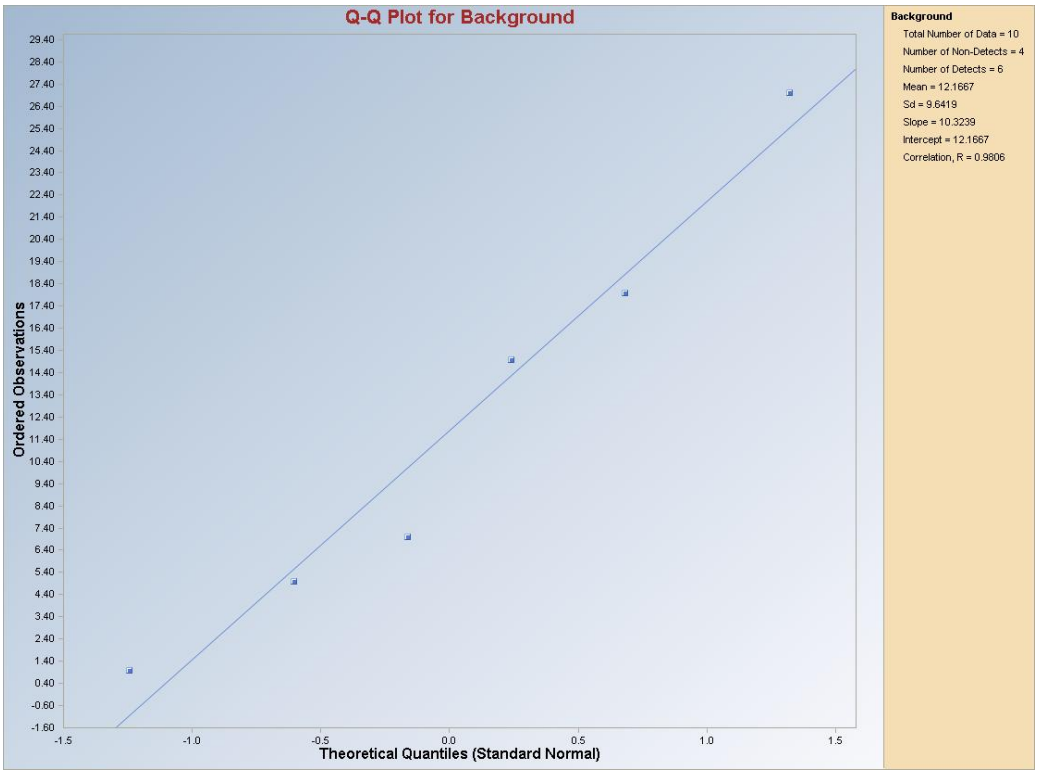
  - The default option for "**Display Regression Lines**" is "**Do Not Display**." If you want to see regression lines, then check the radio button next to "**Display Regression Lines**."

  - The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to use and import graphs in black and white into a document or report, then check the radio button next to "**For Export (BW Printers)**."

- Click "**OK**" to continue or "**Cancel**" to cancel the option.

- Click "**OK**" to continue or "**Cancel**" to cancel the Q-Q Plot.

Q-Q Plot Output Screen
Selected options: Do not Display Non-detects and Color Gradient.

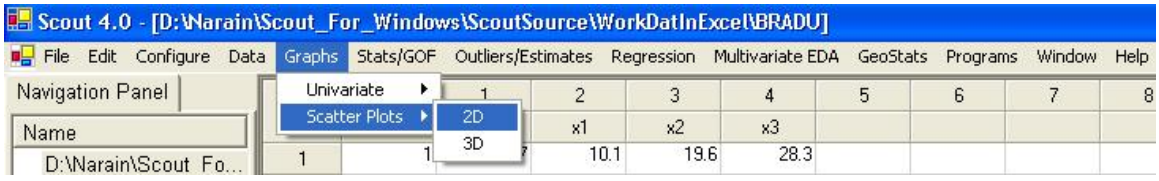


5.2 Scatter Plots

Two-dimensional (2D) and three-dimensional (3D) Scatter Plots displays are available under the Graphs Scatter Plots menu. Those graphs can be numbered according to observations or by groups if a group variable exists in the data set.

5.2.1 2D Scatter Plots

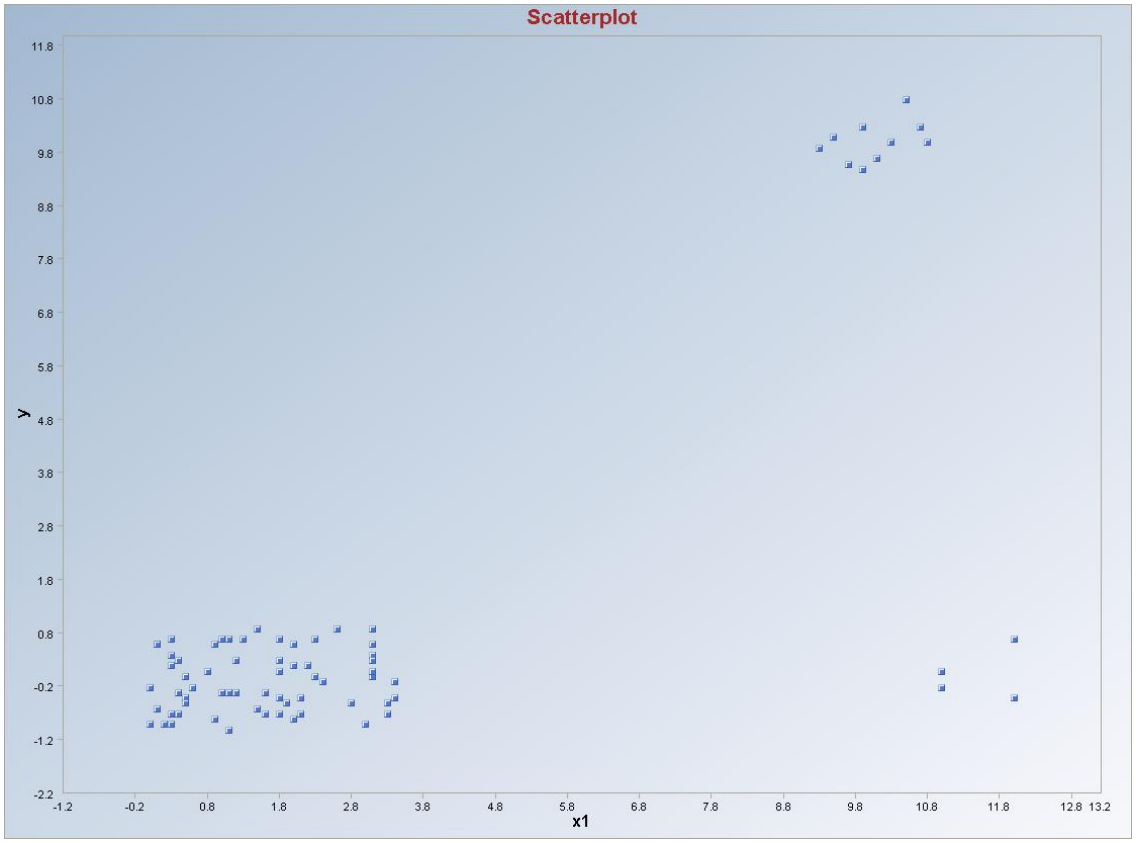
1. Click **Graphs ► Scatter Plots►2D.**



2. The "**Select Variables**" **screen** (Section 3.2) will appear.

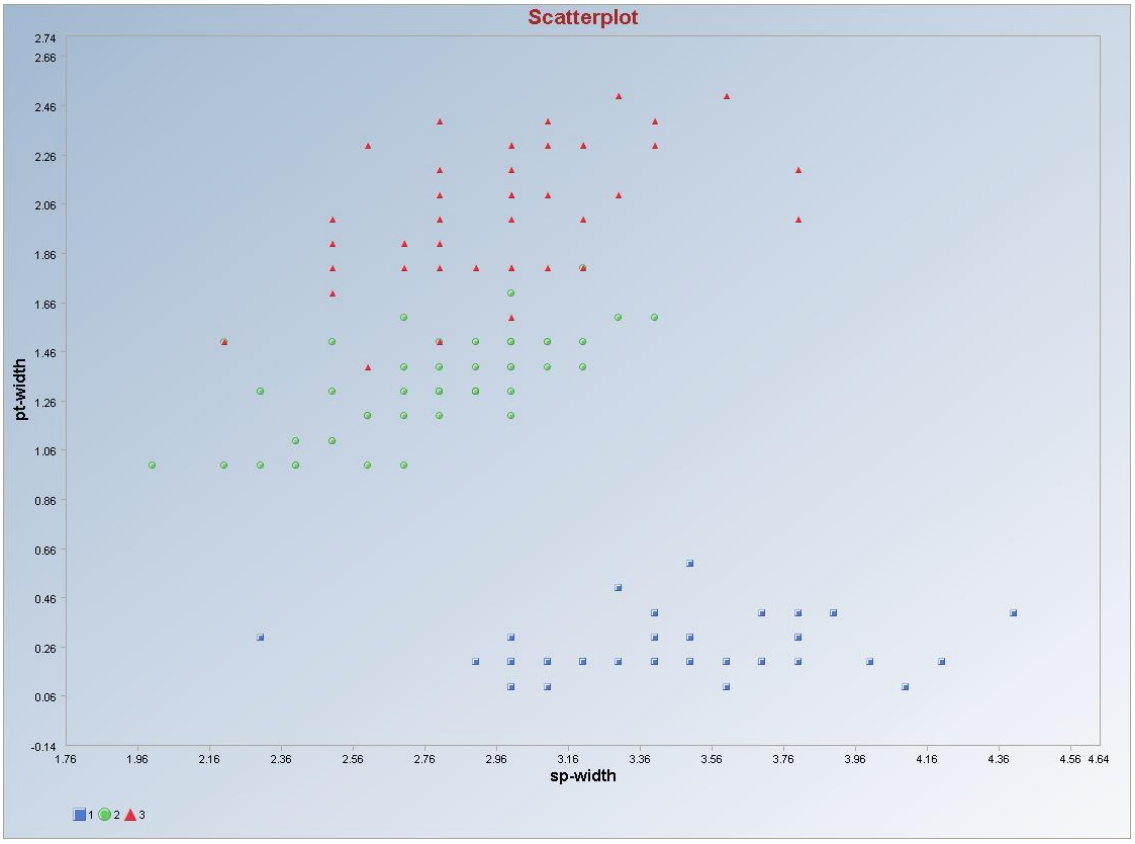
- Select two or more variables from the "**Select Variables**" screen.
- If the graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click "**OK**" to continue or "**Cancel**" to cancel the Graphs.

2D Scatter Plot.
Data Set Used: Bradu (4 variables).



The data set Bradu has four variables. The user can choose any one of the four variables for the X-axis and one of the remaining three for the Y-axis using the drop-down bars in the graphics toolbar as explained in Chapter 2. The observation numbers of the various points on the graph can be viewed by right-clicking of the mouse and using the "**Point Labels**" option.

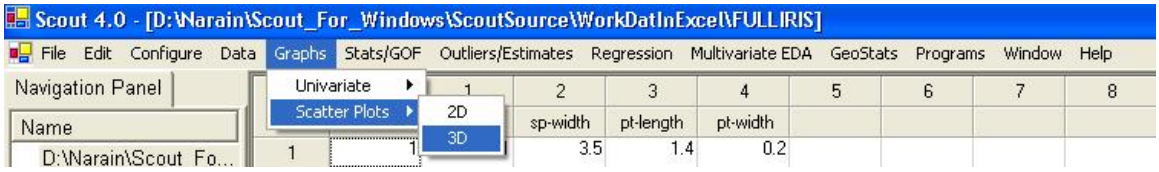
2D Scatter Plot.
Data Set Used: Iris (4 variables, 3 groups).



The user can choose any one of the four variables for the X-axis and one of the remaining three for the Y-axis using the drop-down bars in the graphics toolbar as explained in Chapter 2.

5.2.2 3D Scatter Plots

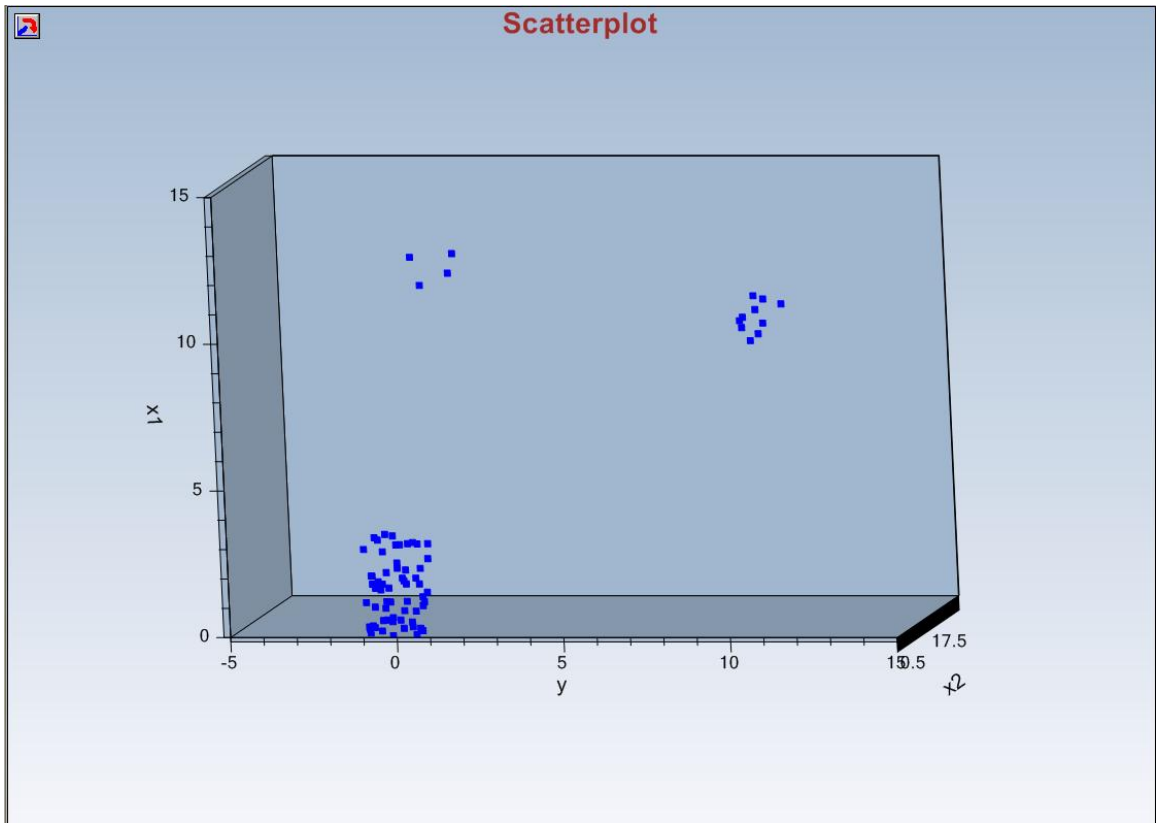
1. Click **Graphs ► Scatter Plots►3D.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

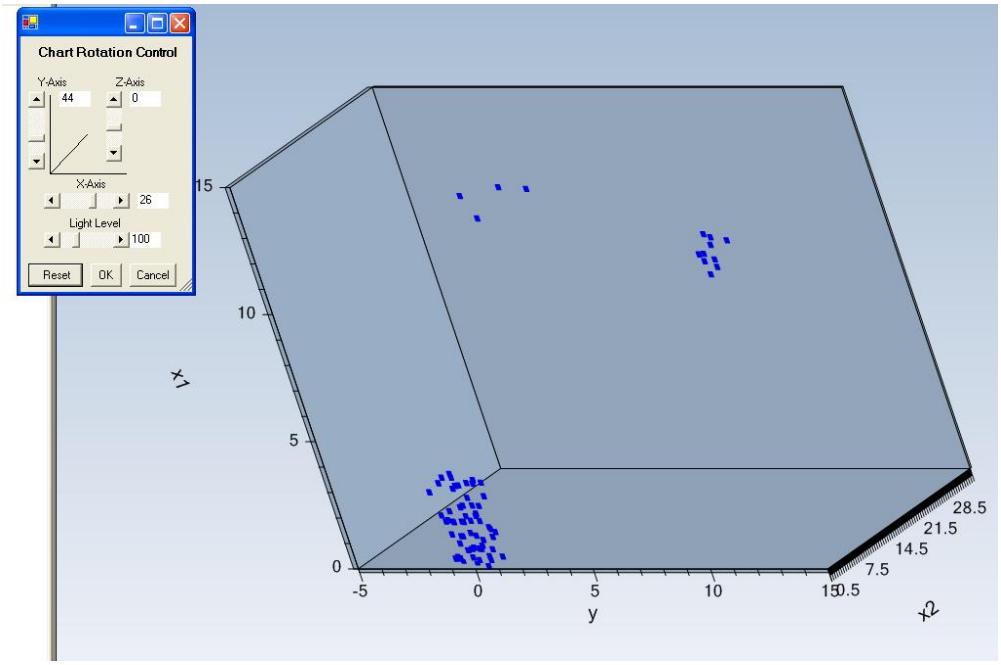
- Select two or more variables from the "**Select Variables**" screen.
- If the graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click "**OK**" to continue or "**Cancel**" to cancel the Graphs.

3D Scatter Plot.
Data Set used: Bradu.

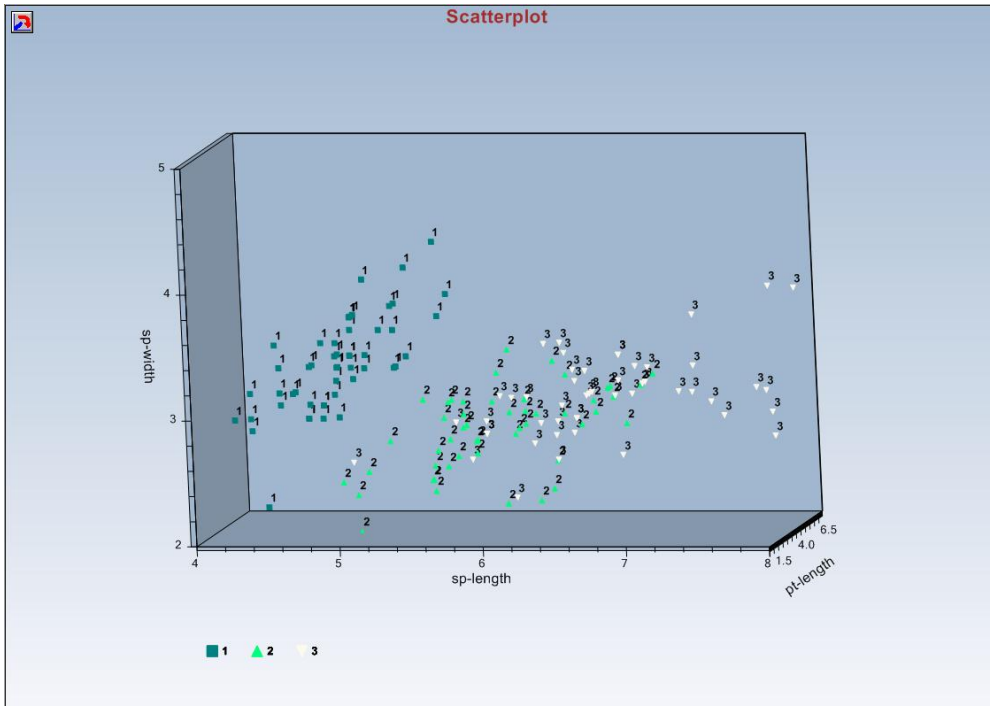


The user can choose different variables for the three axes using the drop-down bars in the graphics toolbar as explained in Chapter 2.

Rotation of axes using the Chart Rotation Control.



3D Scatter Plot using groups.
Data Set Used: Iris (4 variables, 3 groups).



Chapter 6

Goodness-of-Fit and Descriptive Statistics

6.1 Descriptive Statistics of Univariate Data

This option is used to compute general summary statistics for any or all of the variables in the data file. Summary statistics can be generated for full data sets without non-detect observations, and for data sets with non-detect observations. Two menu options: No NDs (Full) and with non-detects (NDs) are available.

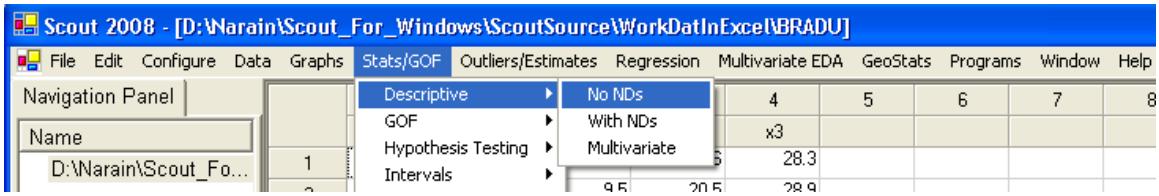
- No NDs (Full) – This option computes summary statistics for any or all of the variables in a data set without any non-detect values.

- With NDs – This option computes simple summary statistics for any or all of the variables in a data set that also have ND observations. For variables with ND observations, simple summary statistics are computed based upon the detected observations only.

- Multivariate – This option computes the mean vector, the median vector, the standard deviation vector, the covariance matrix and the correlation matrix for the multivariate data.

6.1.1 Descriptive (Summary) Statistics for Data Sets with No Non-detects

1. Click **Stats/GOF ► Descriptive ► No NDs.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

   - If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

   - Click "**OK**" to continue or "**Cancel**" to cancel the Descriptive Statistics.

- The following summary statistics are available for the variables selected.
    - Number of Observations
    - Number of Missing Values
    - Minimum Observed Value
    - Maximum Observed Value
    - Mean = Sample Average Value
    - Q1 = 25th Percentile
    - Q2 = Median
    - Q3 = 75th Percentile
    - 90th Percentile
    - 95th Percentile
    - 99th Percentile
    - (Sample) Standard Deviation
    - MAD = Median Absolute Deviation
    - MAD/0.675 = Robust Estimate of Variability, Population Standard Deviation, σ
    - Skewness = Skewness Statistic
    - Kurtosis = Kurtosis Statistic
    - CV = Coefficient of Variation

- The details of these descriptive (summary) statistics are described in the EPA (2006) guidance.

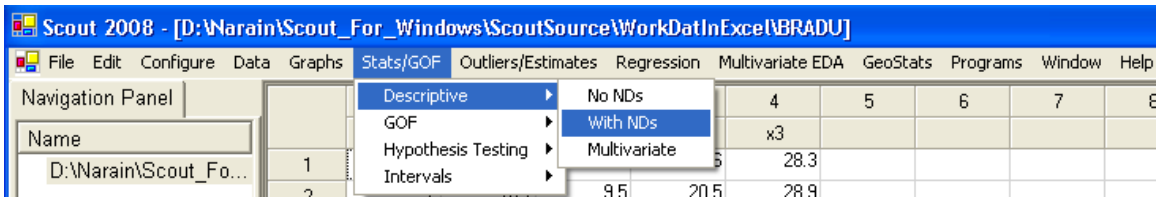
Output for Descriptive Statistics – No Non-detects (NDs).

	Univariate Descriptive Statistics for Datasets with No NDs				
Date/Time of Computation	5/28/2007 5:42:00 PM				
User Selected Options					
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\IRIS.xls				
Full Precision	OFF				
Var 0:	**sp-length**				
Var 2:	**pt-length**				
	Var 0:	**sp-width**	**Var 2:**	**pt-width**	
Number of Observations	50	50	50	50	
Number of Missing Values	0	0	0	0	
Minimum Observed Value	4.3	2.3	1	0.1	
Maximum Observed Value	5.8	4.4	1.9	0.6	
Mean	5.006	3.428	1.462	0.246	
(Q1) 25% Percentile	4.8	3.15	1.4	0.2	
(Q2) Median	5	3.4	1.5	0.2	
(Q3) 75% Percentile	5.2	3.65	1.55	0.3	
90% Percentile	5.4	3.9	1.7	0.4	
95% Percentile	5.6	4.05	1.7	0.4	
99% Percentile	5.75	4.3	1.9	0.55	
Standard Deviation	0.352	0.379	0.174	0.105	
MAD / 0.6745	0.297	0.371	0.148	0	
Skewness	0.12	0.0412	0.106	1.254	
Kurtosis	-0.253	0.955	1.022	1.719	
CV	0.0704	0.111	0.119	0.428	

***Note**: When the variable name is too long to fit in a single cell, then the variable number and its name are printed above the results table. In the above output sheet, the variable, **sp-length**, was chosen as the first variable and variable, **pt-length**, was chosen as the third variable. The names of those two variables cannot fit in individual cells of the descriptive statistics table; hence they are named as **Var 0** and **Var 2**, respectively, in the table.*

6.1.2 Descriptive (Summary) Statistics for Data Sets with Non-detects

1. Click **Stats/GOF ► Descriptive ► With NDs.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select a variable(s) from the list of variables.
   - Only those variables that have non-detect values will be shown.
   - If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
   - Click "**OK**" to continue or "**Cancel**" to cancel the Descriptive Statistics.

   - The following summary statistics are available for the variables selected.

     - Number of Observations
     - Number of Missing Values
     - Number of Detects
     - Number of Non-detects
     - Percentage of Non-detects
     - Minimum Observed Detected Value
     - Maximum Minimum Observed Detected Value
     - Mean of Detected Values
     - Median of Detected Values
     - Standard Deviation of Detected Values
     - MAD/0.675 of Detected Values = Robust Estimate of Variability (standard deviation)
     - Skewness of Detected Values
     - Kurtosis of Detected Values
     - CV = Detected Values Coefficient of Variation
     - Q1 = 25th Percentile of All Observations
     - Q2 = Median of All Observations
     - Q3 = 75th Percentile of All Observations
     - 90th Percentile of All Observations
     - 95th Percentile of All observations

- 99th Percentile of All Observations

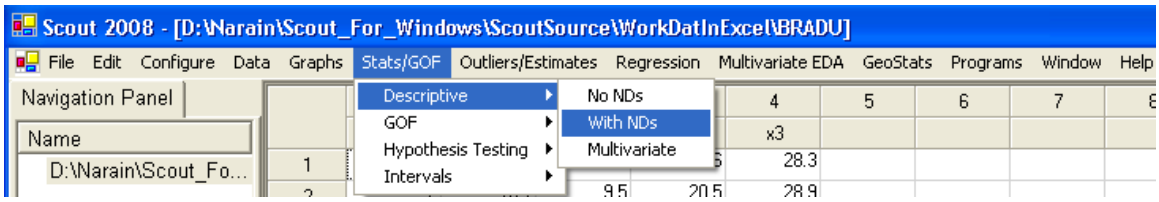
Note: *In Scout, "Descriptive Statistics" for a data set with non-detect observations represent simple summary statistics based upon, and calculated from, the data set without using non-detect observations. The simple "Descriptive Statistics /Univariate/ With NDs" option only provides simple statistics (e.g., % NDs, max ND, Min ND, Mean of detected values) based upon the detected values only. Those statistics may help a user to determine the degree of skewness (e.g., mild, moderate or high) of the data set consisting of detected values. Those statistics may also help the user to choose the most appropriate method (e.g., KM (BCA) UCL or KM (t) UCL) to compute confidence, prediction and tolerance intervals.*

Output for Descriptive Statistics – With Non-detects.

	Univariate Descriptive Statistics for Datasets with NDs
Date/Time of Computation	5/28/2007 5:44:23 PM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1.xls
Full Precision	OFF
	X
Number of Observations	53
Number of Missing Values	0
Number of Detects	49
Number of Non-Detects	4
Percentage of Non-Detects	7.547%
Minimum Observed Detect Value	3.202
Maximum Observed Detect Value	121.1
Mean of Detect values	55.05
Median of Detect values	31.57
Standard Deviation of Detect values	43.2
MAD / 0.6745 of Detect values	46.8
Skewness of Detect values	0.149
Kurtosis of Detect values	-1.758
CV of Detect values	0.785
(Q1) 25% Percentile (All Obs)	9.608
(Q2) Median (All Obs)	31.57
(Q3) 75% Percentile (All Obs)	95.73
90% Percentile (All Obs)	107.6
95% Percentile (All Obs)	112.9
99% Percentile (All Obs)	118.7

6.1.3 Descriptive Statistics for Multivariate Data

1. Click **Stats/GOF ► Descriptive ► Multivariate.**



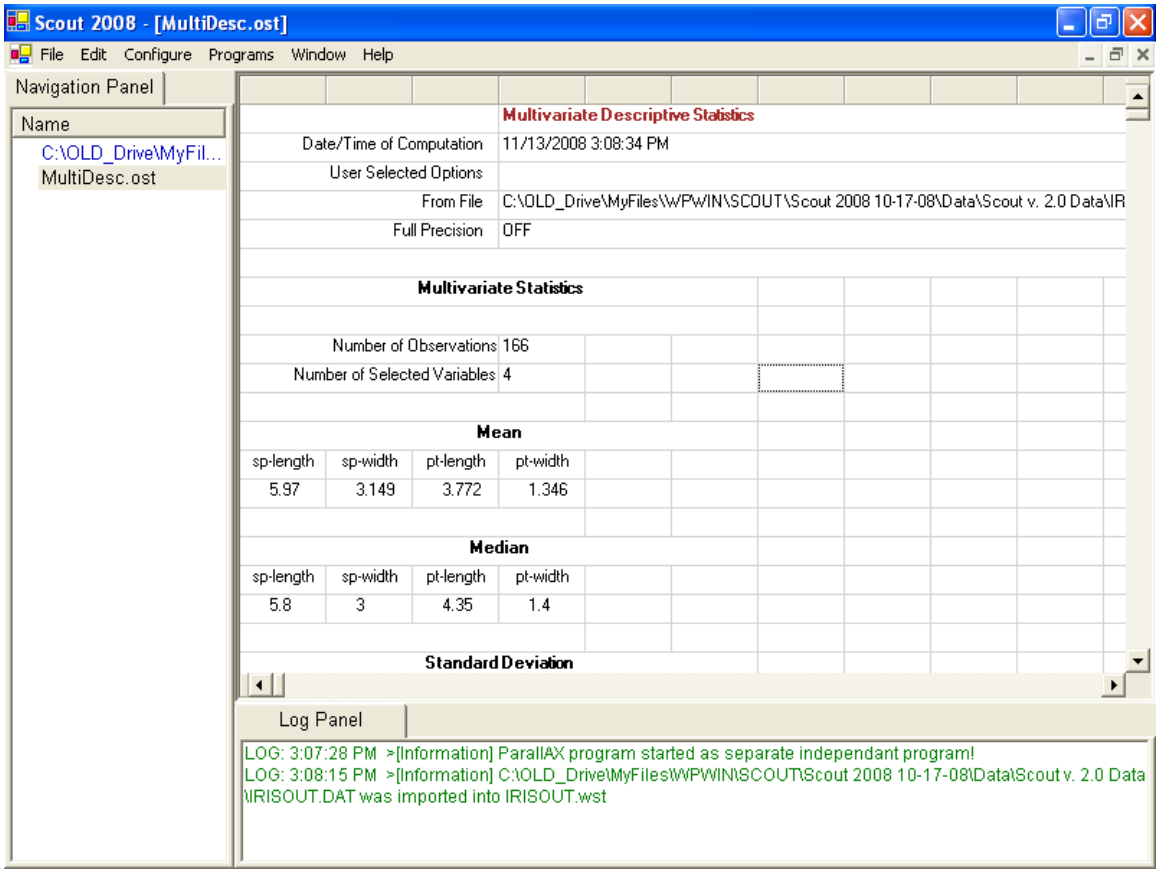
2. The "**Select Variables**" screen (Section 3.2) will appear.

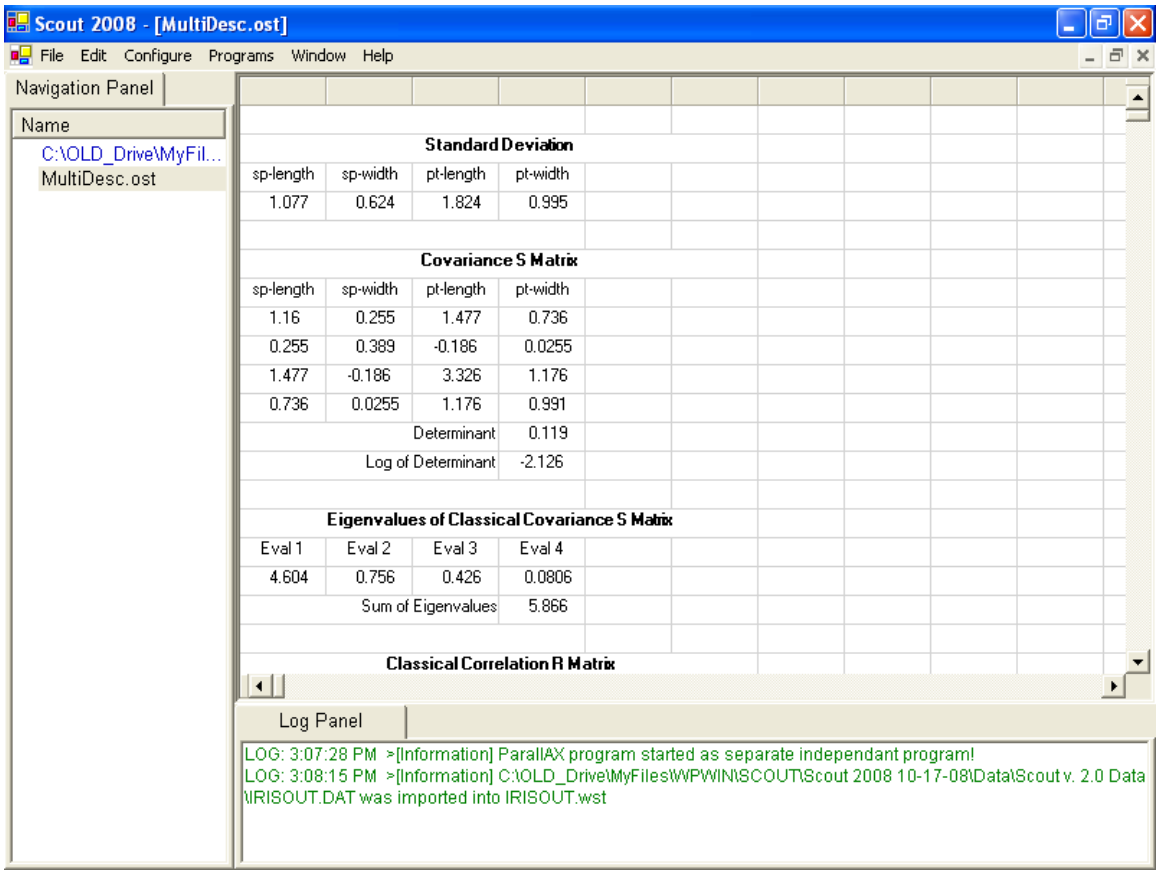
   - Select a variable(s) from the list of variables.

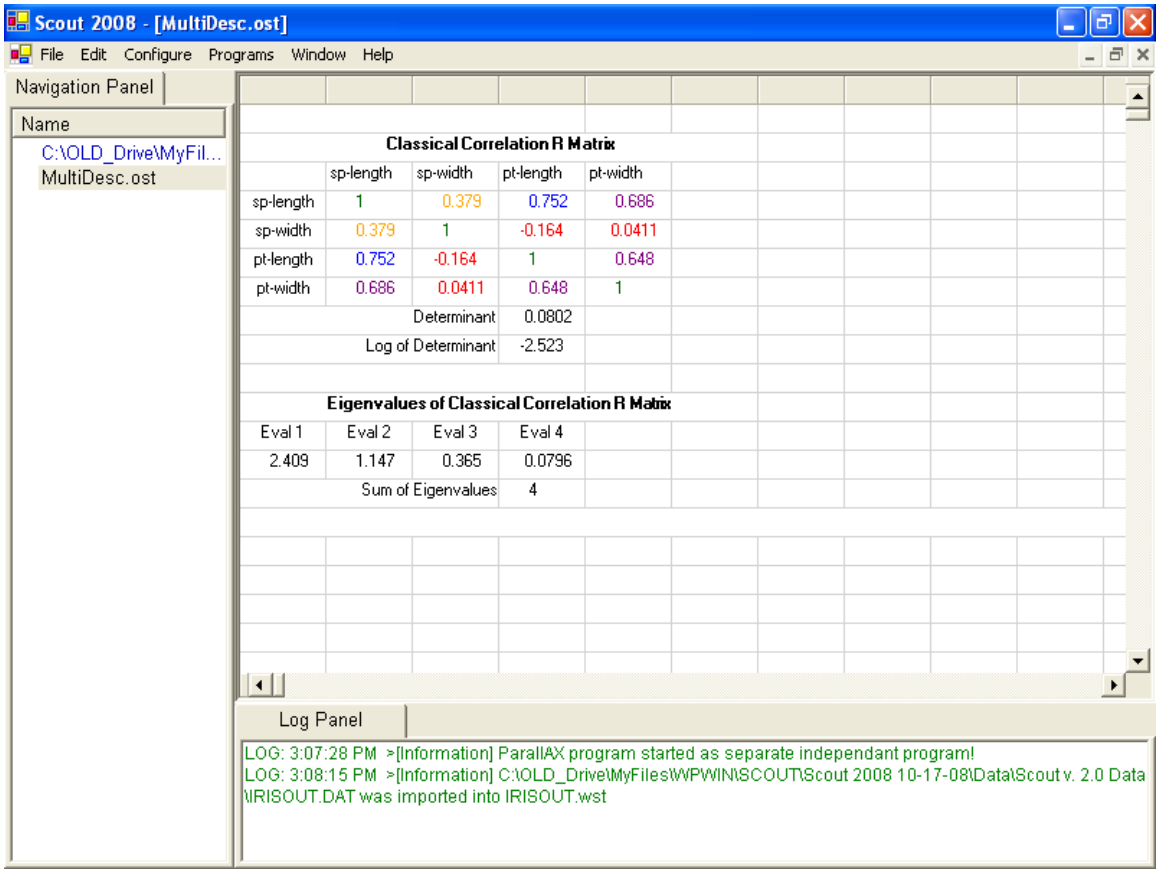
   - If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

   - Click "**OK**" to continue or "**Cancel**" to cancel the Descriptive Statistics.

