Welcome to the CLU-IN Internet Seminar

ProUCL Webinar Part I

Sponsored by: USEPA ORD Site Characterization and Monitoring Technical Support Center (SCMTSC)

Delivered: March 9, 2011, 1:00 PM - 4:00 PM, EST (18:00-21:00 GMT)

Instructors:

Anita Singh, PhD, Operations Researcher Senior Staff Scientist, Information System and Global Services - Civil, Lockheed Martin (asingh428@gmail.com)

Narian Armbya (narania@gmail.com)

Bob Maichle, Computer Scientist, Lockheed Martin (maichlebob@gmail.com)

Moderator:

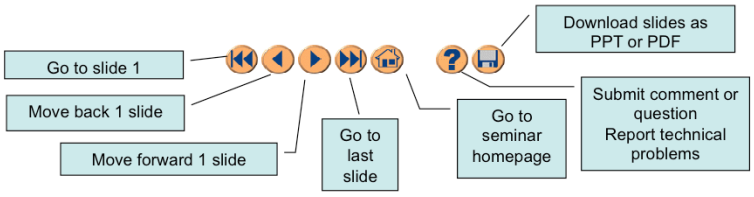
Felicia Barnett, U.S. EPA, ORD Site Characterization and Monitoring Technical Support Center (SCMTSC) (barnett.felicia@epa.gov)

1

Visit the Clean Up Information Network online at www.cluin.org

Housekeeping

- Please mute your phone lines, Do NOT put this call on hold
  - press *6 to mute #6 to unmute your lines at anytime (or applicable instructions)
- Q&A
- Turn off any pop-up blockers
- Move through slides using # links on left or buttons



- This event is being recorded
- Archives accessed for free http://cluin.org/live/archive/

2

Although I'm sure that some of you have these rules memorized from previous CLU-IN events, let's run through them quickly for our new participants.

Please mute your phone lines during the seminar to minimize disruption and background noise. If you do not have a mute button, press *6 to mute #6 to unmute your lines at anytime. Also, please do NOT put this call on hold as this may bring delightful, but unwanted background music over the lines and interupt the seminar.

You should note that throughout the seminar, we will ask for your feedback. You do not need to wait for Q&A breaks to ask questions or provide comments. To submit comments/questions and report technical problems, please use the ? Icon at the top of your screen. You can move forward/backward in the slides by using the single arrow buttons (left moves back 1 slide, right moves advances 1 slide). The double arrowed buttons will take you to 1st and last slides respectively. You may also advance to any slide using the numbered links that appear on the left side of your screen. The button with a house icon will take you back to main seminar page which displays our agenda, speaker information, links to the slides and additional resources. Lastly, the button with a computer disc can be used to download and save today's presentation materials.

With that, please move to slide 3.

ProUCL 4.1.00

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

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

Minimum and Preferred Hardware Requirements

- Hardware:
  - Intel Pentium 1.0 GHz
    - Preferred –something current and more powerful
  - 75 MB of hard drive space
    - Preferred hard drive space– a couple of gigs
  - 512 MB of memory (RAM)
    - Preferred memory – a couple of gigs
  - CD-ROM drive or USB drive
    - Some method to get data off the computer

LOCKHEED MARTIN

Minimum and Preferred Graphics

- Graphics:
  - Minimum graphics display of 800 by 600 pixel
    - Required for display of some graphical user interfaces
    - Main menu bar will wrap around and other less than aesthetic appearances
  - Preferred graphics display
    - 1152 by 864 pixels if old style display
    - 1280 by 768 pixels if wide screen display

LOCKHEED MARTIN

Minimum Software Requirements

- Software
  - Windows 95, 98, XP operating system
    - XP, Vista, or Windows 7
  - .NET Framework 1.1
    - Windows Vista came with .NET Framework 3
    - Windows 7 comes with Net Framework 3.5 or 4.0
    - .NET Framework 1.1 and other (e.g., 2, 3, 3.5, 4.0) .NET Framework can be installed simultaneously on a computer
  - Microsoft Excel
    - Not required but useful especially if editing data files

LOCKHEED MARTIN

Downloading and Installing ProUCL 4.1.00

- ProUCL 4.1.00 can be downloaded from EPA website:

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

- ProUCL Version 4.1.00 zipped file includes ProUCL program files, data files, resource files, and .NET Framework 1.1 setup file
  - Double-clicking .NET Framework 1.1 setup file will install .NET Framework on your computer

- Installation Instructions:
  - Create a new folder called ProUCL 4.1.00, and copy zipped ProUCL Version 4.1.00 in this folder
  - Unzip (extract) ProUCL 4.1.00 in this folder
    - ProUCL cannot be installed and used on a network drive

LOCKHEED MARTIN

Some Common Mistakes

- Trying to run ProUCL without installing .NET Framework 1.1
  - .NET Framework 1.1 setup file is provided in the zipped ProUCL folder. Double-clicking this setup file will install the .NET Framework on your computer
  - A reboot of the system may be required once the installation is complete
- Trying to run ProUCL over the network
  - ProUCL cannot be installed on the server or a network drive

LOCKHEED MARTIN

Some Reoccurring Problems

- Overzealous IT Folks
  - My Help file stopped working!
- Microsoft's Updates
  - Why the program stopped working?
- Missing Data Blues
  - I can get different results with the same data!
- Never thought that this can happen!
  - Standard Deviation of Zero!
  - Balancing functionality and efficiency!

LOCKHEED MARTIN

ProUCL 4.1.00

Input / Output Operations
Data File Creation and Management

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

Focus of ProUCL 4.1 Webinar I

- Focus of Webinar I is to make participants familiar with Statistical and graphical capabilities of ProUCL 4.1

- Emphasis will be placed on showing how to use ProUCL4.1 to:
  - Identify Outliers;
  - Perform Goodness-of-Fit (GOF) tests for Normal, Lognormal, and Gamma distributions;
  - Compute DQOs based Minimum Sample Sizes needed to address project objectives;
  - Compute 95% Upper Confidence Limits (UCL95) to estimate exposure point concentration (EPC) terms;
  - Interpret results generated by ProUCL

- Due to time limitation-statistical details will not be covered

LOCKHEED MARTIN

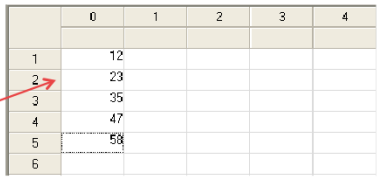
Input /output File Operations

- File operations in ProUCL are similar to Excel 2003 (and older versions)
- It is assumed that participants are familiar with file operations (e.g., create, open, save, copy, cut, paste) available in Excel 2003
- To save time for statistical modules, will go over the following file manipulation slides quickly

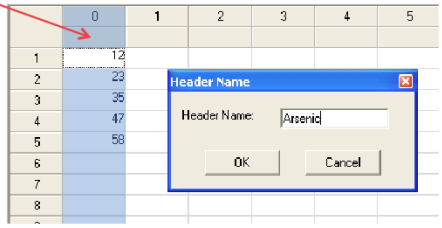
LOCKHEED MARTIN

Creating Data Files

- Left-clicking twice on the ProUCL.exe icon in the ProUCL folder (or ProUCL shortcut on your desktop) will start the ProUCL program with an empty worksheet



- Data can be typed into this work-sheet
- By right-clicking on the numbers on top of the worksheet, column headings, i.e., variable names can be assigned



- Data should be present in the worksheet for drop-down menus to be active

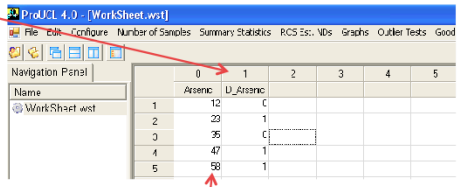
LOCKHEED MARTIN

Creating Data with Nondetects (NDs)

- For each variable with NDs, an indicator column with '0' and '1' is needed. NDs are represented by '0' and detects are represented by '1'