Output for Descriptive Statistics – Multivariate.




Scout 2008 - [MultiDesc.ost]
File Edit Configure Programs Window Help
Navigation Panel
Name
C:\OLD_Drive\MyFil...
MultiDesc.ost
Standard Deviation
sp-length sp-width pt-length pt-width
1.077 0.624 1.824 0.995
Covariance S Matrix
sp-length sp-width pt-length pt-width
1.16 0.255 1.477 0.736
0.255 0.389 -0.186 0.0255
1.477 -0.186 3.326 1.176
0.736 0.0255 1.176 0.991
Determinant 0.119
Log of Determinant -2.126
Eigenvalues of Classical Covariance S Matrix
Eval 1 Eval 2 Eval 3 Eval 4
4.604 0.756 0.426 0.0806
Sum of Eigenvalues 5.866
Classical Correlation R Matrix
Log Panel
LOG: 3:07:28 PM >[Information] ParallAX program started as separate independant program!
LOG: 3:08:15 PM >[Information] C:\OLD_Drive\MyFiles\WPWIN\SCOUT\Scout 2008 10-17-08\Data\Scout v. 2.0 Data \IRISOUT.DAT was imported into IRISOUT.wst


Scout 2008 - [MultiDesc.ost]
File Edit Configure Programs Window Help
Navigation Panel
Name
C:\OLD_Drive\MyFil...
MultiDesc.ost
Classical Correlation R Matrix
sp-length sp-width pt-length pt-width
sp-length 1 0.379 0.752 0.686
sp-width 0.379 1 -0.164 0.0411
pt-length 0.752 -0.164 1 0.648
pt-width 0.686 0.0411 0.648 1
Determinant 0.0802
Log of Determinant -2.523
Eigenvalues of Classical Correlation R Matrix
Eval 1 Eval 2 Eval 3 Eval 4
2.409 1.147 0.365 0.0796
Sum of Eigenvalues 4
Log Panel
LOG: 3:07:28 PM >[Information] ParallAX program started as separate independant program!
LOG: 3:08:15 PM >[Information] C:\OLD_Drive\MyFiles\WPWIN\SCOUT\Scout 2008 10-17-08\Data\Scout v. 2.0 Data \IRISOUT.DAT was imported into IRISOUT.wst

	Multivariate Descriptive Statistics
Date/Time of Computation	3/13/2008 6:27:08 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BRADU
Full Precision	OFF

Multivariate Statistics

Number of Observations	75
Number of Selected Variables	4

Mean

y	x1	x2	x3
1.279	3.207	5.597	7.231

Median

y	x1	x2	x3
0.1	1.8	2.2	2.1

Standard Deviation

y	x1	x2	x3
3.493	3.653	8.239	11.74

Covariance S Matrix

y	x1	x2	x3
12.2	9.477	20.39	31.03
9.477	13.34	28.47	41.24
20.39	28.47	67.88	94.67
31.03	41.24	94.67	137.8
		Determinant	1906
		Log of Determinant	7.553

Eigenvalues of Classical Covariance S Matrix

Eval 1	Eval 2	Eval 3	Eval 4
0.914	1.688	5.538	223.1
		Sum of Eigenvalues	231.3

6.2 Goodness-of-Fit (GOF)

Several goodness-of-fit (GOF) tests for univariate data (both for full data sets, i.e., without non-detects, and for data sets with NDs) and multivariate data are available in Scout. In this user guide, those tests and available options have been illustrated using screen shots generated by Scout. For more details about those tests, refer to the ProUCL 4.00.04 Technical Guide and the Scout Technical Guide (in preparation).

6.2.1 Univariate GOF

Two choices are available for the goodness-of-fit menu: No NDs (Full) and With NDs.

- **No NDs (Full)**
  - This option is used to analyze full data sets without any non-detect observations.
  - This option tests for the normal, gamma, or lognormal distribution of the variables selected using the Select Variables option.
  - GOF Statistics: this option simply generates an output log of the GOF test statistics and any derived conclusions about the data distributions of all selected variables.

- **With NDs**
  - Analyzes data sets that have both non-detected and detected values.
  - Six sub-menu items listed and shown below are available for this option.

    1. Exclude NDs
    2. Normal ROS Estimates
    3. Gamma ROS Estimates
    4. Lognormal ROS Estimates
    5. DL/2 Estimates
    6. GOF Statistics

Scout handles Univariate GOF tests in the same way as ProUCL 4.00.04. More information can be obtained from the ProUCL 4.00.04 Technical Guide and User Guide (Chapter 8). The major upgrade in Scout for the GOF test of univariate data from ProUCL 4.00.04 is the presence of Shapiro-Wilk's test for observations greater than 50 and less than 2000 (Royston 1982).

Classical Correlation R Matrix

	y	x1	x2	x3
y	1	0.743	0.708	0.757
x1	0.743	1	0.946	0.962
x2	0.708	0.946	1	0.979
x3	0.757	0.962	0.979	1
		Determinant	0.00125	
		Log of Determinant	-6.683	

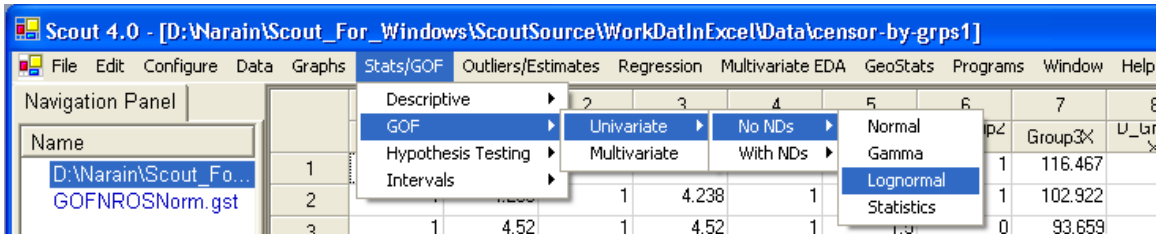
Eigenvalues of Classical Correlation R Matrix

Eval 1	Eval 2	Eval 3	Eval 4
0.0172	0.0556	0.368	3.559
		Sum of Eigenvalues	4

6.2.1.1 GOF Tests for Data Sets with No NDs

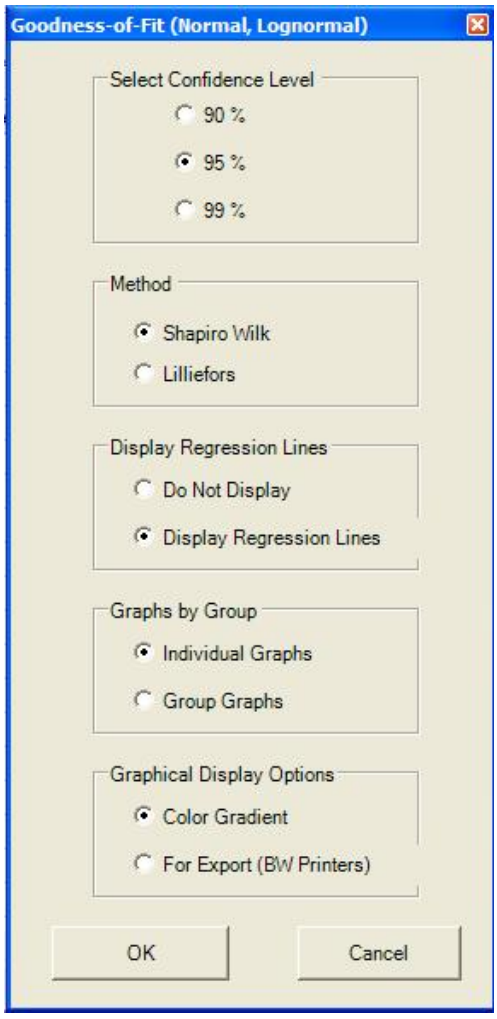
6.2.1.1.1 GOF Tests for Normal and Lognormal Distribution

Click **Stats/GOF ► GOF ► Univariate ► No NDs ► Normal or Lognormal.**



The "**Select Variables**" screen (section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click "**Options**" for GOF options.



- The default option for the "**Select Confidence Level**" is "**95%.**"

- The default GOF method is "**Shapiro Wilk**." If the sample size is greater than 50, the program automatically uses the "**Lilliefors**" test.

- The default method for "**Display Regression Lines**" **is** "**Do Not Display**." If you want to see regression lines on a Q-Q plot, then check the radio button next to Display Regression Lines.

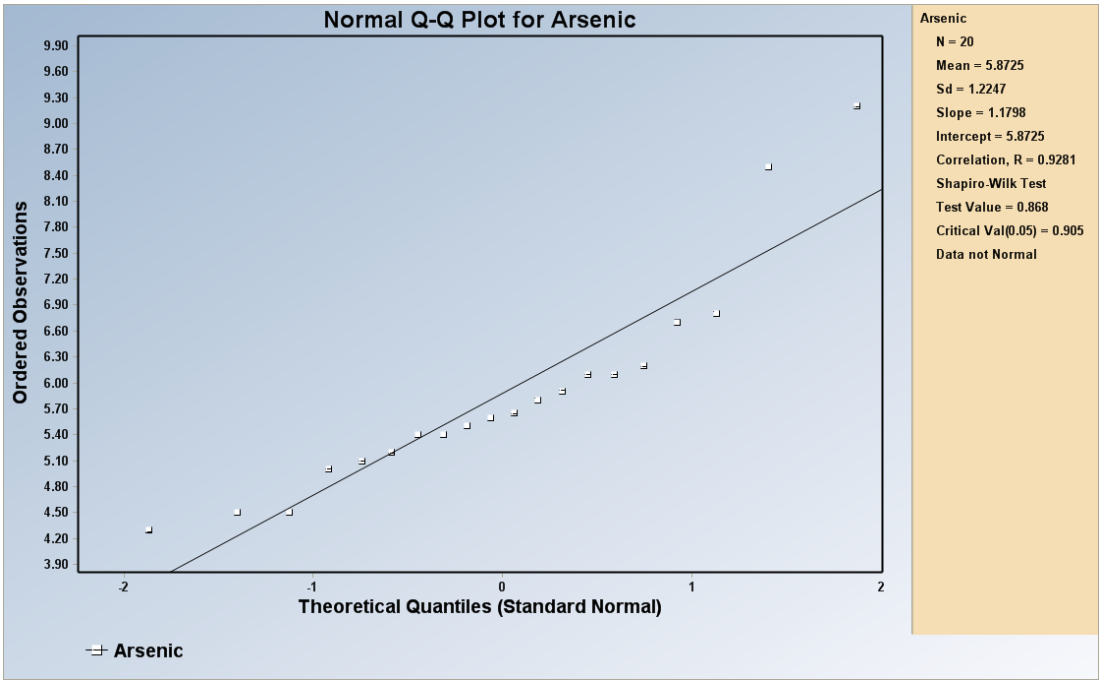
- The default option for "**Graphs by Group**" is "**Individual Graphs**." If you want to see the plots for all selected variables on a single graph, then check the radio button next to Group Graphs.

Note: *This option for Graphs by Group is specifically provided when the user wants to display multiple graphs for a variable by a group variable (e.g., site AOC1, site AOC2, and background). This kind of display represents a useful visual comparison of the values of a variable (e.g., concentrations of COPC-Arsenic) collected from two or more groups (e.g., upgradient wells, monitoring wells, residential wells).*

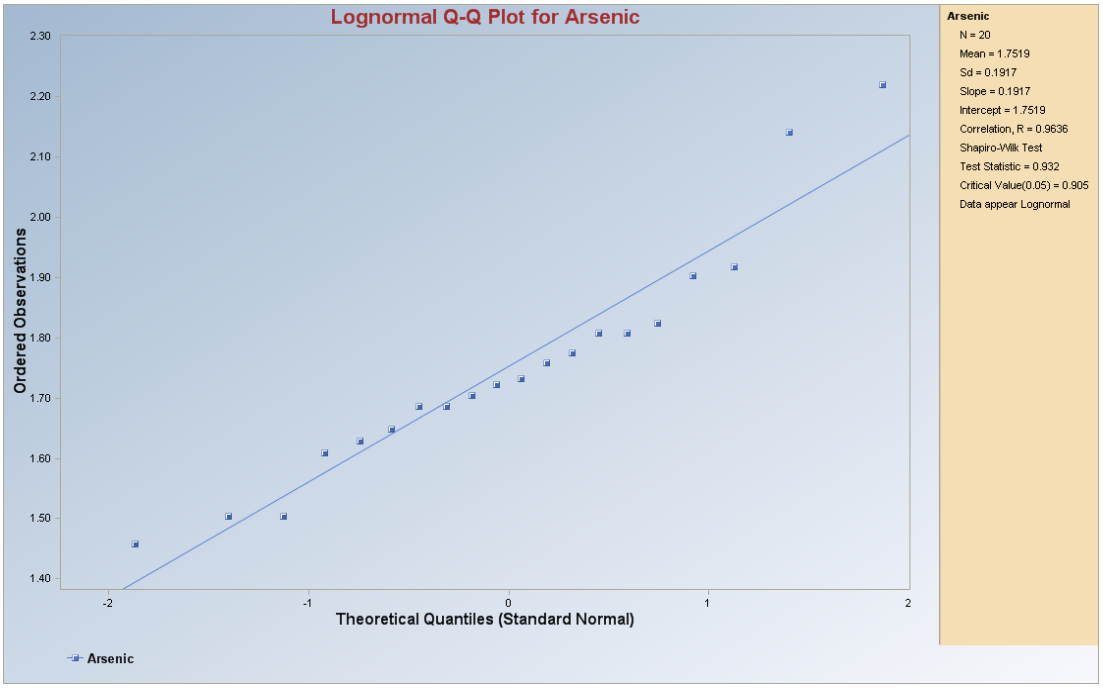
- The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to see the graphs in black and white to be included in reports for later use, then check the radio button next to For Export (BW Printers).

- Click "**OK**" to continue or "**Cancel**" to cancel the goodness-of-fit tests.

Output Screen for Normal Distribution (Full).
Selected options: Shapiro Wilk, Display Regression Line, and For Export (BW Printers).

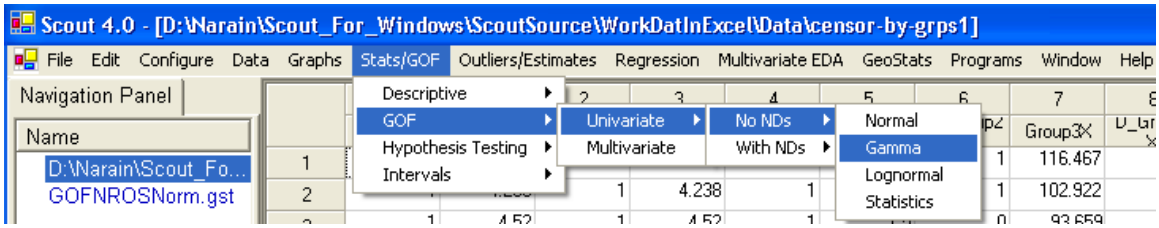


Output Screen for Lognormal Distribution (Full).
Selected options: Shapiro Wilk, Display Regression Lines, and Color Gradient.



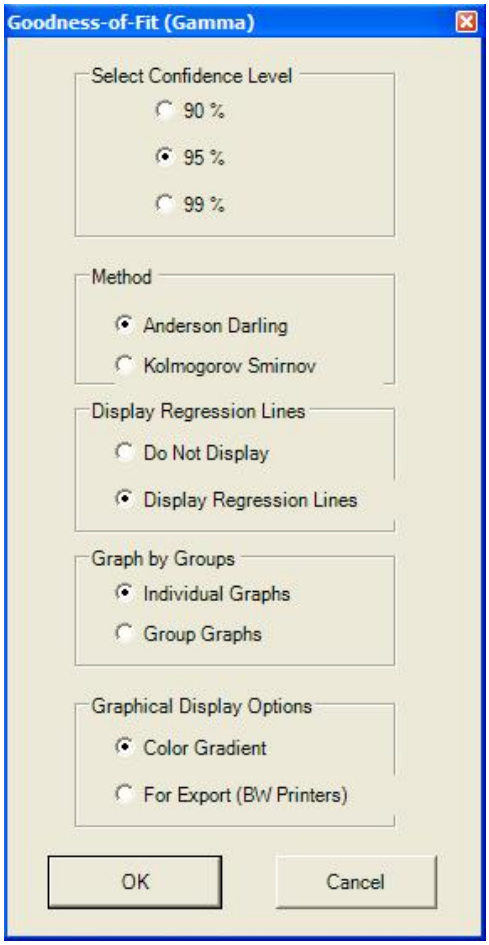
6.2.1.1.2 GOF Tests for Gamma Distribution

Click **Stats/GOF ► GOF ► Univariate ► No NDs ► Gamma.**



The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click "**Options**" for GOF options.



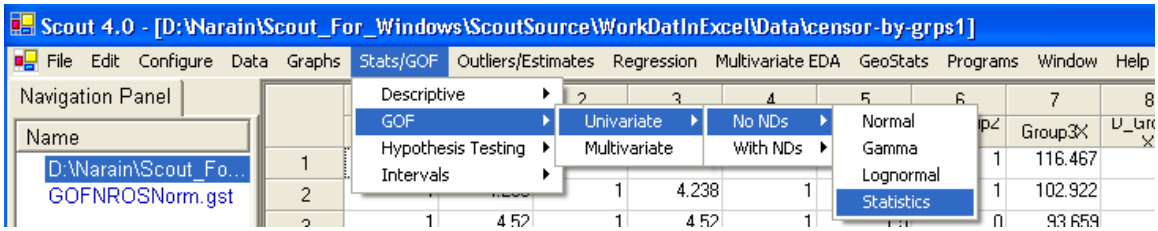
- The default option for the "**Confidence Level**" is "**95%**."
- The default GOF method is "**Anderson Darling**."
- The default option for "**Display Regression Lines**" is "**Do Not Display**." If you want to see regression lines on the Gamma Q-Q plot, then check the radio button next to "**Display Regression Lines**."
- The default option for "**Graph by Groups**" is "**Individual Graphs**." If you want to see the graphs for all the selected variables into a single graph, then check the radio button next to "**Group Graphs**."
- The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to see the graphs in black and white, check the radio button next to "**For Export (BW Printers)**."
- Click "**OK**" to continue or "**Cancel**" to cancel the option.

- Click "**OK**" to continue or "**Cancel**" to cancel the goodness-of-fit tests.

Output Screen for Gamma Distribution (Full).
Selected options: Anderson Darling, Display Regression Lines, Individual Graphs, and Color Gradient.

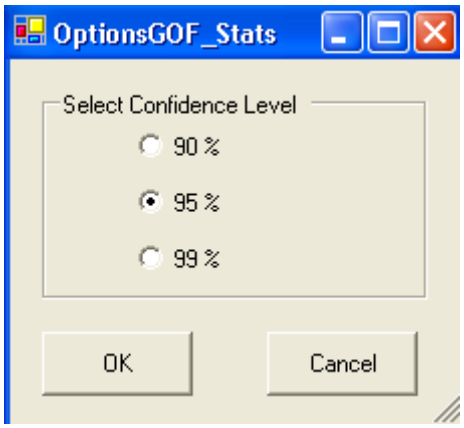
6.2.1.1.3 GOF Statistics

1. Click **Stats/GOF ► GOF ► Univariate ► No NDs ► Statistics.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click "**Options**" for GOF options.



- The default option for the "**Confidence Level**" is "**95%**."
- Click "OK" to continue or "Cancel" to cancel the option.

- Click "OK" to continue or "Cancel" to cancel the Goodness-of-Fit Statistics.

Output for GOF Statistics for univariate data without Non-detects.

	Goodness-of-Fit Test Statistics for Full Data Sets without Non-Detects
Date/Time of Computation	1/14/2008 4:05:46 PM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BEETLES
Full Precision	OFF
Confidence Coefficient	0.95

x2

Raw Statistics	
Number of Valid Samples	74
Number of Distinct Samples	9
Minimum	8
Maximum	16
Mean of Raw Data	12.99
Standard Deviation of Raw Data	2.142
Kstar	32.67
Mean of Log Transformed Data	2.549
Standard Deviation of Log Transformed Data	0.177

Normal Distribution Test Results	
Shapiro Wilk Test Statistic	0.894
Shapiro Wilk Critical (0.95) Value	0.95
Lilliefors Test Statistic	0.195
Lilliefors Critical (0.95) Value	0.103

Data not Normal at (0.05) Significance Level

Gamma Distribution Test Results	
A-D Test Statistic	3.183
A-D Critical (0.95) Value	0.749
K-S Test Statistic	0.214
K-S Critical(0.95) Value	0.103

Data not Gamma Distributed at (0.05) Significance Level

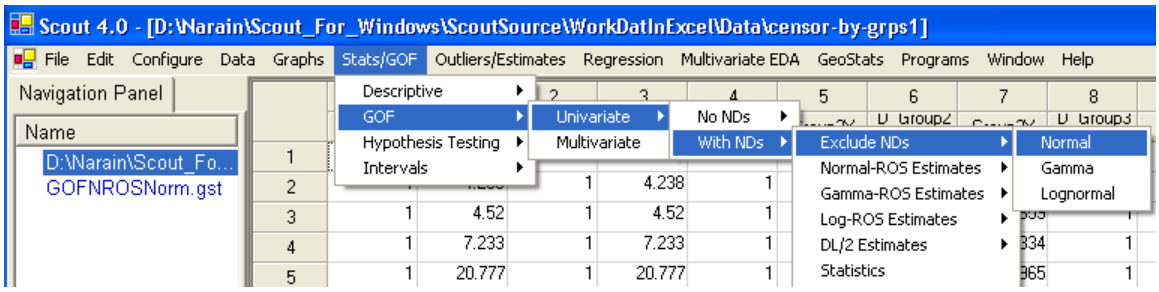
Lognormal Distribution Test Results	
Shapiro Wilk Test Statistic	0.872
Shapiro Wilk Critical (0.95) Value	0.95
Lilliefors Test Statistic	0.225
Lilliefors Critical (0.95) Value	0.103

Data not Lognormal at (0.05) Significance Level

6.2.1.2 GOF Tests for Data Sets With NDs

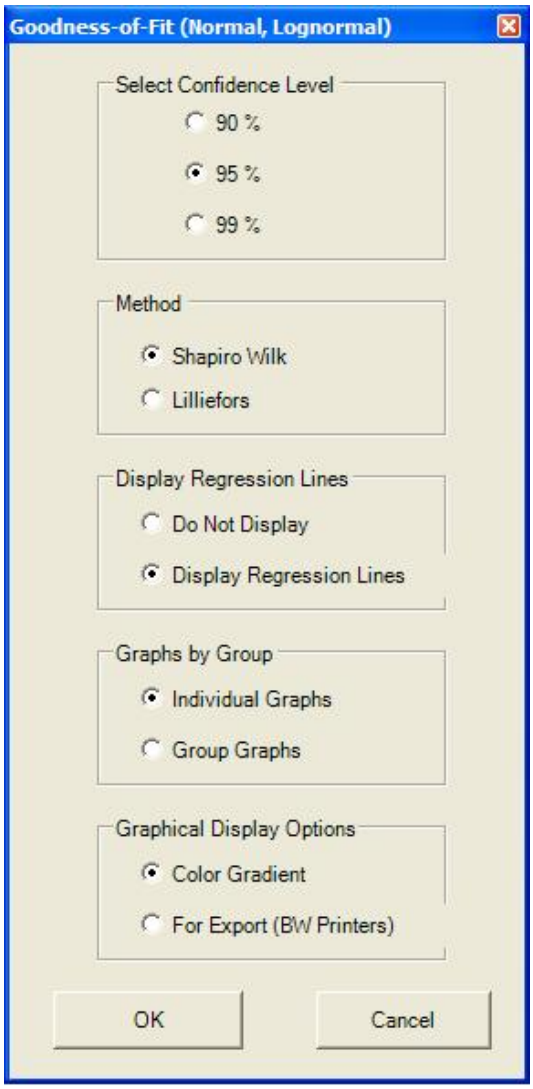
6.2.1.2.1 GOF Tests Using Exclude NDs for Normal and Lognormal Distribution

1. Click **Stats/GOF ► GOF ► Univariate ► With NDs ► Exclude NDs ► Normal or Lognormal.**



2. The "**Select Variables**" screen (Chapter 3) will appear.

   - Select one or more variables from the "**Select Variables**" screen.
   - If graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
   - Click "**Options**" for GOF options.



- The default option for the "**Confidence Level**" is "**95%**."

- The default GOF method is "**Shapiro Wilk**." If the sample size is greater than 50, the program defaults to "**Lilliefors"** test.

- The default for "**Display Regression Lines**" is "**Do Not Display**." If you want to see regression lines on the associated Q-Q plot, check the radio button next to "**Display Regression Lines**."

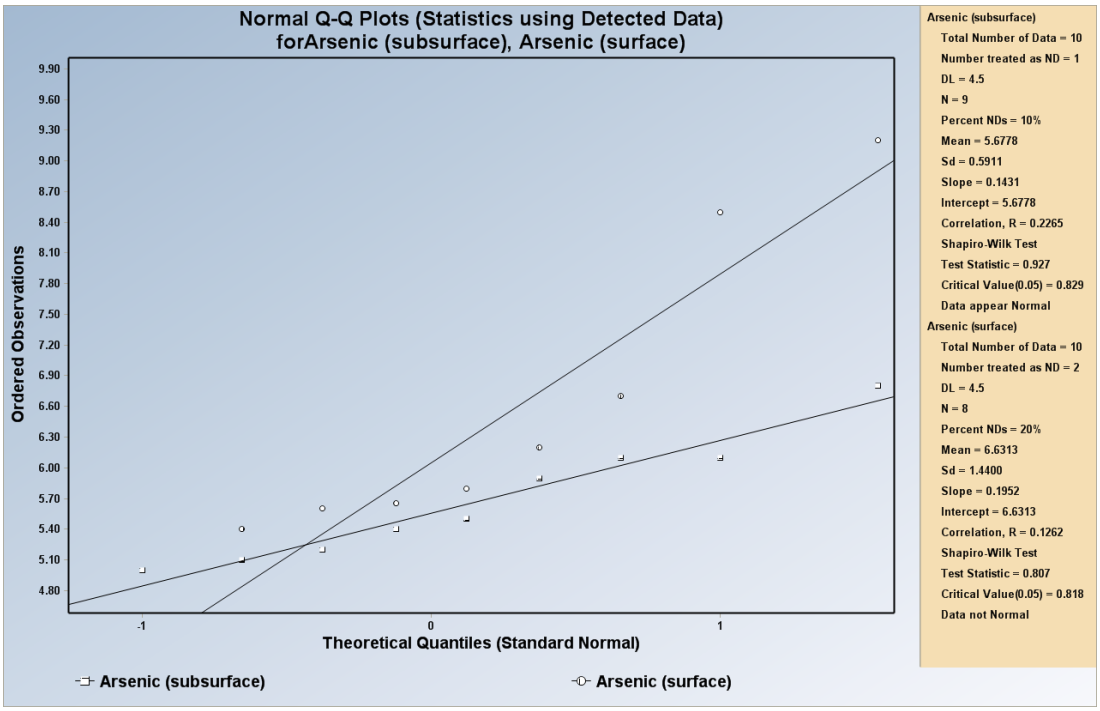
- The default option for "**Graphs by Group**" is "**Individual Graphs**." If you want to see the plots for all selected variables on a single graph, check the radio button next to "**Group Graphs**."

Note: *This option for Graphs by Group is specifically useful when the user wants to display multiple graphs for a variable by a group variable (e.g., site AOC1, Site AOC2, and background). This kind of display represents a useful visual comparison of the values of a variable (e.g., concentrations of COPC-Arsenic) collected from two or more groups (e.g., upgradient wells, monitoring wells, and residential wells).*

- The default option for Graphical Display Option is "**Color Gradient**." If you want to see the graphs in black and white, check the radio button next to "**For Export (BW Printers)**."
- Click "**OK**" to continue or "**Cancel**" to cancel the option.

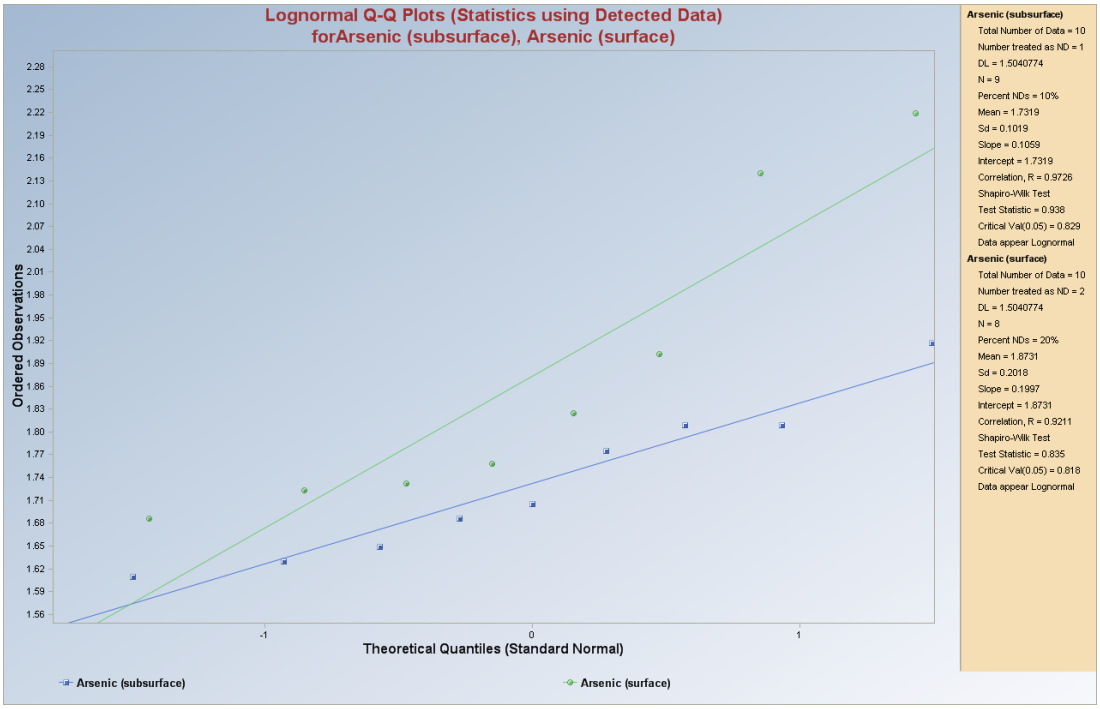
- Click "**OK**" to continue or "**Cancel**" to cancel the goodness-of-fit tests.

Output Screen for Normal Distribution (Exclude NDs).
Selected options: Shapiro Wilk, Display Regression Lines, Group Graphs, and For Export (BW Printers).



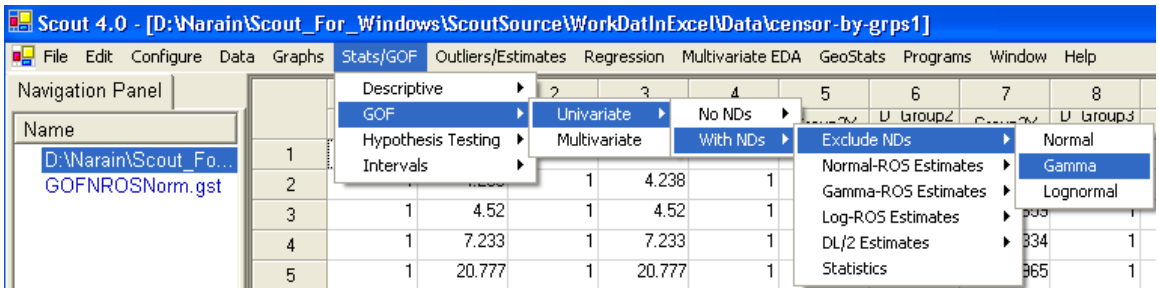
Output Result for Lognormal Distribution (Exclude NDs).

Selected options: Shapiro Wilk, Display Regression Lines, Group Graphs, and Color Gradient.



6.2.1.2.2 GOF Tests Using Exclude NDs for Gamma Distribution

1. Click **Stats/GOF ► GOF ► Univariate ► With NDs ► Exclude NDs ► Gamma.**

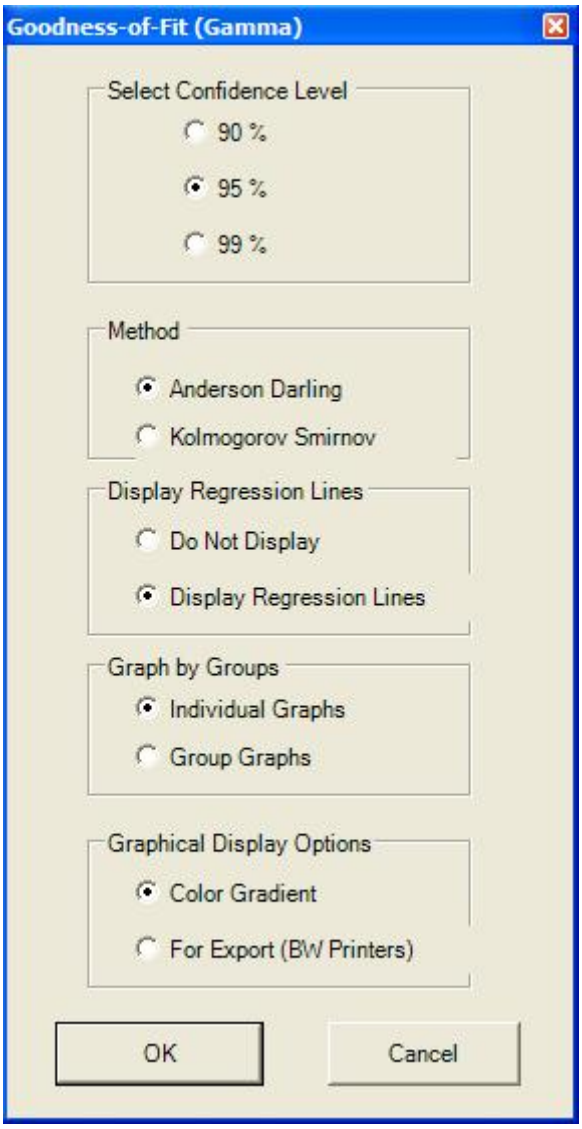


2. The "**Select Variables**" screen (Chapter 3) will appear.

   - Select one or more variables from the "**Select Variables**" screen.
   - If graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The

user should select and click on an appropriate variable representing a group variable.

- Click "**Options**" for GOF options.



  - The default option for the "**Confidence Level**" is "**95%**."
  - The default GOF test method is "**Anderson Darling**."
  - The default method for "**Display Regression Lines**" is "**Do Not Display**." If you want to see regression lines on the normal Q-Q plot, check the radio button next to "**Display Regression Lines**."

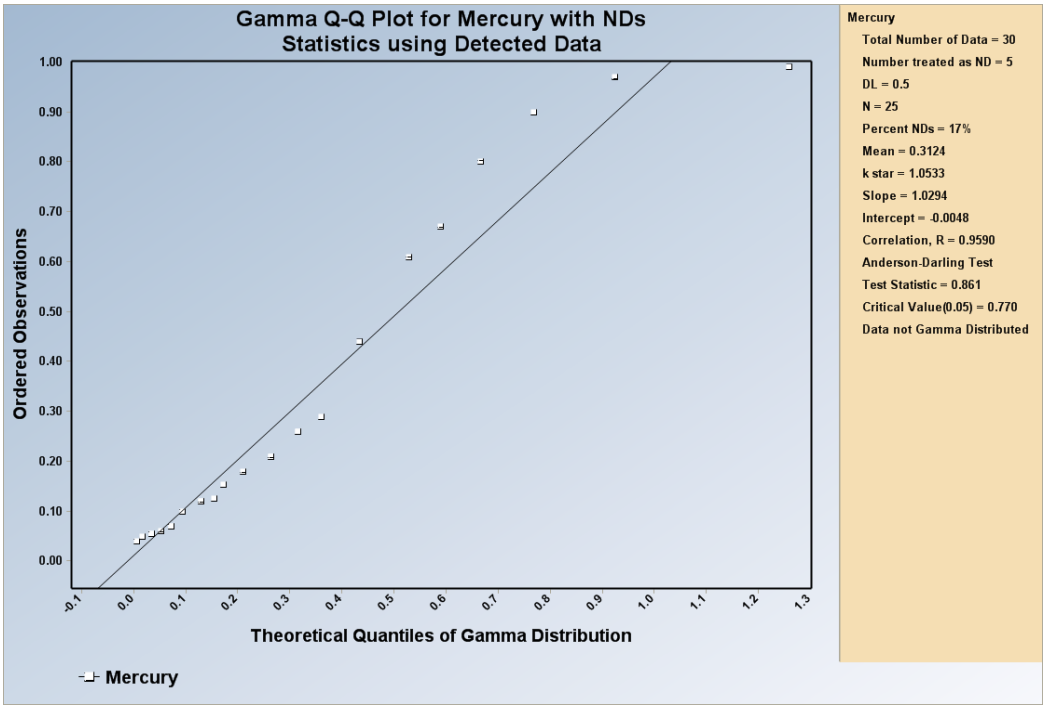
- The default option for "**Graph by Groups**" is "**Individual Graphs**." If you want to display all selected variables on a single graph, check the radio button next to "**Group Graphs**."

- The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to see the graphs in black and white, check the radio button next to "**For Export (BW Printers)**."

- Click "**OK**" to continue or "**Cancel**" to cancel the option.

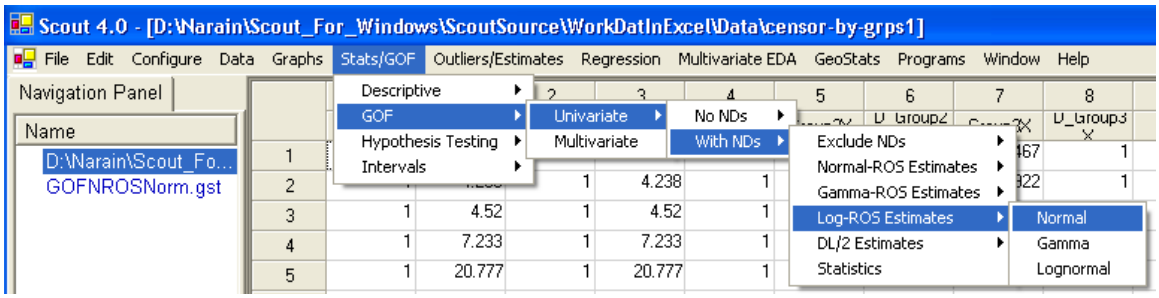
- Click "**OK**" to continue or "**Cancel**" to cancel the goodness-of-fit tests.

Output Screen for Gamma Distribution (Exclude NDs).
Selected options: Anderson Darling, Do Not Display, Individual Graphs, and For Export (BW Printers).



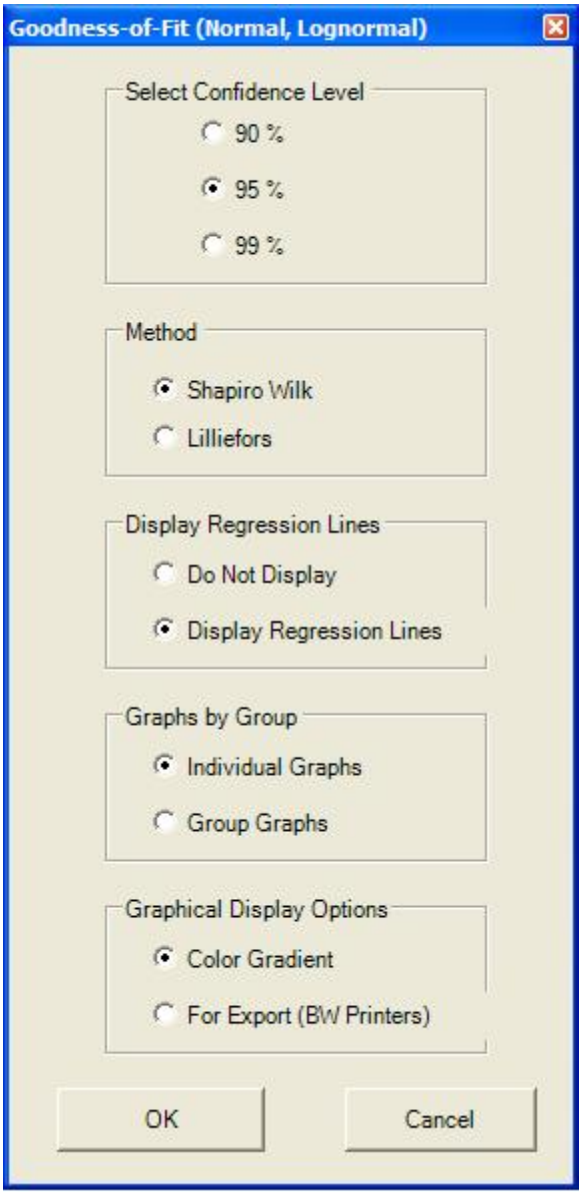
6.2.1.2.3 GOF Tests Using Log-ROS Estimates for Normal and Lognormal Distribution

1. Click **Stats/GOF ► GOF ► Univariate ► With NDs ► Log-ROS Estimates ► Normal or Lognormal.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.
   - If graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
   - Click "**Options**" for GOF options.



- The default option for the "**Confidence Level**" is "**95%**."

- The default GOF test method is "**Shapiro Wilk**." If the sample size is greater than 50, the program defaults to use the "**Lilliefors**" test.

- The default method for "**Display Regression Lines**" is "**Do Not Display**." If you want to see regression lines on the normal Q-Q plot, check the radio button next to "**Display Regression Lines**."

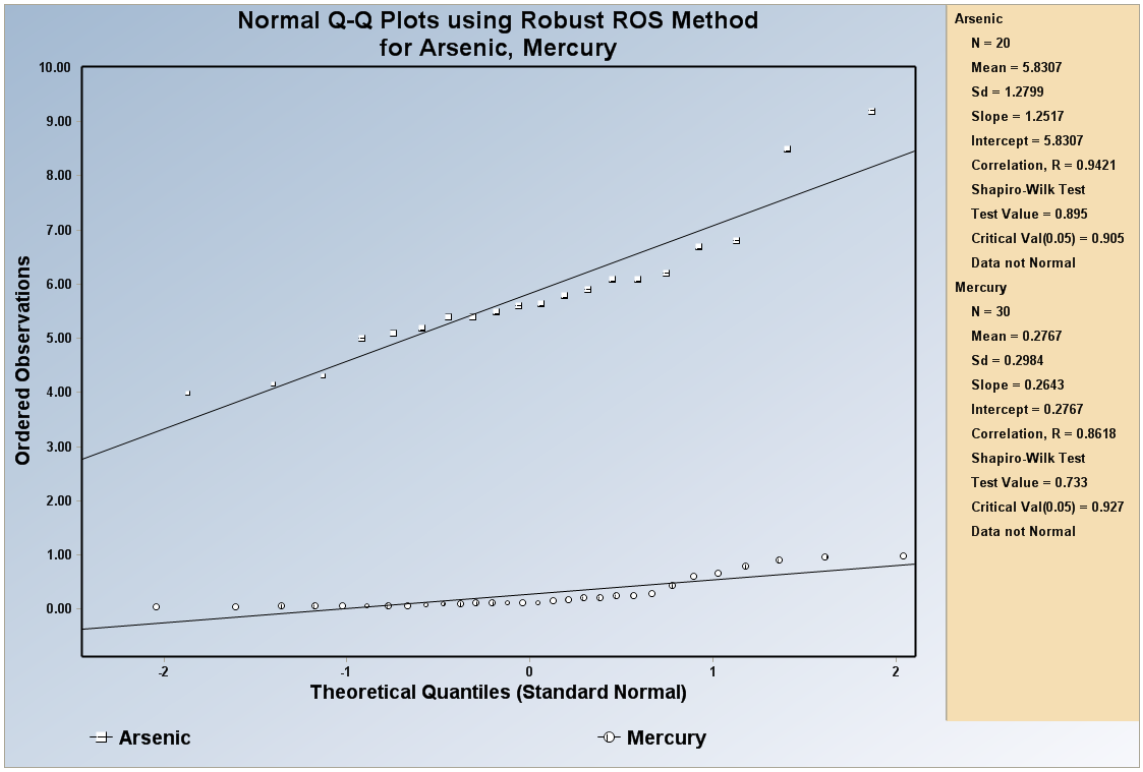
- The default option for "**Graphs by Group**" is '**Individual Graphs**." If you want to display all selected variables into a single graph, check the radio button next to "**Group Graphs**."

- The default option for "**Graphical Display Options**" is "**Color Gradient."** If you want to see the graphs in black and white, check the radio button next to "**For Export (BW Printers)**."

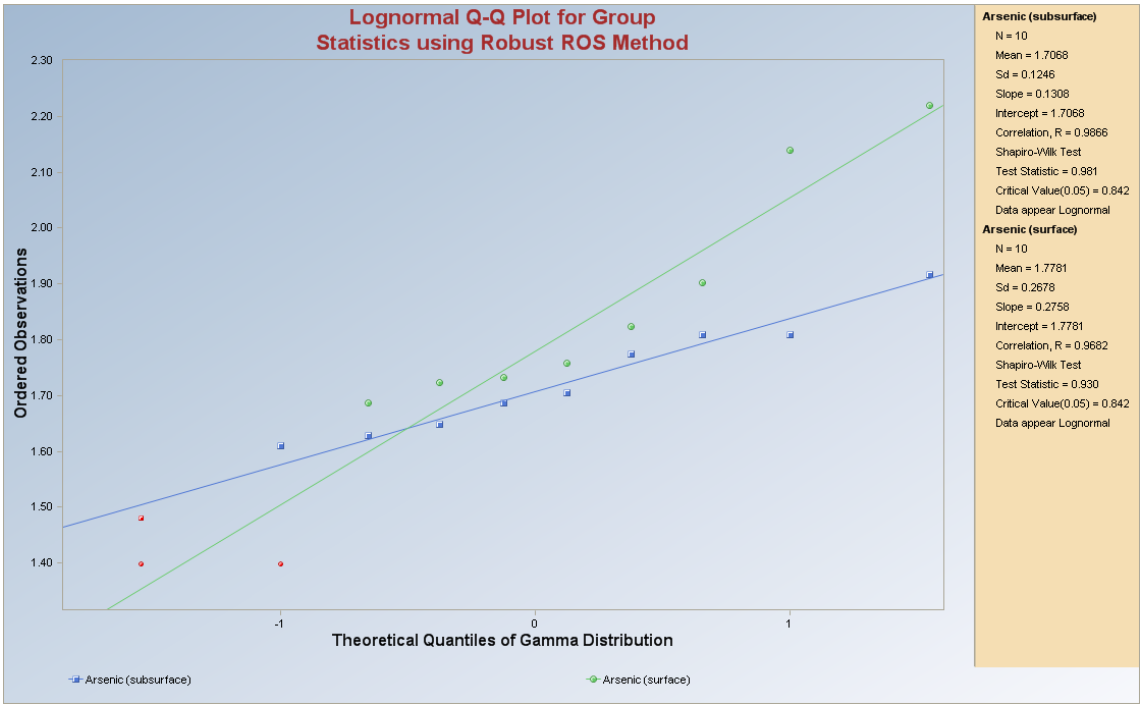
- Click "OK" to continue or "**Cancel**" to cancel the option.

- Click "**OK**" to continue or "**Cancel**" to cancel the goodness-of-fit tests.

Output Screen for Normal Distribution (Log-ROS Estimates).
Selected options: Shapiro Wilk, Display Regression Lines, Group Graphs, and For Export (BW Printers).

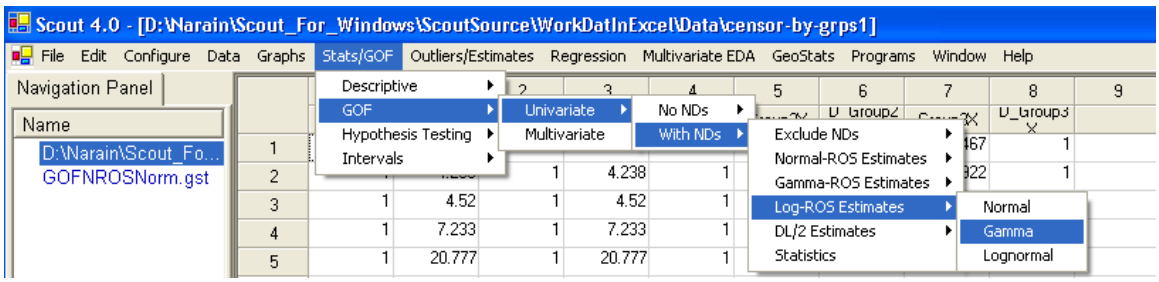


Output Screen for Lognormal Distribution (Log-ROS Estimates).
Selected options: Shapiro Wilk, Display Regression Lines, Group Graphs, and Color Gradient.



6.2.1.2.4 GOF Tests Using Log-ROS Estimates for Gamma Distribution

1. Click **Stats/GOF ► GOF ► Univariate ► With NDs ► Log-ROS Estimates ► Gamma.**

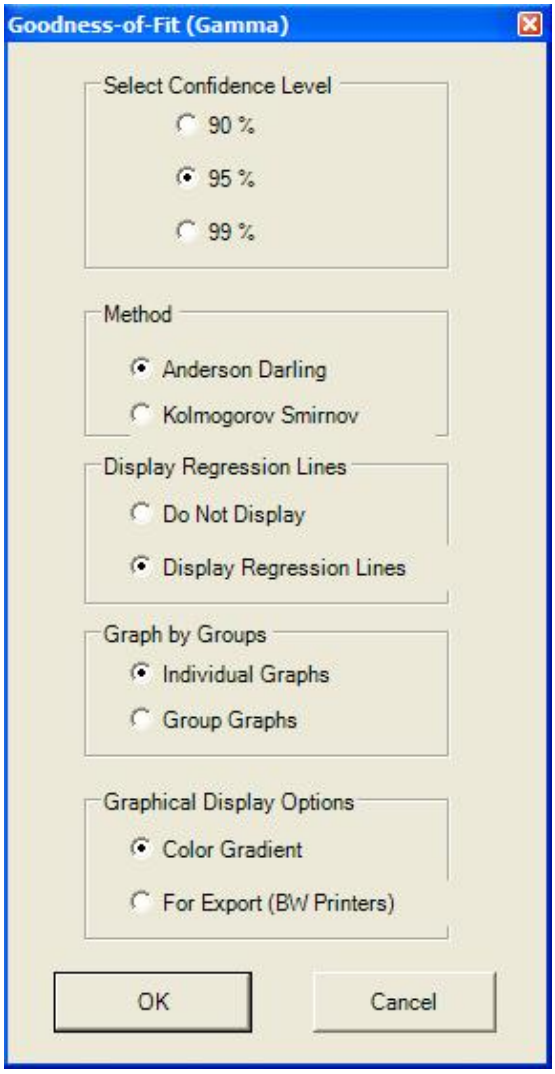


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

   - If graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- Click "**Options**" for GOF options.



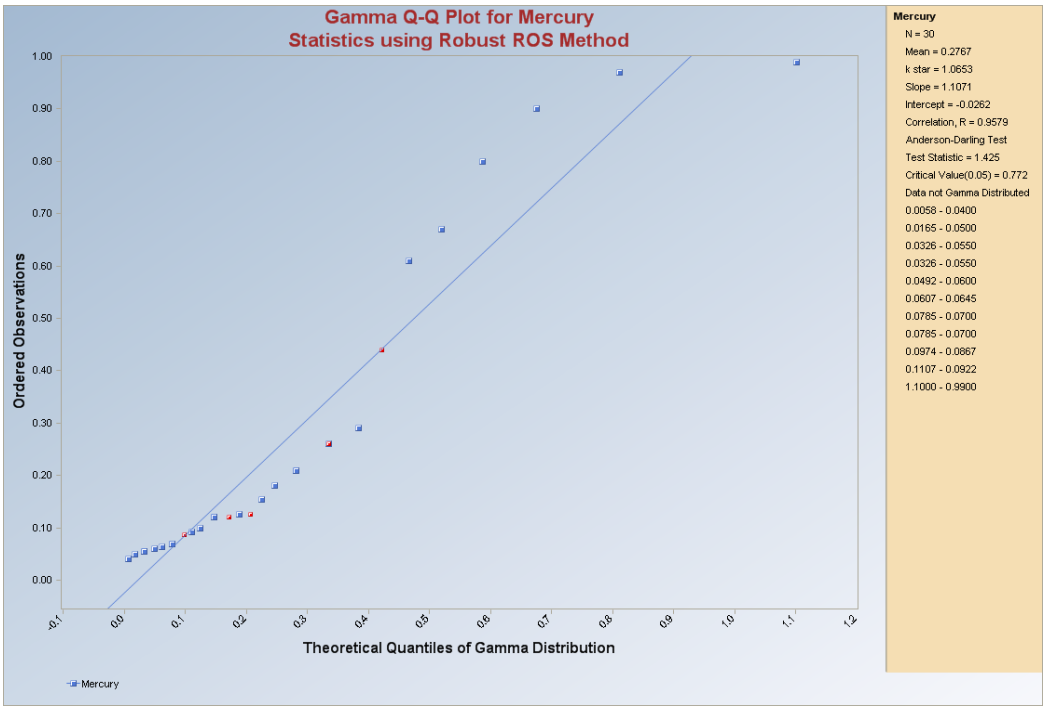
  - The default option for the "**Confidence Level**" is "**95%**."
  - The default GOF test method is "**Anderson Darling**."
  - The default method for "**Display Regression Lines**" is "**Do Not Display**." If you want to see regression lines on the normal Q-Q plot, check the radio button next to "**Display Regression Lines**."
  - The default option for "**Graph by Groups**" is "**Individual Graphs**." If you want to put all of the selected variables into a single graph, check the radio button next to "**Group Graphs**."

- The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to see the graphs in black and white, check the radio button next to "**For Export (BW Printers)**."

- Click "**OK**" **to continue** or "**Cancel**" to cancel the options.

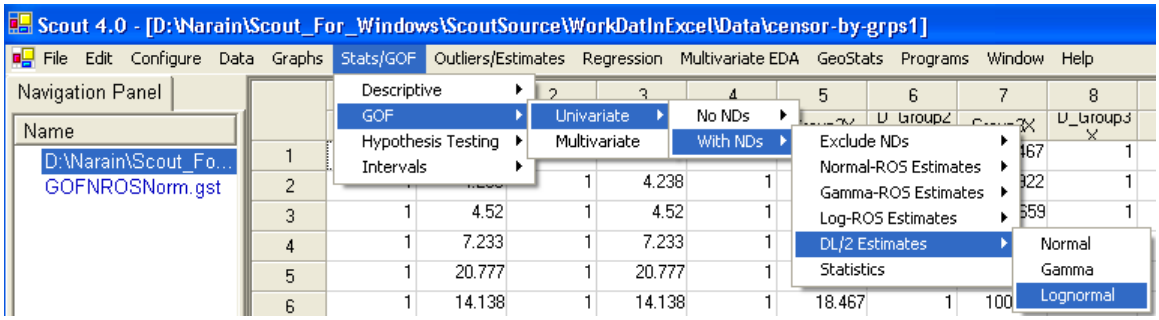
- Click "**OK**" to continue or "**Cancel**" to cancel the goodness-of-fit tests.

Output Screen for Gamma Distribution (Log-ROS Estimates).
Selected options: Anderson Darling, Display Regression Lines, Individual Graphs, and Color Gradient.



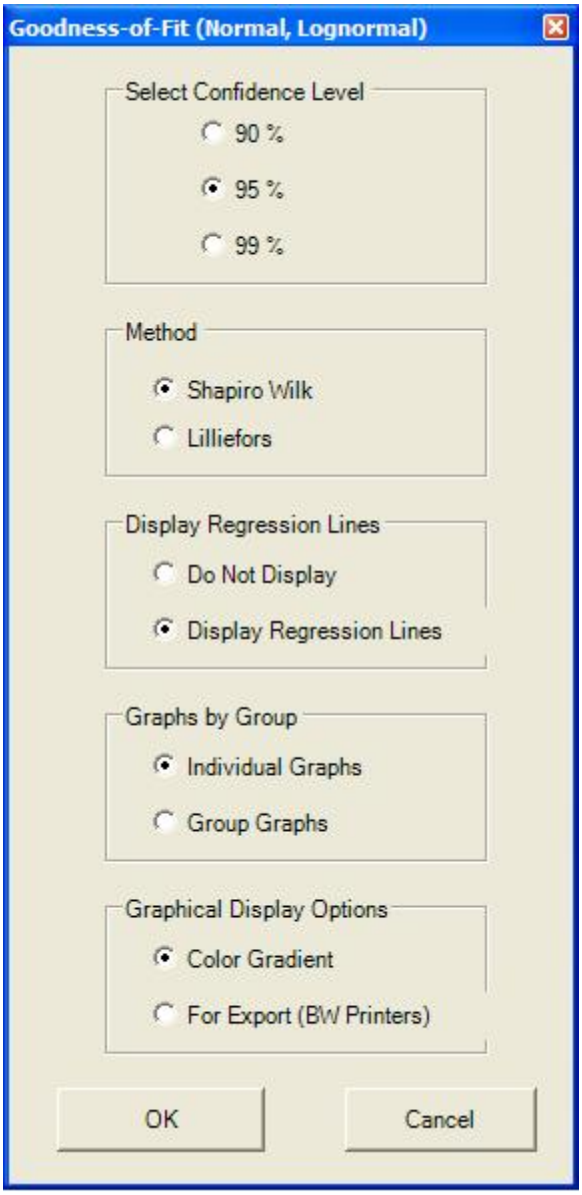
6.2.1.2.5 GOF Tests Using DL/2 Estimates for Normal or Lognormal Distribution

1. Click **Stats/GOF ► GOF ► Univariate ► With NDs ► DL/2 Estimates ► Normal or Lognormal.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.
   - If graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
   - Click "**Options**" for GOF options.



- The default option for the "**Confidence Level**" is "**95%**."

- The default method is "**Shapiro Wilk**." If the sample size is greater than 50, the program defaults to the "**Lilliefors**" test.

- The default method for "Display Regression Lines" is "**Do Not Display**." If you want to see regression lines on the normal Q-Q plot, check the radio button next to "**Display Regression Lines**."

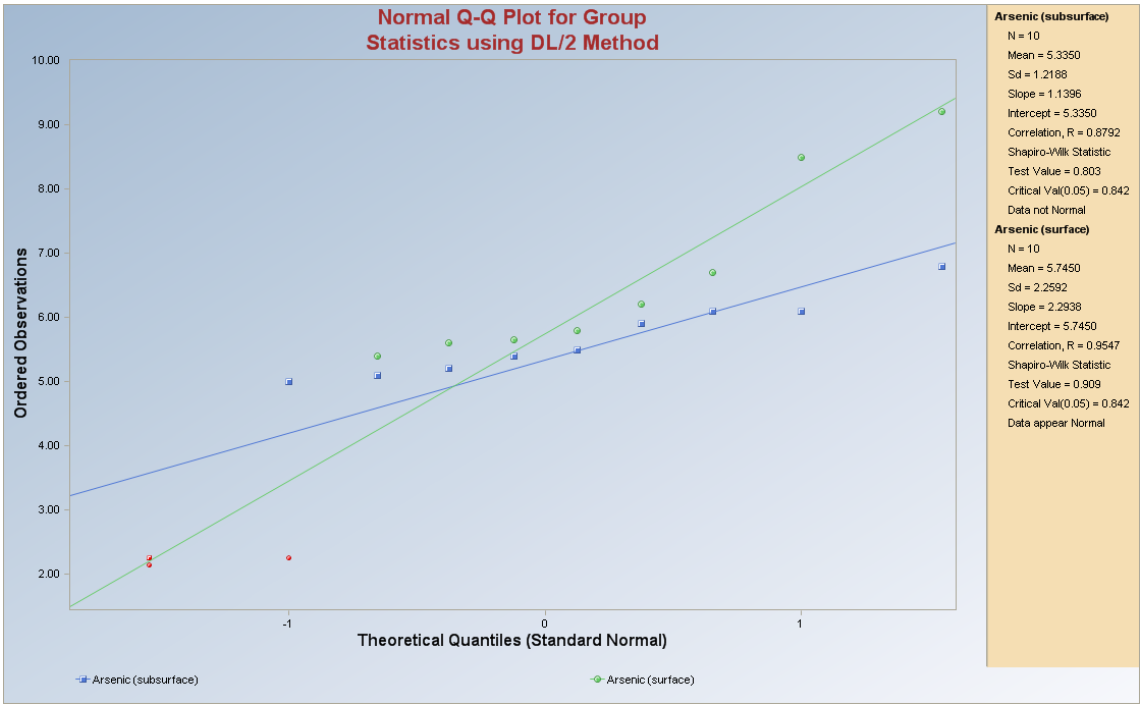
- The default option for "**Graphs by Group**" is "**Individual Graphs**." If you want to put all of the selected variables into a single graph, check the radio button next to "**Group Graphs**."

- The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to see the graphs in black and white, check the radio button next to "**For Export (BW Printers)**."

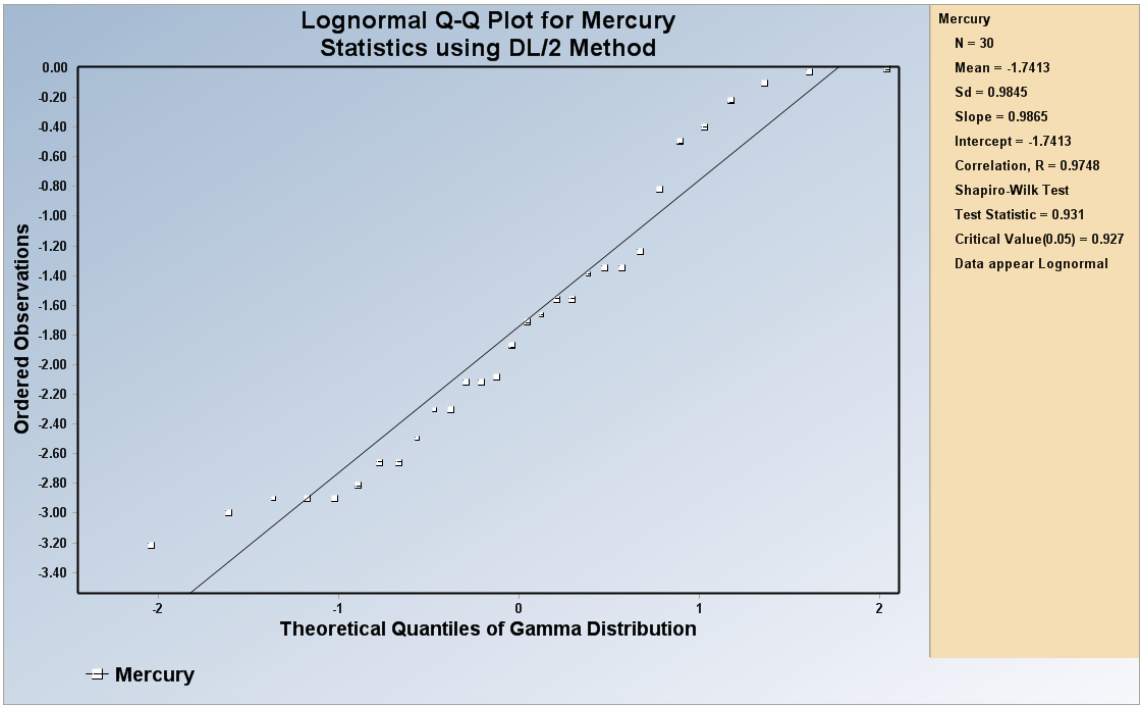
- Click "**OK**" to continue or "**Cancel**" to cancel the option.

- Click "**OK**" to continue or "**Cancel**" to cancel the goodness-of-fit tests.

Output Screen for Normal Distribution (DL/2 Estimates).
Selected options: Shapiro Wilk, Display Regression Lines, Group Graphs, and Color Gradient.

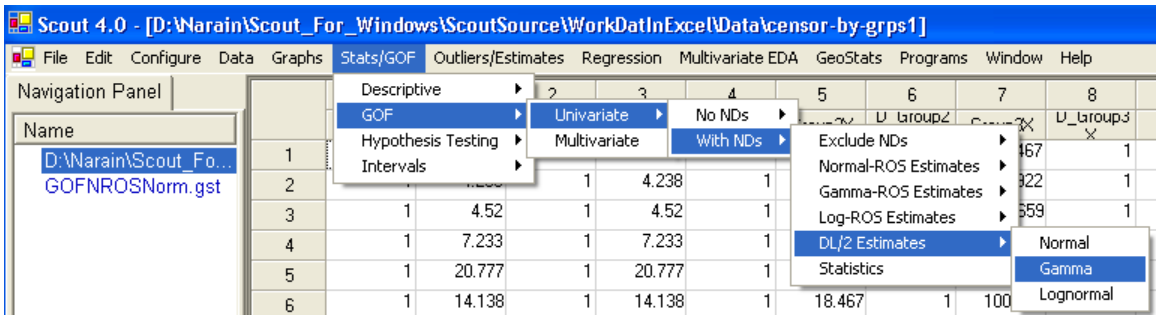


Output Screen for Lognormal Distribution (DL/2 Estimates).
Selected options: Shapiro Wilk, Display Regression Lines, Individual Graphs, and For Export (BW Printers).



6.2.1.2.6 GOF Tests Using DL/2 Estimates for Gamma Distribution

1. Click **Stats/GOF ► GOF ► Univariate ► With NDs ► DL/2 Estimates ► Gamma.**



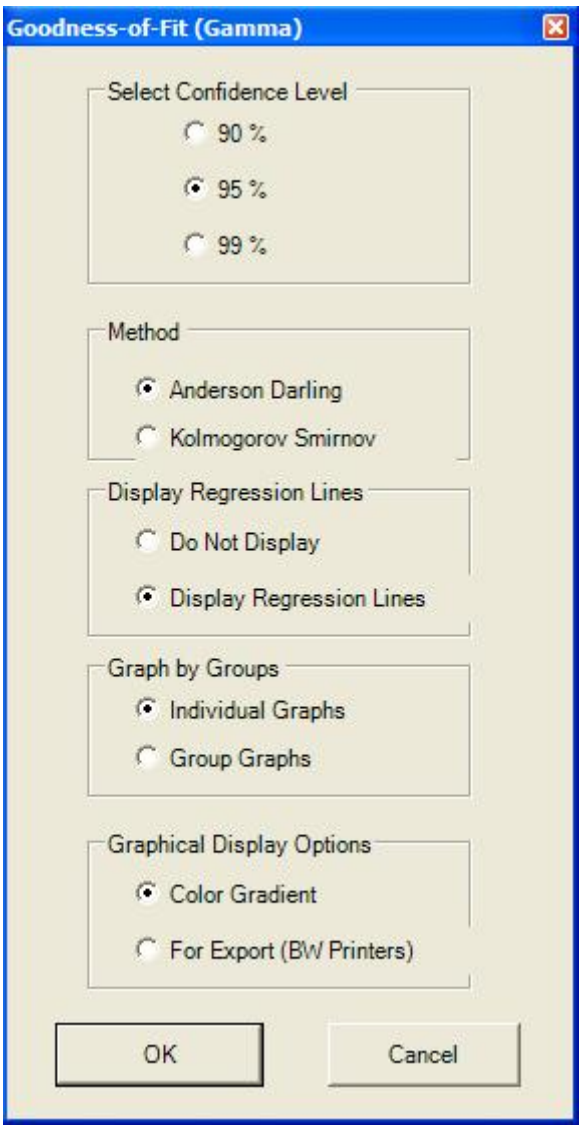
2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

   - If graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The

user should select and click on an appropriate variable representing a group variable.

- Click "**Options**" for GOF options.



  - The default option for the "**Confidence Level**" is "**95%**."
  - The default method is "**Anderson Darling**."
  - The default method for "**Display Regression Lines**" is "**Do Not Display**." If you want to see regression lines on the normal Q-Q plot, check the radio button next to "**Display Regression Lines**."

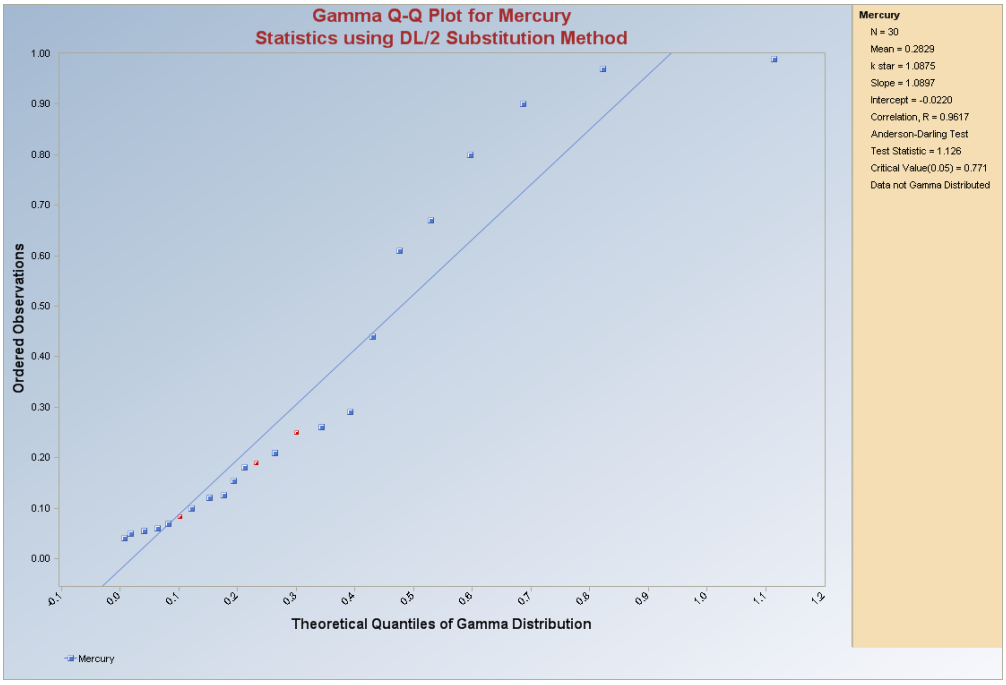
- The default option for "**Graph by Groups**" is "**Individual Graphs**." If you want to put all of the selected variables into a single graph, check the radio button next to "**Group Graphs**."

- The default option for "**Graphical Display Options**" is "**Color Gradient**." If you want to see the graphs in black and white, check the radio button next to "**For Export (BW Printers)**."

- Click "**OK**" to continue or "**Cancel**" to cancel the options.

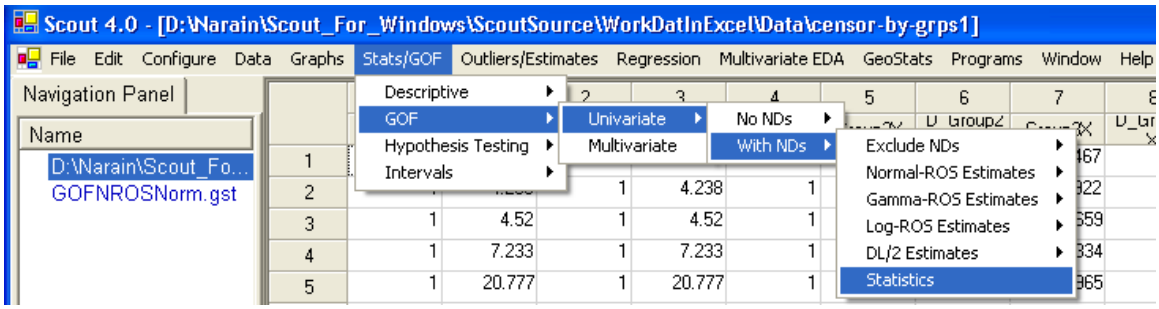
- Click "**OK**" to continue or "**Cancel**" to cancel the goodness-of-fit tests.

Output Screen for Gamma Distribution (DL/2 Estimates).
Selected options: Anderson Darling, Display Regression Lines, Individual Graphs, and Color Gradient.

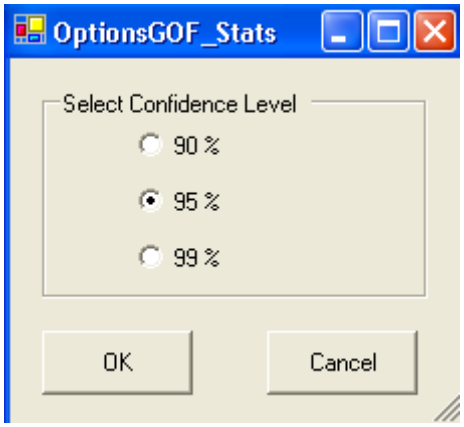


6.2.1.2.7 GOF Statistics

1. Click **Stats/GOF ► GOF ► Univariate ► With NDs ► Statistics.**



2. The "**Select Variables**" screen (Section 3.2) will appear.
   - Select one or more variables from the "**Select Variables**" screen.
   - If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
   - Click "**Options**" for GOF options.



     - The default option for the "**Confidence Level**" is "**95%**."
     - Click "**OK**" to continue or "**Cancel**" to cancel the option.
   - Click "**OK**" to continue or "**Cancel**" to cancel the Goodness-of-Fit Statistics.

Output for GOF Statistics for univariate data with Non-detects.

	Goodness-of-Fit Test Statistics for Data Sets with Non-Detects
Date/Time of Computation	1/25/2008 1:01:29 PM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Confidence Coefficient	0.95

Group1X

	Obs No.	Num Miss	Num Valid	Detects	NDs	% NDs
Group1X Data	10	0	10	8	2	20.00%

	Number	Minimum	Maximum	Mean	Median	SD
Statistics (Non-Detects Only)	2	4	4	4	4	0
Statistics (Detects Only)	8	3.202	20.78	9.277	6.704	6.283
Statistics (All: NDs treated as DL value)	10	3.202	20.78	8.222	5.347	5.971
Statistics (All: NDs treated as DL/2 value)	10	2	20.78	7.822	5.347	6.334
Statistics (Normal ROS Estimated Data)	10	-2.508	20.78	7.256	5.347	7.034
Statistics (Gamma ROS Estimated Data)	10	1.421	20.78	8.027	5.405	6.182
Statistics (Lognormal ROS Estimated Data)	10	2.011	20.78	7.917	5.347	6.243

	K Hat	K Star	Theta Hat	Log Mean	Log Stdv	Log C.V.
Statistics (Detects Only)	2.674	1.938	3.469	2.029	0.673	0.332
Statistics (NDs = DL)	2.578	1.872	3.189	1.901	0.652	0.343
Statistics (NDs = DL/2)	1.844	1.357	4.242	1.762	0.818	0.464
Statistics (Gamma ROS Estimates)	1.995	1.463	4.024	--	--	--
Statistics (Lognormal ROS Estimates)	--	--	--	1.801	0.769	0.427

Output for GOF Statistics for univariate data with Non-detects (continued).