	0	1	2	3	4
	Arsenic	D_Arsenic			
1	12	0			
2	23	1			
3	35	0			
4	47	1			
5	58	1			

- The name of ND column starts with "D_" or "d_" followed by variable name



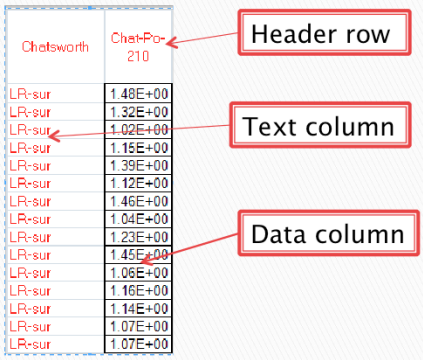
- To access drop-down menus which process ND data, nondetect columns ("D_variable name" or "d_variable name") should be present in data file

Two columns should have same number of values

LOCKHEED MARTIN

Data Entry Requirements

- Data files with text and numerical data can be read → Data File
- Text data: Region, text 9/21/2009.. are used for labels, group ID and sampling events
- Numerical data: used in computations – contains no characters



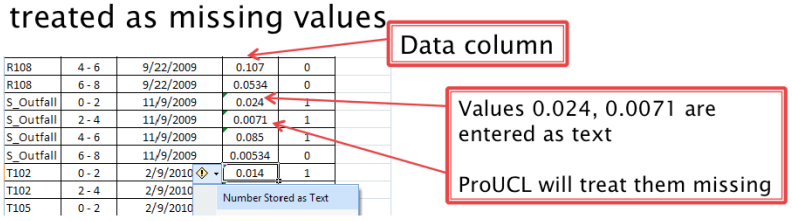
Strings and characters in data column are treated missing values

Data column has numerical values
Text column has characters, strings

15

Missing Values in Data Files

- Entries in data column not entered as numerical values are treated as missing values



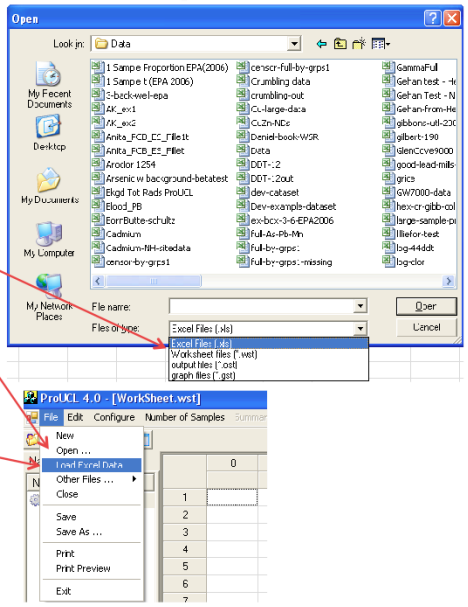
R108	4 - 6	9/22/2009	0.107	0
R108	6 - 8	9/22/2009	0.0534	0
S_Outfall	0 - 2	11/9/2009	0.024	1
S_Outfall	2 - 4	11/9/2009	0.0071	1
S_Outfall	4 - 6	11/9/2009	0.085	1
S_Outfall	6 - 8	11/9/2009	0.00534	0
T102	0 - 2	2/9/2010	0.014	1
T102	2 - 4	2/9/2010		
T105	0 - 2	2/9/2010		

- Large value = 1E31 (= $1x10^{31}$) can be used to represent missing data values
  - Entries with this value are ignored from the computations and counted as missing values

LOCKHEED MARTIN

Accessing Input Data/Output Files

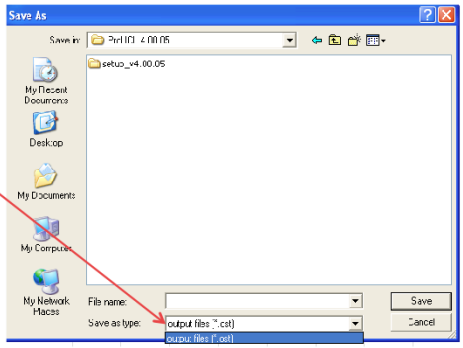
- ProUCL reads following files using "File – Open" option:
  - *.xls – 1995-2003 MS-Excel spreadsheet
  - *.wst – ProUCL worksheet
  - *.ost – ProUCL output sheet
  - *.gst – ProUCL graph sheet
- ProUCL can load an Excel worksheet using "File – Open Excel Data" option