Normal Distribution Test Results

	Test value	Crit. (0.95)	Conclusion with Alpha(0.05)
Shapiro-Wilks (Detects Only)	0.866	0.818	Data Appear Normal
Lilliefors (Detects Only)	0.253	0.313	Data Appear Normal
Shapiro-Wilks (NDs = DL)	0.796	0.842	Data Not Normal
Lilliefors (NDs = DL)	0.266	0.28	Data Appear Normal
Shapiro-Wilks (NDs = DL/2)	0.848	0.842	Data Appear Normal
Lilliefors (NDs = DL/2)	0.237	0.28	Data Appear Normal
Shapiro-Wilks (Normal ROS Estimates)	0.941	0.842	Data Appear Normal
Lilliefors (Normal ROS Estimates)	0.201	0.28	Data Appear Normal

Gamma Distribution Test Results

	Test value	Crit. (0.95)	Conclusion with Alpha(0.05)
Anderson-Darling (Detects Only)	0.404	0.722	
Kolmogorov-Smirnov (Detects Only)	0.197	0.297	Data Appear Gamma Distributed
Anderson-Darling (NDs = DL)	0.737	0.734	
Kolmogorov-Smirnov (NDs = DL)	0.244	0.269	Data appear Approximate Gamma Distribution
Anderson-Darling (NDs = DL/2)	0.367	0.737	
Kolmogorov-Smirnov (NDs = DL/2)	0.165	0.27	Data Appear Gamma Distributed
Anderson-Darling (Gamma ROS Estimates)	0.355	0.736	
Kolmogorov-Smirnov (Gamma ROS Est.)	0.178	0.27	Data Appear Gamma Distributed

Lognormal Distribution Test Results

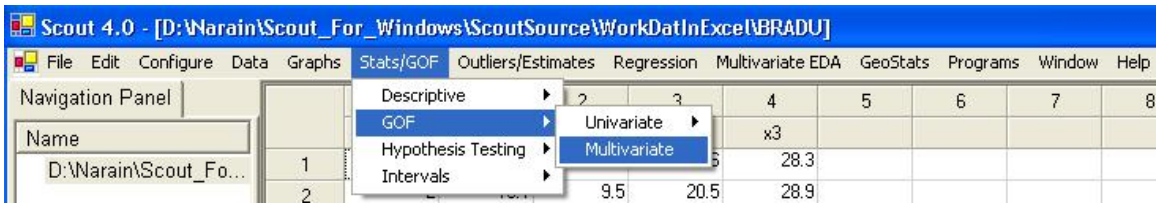
	Test value	Crit. (0.95)	Conclusion with Alpha(0.05)
Shapiro-Wilks (Detects Only)	0.932	0.818	Data Appear Lognormal
Lilliefors (Detects Only)	0.191	0.313	Data Appear Lognormal
Shapiro-Wilks (NDs = DL)	0.878	0.842	Data Appear Lognormal
Lilliefors (NDs = DL)	0.226	0.28	Data Appear Lognormal
Shapiro-Wilks (NDs = DL/2)	0.94	0.842	Data Appear Lognormal
Lilliefors (NDs = DL/2)	0.157	0.28	Data Appear Lognormal
Shapiro-Wilks (Lognormal ROS Estimates)	0.951	0.842	Data Appear Lognormal
Lilliefors (Lognormal ROS Estimates)	0.161	0.28	Data Appear Lognormal

Note: DL/2 is not a recommended method.

6.2.2 *Multivariate GOF*

The multivariate goodness-of-fit test to test for multinormality of a data set can be performed using Scout. Several test statistics, including the correlation coefficient based upon ordered Mahalanobis distances (MDs) versus beta distribution quantiles (and also approximate chi-square quantiles), multivariate kurtosis, and multivariate skewness, are available in Scout. The details of those statistics can be found in Singh (1993) and Mardia (1970).

1. Click **Stats/GOF ► GOF ► Multivariate.**

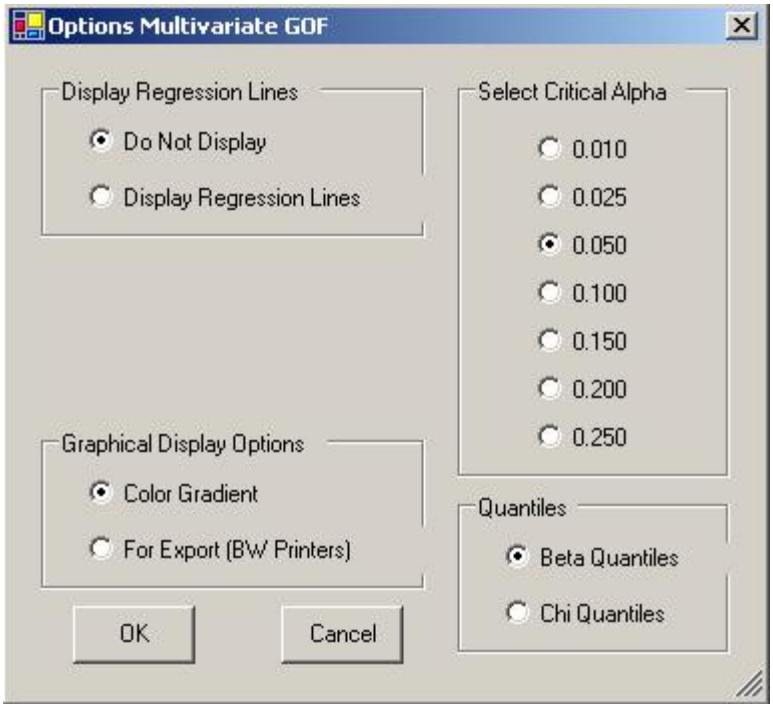


2. The "**Select Variables**" screen (Section 3.4) will appear.

   - Select two or more variables from the "**Select Variables**" screen.

   - If graphs have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

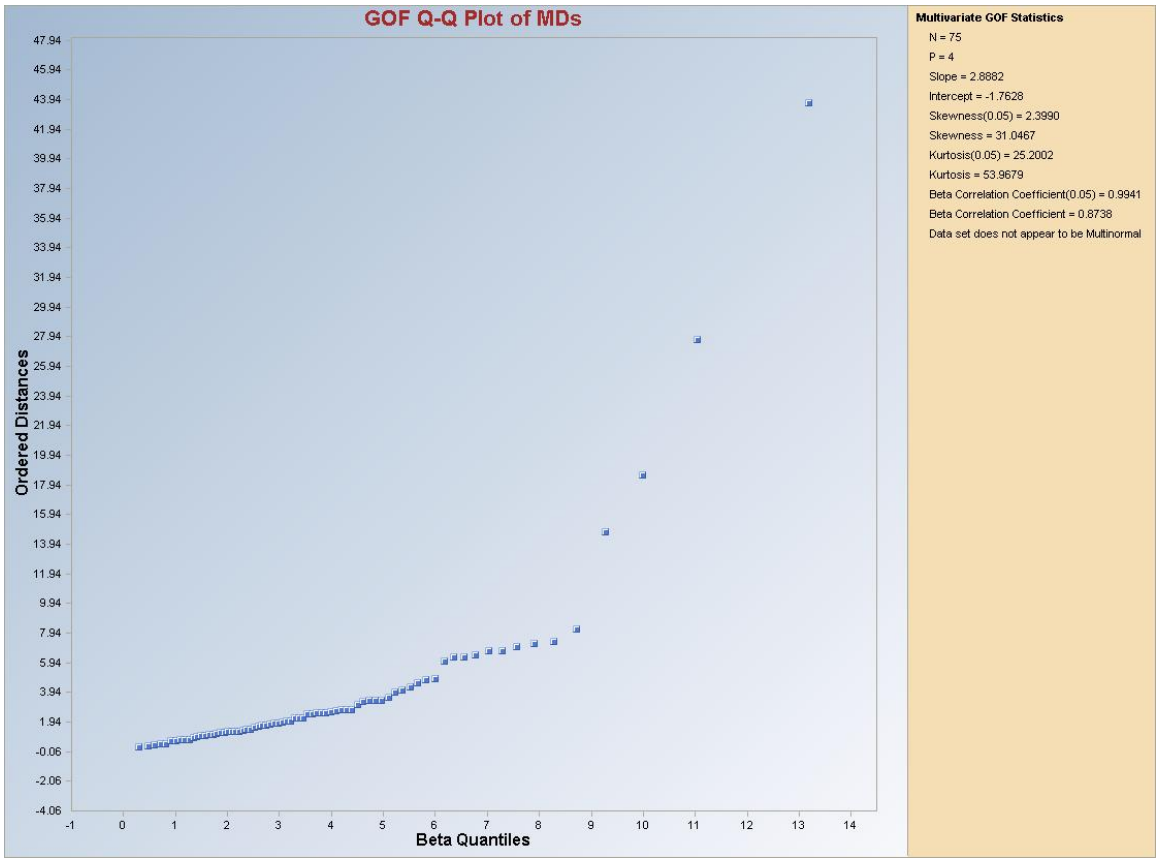
   - Click "**Options**" for the multivariate GOF options.



- Specify the preferred "**Critical Alpha**." The default is "**0.05**."
- Specify the distribution (scaled beta or approximate chi-square) of the MDs used to compute the quantiles. The default is a "**Beta**" distribution.
- The default option for Display Regression Lines is "**Do Not Display**", and the default option for "**Graphical Display Options**" is "**Color Gradient**."
- Click on "**OK**" to continue or "**Cancel**" to cancel the GOF options.

- Click on "**OK**" to continue or "**Cancel**" to cancel the GOF computations.

Output Screen for Multivariate GOF.



***Note**: Several test statistics (correlation coefficient, skewness, and kurtosis) are shown in the above GOF display. Singh (1993) has outlined some of these procedures to assess multivariate normality. Critical values for these three statistics have been computed using extensive Monte Carlo simulations. Critical values are still being simulated at the time of publishing this document. These values will be available in the Q-Q plots in the near future. The developers of Scout may be contacted to obtain these critical values. They do plan to publish them in the near future.*

6.3 Hypothesis Testing

Scout can perform hypothesis tests on data sets with and without ND observations. When one wants to use two-sample hypothesis tests on data sets with NDs, Scout assumes that samples from both of the groups have non-detect observations. This means is that a ND column (with 0 or 1 entries only) needs to be provided for the variable in each of the two groups. This has to be done even if one of the groups has all detected entries; in this case, the associated ND column will have all entries equal to "1." This will allow the user to compare two groups (e.g., arsenic in background vs. site samples) with one group having NDs and the other group having all detected data.

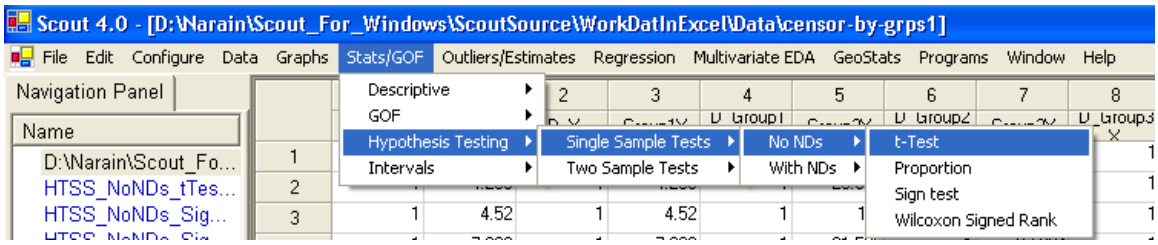
The hypothesis testing module of Scout is exactly same as the one available in ProUCL 4.00.04. ProUCL 4.00.04 has been developed to address several environmental applications. More information on those methods can be obtained from the ProUCL 4.00.04 Technical Guide and User Guide (Chapter 9), respectively.

***Note**: Since the hypothesis testing module of Scout is imported from ProUCL 4.00.04, most of the terminology used (site concentration, background concentration, background threshold values, etc.) are borrowed from various environmental applications. However, all of those tools (e.g., t-test, Gehan test) can be used in various other applications. For an example, a two-sample t-test can be used to compare the means of distributions of any two variables. Similarly, the Gehan test may be used to compare the measures of central tendency of two distributions based upon data sets with below detection limit observations.*

6.3.1.1 Single Sample Hypothesis Tests for Data Sets with No Non-detects

6.3.1.1.1 Single Sample t-Test

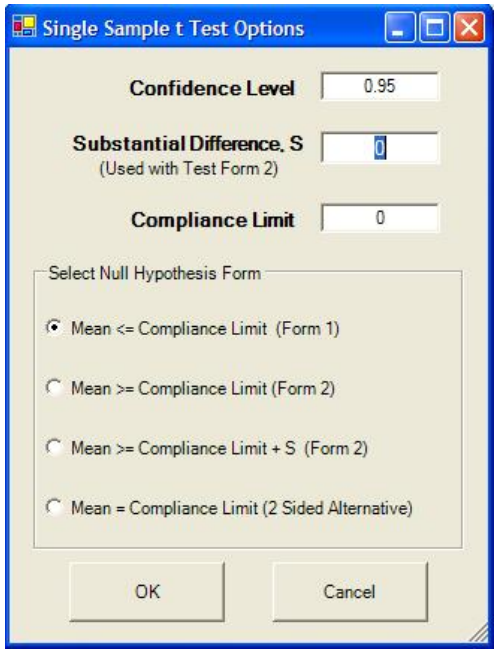
1. Click **Stats/GOF ► Hypothesis Testing ► Single Sample ► No NDs ► t-Test.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.
   - If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- When the options button is clicked, the following window will be shown.



  - Specify the "**Confidence Level**." The default is "**0.95**."

  - Specify meaningful values for "**Substantial Difference, S**" and the "**Compliance Limit**." The default choice for S is "**0**."

  - Select the form of Null Hypothesis. The default is Mean <= Compliance Limit (Form 1).

  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**OK**" to continue or "**Cancel**" to cancel the test.

Output for Single Sample t-Test (Full Data without NDs).

1 Sample- t

Single Sample t-Test

Raw Statistics

Number of Valid Samples	9
Number of Distinct Samples	9
Minimum	82.39
Maximum	113.2
Mean	99.38
Median	103.5
SD	10.41
SE of Mean	3.468

H0: Site Mean = 100

Test Value	-0.178
Two Sided Critical Value (0.05)	2.306
P-Value	0.863

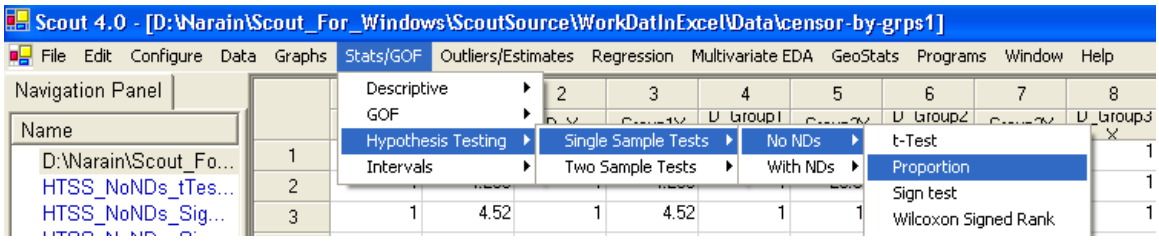
Conclusion with Alpha = 0.05

Do Not Reject H0, Conclude Mean = 100

P-Value > Alpha (0.05)

6.3.1.1.2 Single Sample Proportion Test

1. Click **Stats/GOF ► Hypothesis Testing ► Single Sample ► No NDs ► Proportion.**

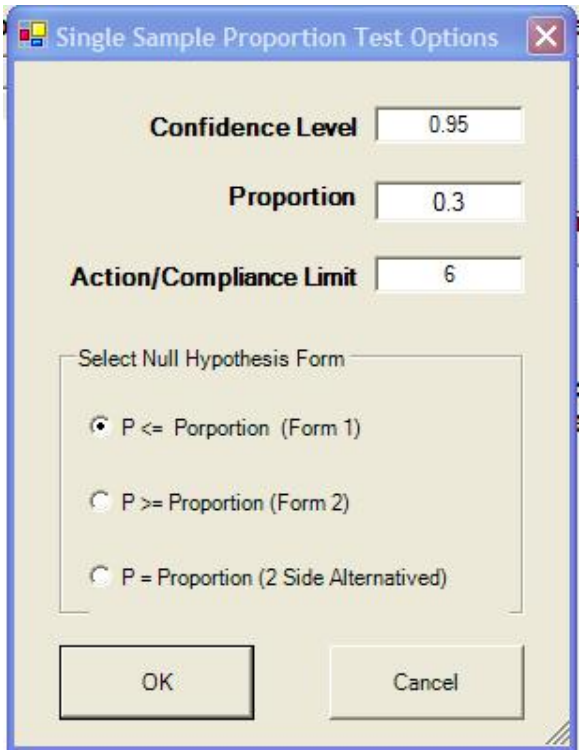


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

- If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- When the options button is clicked, the following window will be shown.



  - Specify the "**Confidence Level**." The default is "**95**."

  - Specify the "**Proportion**" level and a meaningful "**Action/Compliance Limit**."

  - Select the form of Null Hypothesis. The default is P <= Proportion (Form 1).

  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**OK**" to continue or "**Cancel**" to cancel the test.

Output for Single Sample Proportion Test (Full Data without NDs).

One-Sample Proportion Test

Raw Statistics

Number of Valid Samples	85
Number of Distinct Samples	83
Minimum	0.598
Maximum	7.676
Mean	5.183
Median	5.564
SD	1.588
SE of Mean	0.172
Number of Exceedances	27
Sample Proportion of Exceedances	0.318

H0: Site Proportion <= 0.3 (Form 1)

Large Sample z-Test Value	0.237
Critical Value (0.05)	1.645
P-Value	0.406

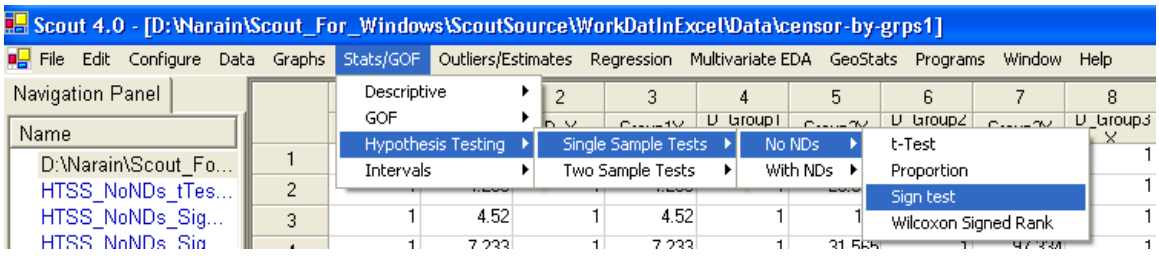
Conclusion with Alpha = 0.05

Do Not Reject H0, Conclude Site Proportion <= 0.3

P-Value > Alpha (0.05)

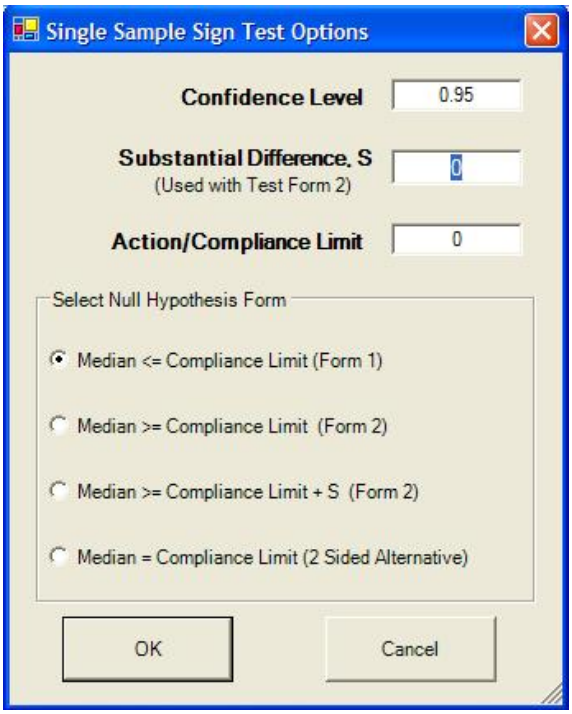
6.3.1.1.3 Single Sample Sign Test

1. Click **Stats/GOF ► Hypothesis Testing ► Single Sample ► No NDs ► Sign test.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- When the options button is clicked, the following window will be shown.



  - Specify the "**Confidence Level**." The default choice is "**0.95.**"
  - Specify meaningful values for "**Substantial Difference, S**" and "**Action/Compliance Limit**."
  - Select the form of Null Hypothesis. The default is Median <= Compliance Limit (Form 1).
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.
- Click "**OK**" to continue or "**Cancel**" to cancel the test.

Output for Single Sample Proportion Test (Full Data without NDs).

Single Sample Sign Test

Raw Statistics

Number of Valid Samples	10
Number of Distinct Samples	10
Minimum	750
Maximum	1161
Mean	925.7
Median	888
SD	136.7
SE of Mean	43.24
Number Above Limit	3
Number Equal Limit	0
Number Below Limit	7

H0: Site Median >= 1000 (Form 2)

Test Value	3
Lower Critical Value (0.05)	1
P-Value	0.172

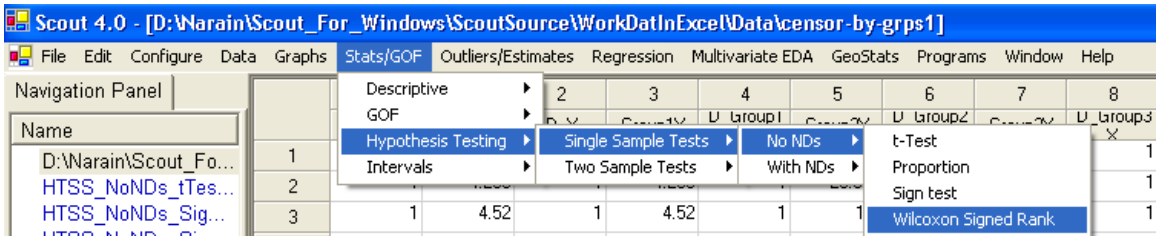
Conclusion with Alpha = 0.05

Do Not Reject H0, Conclude Median >= 1000

P-Value > Alpha (0.05)

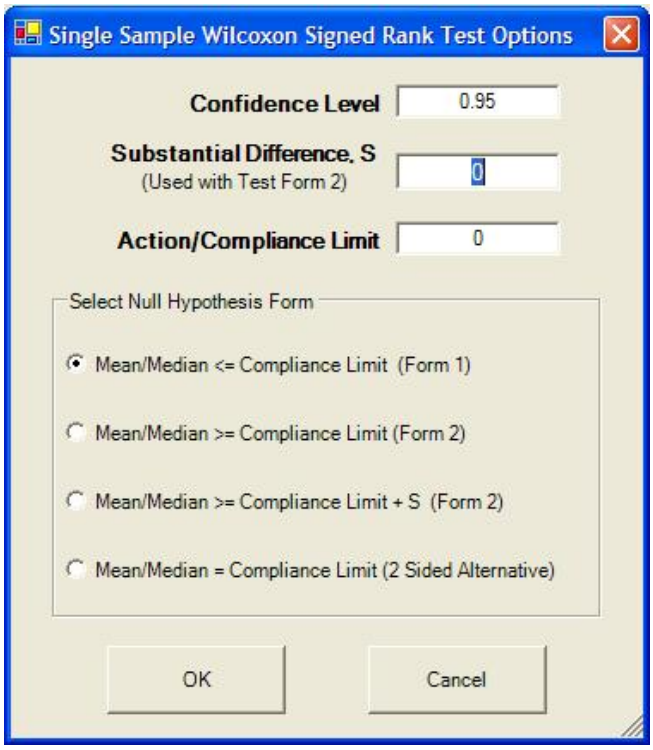
6.3.1.1.4 Single Sample Wilcoxon Signed Rank Test

1. Click **Stats/GOF ► Hypothesis Testing ► Single Sample ► No NDs ► Wilcoxon Signed Rank test.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- When the options button is clicked, the following window will be shown.



  - Specify the "**Confidence Level**." The default is "**0.95**."
  - Specify meaningful values for "**Substantial Difference**, **S**," and "**Action/Compliance Limit**."
  - Select the form of Null Hypothesis. The default is Mean/Median <= Compliance Limit (Form 1).
  - Click "**OK**" to continue or "**Cancel**" to cancel the option.
- Click "**OK**" to continue or "**Cancel**" to cancel the test.

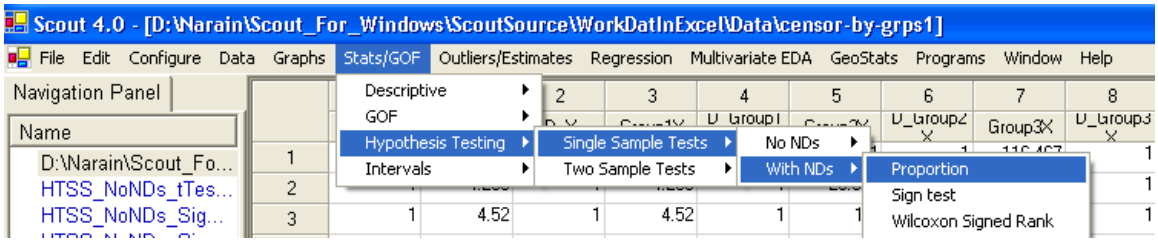
Output for Single Sample Wilcoxon Signed Rank Test (Full Data without NDs)

Single Sample Wilcoxon Signed Rank Test	
Raw Statistics	
Number of Valid Samples	10
Number of Distinct Samples	10
Minimum	750
Maximum	1161
Mean	925.7
Median	888
SD	136.7
SE of Mean	43.24
Number Above Limit	3
Number Equal Limit	0
Number Below Limit	7
T-plus	11.5
T-minus	43.5
H0: Site Median <= 1000 (Form 1)	
Test Value	11.5
Critical Value (0.05)	45
P-Value	0.947
Conclusion with Alpha = 0.05	
Do Not Reject H0, Conclude Mean/Median <= 1000	
P-Value > Alpha (0.05)	

6.3.1.2 Single Sample Hypothesis Tests for Data Sets With Non-detects

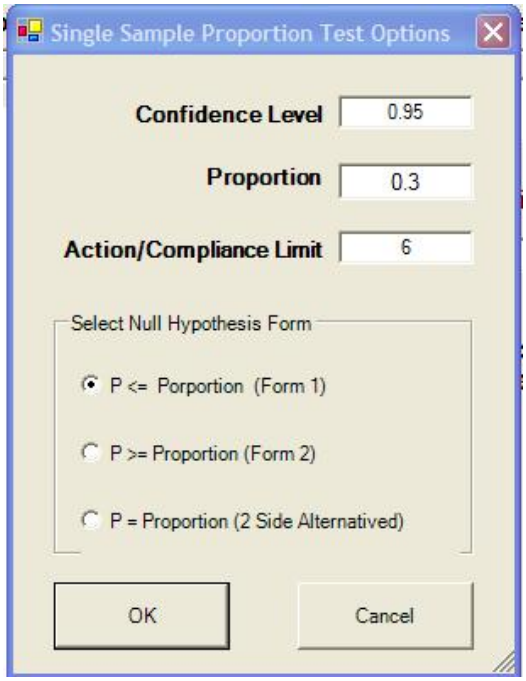
6.3.1.2.1 Single Sample Proportion Test

1. Click **Stats/GOF ► Hypothesis Testing ► Single Sample ► With NDs ► Proportion test.**



3. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- When the options button is clicked, the following window will be shown.



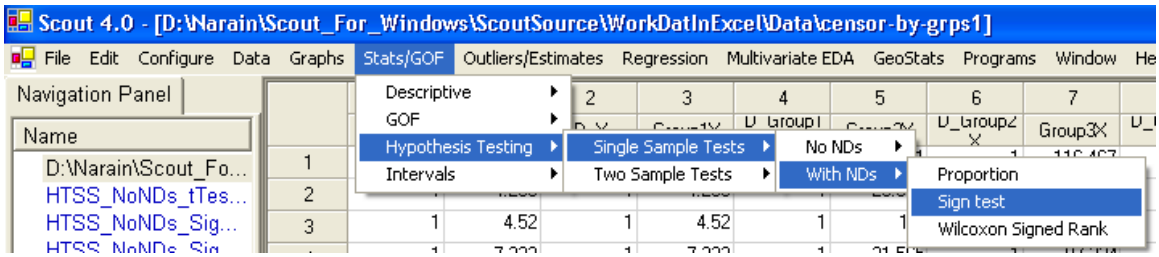
  - Specify the "**Confidence Level**." The default is "**0.95**."
  - Specify meaningful values for "**Proportion**" and the "**Action/Compliance Limit**."
  - Select the form of Null Hypothesis. The default is P <= Proportion (Form 1).
  - Click "**OK**" to continue or "**Cancel**" to cancel the option.
- Click "**OK**" to continue or "**Cancel**" to cancel the test.

Output for Single Sample Proportion Test (with NDs).

Arsenic	
Single Sample Proportion Test	
Raw Statistics	
Number of Valid Samples	24
Number of Distinct Samples	10
Number of Non-Detect Data	13
Number of Detected Data	11
Percent Non-Detects	54.17%
Minimum Non-detect	0.9
Maximum Non-detect	2
Minimum Detected	0.5
Maximum Detected	3.2
Mean of Detected Data	1.236
Median of Detected Data	0.7
SD of Detected Data	0.965
Number of Exceedances	2
Sample Proportion of Exceedances	0.0833
Some Non-Detect Values Exceed	
The User Selected Action/Compliance Limit	
Unable to do Proportion Test with such parameters	

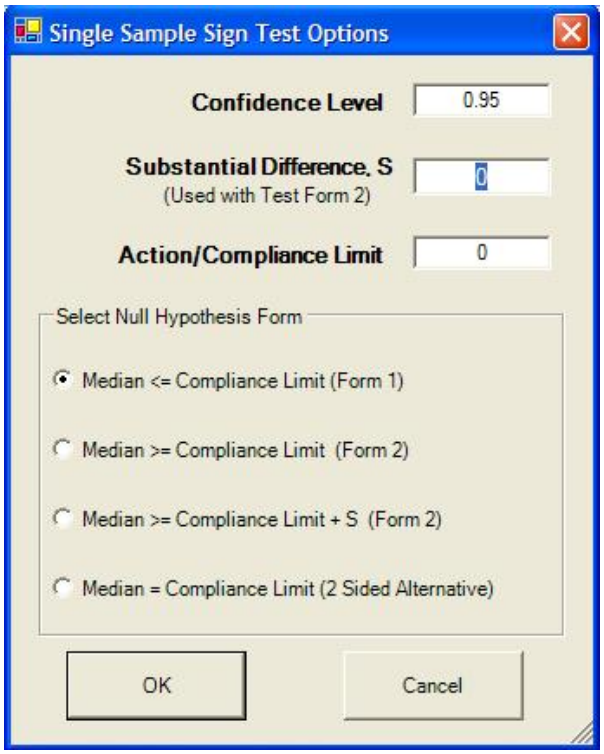
6.3.1.2.2 Single Sample Sign Test

1. Click **Stats/GOF ► Hypothesis Testing ► Single Sample ► With NDs ► Sign test.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.
   - If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
   - When the options button is clicked, the following window will be shown.



- Specify the "**Confidence Level**." The default is "**0.95**."
- Specify meaningful values for "**Substantial Difference, S**" and "**Action/Compliance Limit**."
- Select the form of Null Hypothesis. The default is Median <= Compliance Limit (Form 1).
- Click "**OK**" to continue or "**Cancel**" to cancel the option.

- Click "**OK**" to continue or "**Cancel**" to cancel the test.

Output for Single Sample Sign Test (Data with Non-detects).

Arsenic

Single Sample Sign Test

Raw Statistics

Number of Valid Samples	24
Number of Distinct Samples	10
Number of Non-Detect Data	13
Number of Detected Data	11
Percent Non-Detects	54.17%
Minimum Non-detect	0.9
Maximum Non-detect	2
Minimum Detected	0.5
Maximum Detected	3.2
Mean of Detected Data	1.236
Median of Detected Data	0.7
SD of Detected Data	0.965
Number Above Limit	0
Number Equal Limit	0
Number Below Limit	24

H0: Site Median <= 5 (Form 1)

Test Value	0
Upper Critical Value (0.05)	17
P-Value	1

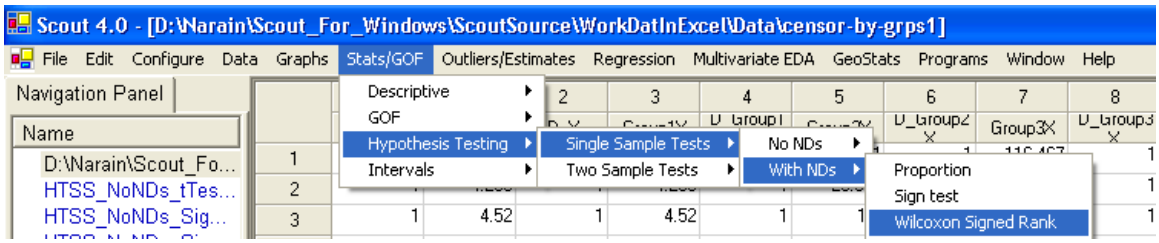
Conclusion with Alpha = 0.05

Do Not Reject H0, Conclude Median <= 5

P-Value > Alpha (0.05)

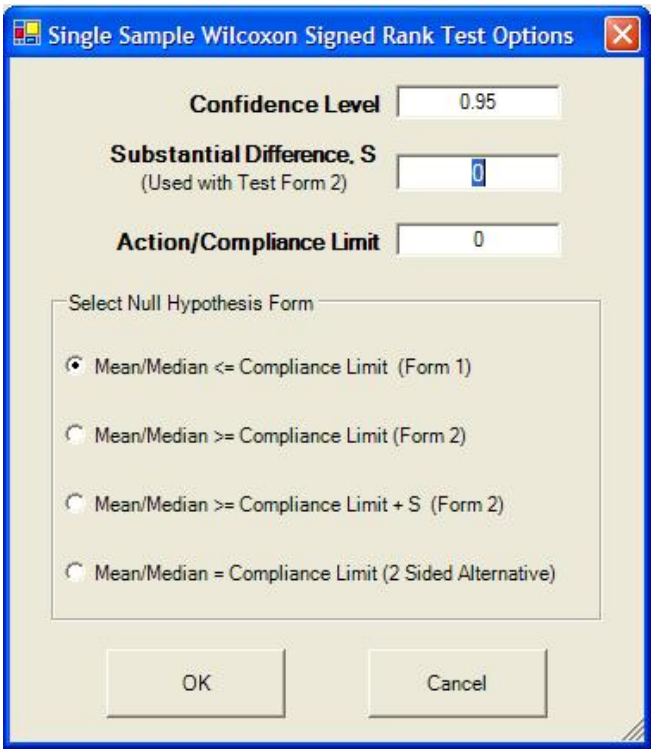
6.3.1.2.3 Single Sample Wilcoxon Signed Rank Test

1. Click **Stats/GOF ► Hypothesis Testing ► Single Sample ► With NDs ► Wilcoxon Signed Rank test.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

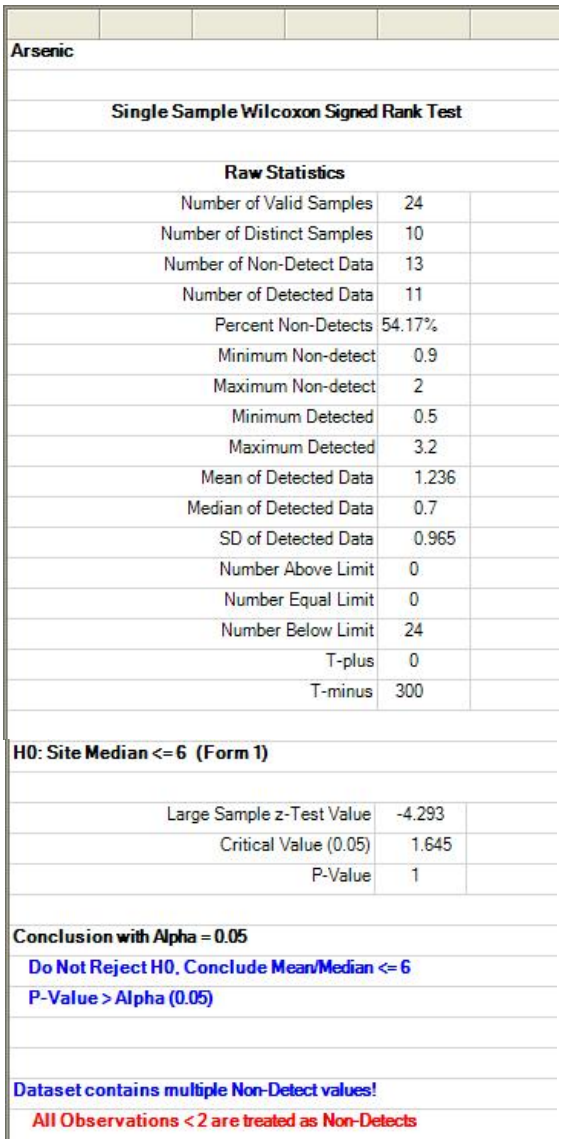
   - Select one or more variables from the "**Select Variables**" screen.
   - If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
   - When the options button is clicked, the following window will be shown.



- Specify the "**Confidence Level**." The default is "**0.95**."
- Specify meaningful values for "**Substantial Difference, S**" and "**Action/Compliance Limit**."
- Select the form of Null Hypothesis. The default is Mean/Median <= Compliance Limit (Form 1).
- Click "**OK**" to continue or "**Cancel**" to cancel the option.

- Click "**OK**" to continue or "**Cancel**" to cancel the test.

Output for Single Sample Wilcoxon Signed Rank Test (Data with Non-detects).



Arsenic

Single Sample Wilcoxon Signed Rank Test

Raw Statistics

Number of Valid Samples	24
Number of Distinct Samples	10
Number of Non-Detect Data	13
Number of Detected Data	11
Percent Non-Detects	54.17%
Minimum Non-detect	0.9
Maximum Non-detect	2
Minimum Detected	0.5
Maximum Detected	3.2
Mean of Detected Data	1.236
Median of Detected Data	0.7
SD of Detected Data	0.965
Number Above Limit	0
Number Equal Limit	0
Number Below Limit	24
T-plus	0
T-minus	300

H0: Site Median <= 6 (Form 1)

Large Sample z-Test Value	-4.293
Critical Value (0.05)	1.645
P-Value	1

Conclusion with Alpha = 0.05

Do Not Reject H0, Conclude Mean/Median <= 6

P-Value > Alpha (0.05)

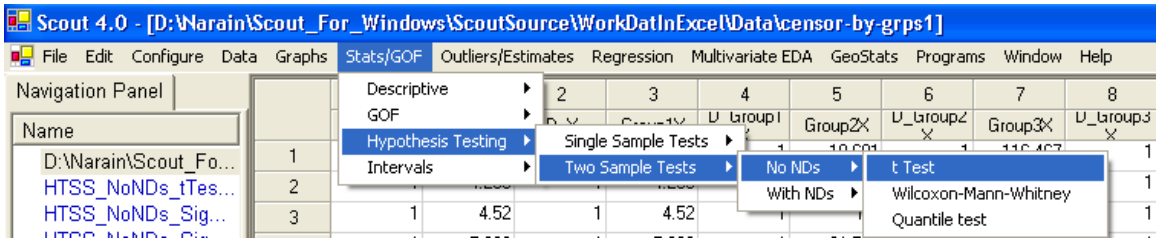
Dataset contains multiple Non-Detect values!

All Observations < 2 are treated as Non-Detects

6.3.2.1 Two-Sample Hypothesis Tests for Data Sets With No Non-detects

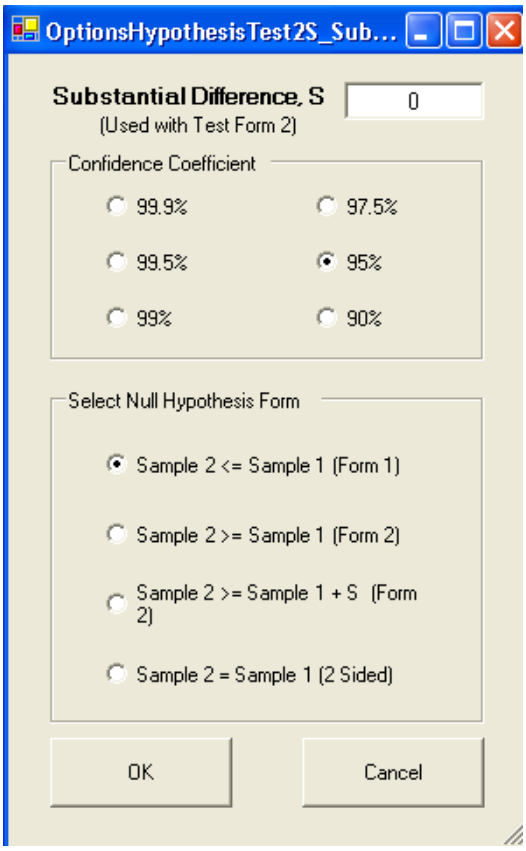
6.3.2.1.1 Two-Sample t-Test

1. Click **Stats/GOF ► Hypothesis Testing ► Two-Sample Tests ► No NDs ► t-Test.**



2. The "**Select Variables**" screen (Section 3.2.2) will appear.

   - Select the variables for testing.
   - When the options button is clicked, the following window will be shown.



- Specify a useful "**Substantial Difference, S**" value. The default choice is "**0**."
- Choose the "**Confidence Level**." The default choice is "**95%**."
- Select the form of Null Hypothesis. The default is AOC <= Background (Form 1).
- Click on "**OK**" to continue or on "**Cancel**" to cancel the option.

- Click on the "**OK**" to continue or on "**Cancel**" to cancel the test.

Output for Two-Sample t-Test (Full Data without NDs).

Raw Statistics

	Sample 1	Sample 2	
Number of Valid Samples	10	20	
Number of Distinct Samples	9	19	
Minimum	3.202	1.5	
Maximum	20.78	37.87	
Mean	8.222	17.09	
Median	5.347	18.79	
SD	5.971	9.713	
SE of Mean	1.888	2.172	

Sample 1 vs Sample 2 Two-Sample t-Test

H0: Mu of Sample 2 - Mu of Sample 1 <= 0

		t-Test	Critical	
Method	DF	Value	t (0.050)	P-Value
Pooled (Equal Variance)	28	2.637	1.701	0.007
Satterthwaite (Unequal Variance)	26.6	3.083	1.703	0.002

Pooled SD 8.688

Conclusion with Alpha = 0.050

* Student t (Pooled) Test: Reject H0, Conclude Sample 2 > Sample 1

* Satterthwaite Test: Reject H0, Conclude Sample 2 > Sample 1

Test of Equality of Variances

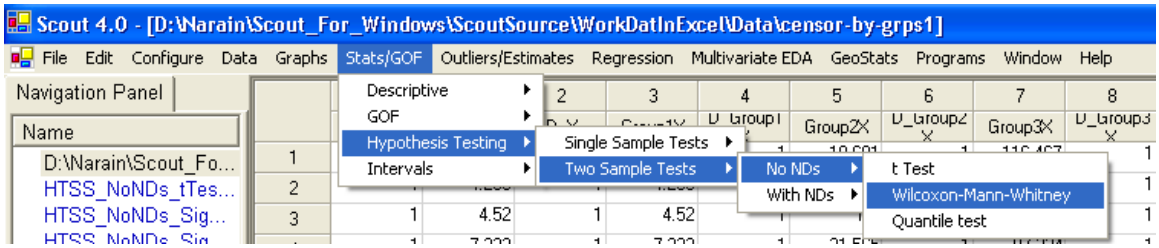
Numerator DF	Denominator DF	F-Test Value	P-Value
19	9	2.646	0.137

Conclusion with Alpha = 0.05

* Two variances appear to be equal

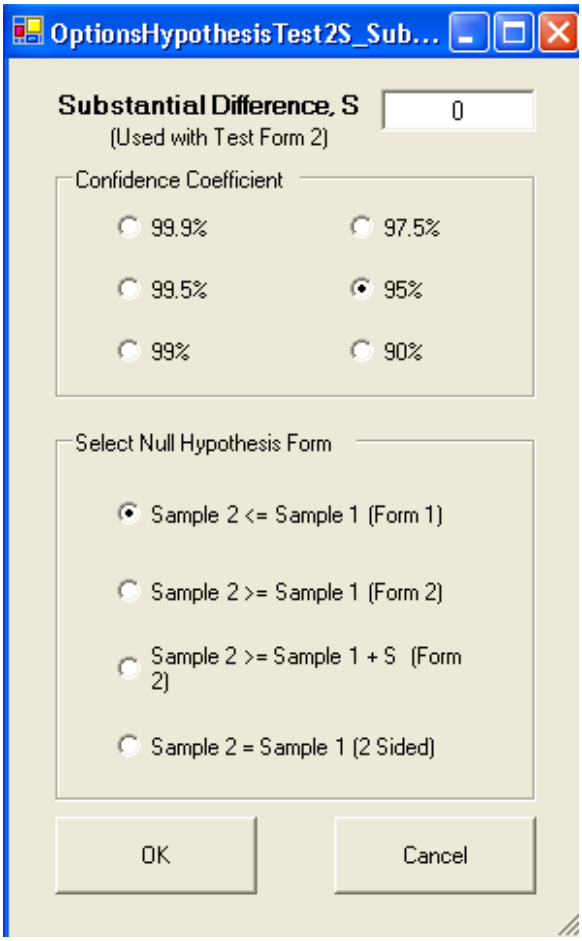
6.3.2.1.2 Two-Sample Wilcoxon Mann Whitney Test

1. Click **Stats/GOF ► Hypothesis Testing ► Two-Sample Tests ► No NDs ► Wilcoxon Mann Whitney test.**



2. The "**Select Variables**" screen (Section 3.2.2) will appear.

   - Select the variables for testing.
   - When the options button is clicked, the following window will be shown.



- Specify a "**Substantial Difference**, **S**" value. The default choice is "**0**."
- Choose the "**Confidence Level**." The default choice is "**95%**."
- Select the form of Null Hypothesis. The default is AOC <= Background (Form 1).
- Click on "**OK**" button to continue or on "**Cancel**" button to cancel the selected options.

- Click on the "**OK**" button to continue or on the "**Cancel**" button to cancel test.

Output for Two-Sample Wilcoxon-Mann-Whitney Test (Full Data).

Sample 2 Data: X(2)
Sample 1 Data: X(1)

Raw Statistics

	Sample 1	Sample 2
Number of Valid Samples	10	20
Number of Distinct Samples	9	19
Minimum	3.202	1.5
Maximum	20.78	37.87
Mean	8.222	17.09
Median	5.347	18.79
SD	5.971	9.713
SE of Mean	1.888	2.172

Wilcoxon-Mann-Whitney (WMW) Test

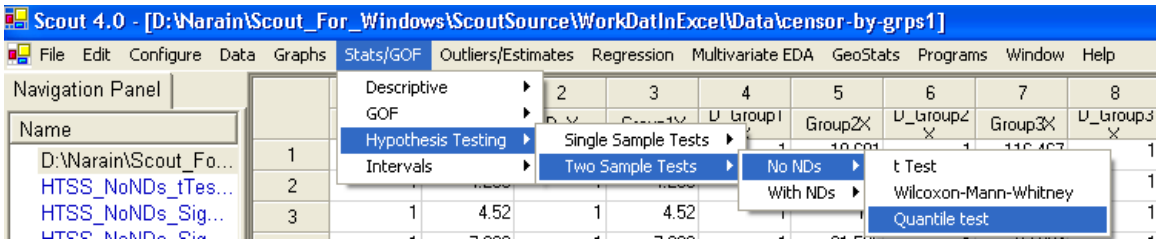
H0: Mean/Median of Sample 2 <= Mean/Median of Sample 1

Sample 2 Rank Sum W-Stat	366
WMW Test U-Stat	156
WMW Critical Value (0.050)	137
Approximate P-Value	0.00731

Conclusion with Alpha = 0.05
Reject H0, Conclude Sample 2 > Sample 1

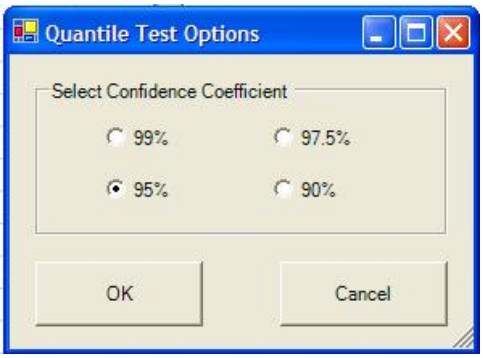
6.3.2.1.3 Two-Sample Quantile Test

1. Click **Stats/GOF ► Hypothesis Testing ► Two-Sample Tests ► No NDs ► Quantile Test.**



2. The "**Select Variables**" screen (Section 3.2.2) will appear.

   - Select the variables for testing.
   - When the options button is clicked, the following window will be shown.



     - Choose the "**Confidence Level**." The default choice is "**95%**."
     - Click on "**OK**" button to continue or on "**Cancel**" button to cancel the option.

   - Click on the "**OK**" button to continue or on the "**Cancel**" button to cancel the test.

Output for Two-Sample Quantile Test (Full Data).

	Non-parametric Quantile Hypothosis Test for Full Dataset (No Non-Detects)
Date/Time of Computation	3/4/2008 6:52:32 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Confidence Coefficient	95%
Null Hypothesis	Sample 2 Concentration Less Than or Equal to Sample 1 Concentration (Form 1)
Alternative Hypothesis	Sample 2 Concentration Greater Than Sample 1 Concentration

Sample 1 Data: Group1X

Sample 2 Data: Group2X

Raw Statistics

	Sample 1	Sample 2
Number of Valid Samples	10	20
Number of Distinct Samples	9	19
Minimum	3.202	1.5
Maximum	20.78	37.87
Mean	8.222	17.09
Median	5.347	18.79
SD	5.971	9.713
SE of Mean	1.888	2.172

Quantile Test

H0: Sample 2 Concentration <= Sample 1 Concentration (Form 1)

Approximate R Value (0.045)	14
Approximate K Value (0.045)	12
Number of Sample 2 Observations in 'R' Largest	13
Calculated Alpha	0.0446

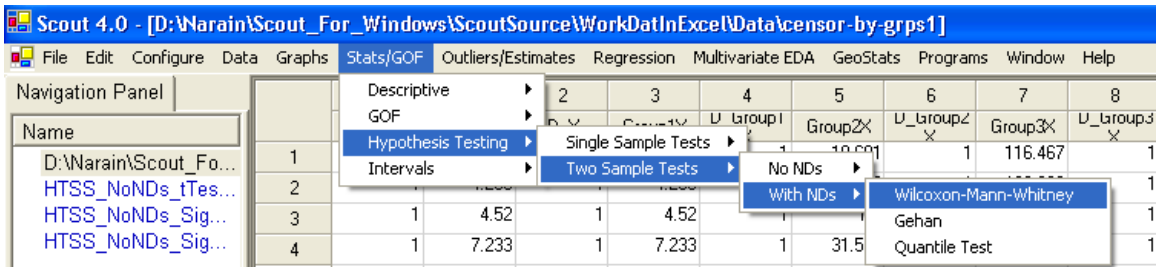
Conclusion with Alpha = 0.045

Reject H0, Conclude Sample 2 Concentration > Sample 1 Concentration

6.3.2.2 Two-Sample Hypothesis Tests for Data Sets With Non-detects

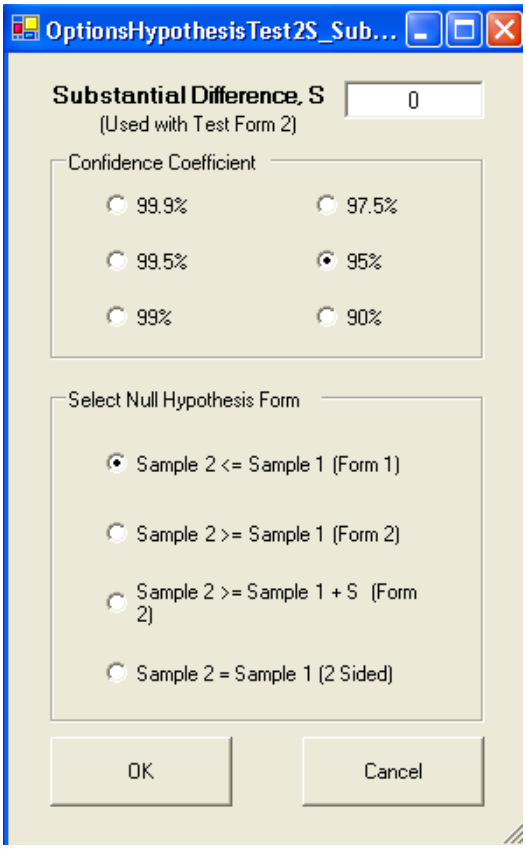
6.3.2.2.1 Two-Sample Wilcoxon Mann Whitney Test

1. Click **Stats/GOF ► Hypothesis Testing ► Two-Sample Tests ► With NDs ► Wilcoxon Mann Whitney test.**



2. The "**Select Variables**" screen (Section 3.2.2) will appear.

   - Select the variables for testing.
   - When the options button is clicked, the following window will be shown.



- Specify a meaningful "**Substantial Difference**, **S**" value. The default choice is "**0.**"

- Choose the "**Confidence level**." The default choice is "**95%.**"

- Select the form of Null Hypothesis. The default is AOC <= Background (Form 1).

- Click on the "**OK**" button to continue or on the "**Cancel**" button to cancel the selected options.

- Click on "**OK**" button to continue or on "**Cancel**" button to cancel the test.

Output for Two-Sample Wilcoxon-Mann-Whitney Test (with Non-detects).

User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Confidence Coefficient	95%
Substantial Difference (S)	0.000
Selected Null Hypothesis	Sample 2 Mean/Median Less Than or Equal to Sample 1 Mean/Median (Form 1)
Alternative Hypothesis	Sample 2 Mean/Median Greater Than Sample 1 Mean/Median

Sample 1 Data: Group1X
Sample 2 Data: Group2X

Raw Statistics

	Sample 1	Sample 2
Number of Valid Samples	10	20
Number of Non-Detect Data	2	2
Number of Detect Data	8	18
Minimum Non-Detect	4	1.5
Maximum Non-Detect	4	1.5
Percent Non detects	20.00%	10.00%
Minimum Detected	3.202	6.316
Maximum Detected	20.78	37.87
Mean of Detected Data	9.277	18.83
Median of Detected Data	6.704	19.36
SD of Detected Data	6.283	8.582

Wilcoxon-Mann-Whitney Sample 1 vs Sample 2 Test
All observations <= 4 (Max DL) are ranked the same
Wilcoxon-Mann-Whitney (WMW) Test

H0: Mean/Median of Sample 2 <= Mean/Median of Sample 1

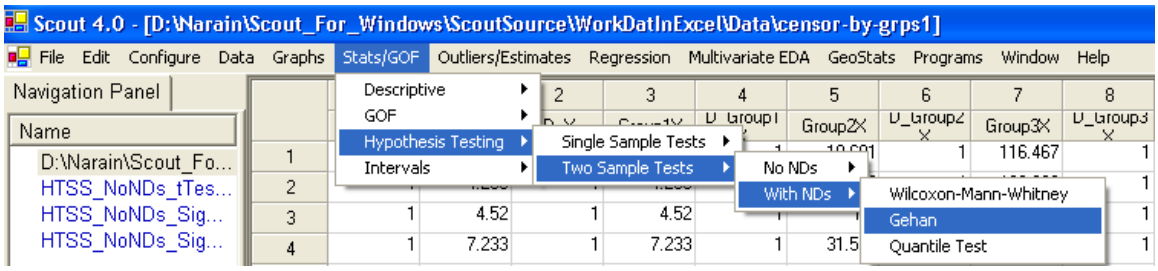
Sample 2 Rank Sum W-Stat	369
WMW Test U-Stat	159
WMW Critical Value (0.050)	137
Approximate P-Value	0.00503

Conclusion with Alpha = 0.05
Reject H0, Conclude Sample 2 > Sample 1

Note: *In the WMW test, all observations below the largest detection limit are considered to be NDs (potentially including detected values) and hence they all receive the same average rank. This action may reduce the associated power of the WMW test considerably. This in turn may lead to incorrect conclusion. All of the hypothesis testing approaches should be supplemented with graphical displays such as Q-Q plots and box plots. When multiple detection limits are present, the use of the Gehan test is preferable.*

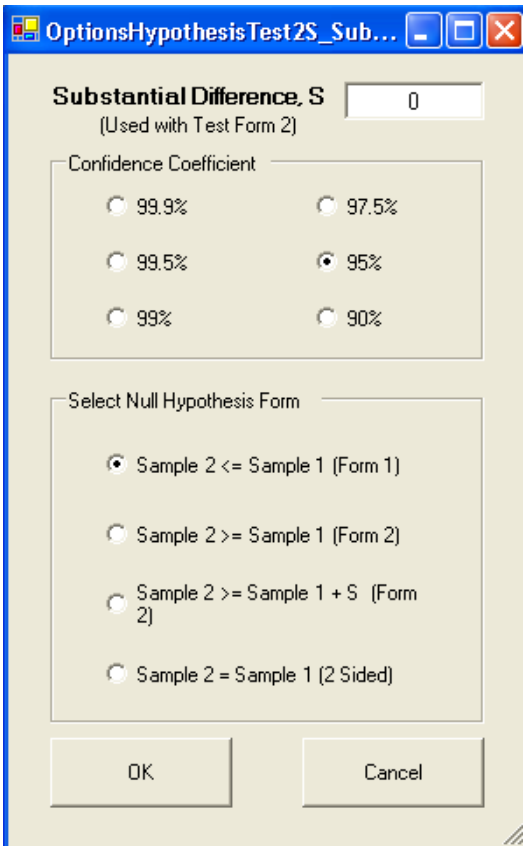
6.3.2.2.2 Two-Sample Gehan Test

1. Click **Stats/GOF ► Hypothesis Testing ► Two-Sample Tests ► With NDs ► Gehan test.**



2. The "**Select Variables**" screen (Section 3.2.2) will appear.

   - Select the variables for testing.
   - When the options button is clicked, the following window will be shown.



- Specify a "**Substantial Difference**, **S**" value. The default choice is "**0**."

- Choose the "**Confidence Level**." The default choice is "**95%**."

- Select the form of Null Hypothesis. The default is AOC <= Background (Form 1).

- Click on "**OK**" button to continue or on "**Cancel**" button to cancel selected options.

- Click on the "**OK**" button to continue or on the "**Cancel**" button to cancel the test.

Output for Two-Sample Gehan Test (with Non-detects).

	Gehan Sample 1 vs Sample 2 Comparison Hypothesis Test for Data Sets with Non-Detects
Date/Time of Computation	3/4/2008 7:10:37 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Confidence Coefficient	95%
Substantial Difference	0.000
Selected Null Hypothesis	Sample 2 Mean/Median Less Than or Equal to Sample 1 Mean/Median (Form 1)
Alternative Hypothesis	Sample 2 Mean/Median Greater Than Sample 1 Mean/Median

Sample 1 Data: Group1X

Sample 2 Data: Group2X

Raw Statistics

	Sample 1	Sample 2
Number of Valid Samples	10	20
Number of Non-Detect Data	2	2
Number of Detect Data	8	18
Minimum Non-Detect	4	1.5
Maximum Non-Detect	4	1.5
Percent Non detects	20.00%	10.00%
Minimum Detected	3.202	6.316
Maximum Detected	20.78	37.87
Mean of Detected Data	9.277	18.83
Median of Detected Data	6.704	19.36
SD of Detected Data	6.283	8.582

Sample 1 vs Sample 2 Gehan Test

H0: Mean/Median of Sample 2 <= Mean/Median of background

Gehan z Test Value	2.556
Critical z (0.95)	1.645
P-Value	0.00529

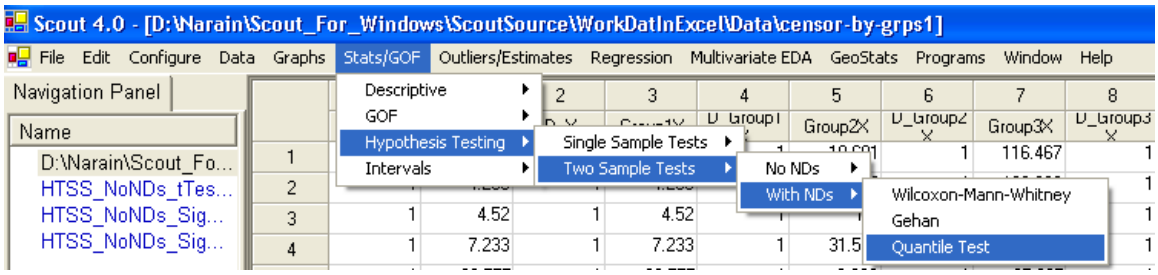
Conclusion with Alpha = 0.05

Reject H0, Conclude Sample 2 > Sample 1

P-Value < alpha (0.05)

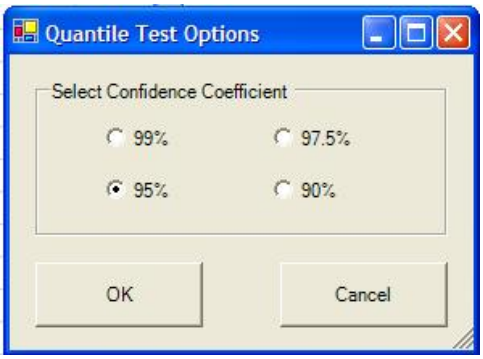
6.3.2.2.3 Two-Sample Quantile Test

1. Click **Stats/GOF ► Hypothesis Testing ► Two-Sample Tests ► With NDs ► Quantile Test.**



2. The "**Select Variables**" screen (Section 3.2.2) will appear.

   - Select the variables for testing.
   - When the options button is clicked, the following window will be shown.



     - Choose the "**Confidence Level**." The default choice is "**95%**."
     - Click on "**OK**" button to continue or on "**Cancel**" button to cancel the option.
   - Click on the "**OK**" button to continue or on the "**Cancel**" button to cancel the test.

Output for Two-Sample Quantile Test (with Non-detects).

	Gehan Sample 1 vs Sample 2 Comparison Hypothesis Test for Data Sets with Non-Detects
Date/Time of Computation	3/4/2008 7:10:37 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Confidence Coefficient	95%
Substantial Difference	0.000
Selected Null Hypothesis	Sample 2 Mean/Median Less Than or Equal to Sample 1 Mean/Median (Form 1)
Alternative Hypothesis	Sample 2 Mean/Median Greater Than Sample 1 Mean/Median

Sample 1 Data: Group1X

Sample 2 Data: Group2X

Raw Statistics

	Sample 1	Sample 2
Number of Valid Samples	10	20
Number of Non-Detect Data	2	2
Number of Detect Data	8	18
Minimum Non-Detect	4	1.5
Maximum Non-Detect	4	1.5
Percent Non detects	20.00%	10.00%
Minimum Detected	3.202	6.316
Maximum Detected	20.78	37.87
Mean of Detected Data	9.277	18.83
Median of Detected Data	6.704	19.36
SD of Detected Data	6.283	8.582

Sample 1 vs Sample 2 Gehan Test

H0: Mean/Median of Sample 2 <= Mean/Median of background

Gehan z Test Value	2.556
Critical z (0.95)	1.645
P-Value	0.00529

Conclusion with Alpha = 0.05

Reject H0, Conclude Sample 2 > Sample 1

P-Value < alpha (0.05)

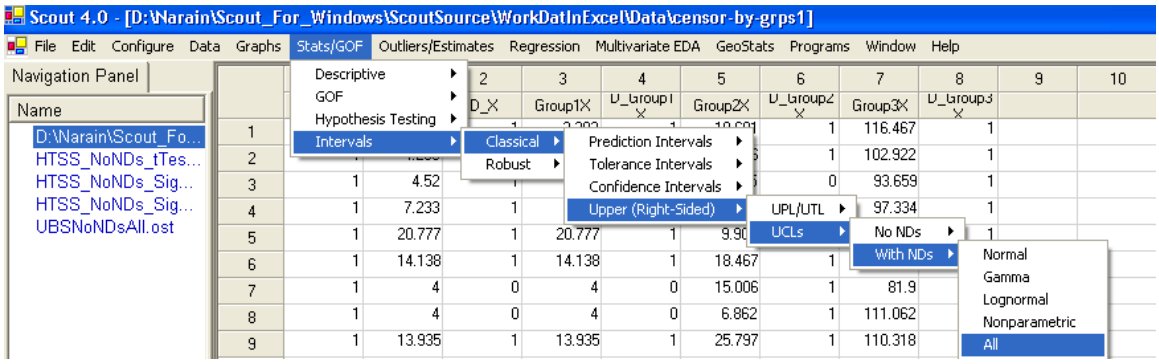
6.4 Classical Intervals

This section illustrates the computations of various parametric and nonparametric lower and upper limits for the confidence, prediction and tolerance intervals. The data used is univariate and can be with or with out non-detects. A detailed description of those limits can be found in the ProUCL 4.00.04 Technical Guide.

6.4.1 Upper (Right Sided) Limits

This module in Scout computes various parametric and nonparametric statistics and upper limits that can be used as background threshold values and other not-to-exceed values. The detailed illustrations of the computing of those statistics can be found in the ProUCL 4.00.04 Technical Guide and User Guide (Chapter 10 and Chapter 11).

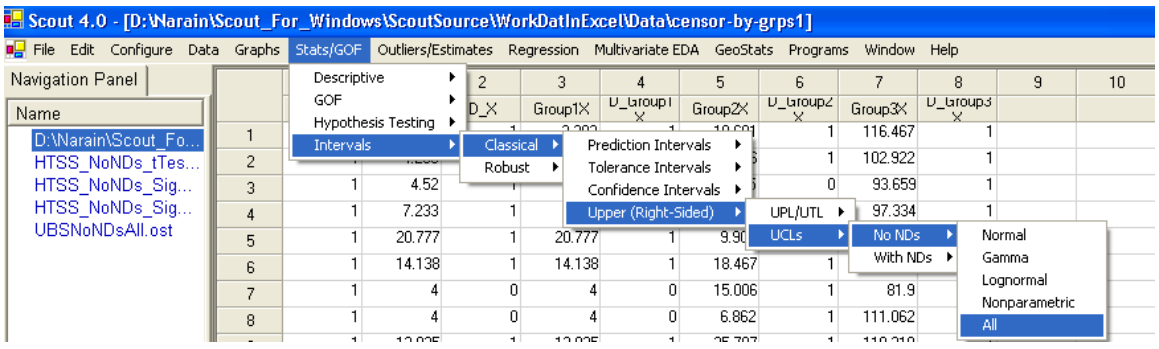
Right sided limits can be obtained separately, for the data following normal, gamma lognormal or nonparametric distributions, using any of the four options ("**Normal**," "**Gamma**," "**Lognormal**" or "**Nonparametric**") from the drop-down menu. If the "**All**" option in the drop-down menu is used, then the limits for all four distributions are printed on single output sheet. Examples illustrated for the Upper (Right Sided) limits are shown using the "**All**" option.



6.4.1.1 Upper (Right Sided) Confidence Limits (UCLs)

6.4.1.1.1 No NDs

1. Click **Stats/GOF ► Intervals ► Upper (Right Sided) ► UCLs ► No NDs ► All.**

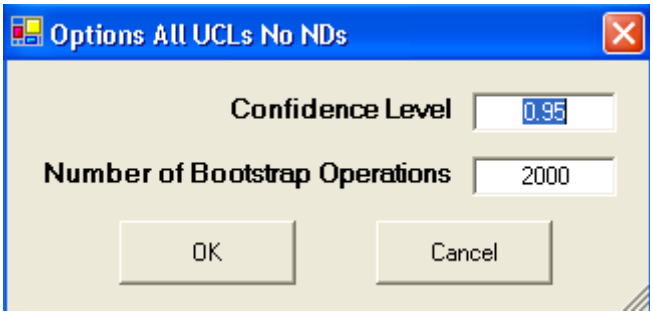


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

- If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- When the option button is clicked, the following window will be shown.



  - Choose the "**Confidence Level**." The default choice is "**95%**."

  - Choose "**Number of Bootstrap Operations**." The default is "**2000**."

  - Click on "**OK**" button to continue or on "**Cancel**" button to cancel the option.

- Click on the "**OK**" button to continue or on the "**Cancel**" button to cancel the UCLs.

Output Screen for UCL for Data Sets with No Non-detects (All option).

General UCL Statistics for Full Data Sets	
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Confidence Coefficient	95%
Number of Bootstrap Operations	2000

X

General Statistics

Number of Valid Observations	53	Number of Distinct Observations	51

Raw Statistics		**Log-transformed Statistics**	
Minimum	1.5	Minimum of Log Data	0.405
Maximum	121.1	Maximum of Log Data	4.797
Mean	51.1	Mean of log Data	3.325
Median	24.56	SD of log Data	1.298
SD	43.78		
Coefficient of Variation	0.857		
Skewness	0.277		

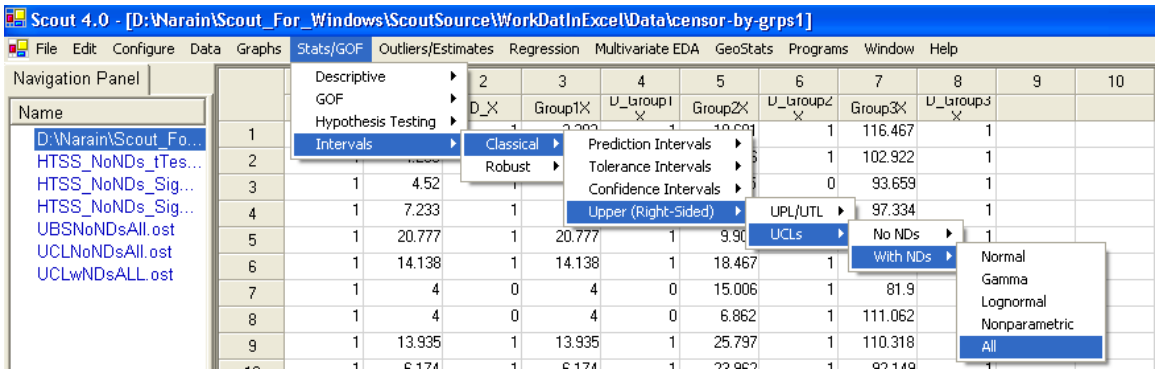
Relevant UCL Statistics

Normal Distribution Test		**Lognormal Distribution Test**	
Lilliefors Test Statistic	0.247	Lilliefors Test Statistic	0.225
Lilliefors Critical Value	0.122	Lilliefors Critical Value	0.122
Data Not Normal at 5% Significance Level		**Data Not Lognormal at 5% Significance Level**	
Assuming Normal Distribution		**Assuming Lognormal Distribution**	
95% Student's-t UCL	61.18	95% H-UCL	100.5
95% UCLs (Adjusted for Skewness)		95% Chebyshev (MVUE) UCL	124.7
95% Adjusted-CLT UCL	61.24	97.5% Chebyshev (MVUE) UCL	151.5
95% Modified-t UCL	61.21	99% Chebyshev (MVUE) UCL	204.1

Gamma Distribution Test		**Data Distribution**	
k star (bias corrected)	0.912	**Data do not follow a Discernable Distribution (0.05)**	
Theta star	56.04		
nu star	96.66		
Approximate Chisquare Value (.05)	74.98	**Nonparametric Statistics**	
Adjusted Level of Significance	0.0455	95% CLT UCL	61
Adjusted Chisquare Value	74.45	95% Jackknife UCL	61.18
		95% Standard Bootstrap UCL	60.9
Anderson-Darling Test Statistic	2.591	95% Bootstrap-t UCL	61.13
Anderson-Darling 5% Critical Value	0.782	95% Hall's Bootstrap UCL	61.15
Kolmogorov-Smirnov Test Statistic	0.222	95% Percentile Bootstrap UCL	61.33
Kolmogorov-Smirnov 5% Critical Value	0.126	95% BCA Bootstrap UCL	61.03
Data Not Gamma Distributed at 5% Significance Level		95% Chebyshev(Mean, Sd) UCL	77.32
		97.5% Chebyshev(Mean, Sd) UCL	88.66
Assuming Gamma Distribution		99% Chebyshev(Mean, Sd) UCL	110.9
95% Approximate Gamma UCL	65.88		
95% Adjusted Gamma UCL	66.35		
Potential UCL to Use		Use 97.5% Chebyshev (Mean, Sd) UCL	88.66

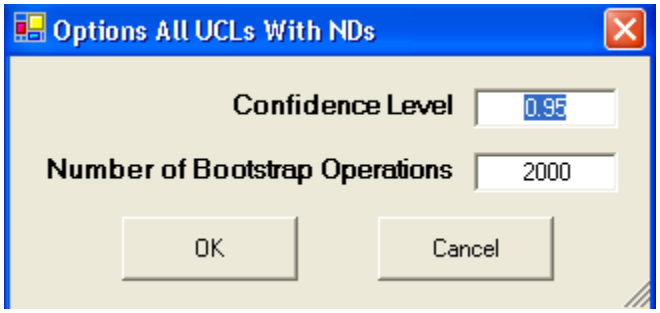
6.4.1.1.2 With NDs

1. Click **Stats/GOF ► Intervals ► Upper (Right Sided) ► UCLs ► With NDs ► All.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- When the option button is clicked, the following window will be shown.



  - Choose the "**Confidence Level**." The default choice is "**95%**."
  - Choose "**Number of Bootstrap Operations**." The default is "**2000**."
  - Click on "**OK**" button to continue or on "**Cancel**" button to cancel the option.

- Click on the "**OK**" button to continue or on the "**Cancel**" button to cancel the UCLs.

Output Screen for UCL for Data Sets with Non-detects (All option).

	General UCL Statistics for Data Sets with Non-Detects			
User Selected Options				
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1			
Full Precision	OFF			
Confidence Coefficient	95%			
Number of Bootstrap Operations	2000			
X				
	General Statistics			
Number of Valid Data	53		Number of Detected Data	49
Number of Distinct Detected Data	49		Number of Non-Detect Data	4
			Percent Non-Detects	7.55%
Raw Statistics			**Log-transformed Statistics**	
Minimum Detected	3.202		Minimum Detected	1.164
Maximum Detected	121.1		Maximum Detected	4.797
Mean of Detected	55.05		Mean of Detected	3.523
SD of Detected	43.2		SD of Detected	1.128
Minimum Non-Detect	1.5		Minimum Non-Detect	0.405
Maximum Non-Detect	4		Maximum Non-Detect	1.386
Note: Data have multiple DLs - Use of KM Method is recommended			Number treated as Non-Detect	5
For all methods (except KM, DL/2, and ROS Methods),			Number treated as Detected	48
Observations < Largest ND are treated as NDs			Single DL Non-Detect Percentage	9.43%
	UCL Statistics			
Normal Distribution Test with Detected Values Only			**Lognormal Distribution Test with Detected Values Only**	
Lilliefors Test Statistic	0.802		Lilliefors Test Statistic	0.856
5% Lilliefors Critical Value	0.947		5% Lilliefors Critical Value	0.947
Data Not Normal at 5% Significance Level			**Data Not Lognormal at 5% Significance Level**	

Output Screen for UCL for Data Sets with Non-detects (All option) (continued).