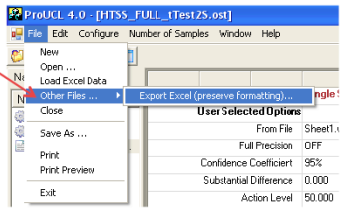


LOCKHEED MARTIN

Saving Files

- ProUCL saves data files (*.wst), output files (*.ost) and graphs (*.gst) using “File – Save as” option
- ProUCL can save data files and output sheets as an Excel file using “File – Other Files – Export Excel (preserve formatting)” option
- The output saved using the “Export Excel” option can be copied from MS-Excel to any other document without losing the format of the output

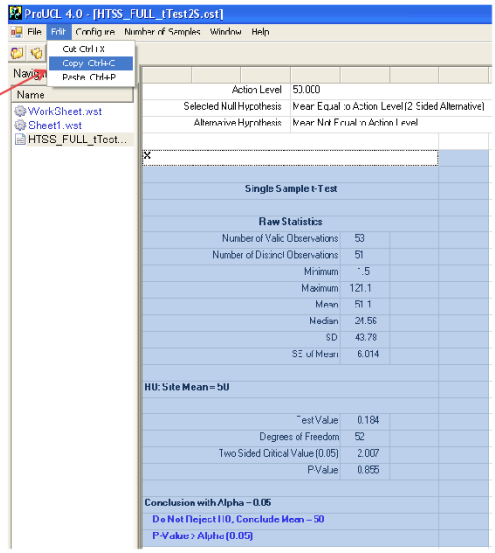




LOCKHEED MARTIN

Editing Data

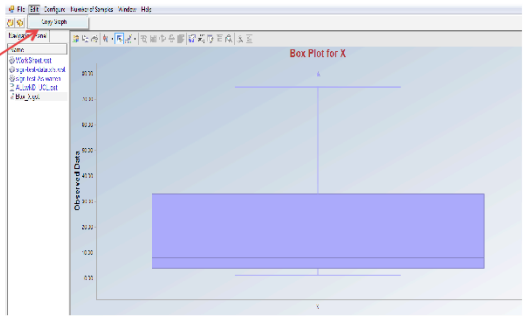
- Data or parts of an output can be cut, copied and/or pasted to a different worksheet or a MS-Word document using the options in the "Edit" drop-down menu
- The figure on right illustrates copying part of a single sample t-Test output
- Note: Copying from an output sheet generated by ProUCL to a document without saving it in Excel may result in the output format being changed



LOCKHEED MARTIN

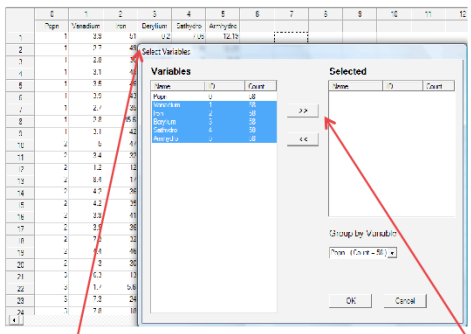
Copying Graphs

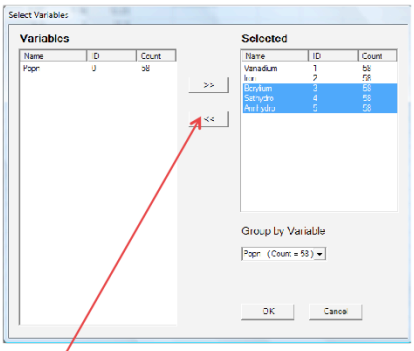
- ProUCL generated graphs can be copied in a MS-Word document or in an image processing software using the "Edit – Copy Graph" option from the drop down menu (shown in figure) or the copy icon as shown in the figure
- Note: Graphs saved using "File – Save as" (.gst) option can be read by ProUCL only