Assuming Normal Distribution	
DL/2 Substitution Method	
Mean	51
SD	43.9
95% DL/2 (t) UCL	61.1
Maximum Likelihood Estimate(MLE) Method	
Mean	48.86
SD	46.77
95% MLE (t) UCL	59.62
95% MLE (Tiku) UCL	59.4

Assuming Lognormal Distribution	
DL/2 Substitution Method	
Mean	3.273
SD	1.406
95% H-Stat (DL/2) UCL	105.5
Log ROS Method	
Mean in Log Scale	3.34
SD in Log Scale	1.264
Mean in Original Scale	51.13
SD in Original Scale	43.75
95% Percentile Bootstrap UCL	61.06
95% BCA Bootstrap UCL	60.82

Gamma Distribution Test with Detected Values Only	
k star (bias corrected)	1.111
Theta star	49.54
nu star	108.9
A-D Test Statistic	2.882
5% A-D Critical Value	0.775
K-S Test Statistic	0.775
5% K-S Critical Value	0.13
Data Not Gamma Distributed at 5% Significance Level	

Data Distribution Test with Detected Values Only
Data do not follow a Discernable Distribution (0.05)

Assuming Gamma Distribution	
Gamma ROS Statistics using Extrapolated Data	
Minimum	1.0000E-9
Maximum	121.1
Mean	50.9
Median	24.56
SD	44.02
k star	0.302
Theta star	168.3
Nu star	32.05
AppChi2	20.11
95% Gamma Approximate UCL	81.11
95% Adjusted Gamma UCL	82.2

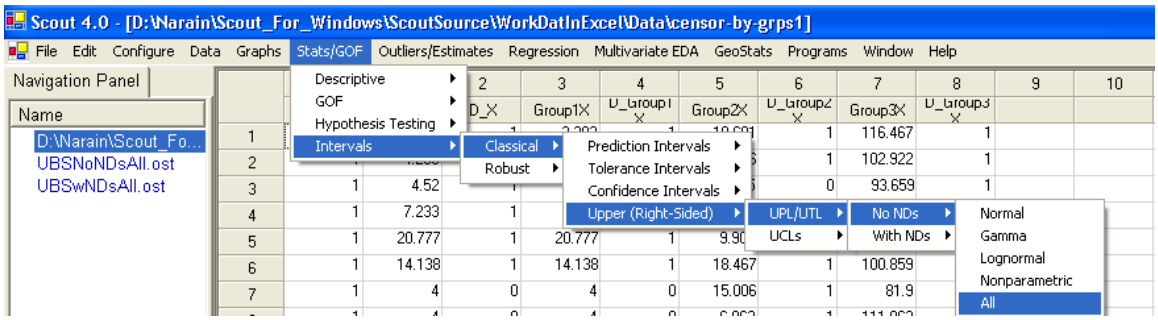
Nonparametric Statistics	
Kaplan-Meier (KM) Method	
Mean	51.14
SD	43.33
SE of Mean	6.013
95% KM (t) UCL	61.21
95% KM (z) UCL	61.03
95% KM (jackknife) UCL	61.14
95% KM (bootstrap t) UCL	62.07
95% KM (BCA) UCL	60.58
95% KM (Percentile Bootstrap) UCL	60.92
95% KM (Chebyshev) UCL	77.35
97.5% KM (Chebyshev) UCL	88.69
99% KM (Chebyshev) UCL	111
Potential UCLs to Use	
95% KM (Chebyshev) UCL	77.35

Note: DL/2 is not a recommended method.

6.4.1.2 Upper Prediction Limits (UPL) / Upper Tolerance Limits (UTL)

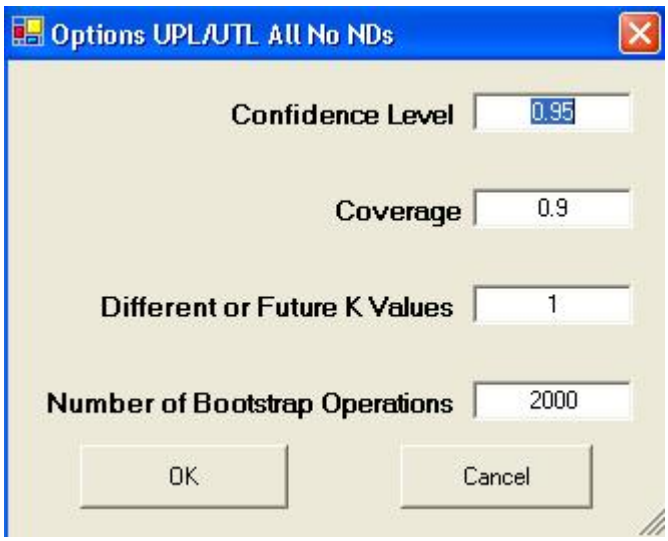
6.4.1.2.1 No NDs

1. Click **Stats/GOF ► Intervals ► Upper (Right Sided) ► UPL/UTL ► No NDs ► All.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- When the option button is clicked, the following window will be shown.



- Specify the "**Confidence Level**"; a number in the interval [0.5, 1), 0.5 inclusive. The default choice is "**0.95.**"

- Specify the "**Coverage**" level; a number in the interval (0.0, 1). Default is "**0.9**."

- Specify the next "**K.**" The default choice is "**1**."

- Specify the "**Number of Bootstrap Operations**." The default choice is "**2000**."

- Click on "**OK**" button to continue or on "**Cancel**" button to cancel the option.

- Click on "**OK**" button to continue or on "**Cancel**" button to cancel the UPLs and UTLs.

Output Screen for UPL/UTL for Data Sets with No Non-detects (All option).

	General Background Statistics for Full Data Sets
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Confidence Coefficient	95%
Coverage	90%
Different or Future K Values	1
Number of Bootstrap Operations	2000

X

General Statistics

Total Number of Observations	53	Number of Distinct Observations	51

Raw Statistics		**Log-Transformed Statistics**	
Minimum	1.5	Minimum	0.405
Maximum	121.1	Maximum	4.797
Second Largest	116.5	Second Largest	4.758
First Quartile	9.708	First Quartile	2.273
Median	24.56	Median	3.201
Third Quartile	96.88	Third Quartile	4.573
Mean	51.1	Mean	3.325
SD	43.78	SD	1.298
Coefficient of Variation	0.857		
Skewness	0.277		

Background Statistics

Normal Distribution Test		**Lognormal Distribution Test**	
Lilliefors Test Statistic	0.247	Lilliefors Test Statistic	0.225
Lilliefors Critical Value	0.122	Lilliefors Critical Value	0.122
Data Not Normal at 5% Significance Level		**Data Not Lognormal at 5% Significance Level**	

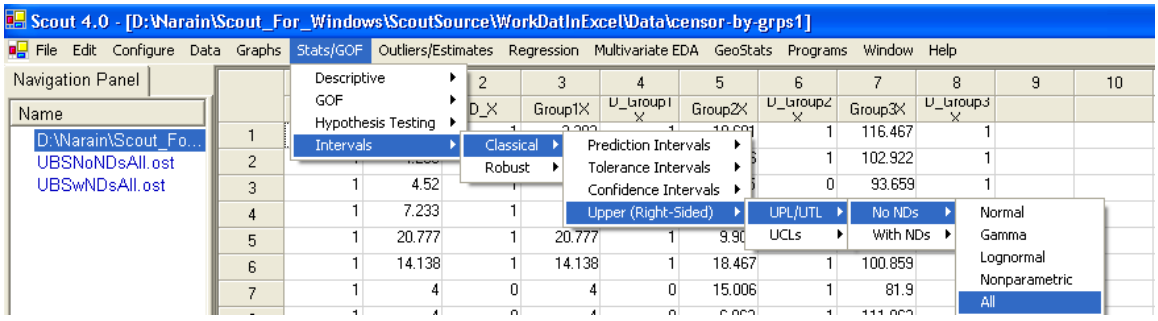
Output Screen for UPL/UTL for Data Sets with No Non-detects (All option) (continued).

Assuming Normal Distribution		Assuming Lognormal Distribution	
95% UTL with 90% Coverage	122.4	95% UTL with 90% Coverage	229.8
95% UPL (t)	125.1	95% UPL (t)	249.3
90% Percentile (z)	107.2	90% Percentile (z)	146.6
95% Percentile (z)	123.1	95% Percentile (z)	234.9
99% Percentile (z)	153	99% Percentile (z)	568.9
Gamma Distribution Test		**Data Distribution Test**	
k star	0.912	Data do not follow a Discernable Distribution (0.05)	
Theta star	56.04		
nu star	96.66		
A-D Test Statistic	2.591	**Nonparametric Statistics**	
5% A-D Critical Value	0.782	90% Percentile	110
K-S Test Statistic	0.222	95% Percentile	116.4
5% K-S Critical Value	0.126	99% Percentile	121.1
Data Not Gamma Distributed at 5% Significance Level			
Assuming Gamma Distribution		95% UTL with 90% Coverage	116.4
90% Percentile	120.4	95% Percentile Bootstrap UTL with 90% Coverage	114.8
95% Percentile	158.2	95% BCA Bootstrap UTL with 90% Coverage	114.8
99% Percentile	246.6	95% UPL	116.4
		95% Chebyshev UPL	243.7
		Upper Threshold Limit Based upon IQR	227.6

Note: UPL (or upper percentile for gamma distributed data) represents a preferred estimate of BTV

6.4.1.2.2 With NDs

1. Click **Stats/GOF ► Intervals ► Upper (Right Sided) ► UPL/UTL ► With NDs ► All.**

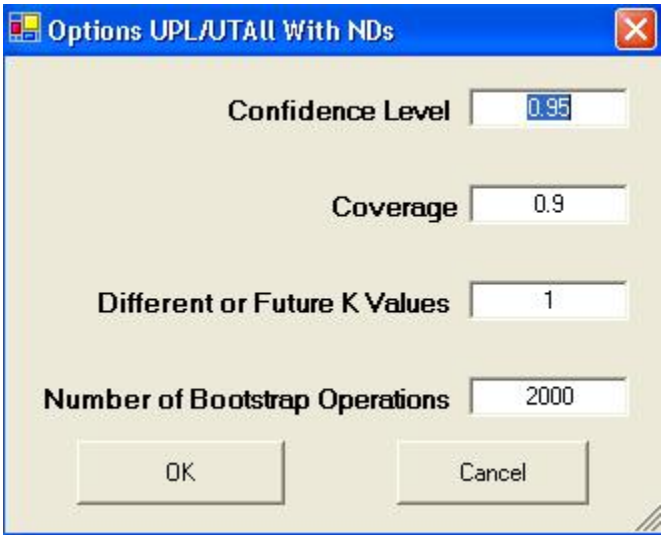


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

- If the statistics have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- When the option button is clicked, the following window will be shown.



  - Specify the "**Confidence Level**"; a number in the interval [0.5, 1), 0.5 inclusive. The default choice is "**0.95**."

  - Specify the "**Coverage**" level; a number in the interval (0.0, 1). Default is "**0.9**."

  - Specify the next "**K**." The default choice is "**1**."

  - Specify the "**Number of Bootstrap Operations**." The default choice is "**2000**."

  - Click on "**OK**" button to continue or on "**Cancel**" button to cancel the option.

- Click on "**OK**" button to continue or on "**Cancel**" button to cancel the UPLs and UTLs.

Output Screen for UPL/UTL for Data Sets With Non-detects (All option).

	General Background Statistics for Data Sets with Non-Detects
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Confidence Coefficient	95%
Coverage	90%
Different or Future K Values	1
Number of Bootstrap Operations	2000

×

General Statistics

Number of Valid Data	53	Number of Detected Data	49
Number of Distinct Detected Data	49	Number of Non-Detect Data	4
		Percent Non-Detects	7.55%

Raw Statistics		**Log-transformed Statistics**	
Minimum Detected	3.202	Minimum Detected	1.164
Maximum Detected	121.1	Maximum Detected	4.797
Mean of Detected	55.05	Mean of Detected	3.523
SD of Detected	43.2	SD of Detected	1.128
Minimum Non-Detect	1.5	Minimum Non-Detect	0.405
Maximum Non-Detect	4	Maximum Non-Detect	1.386

Data with Multiple Detection Limits	**Single Detection Limit Scenario**	
Note: Data have multiple DLs - Use of KM Method is recommended	Number treated as Non-Detect with Single DL	5
For all methods (except KM, DL/2, and ROS Methods),	Number treated as Detected with Single DL	48
Observations < Largest ND are treated as NDs	Single DL Non-Detect Percentage	9.43%

Background Statistics

Normal Distribution Test with Detected Values Only		**Lognormal Distribution Test with Detected Values Only**	
Lilliefors Test Statistic	0.802	Lilliefors Test Statistic	0.856
5% Lilliefors Critical Value	0.947	5% Lilliefors Critical Value	0.947
Data Not Normal at 5% Significance Level		**Data Not Lognormal at 5% Significance Level**	

Output Screen for UPL/UTL for Data Sets With Non-detects (All option) (continued).

Assuming Normal Distribution		Assuming Lognormal Distribution	
DL/2 Substitution Method		DL/2 Substitution Method	
Mean	51	Mean (Log Scale)	3.273
SD	43.9	SD (Log Scale)	1.406
95% UTL 90% Coverage	122.5	95% UTL 90% Coverage	260.2
95% UPL (t)	125.2	95% UPL (t)	284.1
90% Percentile (z)	107.3	90% Percentile (z)	159.9
95% Percentile (z)	123.2	95% Percentile (z)	266.5
99% Percentile (z)	153.1	99% Percentile (z)	694.8
Maximum Likelihood Estimate(MLE) Method		Log ROS Method	
Mean	48.86	Mean in Original Scale	51.13
SD	46.77	SD in Original Scale	43.75
95% UTL with 90% Coverage	125	95% UTL with 90% Coverage	220.8
		95% BCA UTL with 90% Coverage	114.5
		95% Bootstrap (%) UTL with 90% Coverage	114.8
95% UPL (t)	127.9	95% UPL (t)	238.9
90% Percentile (z)	108.8	90% Percentile (z)	142.5
95% Percentile (z)	125.8	95% Percentile (z)	225.6
99% Percentile (z)	157.7	99% Percentile (z)	533.6
Gamma Distribution Test with Detected Values Only		**Data Distribution Test with Detected Values Only**	
k star (bias corrected)	1.111	Data do not follow a Discernable Distribution (0.05)	
Theta star	49.54		
nu star	108.9		
A-D Test Statistic	2.882	**Nonparametric Statistics**	
5% A-D Critical Value	0.775	Kaplan-Meier (KM) Method	
K-S Test Statistic	0.236	Mean	51.14
5% K-S Critical Value	0.13	SD	43.33
Data Not Gamma Distributed at 5% Significance Level		SE of Mean	6.013
		95% KM UTL with 90% Coverage	121.7
Assuming Gamma Distribution		95% KM Chebyshev UPL	241.8
Gamma ROS Statistics with extrapolated Data		95% KM UPL (t)	124.4
Mean	50.9	90% Percentile (z)	106.7
Median	24.56	95% Percentile (z)	122.4
SD	44.02	99% Percentile (z)	151.9
k star	0.302		
Theta star	168.3		
Nu star	32.05		
95% Percentile of Chisquare (2k)	2.759		
90% Percentile	150		
95% Percentile	232.3		
99% Percentile	445.9		
Note: UPL (or upper percentile for gamma distributed data) represents a preferred estimate of BTV			
For an Example: KM-UPL may be used when multiple detection limits are present			
Note: DL/2 is not a recommended method.			

6.4.2 Classical Confidence Intervals

6.4.2.1 Without Non-detects

The confidence intervals for data with no non-detects available in Scout are:

- Normal:
  - Student's t

$$\bar{x} \pm t_{(\alpha/2, n-1)} \frac{s}{\sqrt{n}}$$

- Gamma:
  - Approximate Gamma
  - Adjusted Gamma
- Lognormal
  - Land's H

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

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

  - Chebyshev MVUE

$$\bar{x}_{mvue} \pm \frac{1}{\sqrt{\alpha}} \frac{\sigma_{mvue}}{\sqrt{n}}$$

- Nonparametric
  - CLT

$$\bar{x} \pm z_{(\alpha/2)} \frac{s}{\sqrt{n}}$$

  - Jackknife

$$J\left(\hat{\theta}\right) \pm t_{(\alpha/2, n-1)} \hat{\sigma}_{J(\hat{\theta})}$$

  - Standard Bootstrap

$$\hat{\theta} \pm z_{(\alpha/2)} \hat{\sigma}_B$$

- Bootstrap t

$$LCL = \bar{x} - t_{(1-\alpha/2,n-1)}\frac{s}{\sqrt{n}} \qquad UCL = \bar{x} - t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}}$$

- Percentile Bootstrap

$$\text{LCL} = \frac{\alpha}{2} \text{ percentile of } \bar{x}$$

$$\text{UCL} = 1 - \frac{\alpha}{2} \text{ percentile of } \bar{x}$$

- Chebyshev

$$\bar{x} \pm \frac{1}{\sqrt{\alpha}}\frac{s}{\sqrt{n}}$$

- Modified (t)

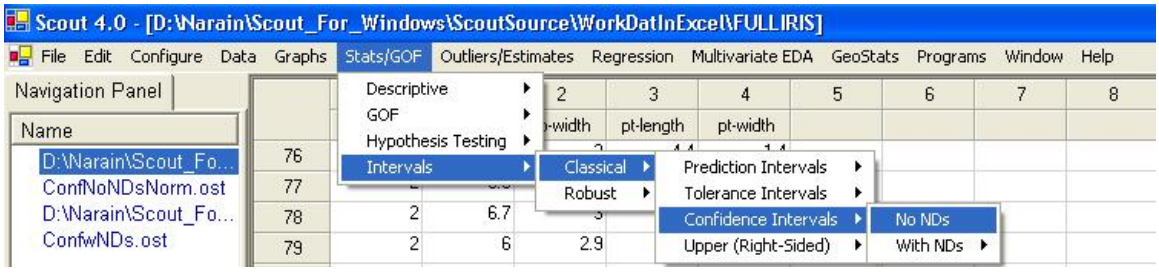
$$\bar{x} + \frac{\hat{\mu}}{6s^2 n} \pm t_{(\alpha/2,n-1)}\frac{s}{\sqrt{n}}$$

- Adjusted CLT

$$\bar{x} \pm \left(z_{(\alpha/2)} + \frac{\hat{k}_3\left(1 + 2z_{(\alpha/2)}\right)}{6\sqrt{n}} \right)\frac{s}{\sqrt{n}}$$

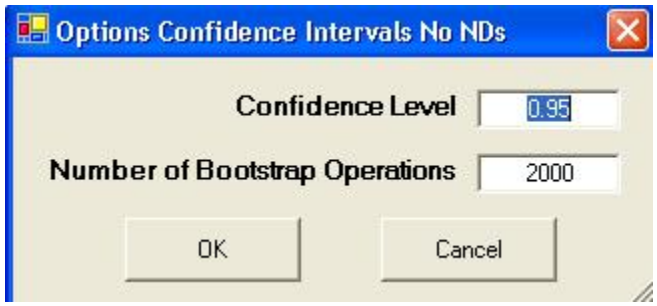
Details of those intervals can be found in the ProUCL 4.00.04 Technical Guide.

1. Click **Stats/GOF ►Intervals ► Classical ► Confidence Intervals ► No NDs.**



The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click on "**Options**" for interval options.



  - Specify the preferred "**Confidence Level**." The default is "**0.95**."
  - Specify the preferred number of bootstrap operations. The default is "**2000**."
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.
- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Classical Confidence Intervals without Non-detects.

	Confidence Intervals for Datasets without Non-Detects
Date/Time of Computation	1/15/2008 12:39:56 PM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\BODYFAT
Full Precision	OFF
Number of Bootstrap Operations	2000
Confidence Coefficient	0.95

Skin(x1)

Number of Valid Observations	20
Number of Distint Observations	20

Raw Statistics

Mean	25.31
Median	25.55
Variance	25.23
Standard Deviation	5.023

Normal Intervals

Normal	Lower Limit	Upper Limit
Student's t	22.95	27.66

Gamma Statistics

k Star (Bias Corrected)	20.54
Theta Star	1.232
nu Star	821.5

Gamma Intervals

Gamma	Lower Limit	Upper Limit
Approximate Gamma	23.03	27.94
Adjusted Gamma	22.82	28.21

Log-Transformed Statistics

Mean of Log-Transformed Data	3.21
Standard Deviation of Log-Transformed Data	0.216
MVU Estimate of Median	24.75
MVU Estimate of Mean	25.34
MVU Estimate of SD	5.509
MVU Estimate of Standard Error of Mean	1.232

Lognormal Intervals

Lognormal	Lower Limit	Upper Limit
Land's H	23.02	28.26
Chebyshev (MVUE)	19.83	30.85

Nonparametric Intervals

Nonparametric	Lower Limit	Upper Limit
Central Limit Theorem	23.1	27.51
Jackknife	22.95	27.66
Standard Bootstrap	23.19	27.42
Bootstrap-t	22.67	27.59
Percentile Bootstrap	23.13	27.31
Chebyshev	20.28	30.33
Modified (t)	22.93	27.63
Adjusted CLT	23.3	27.31

6.4.2.2 With Non-detects

The confidence intervals for data with non-detects available in Scout are:

- Normal:
  - Student's t

$$\hat{\mu}_{mle} \pm t_{(\alpha/2,n-1)}\sqrt{\frac{\hat{\sigma}^2_{mle}}{n}}$$

  - Normal ROS Student's t
- Gamma:
  - Gamma ROS Approximate Gamma
  - Gamma ROS Adjusted Gamma
- Lognormal:
  - Lognormal ROS Land's H
  - Lognormal ROS Chebyshev MVUE
  - Lognormal ROS % Bootstrap
- Nonparametric:
  - Kaplan-Meier (t)

$$\hat{\mu}_{KM} \pm t_{(\alpha/2,n-1)}\sqrt{\sigma^2_{KM-se}}$$

  - Kaplan-Meier (z)

$$\hat{\mu}_{KM} \pm z_{(\alpha/2,n-1)}\sqrt{\sigma^2_{KM-se}}$$

  - Kaplan-Meier % Bootstrap (bootstrapping the KM means)

$$\text{LCL} = \frac{\alpha}{2} \text{ percentile of } \bar{x}$$

$$\text{UCL} = 1-\frac{\alpha}{2} \text{ percentile of } \bar{x}$$

  - Kaplan Meier BCA Bootstrap

- Kaplan Meier Chebyshev

$$\overline{x} \pm \frac{1}{\sqrt{\alpha}} \frac{s}{\sqrt{n}}$$

- Winsor (t)

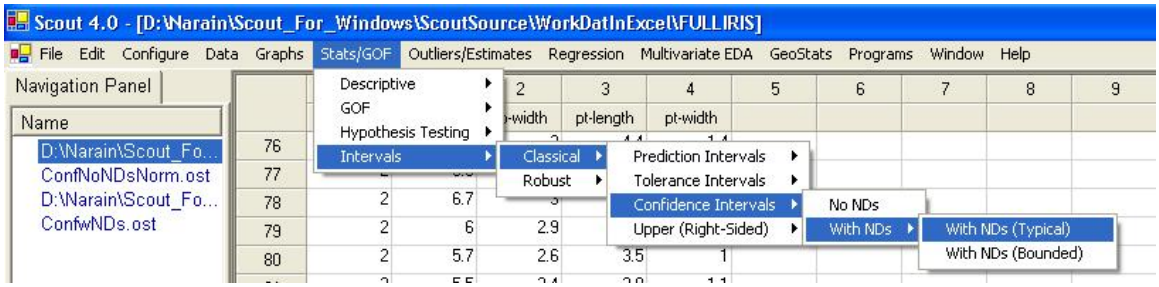
$$\overline{x}_w \pm t_{(\alpha/2,\nu-1)} \frac{s_w}{\sqrt{n}}$$

where ν = n-2k $\qquad s_w = \frac{s(n-k)}{\nu - 1}$

$\overline{x}_w$ = Winsorized mean

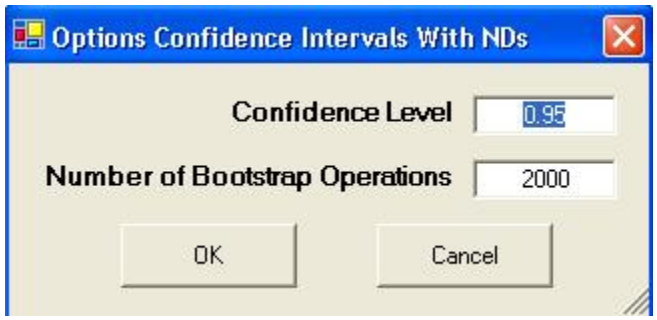
Details of those intervals can be found in the ProUCL 4.00.04 Technical Guide.

1. Click **Stats/GOF ►Intervals ► Classical ► Confidence Intervals ► With NDs (Typical) or With NDs (Bounded).**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click on "**Options**" for interval options.



- Specify the preferred "**Confidence Level**." The default is "**0.95**."
- Specify the preferred number of bootstrap operations. The default is "**2000**."
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Classical Confidence Intervals with Non-detects (Typical).

	Confidence Intervals Datasets with Non-Detects		
Date/Time of Computation	1/21/2008 1:25:37 PM		
User Selected Options			
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1.x		
Full Precision	OFF		
Number of Bootstrap Operations	2000		
Confidence Coefficient	0.95		
X			
General Statistics			
Number of Valid Data	53		
Number of Detected Data	49		
Number of Distinct Detected Data	49		
Minimum Detected	3.202		
Maximum Detected	121.1		
Number of Non-Detect Data	4		
Percent Non-Detects	7.55%		
Minimum Non-detect	1.5		
Maximum Non-detect	4		
Mean of Detected Data	55.05		
SD of Detected Data	43.2		
Maximum Likelihood Statistics			
Maximum Likelihood Estimated Mean	48.86		
Maximum Likelihood Estimated Stdv	46.77		
Normal Confidence Intervals			
Normal	Lower Limit	Upper Limit	
MLE (t)	35.97	61.75	
Normal ROS Statistics			
Mean of Normal ROS Data	48.06		
Stdv of Normal ROS Data	48.36		
ROS Student's t	34.73	61.39	

Output for Classical Confidence Intervals with Non-detects (Typical) (continued).

Gamma ROS Statistics		
k Star of Gamma ROS Data	0.302	
Theta Star of Gamma ROS Data	168.3	
Nu Star of Gamma ROSData	32.05	
Gamma Intervals		
Gamma	Lower Limit	Upper Limit
ROS Approximate Gamma	32.93	89
ROS Adjusted Gamma	43.94	62.09
Log-Transformed Statistics		
Mean of Log-Transformed Detected Data	3.523	
Stdv of Log-Transformed Detected Data	1.128	
Mean of Lognormal ROS Data	51.13	
Stdv of Lognormal ROS Data	43.75	
Lognormal Confidence Intervals		
Lognormal	Lower Limit	Upper Limit
ROS Land's H	41.91	109.5
ROS % Bootstrap	40.11	62.98
ROS BCA Bootstrap	39.71	63.51
Kaplan Meier Distribution Free Statistics		
Kaplan Meier Mean	51.14	
Kaplan Meier Stdv	43.33	
Kaplan Meier SEM	6.013	
Nonparametric Confidence Intervals		
Nonparametric	Lower Limit	Upper Limit
Kaplan Meier (t)	39.07	63.21
Kaplan Meier (z)	39.35	62.92
Kaplan Meier % Bootstrap	40.1	62.95
Kaplan Meier BCA Bootstrap	40.91	63.54
Kaplan Meier Chebyshev	24.25	78.03
Winsorization Statistics		
Winsor Mean	50.72	
Winsor Stdv	42.87	
Winsor (t)	38.83	62.6

Output for Classical Confidence Intervals with Non-detects (Bounded).

	Bounded Confidence Intervals for Datasets with Non-Detects
Date/Time of Computation	1/15/2008 12:45:11 PM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Number of Bounding Operations	1000
Bounding Coefficient	0.9
Number of Bootstrap Operations	2000
Confidence Coefficient	0.9

×

General Statistics	Lower Bound (LB)	Upper Bound (UB)
Mean	50.95	51.06
Standard Deviation	43.84	43.97

Normal Confidence Limits	LB LCL	UB LCL	LB UCL	UB UCL
Student (t)	40.83	40.97	61.06	61.14

Gamma Statistics	Lower Bound (LB)	Upper Bound (UB)
k Star (Bias Corrected)	0.761	0.883
Theta Star	57.8	66.87
nu Star	80.62	93.58

Gamma Confidence Limits	LB LCL	UB LCL	LB UCL	UB UCL
Approximate Gamma	40.04	40.77	66.11	67.4
Adjusted Gamma	39.72	40.51	66.61	68.11

Lognormal Statistics	Lower Bound (LB)	Upper Bound (UB)
Mean of Log-Transformed Data	3.179	3.297
d Deviation of Log-Transformed Data	1.355	1.674

Lognormal Statistics	Lower Bound (LB)	Upper Bound (UB)
Mean of Log-Transformed Data	3.179	3.297
d Deviation of Log-Transformed Data	1.355	1.674

Lognormal Confidence Limits	LB LCL	UB LCL	LB UCL	UB UCL
Land's H	46.22	59.14	106.6	197.8
Chebyshev (MVUE)	-1.432	16.07	114.1	189.1

Nonparametric Confidence Limits	LB LCL	UB LCL	LB UCL	UB UCL
Central Limit Theorem	41.01	41.15	60.88	60.96
Central Limit Theorem	40.83	40.97	61.06	61.14
Standard Bootstrap	40.9	41.43	60.58	61.09
Bootstrap-t	40.64	41.65	60.88	61.88
Percentile Bootstrap	40.76	41.69	60.42	61.36
Chebyshev	31.84	32.01	70.04	70.1
Modified (t)	40.87	41.02	61.1	61.18
Adjusted CLT	40.78	40.9	61.12	61.2

6.4.3 Classical Tolerance Intervals

6.4.3.1 Without Non-detects

The tolerance intervals for data with no non-detects available in Scout are:

- Normal:

$$LTL = \overline{x} - K_{\left(n,\alpha/2,p\right)}s \qquad UTL = \overline{x} + K_{\left(n,1-\alpha/2,p\right)}s$$

- Lognormal:

$$LTL = \exp\left(\overline{y} - K_{\left(n,\alpha/2,p\right)}s_y\right) \qquad UTL = \exp\left(\overline{y} + K_{\left(n,1-\alpha/2,p\right)}s_y\right)$$

- Nonparametric:
  - Percentile Bootstrap
  - BCA Bootstrap

$$\hat{\alpha} = \frac{\sum\left(\overline{x} - \overline{x}_{-i}\right)^3}{6\left[\sum\left(\overline{x} - \overline{x}_{-i}\right)^2\right]^{1.5}} \qquad \hat{z}_0 = \Phi^{-1}\left[\frac{\#\left(\overline{x}_i < \overline{x}\right)}{N}\right]$$

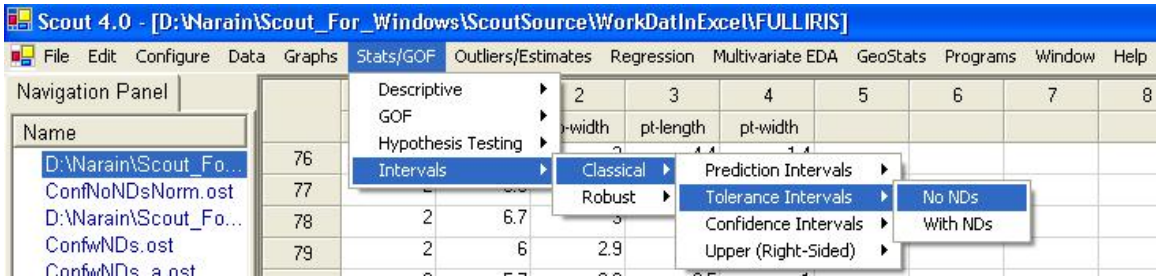
$$\alpha_{2(LOWER)} = \Phi\left[\overline{z}_0 + \frac{\hat{z}_0 + z^{\alpha/2}}{1-\left(\hat{z}_0 + z^{\alpha/2}\right)\hat{\alpha}}\right] \qquad \alpha_{2(UPPER)} = \Phi\left[\overline{z}_0 + \frac{\hat{z}_0 + z^{1-\alpha/2}}{1-\left(\hat{z}_0 + z^{1-\alpha/2}\right)\hat{\alpha}}\right]$$

$$LTL = \overline{x}^{\left(\alpha_{2(LOWER)}\right)} \qquad UTL = \overline{x}^{\left(\alpha_{2(UPPER)}\right)}$$

  - Percentile Tolerance

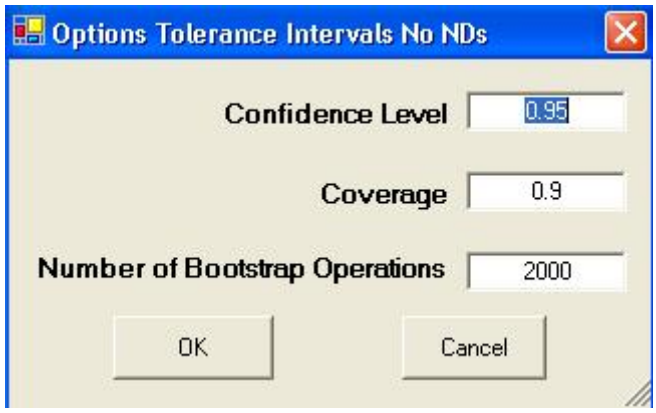
Details of those intervals can be found in the ProUCL 4.00.04 Technical Guide.

1. Click **Stats/GOF ►Intervals ► Classical ► Tolerance Intervals ► No NDs.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click on "**Options**" for interval options.



  - Specify the preferred "**Confidence Level**." The default is "**0.95**."
  - Specify the preferred coverage percentage. The default is "**0.9**."
  - Specify the preferred number of bootstrap operations. The default is "**2000**."
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.
- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Classical Tolerance Intervals without Non-detects.

Tolerance Intervals/Limits (TLs) for Datasets Without Non-Detects	
Date/Time of Computation	2/25/2008 7:51:11 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Number of Bootstrap Operations	2000
Coverage	0.9
Confidence Coefficient	0.95

X

Number of Valid Observations	53
Number of Distinct Observations	51

Raw Statistics

Mean	51.1
Minimum	1.5
5% Percentile	2.606
10% Percentile	4.071
1st Quartile	9.608
Median	24.56
3rd Quartile	95.73
90% Percentile	107.6
95% Percentile	112.9
Maximum	121.1
Standard Deviation	43.78
MAD / 0.6745	30.48
IQR / 1.35	64.57
1% Percentile (z)	-50.75
5% Percentile (z)	-20.91
10% Percentile (z)	-5.006
1st Quartile (z)	21.57
ROS Median (z)	51.1
3rd Quartile (z)	80.64
90% Percentile (z)	107.2
95% Percentile (z)	123.1
99% Percentile (z)	153

Normal Tolerance Limits

Tolerance	Lower Limit	Upper Limit
Normal	-35.74	137.9

Log-Transformed Statistics

Mean of Log-Transformed Data	3.325
Standard Deviation of Log-Transformed Data	1.298

Log-Transformed Tolerance Limits

Lognormal	2.119	364.6

Nonparametric Tolerance Limits

% Bootstrap	98.51	116.4
BCA Bootstrap	97.97	114.8
% TL	2.053	116.4

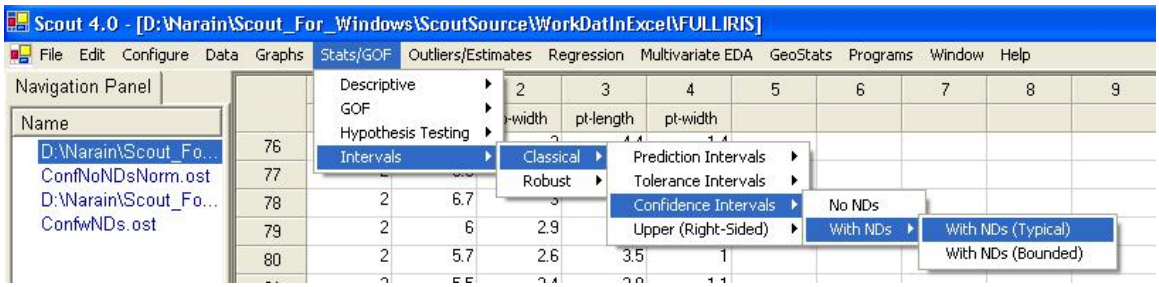
6.4.3.2 With Non-detects

The tolerance intervals for data with non-detects available in Scout are:

- Normal:
  - Using MLE of mean and standard deviation
  - Using Normal ROS methods
- Lognormal ROS
  - Using bootstrap methods based on Lognormal ROS
- Nonparametric:
  - Nonparametric KM

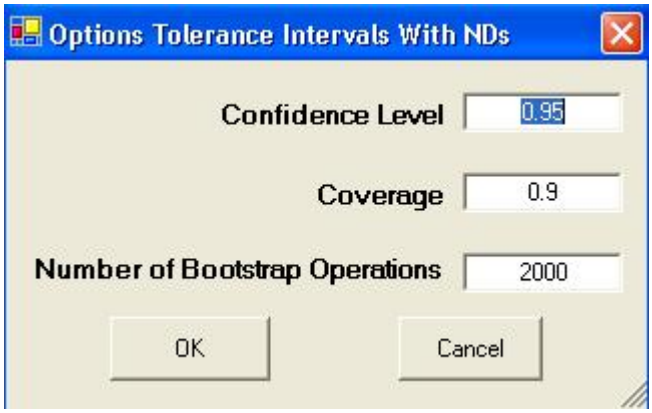
Details of those intervals can be found in the ProUCL 4.00.04 Technical Guide and the Scout Technical Guide.

1. Click **Stats/GOF ►Intervals ► Classical ► Tolerance Intervals ► With NDs.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click on "**Options**" for interval options.



  - Specify the preferred "**Confidence Level**." The default is "**0.95**."
  - Specify the preferred coverage percentage. The default is "**0.9**."
  - Specify the preferred number of bootstrap operations. The default is "**2000**."

- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Classical Tolerance Intervals with Non-detects.

	Tolerance Intervals for Datasets with Non-Detects
Date/Time of Computation	2/25/2008 8:36:35 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Number of Bootstrap Operations	2000
Coverage	0.9
Confidence Coefficient	0.95

K2 represents the two-sided cutoff for tolerance intervals based upon the procedure described in Hahn and Meeker (1991)

X	
Number of Valid Observations	53
Number of Distinct Observations	49
Number of Non-Detect Data	4
Number of Detected Data	49
Minimum Detected	3.202
Maximum Detected	121.1
Percent Non-Detects	7.55%
Minimum Non-detect	1.5
Maximum Non-detect	4
Raw Statistics	
Mean of Detected Data	55.05
SD of Detected Data	43.2
Maximum Likelihood Estimates (MLEs)	
MLE Mean	48.86
1% Percentile (z)	-59.95
5% Percentile (z)	-28.07
10% Percentile (z)	-11.08
1st Quartile (z)	17.31
ROS Median (z)	48.86
3rd Quartile (z)	80.41
90% Percentile (z)	108.8
95% Percentile (z)	125.8
99% Percentile (z)	157.7
MLE Stdv	46.77

Output for Classical Tolerance Intervals with Non-detects (continued).

K2	1.983	
Normal Tolerance Intervals		
	Lower Limit	Upper Limit
MLE	-43.91	141.6
Normal ROS Statistics		
Minimum of ROS Data	-49.39	
Maximum of ROS Data	121.1	
Mean of ROS Data	48.06	
SD of ROS Data	48.36	
K2	1.983	
Nonparamtric Percentiles Using ROS Data		
1% ROS Percentile	-49.39	
5% ROS Percentile	-36.93	
10% ROS Percentile	3.513	
1st ROS Quartile	9.608	
ROS Median	24.26	
3rd ROS Quartile	95.73	
90% ROS Percentile	107.6	
95% ROS Percentile	112.9	
99% ROS Percentile	118.7	
Parametric Percentiles Using Normal Distribution		
1% ROS Percentile (z)	-64.44	
5% ROS Percentile (z)	-31.49	
10% ROS Percentile (z)	-13.92	
1st ROS Quartile (z)	15.44	
ROS ROS Median (z)	48.06	
3rd ROS Quartile (z)	80.68	
90% ROS Percentile (z)	110	
95% ROS Percentile (z)	127.6	
99% ROS Percentile (z)	160.6	
Normal ROS Tolerance Interval		
	Lower Limit	Upper Limit
Normal	-47.86	144

Output for Classical Tolerance Intervals with Non-detects (continued).

Log-Transformed Statistics	
Mean of Log-Transformed Detected Data	3.523
Stdv of Log-Transformed Detected Data	1.128
Minimum of Lognormal ROS Data	2.204
Maximum of Lognormal ROS Data	121.1
Mean of Lognormal ROS Data	51.13
Stdv of Lognormal ROS Data	43.75
K2	1.983

Nonparamtric Percentiles Using ROS Data	
1% ROS Percentile	2.204
5% ROS Percentile	3.041
10% ROS Percentile	4.174
1st ROS Quartile	9.608
ROS Median	24.26
3rd ROS Quartile	95.73
90% ROS Percentile	107.6
95% ROS Percentile	112.9
99% ROS Percentile	118.7

Parametric Percentiles Using Lognormal Distribution	
1% ROS Percentile (z)	1.493
5% ROS Percentile (z)	3.532
10% ROS Percentile (z)	5.589
1st ROS Quartile (z)	12.04
ROS ROS Median (z)	28.22
3rd ROS Quartile (z)	66.19
90% ROS Percentile (z)	142.5
95% ROS Percentile (z)	225.6
99% ROS Percentile (z)	533.6

Lognormal Tolerance Intervals		
	Lower Limit	Upper Limit
ROS Lognormal	2.302	346
ROS % Bootstrap	98.51	116.4
ROS BCA Bootstrap	97.97	116.4

Output for Classical Tolerance Intervals with Non-detects (continued).

Kaplan Meier Distribution Free Statistics	
Mean	51.14
1% Percentile (z)	-49.66
5% Percentile (z)	-20.13
10% Percentile (z)	-4.389
1st Quartile (z)	21.91
Median (z)	51.14
3rd Quartile (z)	80.36
90% Percentile (z)	106.7
95% Percentile (z)	122.4
99% Percentile (z)	151.9
Standard Deviation	43.33
Kaplan Meier SEM	6.013
K2	1.983

Nonparametric Tolerance Intervals		
	Lower Limit	Upper Limit
KM Nonparametric	-34.8	137.1

6.4.4 Classical Prediction Intervals

6.4.4.1 Without Non-detects

The prediction intervals for data with no non-detects available in Scout are (the square root quantity, $[(1/k) + (1/n)]^{1/2}$, in the equations below is given for k = 1 future observation):

- Normal

$$\bar{x} \pm t_{(\alpha/2,n-1)} s \sqrt{1+\frac{1}{n}}$$

- Lognormal

$$\exp\left(\bar{y} \pm t_{(\alpha/2,n-1)} s_y \sqrt{1+\frac{1}{n}} \right)$$

- Chebyshev

$$\bar{x} \pm \frac{1}{\sqrt{\alpha}} s \sqrt{1+\frac{1}{n}}$$

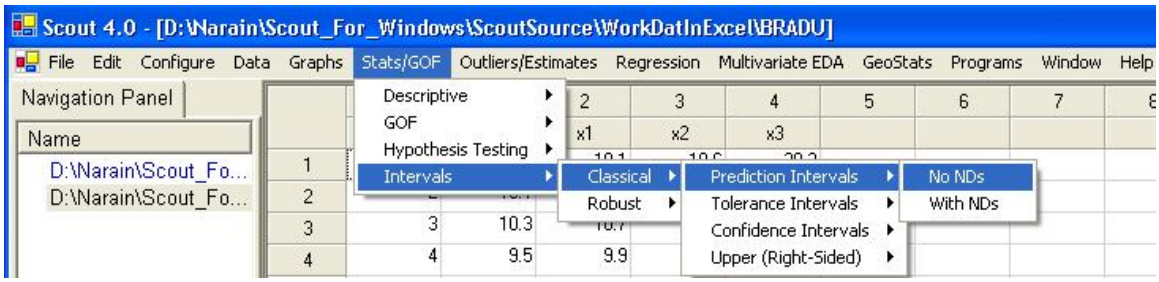
- Nonparametric t

$$LPL = x_{(m)} \qquad m = (n+1)\left(\frac{\alpha}{2}\right)$$

$$UPL = x_{(m)} \qquad m = (n+1)\left(1-\frac{\alpha}{2}\right)$$

Details of those intervals can be found in the ProUCL 4.00.04 Technical Guide and the Scout Technical Guide.

1. Click **Stats/GOF ►Intervals ► Classical ► Prediction Intervals ► No NDs.**

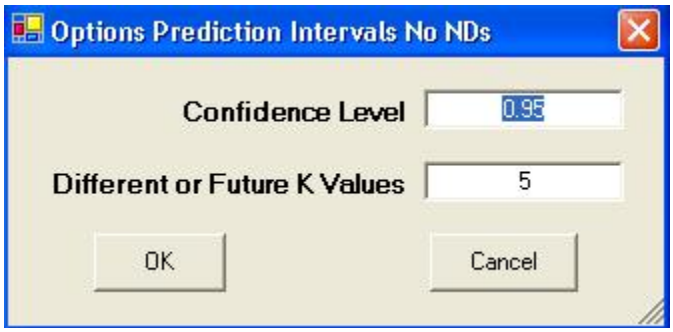


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

   - If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

   - Click on "**Options**" for interval options.



- Specify the preferred "**Confidence Level**." The default is "**0.95**."
- Specify the number of future k values. The default is "**5**."
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Classical Prediction Intervals without Non-detects.

	Prediction Intervals/Limts (PLs) for Datasets Without Non-Detects
User Selected Options	
Date/Time of Computation	2/25/2008 9:03:29 AM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Number of Future K Values	5
Confidence Coefficient	0.95

X

Number of Valid Observations	53
Number of Distinct Observations	51

Raw Statistics

Minimum	1.5
Mean	51.1
Median	24.56
Maximum	121.1
Standard Deviation	43.78

Normal Prediction Intervals

Normal	Lower Limit	Upper Limit
Student's t	-37.58	139.8
For Next 5	-67.06	169.3

Log-Transformed Statistics

Mean of Log-Transformed Data	3.325
Standard Deviation of Log-Transformed Data	1.298

Lognormal	Lower Limit	Upper Limit
Log	2.007	385
For Next 5	0.838	922.5

Chebyshev	Lower Limit	Upper Limit
Chebyshev	-146.5	248.7

Nonparametric	Lower Limit	Upper Limit
Nonparametric	0.394	119.5

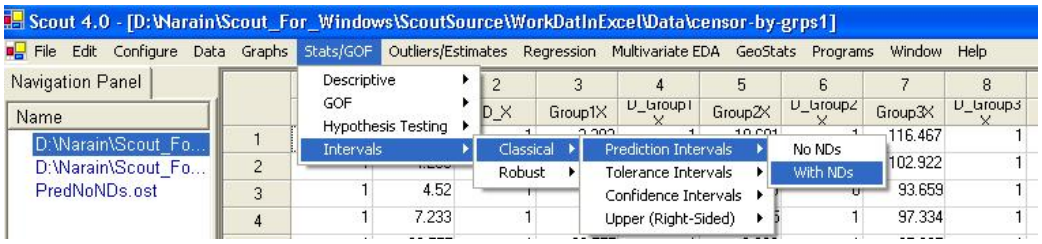
6.4.4.2 With Non-detects

The prediction intervals for data with non-detects available in Scout are:

- MLE – t
- Lognormal ROS - t
- Nonparametric
    - KM Chebyshev
    - KM – t
    - KM – z

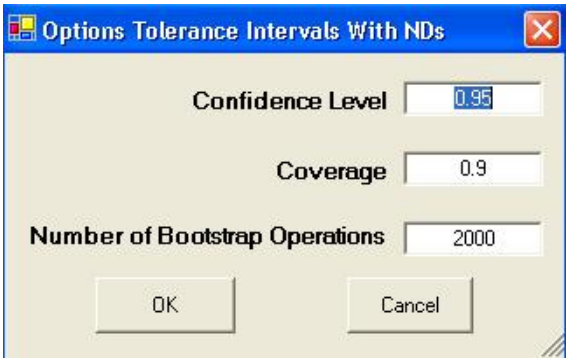
Details of those intervals can be found in the ProUCL 4.00.04 Technical Guide and the Scout Technical Guide.

1. Click **Stats/GOF ►Intervals ► Classical ► Prediction Intervals ► With NDs.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

    - Select one or more variables from the "**Select Variables**" screen.
    - If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
    - Click on "**Options**" for interval options.



- Specify the preferred "**Confidence Level**." The default is "**0.95**."
- Specify the number of future k values. The default is "**5**."
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Classical Prediction Intervals with Non-detects.

	Prediction Intervals for Datasets with Non-Detects
User Selected Options	
Date/Time of Computation	2/25/2008 9:06:12 AM
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Number of Future K Values	5
Confidence Coefficient	0.95

X	
General Statistics	
Number of Valid Observations	53
Number of Distinct Observations	49
Number of Non-Detect Data	4
Number of Detected Data	49
Minimum Detected	3.202
Maximum Detected	121.1
Percent Non-Detects	7.55%
Minimum Non-detect	1.5
Maximum Non-detect	4
Raw Statistics	
Mean of Detected Data	55.05
SD of Detected Data	43.2
Maximum Likelihood Estimates (MLEs)	
MLE Mean	48.86
1% Percentile (z)	-59.95
5% Percentile (z)	-28.07
10% Percentile (z)	-11.08
1st Quartile (z)	17.31
ROS Median (z)	48.86
3rd Quartile (z)	80.41
90% Percentile (z)	108.8
95% Percentile (z)	125.8
99% Percentile (z)	157.7
MLE Stdv	46.77

Output for Classical Prediction Intervals with Non-detects (continued).

Normal Prediction Intervals		
	Lower Limit	Upper Limit
MLE (t)	-45.88	143.6
Prediction Interval for Next 5	-77.37	175.1
Normal ROS Statistics		
Minimum of ROS Data	-49.39	
Mean of ROS Data	48.06	
Maximum of ROS Data	121.1	
SD of ROS Data	48.36	
Nonparamtric Percentiles Using ROS Data		
1% ROS Percentile	-49.39	
5% ROS Percentile	-36.93	
10% ROS Percentile	3.513	
1st ROS Quartile	9.608	
ROS Median	24.26	
3rd ROS Quartile	95.73	
90% ROS Percentile	107.6	
95% ROS Percentile	112.9	
99% ROS Percentile	118.7	
Parametric Percentiles Using Normal Distribution		
1% ROS Percentile (z)	-64.44	
5% ROS Percentile (z)	-31.49	
10% ROS Percentile (z)	-13.92	
1st ROS Quartile (z)	15.44	
ROS ROS Median (z)	48.06	
3rd ROS Quartile (z)	80.68	
90% ROS Percentile (z)	110	
95% ROS Percentile (z)	127.6	
99% ROS Percentile (z)	160.6	
Normal ROS Prediction Intervals		
	Lower Limit	Upper Limit
Normal	-49.89	146
Prediction Interval for Next 5	-82.46	178.6

Output for Classical Prediction Intervals with Non-detects (continued).

Kaplan Meier Distribution Free Statistics	
Mean	51.14
1% Percentile (z)	-49.66
5% Percentile (z)	-20.13
10% Percentile (z)	-4.389
1st Quartile (z)	21.91
Median (z)	51.14
3rd Quartile (z)	80.36
90% Percentile (z)	106.7
95% Percentile (z)	122.4
99% Percentile (z)	151.9
Standard Deviation	43.33
Kaplan Meier SEM	6.013

Nonparametric Prediction Intervals		
	Lower Limit	Upper Limit
KM Chebyshev	-144.5	246.7
KM (t)	-36.62	138.9
KM (z)	-34.58	136.9
Prediction Interval for Next 5	-65.8	168.1

6.5 Robust Intervals

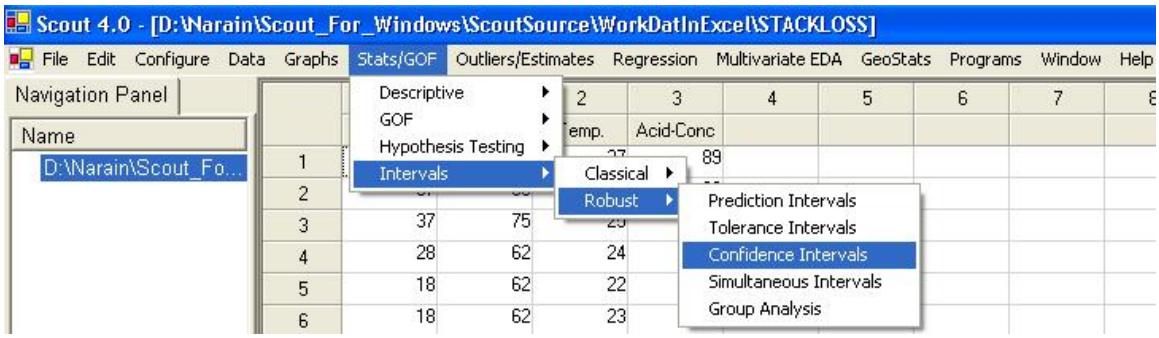
Various robust and resistant univariate intervals (confidence intervals, prediction intervals, tolerance intervals, and simultaneous intervals) can be computed using Scout. For details of those robust intervals, refer to Kafadar (1982) and Singh and Nocerino (1997). Singh and Nocerino (1997) discussed the performance of those intervals. Typically, those robust procedures are iterative requiring initial estimates of location and scale. In Scout, those robust intervals can be computed using the mean and the standard deviation, or median and MAD/0.6745 as the initial estimates of center and location. The different methods for the computation of the robust intervals available in Scout are:

- PROP (using PROP influence function)
- Huber (using Huber influence function)
- Tukey's Biweight as described in Tukey (1977)
- Lax/Kafadar Biweight as described in Kafadar (1982) and Horn (1988)
- MVT (using trimming percentage)

The performance of these intervals can also be compared using the graphics option in the variable selection screen. If the graphics option is selected, then a plot of intervals will be generated for all of the interval methods selected in the options window.

6.5.1 Robust Confidence Intervals

1. Click **Stats/GOF ►Intervals ► Robust ► Confidence Intervals.**

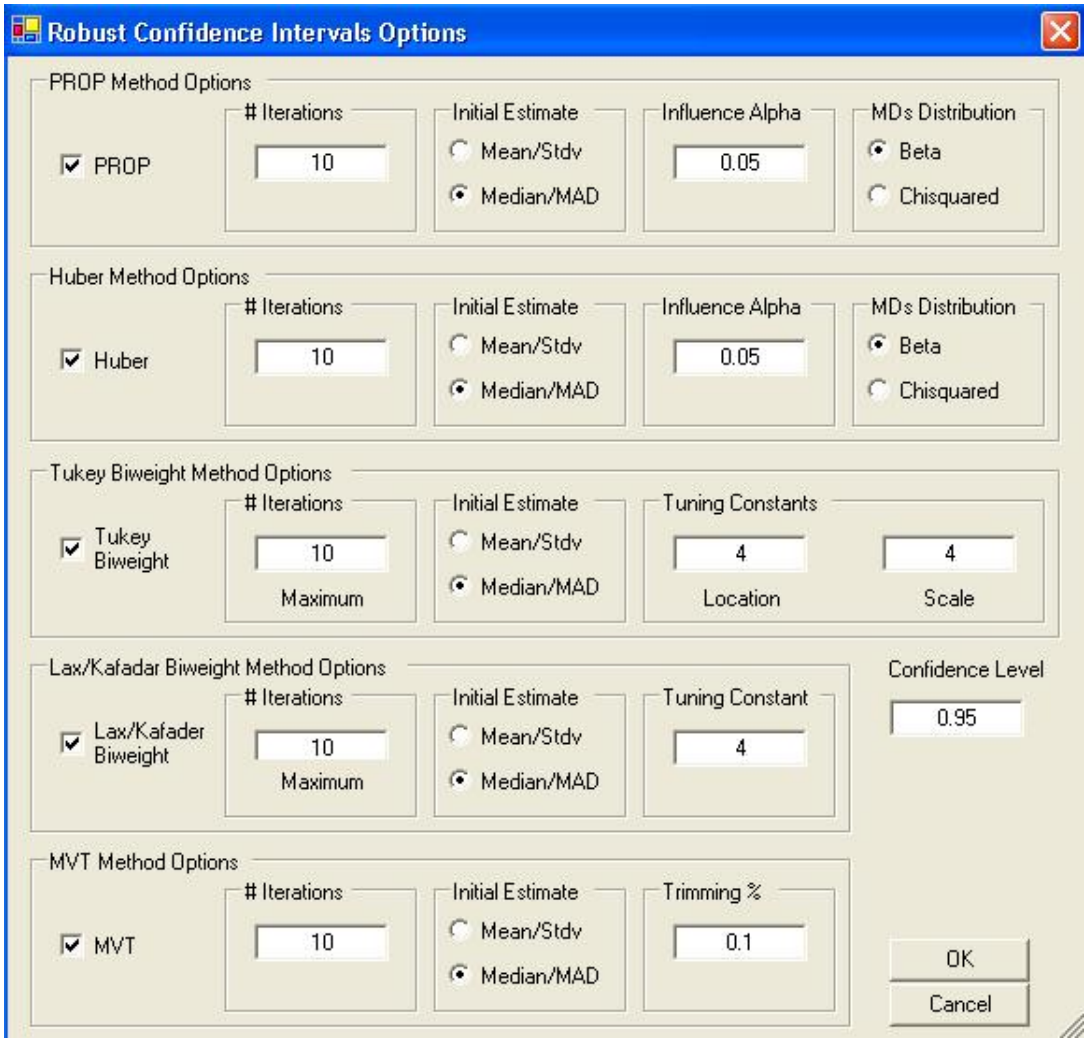


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

   - If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

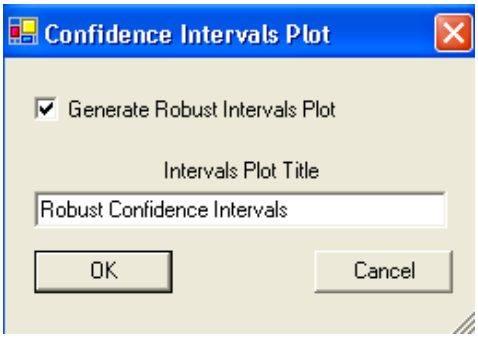
- Click on "**Options**" for interval options.



  - Choose your methods and options. All of the options displayed in the above graphical user interface (GUI) are the default options.

  - Click "**OK**" to continue or "**Cancel**" to cancel selected options.

- Click "**Graphics**" for the graphics option.



- Click "**OK**" to continue or "**Cancel**" to cancel graphics options.

- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Robust Confidence Intervals.

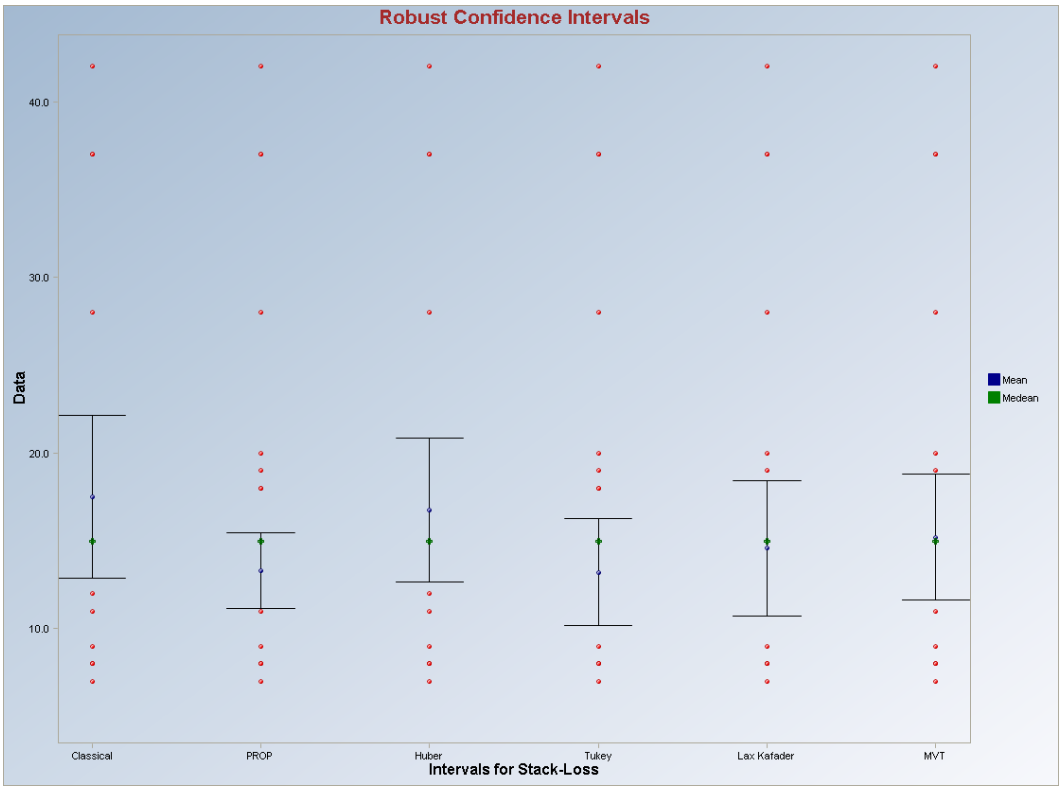
	Robust Confidence Intervals
Date/Time of Computation	1/15/2008 11:48:55 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\STACKLOSS
Full Precision	OFF
Confidence Coefficient	0.95
PROP Method	Influence Function Alpha of 0.05 with MDs following Beta Distribution.
	PROP CLs derived using 10 Iterations and initial estimates of median/MAD.
Huber Method	Influence Function Alpha of 0.05 with MDs following Beta Distribution.
	Huber CLs derived using 10 Iterations and initial estimates of median/MAD.
Tukey Biweight Method	Location Tuning Constant of 4 and a Scale Tuning Constant of 4
	Tukey CLs derived using a Maximum of 10 Iterations and initial estimates of median/MAD.
Lax/Kafader Biweight Method	Tuning Constant of 4
	Lax/Kafader CLs derived using a Maximum of 10 Iterations and initial estimates of median/MAD.
MVT Method	Triming Percentage of 10%
	MVT CLs derived using 10 Iterations and initial estimates of median/MAD.

Stack-Loss

	Number Obs.	Mean	Median	Standard Deviation	MAD/ 0.6745	SE Mean	Critical t	LCL	UCL
Classical	21	17.52	15	10.17	5.93	2.22	2.086	12.89	22.15

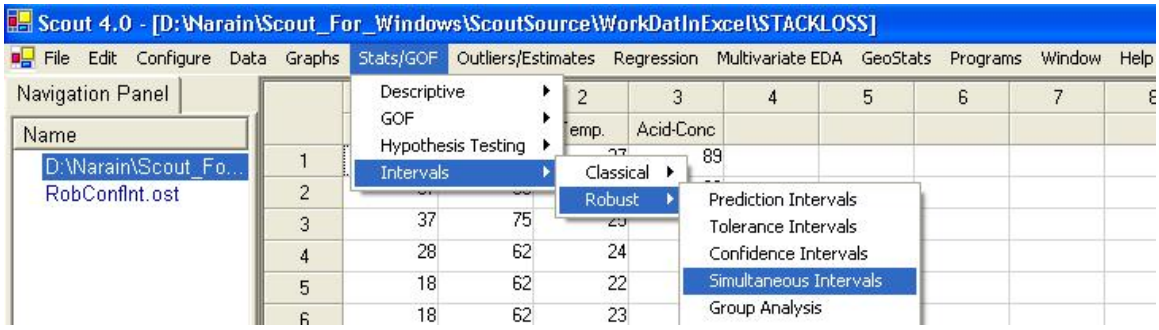
Method	Initial Mean	Initial Stdv	Final Mean	Final Stdv	Wsum	SEM	Critical t	LCL	UCL
PROP	15	5.93	13.3	4.206	17.13	1.016	2.119	11.14	15.45
Huber	15	5.93	16.76	8.79	20.3	1.951	2.091	12.68	20.84
Tukey Biweight	15	5.93	13.21	5.839	16.63	1.432	2.124	10.17	16.25
Lax Kafader Biweight	15	5.93	14.57	7.571	17.42	1.814	2.116	10.74	18.41
MVT	15	5.93	15.21	7.413	19	1.701	2.101	11.64	18.78

Output for Robust Confidence Intervals (continued).



6.5.2 Robust Simultaneous Intervals

1. Click **Stats/GOF ►Intervals ► Robust ► Simultaneous Intervals.**



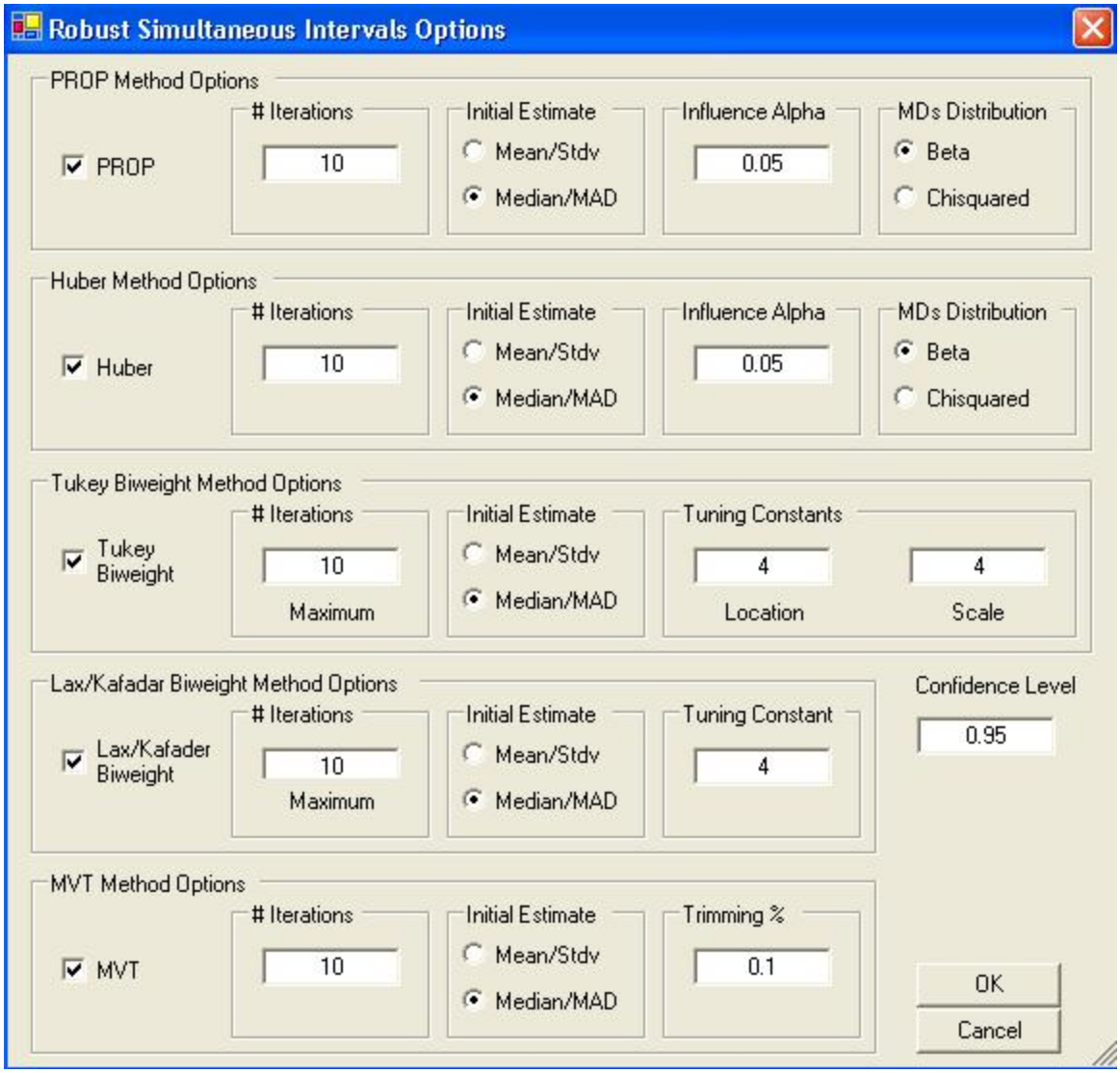
2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

   - If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The

user should select and click on an appropriate variable representing a group variable.

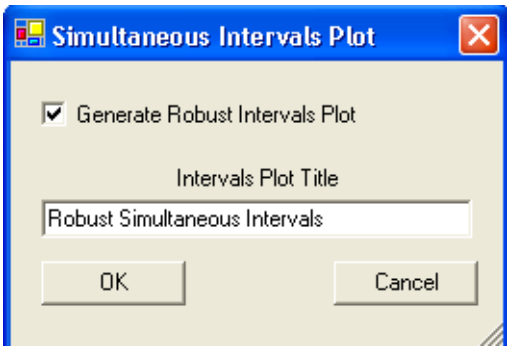
- Click on "**Options**" for interval options.



  - Specify the preferred options. All of the options displayed are defaults.

  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**Graphics**" for the graphics option.



- Click "**OK**" to continue or "**Cancel**" to cancel graphics options.

- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Simultaneous Intervals.

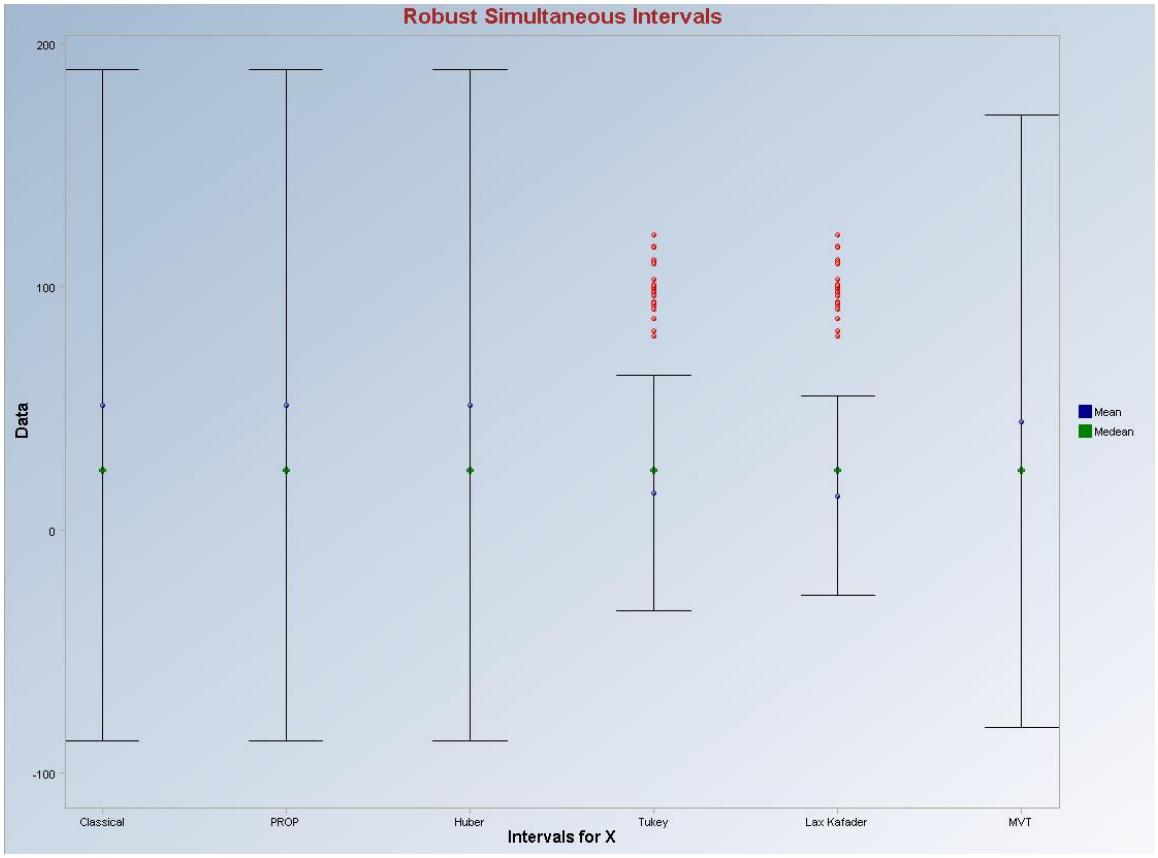
	Robust Simultaneous Intervals/Limits (SLs)
Date/Time of Computation	2/25/2008 9:22:03 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Confidence Coefficient	0.95
PROP Method	Influence Function Alpha of 0.05 with MDs following Beta Distribution.
	PROP SLs derived using 10 Iterations and initial estimates of median/MAD.
Huber Method	Influence Function Alpha of 0.05 with MDs following Beta Distribution.
	Huber SLs derived using 10 Iterations and initial estimates of median/MAD.
Tukey Biweight Method	Location Tuning Constant of 4 and a Scale Tuning Constant of 4
	Tukey SLs derived using a Maximum of 10 Iterations and initial estimates of median/MAD.
Lax/Kafader Biweight Method	Tuning Constant of 4
	Lax/Kafader SLs derived using a Maximum of 10 Iterations and initial estimates of median/MAD.
MVT Method	Triming Percentage of 10.08%
	MVT SLs derived using 10 Iterations and initial estimates of median/MAD.

D2Max represents unsquared critical value of Max-MD (Mahalanobis Distances) computed based upon Wsum Values

	Number Obs.	Mean	Median	Standard Deviation	MAD/ 0.6745	D2Max	LSL	USL
Classical	53	51.1	24.56	43.78	30.48	3.151	-86.88	189.1

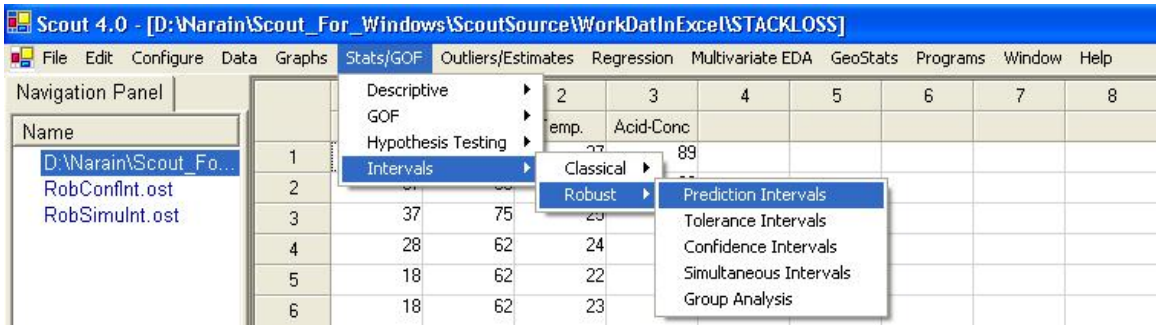
Method	Initial Location	Initial Scale	Final Mean	Final Stdv	Wsum	D2Max	LSL	USL
PROP	24.56	30.48	51.1	43.78	53	3.151	-86.88	189.1
Huber	24.56	30.48	51.1	43.78	53	3.151	-86.88	189.1
Tukey Biweight	24.56	30.48	14.95	15.9	41	3.047	-33.48	63.38
Lax Kafader Biweight	24.56	30.48	14.02	13.09	49.83	3.127	-26.9	54.93
MVT	24.56	30.48	44.44	40.48	48	3.112	-81.52	170.4

Output for Simultaneous Intervals (continued).



6.5.3 Robust Prediction Intervals

1. Click **Stats/GOF ►Intervals ► Robust ► Prediction Intervals.**



2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

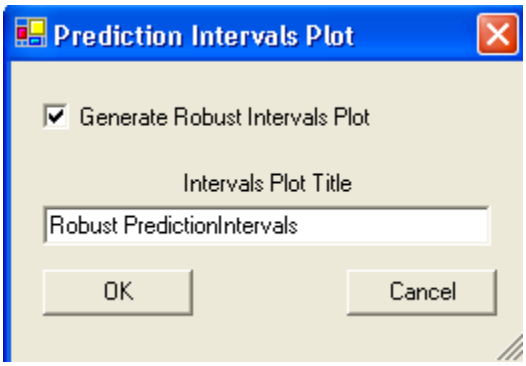
- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- Click on "**Options**" for interval options.

  - Specify the preferred options. All of the options displayed are defaults.

  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**Graphics**" for the graphics option.



- Click "**OK**" to continue or "**Cancel**" to cancel graphics options.

- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Robust Prediction Intervals.

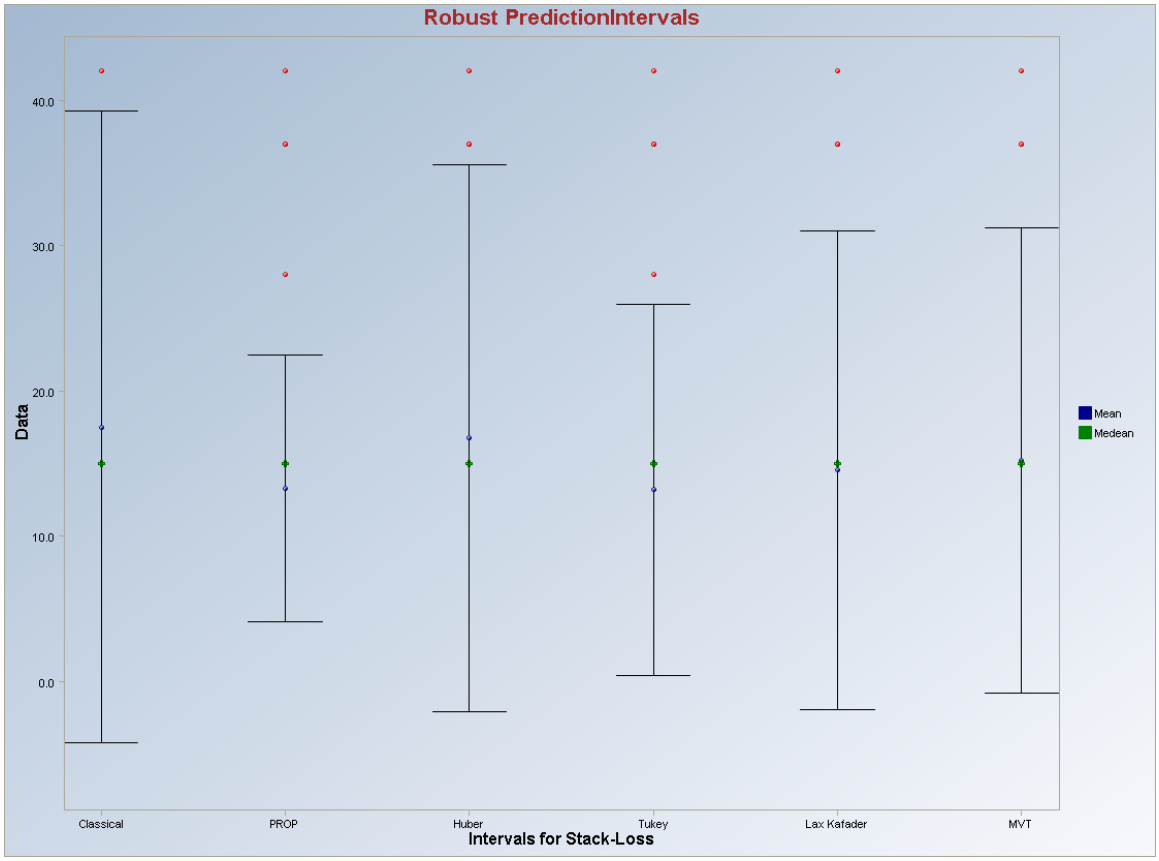
	Robust Prediction Intervals
Date/Time of Computation	1/15/2008 12:13:44 PM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\STACKLOSS
Full Precision	OFF
Confidence Coefficient	0.95
PROP Method	Influence Function Alpha of 0.05 with MDs following Beta Distribution.
	PROP PLs derived using 10 Iterations and initial estimates of median/MAD.
Huber Method	Influence Function Alpha of 0.05 with MDs following Beta Distribution.
	Huber PLs derived using 10 Iterations and initial estimates of median/MAD.
Tukey Biweight Method	Location Tuning Constant of 4 and a Scale Tuning Constant of 4
	Tukey PLs derived using a Maximum of 10 Iterations and initial estimates of median/MAD.
Lax/Kafader Biweight Method	Tuning Constant of 4
	Lax/Kafader PLs derived using a Maximum of 10 Iterations and initial estimates of median/MAD.
MVT Method	Triming Percentage of 10%
	MVT PLs derived using 10 Iterations and initial estimates of median/MAD.

Air-Flow

	Number Obs.	Mean	Median	Standard Deviation	MAD/ 0.6745	SE Mean	Critical t	LPL	UPL
Classical	21	60.43	58	9.168	5.93	2.001	2.086	40.85	80

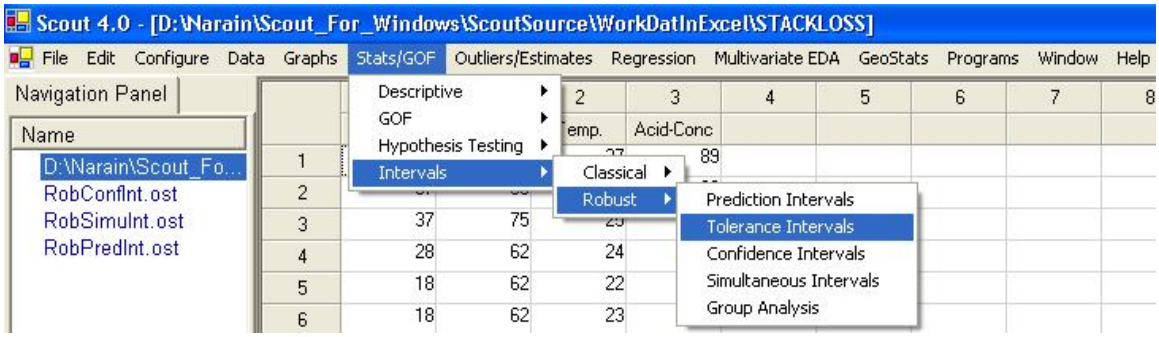
Method	Initial Mean	Initial Stdv	Final Mean	Final Stdv	Wsum	SEM	Critical t	LPL	UPL
PROP	58	5.93	57.18	5.02	17.54	1.199	2.114	46.26	68.09
Huber	58	5.93	60.07	8.546	20.62	1.882	2.089	41.79	78.34
Tukey Biweight	58	5.93	57.48	7.498	17.66	1.784	2.113	41.19	73.76
Lax Kafader Biweight	58	5.93	59.41	4.164	14.84	1.081	2.147	50.18	68.65
MVT	58	5.93	58.37	6.809	19	1.562	2.101	43.69	73.04

Output for Robust Prediction Intervals (continued).



6.5.4 Robust Tolerance Intervals

1. Click **Stats/GOF ►Intervals ► Robust ► Tolerance Intervals.**

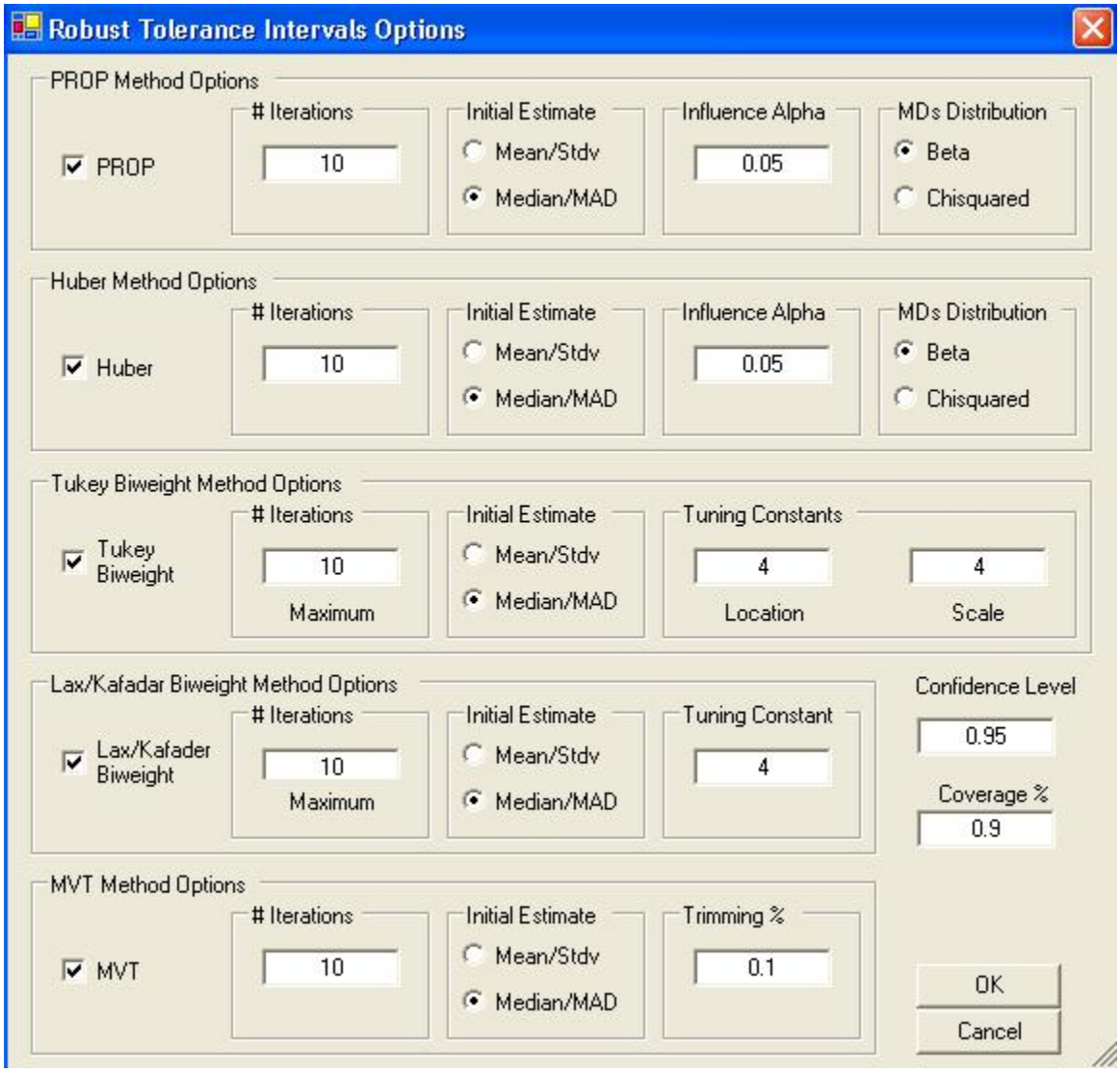


2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

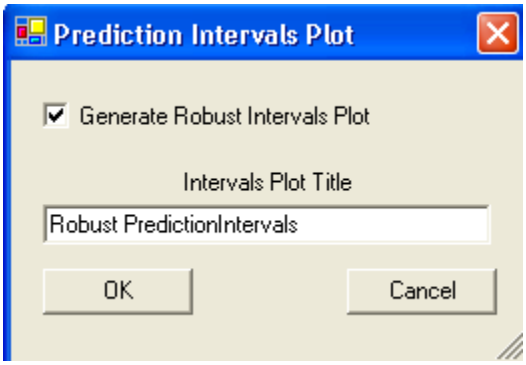
- Click on "**Options**" for interval options.



  - Specify the preferred options. All of the options displayed are defaults.

  - Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**Graphics**" for the graphics option.



    - Click "**OK**" to continue or "**Cancel**" to cancel graphics options.

- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Robust Tolerance Intervals.

	Robust Tolerance Intervals/Limits (TLs)
Date/Time of Computation	2/25/2008 9:23:20 AM
User Selected Options	
From File	D:\Narain\Scout_For_Windows\ScoutSource\WorkDatInExcel\Data\censor-by-grps1
Full Precision	OFF
Confidence Coefficient	0.95
Coverage	0.9
PROP Method	Influence Function Alpha of 0.05 with MDs following Beta Distribution.
	PROP TLs derived using 10 Iterations and initial estimates of median/MAD.
Huber Method	Influence Function Alpha of 0.05 with MDs following Beta Distribution.
	Huber TLs derived using 10 Iterations and initial estimates of median/MAD.
Tukey Biweight Method	Location Tuning Constant of 4 and a Scale Tuning Constant of 4
	Tukey TLs derived using a Maximum of 10 Iterations and initial estimates of median/MAD.
Lax/Kafader Biweight Method	Tuning Constant of 4
	Lax/Kafader TLs derived using a Maximum of 10 Iterations and initial estimates of median/MAD.
MVT Method	Triming Percentage of 10%
	MVT TLs derived using 10 Iterations and initial estimates of median/MAD.

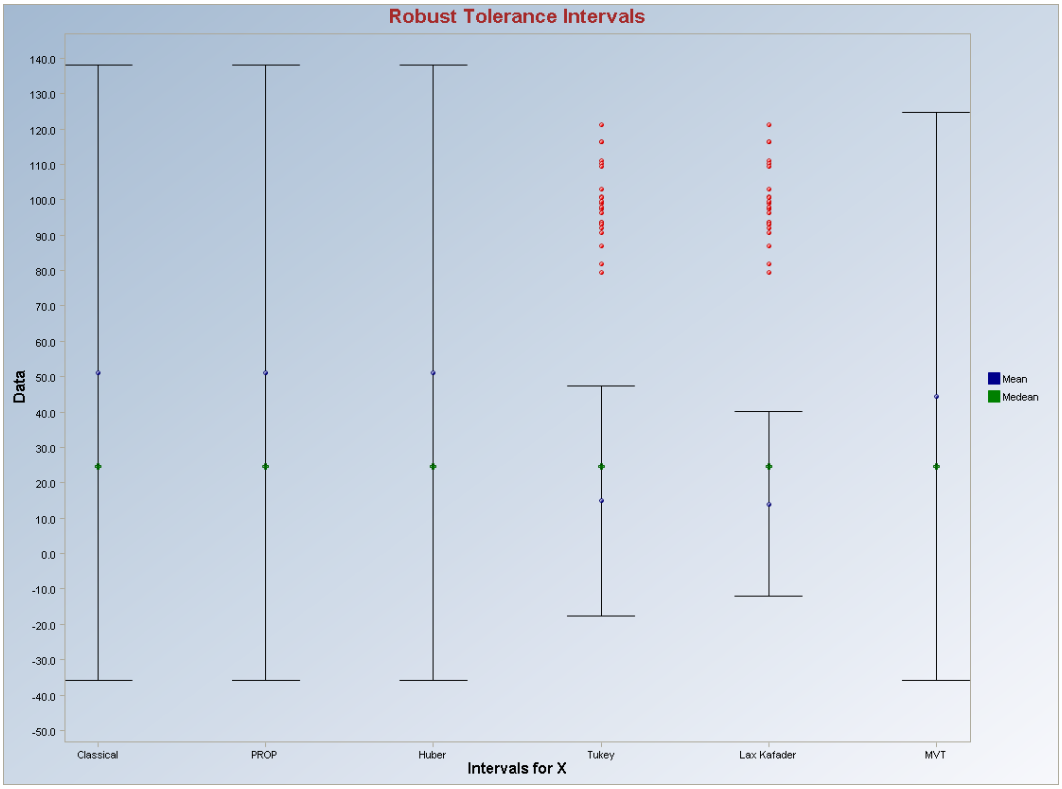
K2 represents the two-sided cutoff for tolerance intervals and is computed based upon Wsum Values following the procedure described in Hahn and Meeker (1991)

X

	Number Obs.	Mean	Median	Standard Deviation	MAD/ 0.6745	k2	LTL	UTL
Classical	53	51.1	24.56	43.78	30.48	1.983	-35.74	137.9

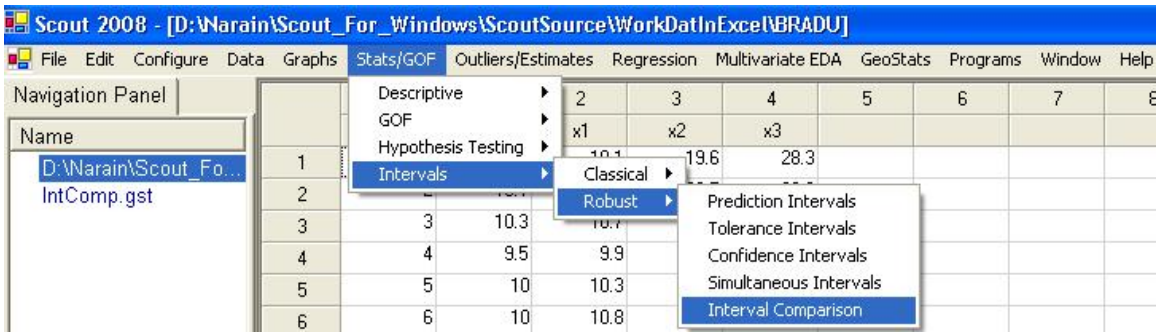
Method	Initial Location	Initial Scale	Final Mean	Final Stdv	Wsum	k2	LTL	UTL
PROP	24.56	30.48	51.1	43.78	53	1.983	-35.74	137.9
Huber	24.56	30.48	51.1	43.78	53	1.983	-35.74	137.9
Tukey Biweight	24.56	30.48	14.95	15.9	41	2.045	-17.56	47.46
Lax Kafader Biweight	24.56	30.48	14.02	13.09	49.83	1.997	-12.12	40.15
MVT	24.56	30.48	44.44	40.48	48	1.983	-35.85	124.7

Output for Robust Tolerance Intervals (continued).



6.5.5 Intervals Comparison

1. Click **Stats/GOF ►Intervals ► Robust ► Intervals Comparison.**

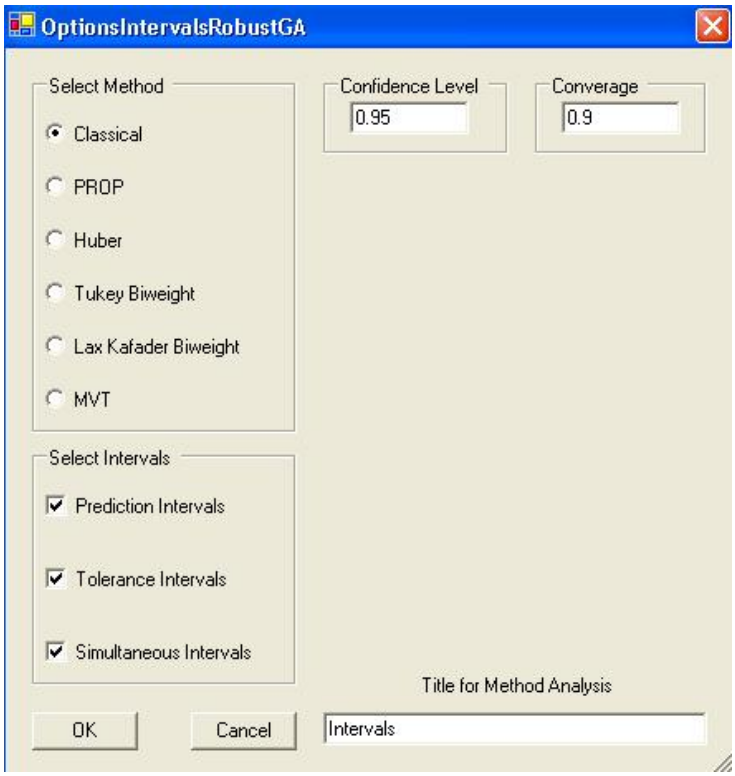


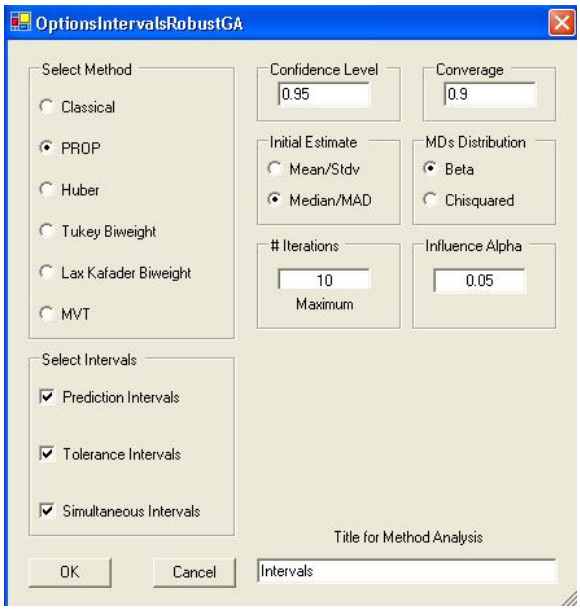
2. The "**Select Variables**" screen (Section 3.2) will appear.

   - Select one or more variables from the "**Select Variables**" screen.

- If the results have to be produced by using a Group variable, then select a group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.

- Click on "**Options**" for interval options. The options screens shown below are the default options screen and the options screen for the PROP method.

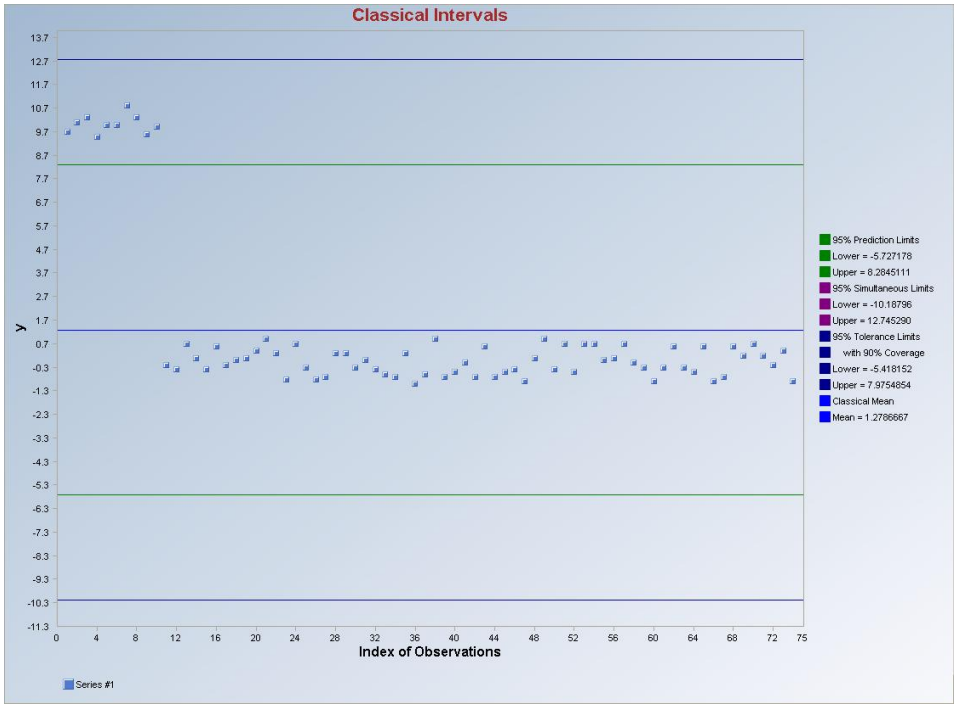




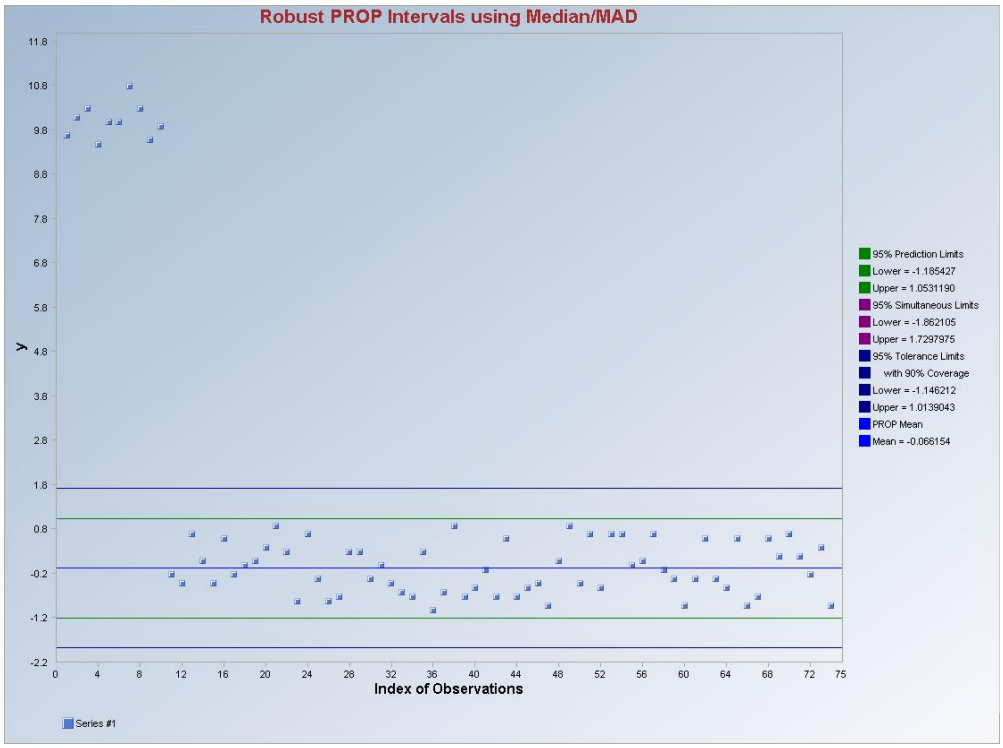
- Specify the preferred options.
- Click "**OK**" to continue or "**Cancel**" to cancel the options.

- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Intervals Comparison (Default Options – Classical on data set BRADU.xls).



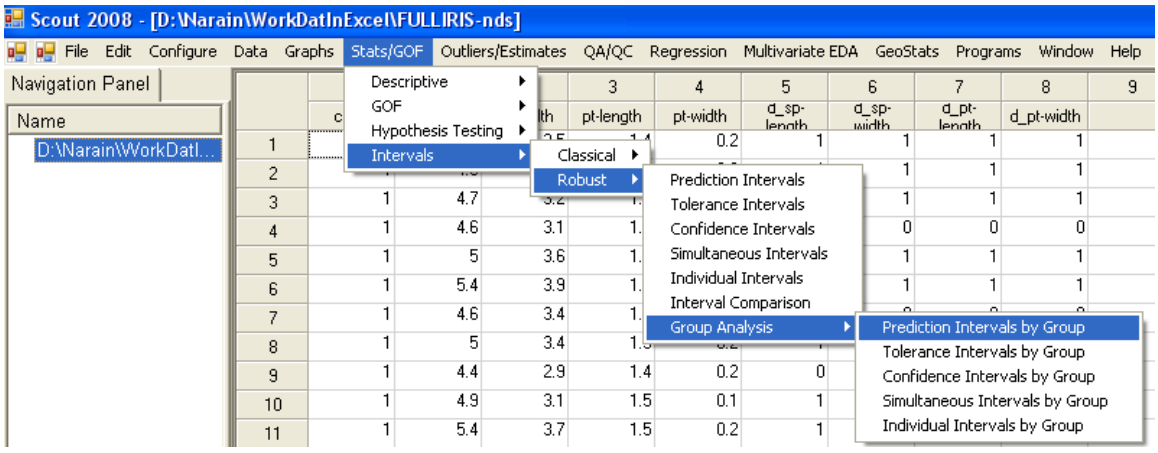
Output for Intervals Comparison (Default Options – PROP on data set BRADU.xls).



6.5.6 Group Analysis

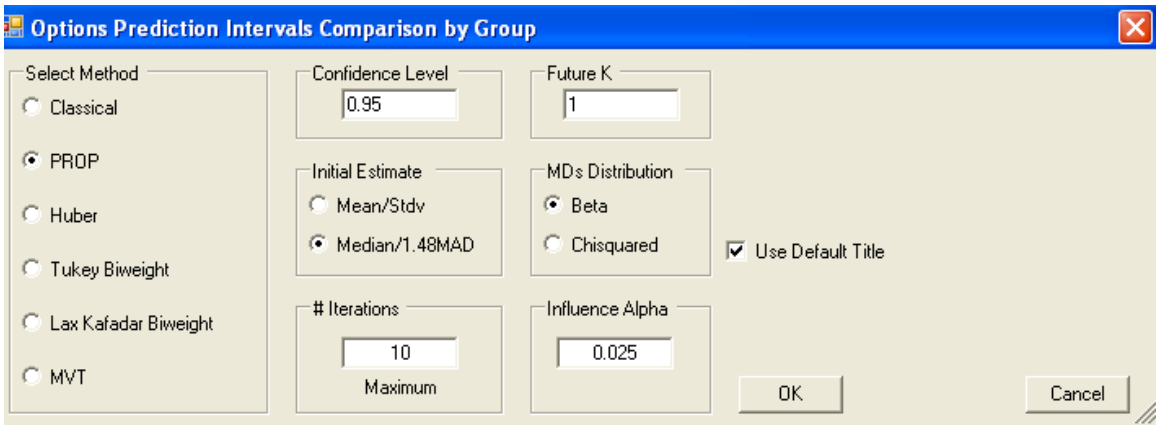
This option in Scout is used for comparing the intervals for each of the groups in a particular variable of the data.

1. Click **Stats/GOF ►Intervals ► Robust ► Intervals Comparison.**



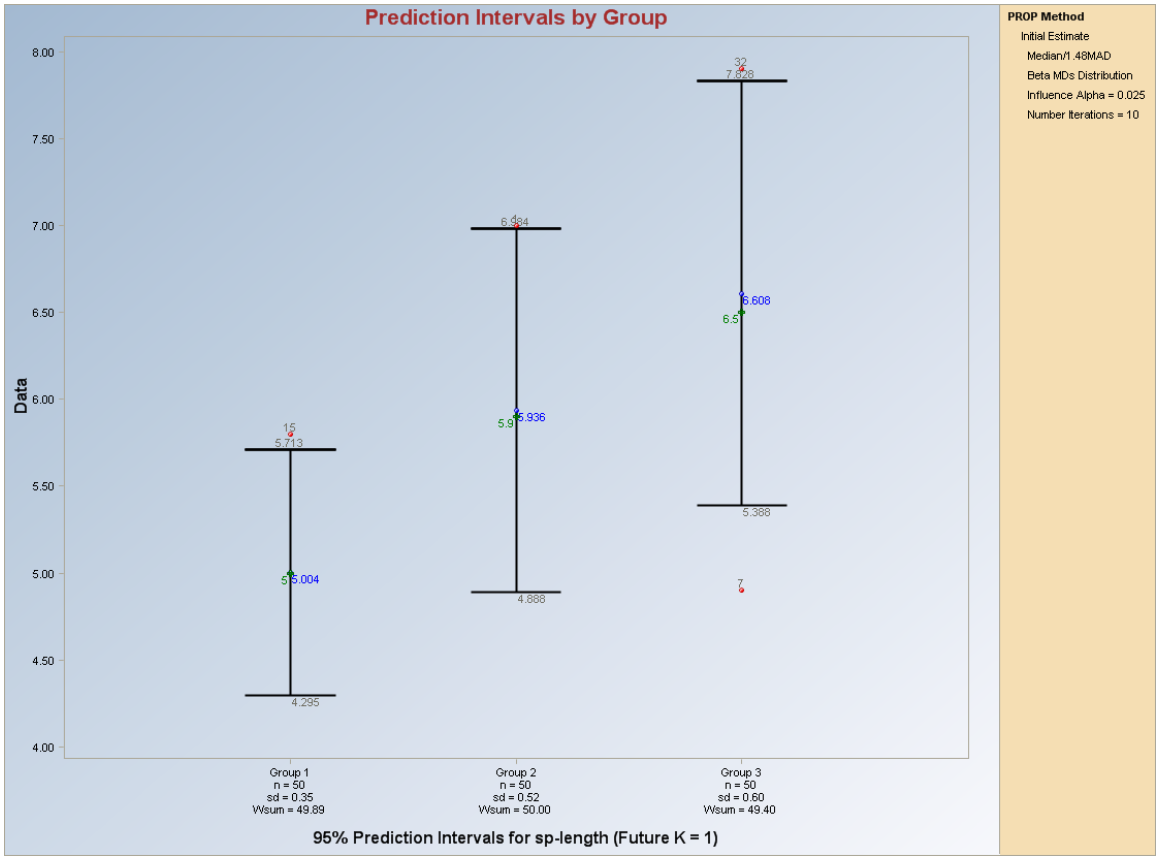
2. The "**Select Variables**" screen (Section 3.2) will appear.

- Select one or more variables from the "**Select Variables**" screen.
- Select the Group variable by clicking the arrow below the "**Group by Variable**" button. This will result in a drop-down list of available variables. The user should select and click on an appropriate variable representing a group variable.
- Click on "**Options**" for interval options. The options screen shown below is the options screen for the PROP method.



  - Specify the preferred input parameters for PROP method.
  - Click "**OK**" to continue or "**Cancel**" to cancel the options.
- Click "**OK**" to continue or "**Cancel**" to cancel the computations.

Output for Group Analysis (PROP Options – FULLIRIS.xls).



References

Dixon, W.J., and Tukey, J.W. (1968). "Approximate Behavior of Winsorized *t* (trimming/Winsorization 2)," Technometrics, 10, 83-98.

Fisher, A. and Horn, P. (1994). "Robust Prediction Intervals in a Regression Setting," Computational Statistics & Data Analysis, 17, pp. 129-140.

Giummol`e, F. and Ventura, L. (2006). "Robust Prediction Limits Based on M-estimators," Statistics and Probability Letters, 76, 1725-1740.

Gross, A.M. (1976). "Confidence Interval Robustness with Long-Tailed Symmetric Distributions," Journal of the American Statistical Association, 71, 409-417.

Horn, P.S., Britton, P.W, and Lewis, D.F. (1988). "On The Prediction of a Single Future Observation from a Possibly Noisy Sample," *The Statistician*, 37, 165-172.

Huber, P.J. (1981). Robust Statistics, John Wiley and Sons, NY.

Kafadar, K. (1982). "A Biweight Approach to the One-Sample Problem," Journal of the American Statistical Association, 77, 416-424.

Mardia, K.V. (1970). "Measures of Multivariate Skewness and Kurtosis with Applications," Biometrika, 57, 519-530.

ProUCL 4.00.04. (2009). "ProUCL Version 4.00.04 User Guide." The software ProUCL 4.00.04 can be downloaded from the web site at: http://www.epa.gov/esd/tsc/software.htm.

ProUCL 4.00.04. (2009). "ProUCL Version 4.00.04 Technical Guide." The software ProUCL 4.00.04 can be downloaded from the web site at: http://www.epa.gov/esd/tsc/software.htm.

Royston, J. P. (1982). "The W test for Normality," Applied Statistics, 31, 2, 176-180.

Scout. 2002. A Data Analysis Program, Technology Support Project, USEPA, NERL-LV, Las Vegas, Nevada.

Scout. 2008. Technical Guide under preparation.

Singh, A., and Nocerino, J.M. 1997. "Robust Intervals in Some Chemometric Applications," Chemometrics and Intelligent Laboratory Systems, 37, pp. 55-69.

Singh, A. and Nocerino, J.M. 2002. "Robust Estimation of the Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations," Chemometrics and Intelligent Laboratory Systems Vol. 60, pp. 69-86.

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.

Tukey, J.W. (1977). Exploratory Data Analysis, Addison-Wesley Publishing Company, Reading, MA.

USEPA. 2006. Data Quality Assessment: Statistical Methods for Practitioners, EPA QA/G-9S. EPA/240/B-06/003. Office of Environmental Information, Washington, D.C. Download from: http://www.epa.gov/quality/qs-docs/g9s-final.pdf.