Edit-> Copy Graph Option

Processing Multiple Variables





Highlight variables using Select Variables option

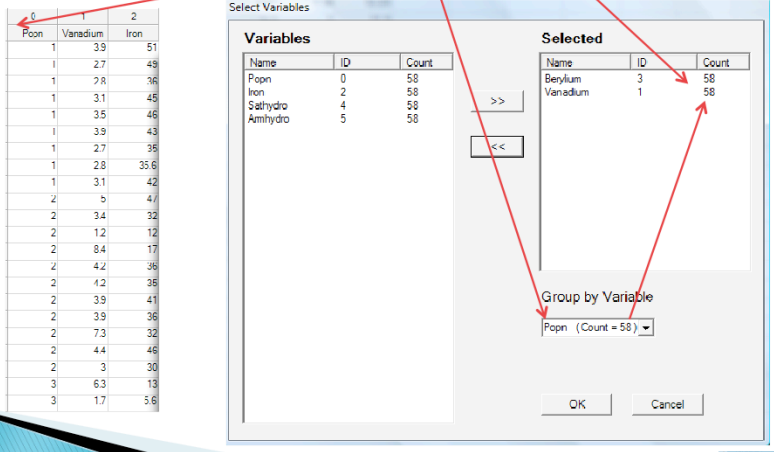
Click on right double arrow to process selected variables

One can deselect variables by clicking on left double arrow

LOCKHEED MARTIN

Processing Data by Groups

- ProUCL can perform statistical analysis for several groups by choosing "Group by Variable" Option
- Data file should have a Group column
- Number of values in group and variable columns should be same



0	1	2
Popn	Vanadium	Iron
1	3.9	51
1	2.7	49
1	2.8	36
1	3.1	45
1	3.5	46
1	3.9	43
1	2.7	35
1	2.8	35.6
1	3.1	42
2	5	47
2	3.4	32
2	1.2	12
2	8.4	17
2	4.2	36
2	4.2	35
2	3.9	41
2	3.9	36
2	7.3	32
2	4.4	46
2	3	30
3	6.3	13
3	1.7	5.6

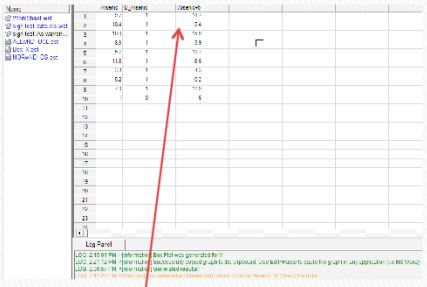
LOCKHEED MARTIN

Error and Warning Messages

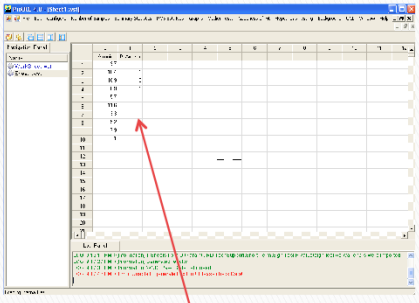
- Suggestions, conclusions and recommendations are displayed in blue on ProUCL outputs
- Error and warning messages are displayed in red
  - Error messages are also displayed in the log panel at the bottom of ProUCL screen window
  - Error messages are fatal and require users to fix data set to be able to use ProUCL without errors
  - Warning messages are displayed to caution users about the reliability of computed statistics and conclusions derived (e.g., due to not enough data)

LOCKHEED MARTIN

Error Messages in Log Panel



Error message when gamma or lognormal distribution used on data with negative values



Error message in box plot when ND column is not of same length as Arsenic column

Warning Messages on the Output Sheet

- Warning messages: generated when statistics cannot be computed or computed statistics may not be reliable
- Some of these warning messages include:
  - Dataset is too small to compute statistics
  - Inadequate amount of detected values in the data
  - Too few distinct values in data set
  - Negative values in data when computing certain statistics (e.g., Gamma statistics)
- Following slides display some warning messages:

LOCKHEED MARTIN

Computed statistics may not be meaningful and reliable

- When a dataset has less than 8 values, computed statistics may not be meaningful and reliable
- A message suggesting the minimum number of observations is also displayed

Warning: A sample size of 'n' = 5 may not adequate enough to compute meaningful and reliable test statistics and estimates!
It is suggested to collect at least 8 to 10 observations using these statistical methods!
If possible compute and collect Data Quality Objectives (DQO) based sample size and analytical results.
Warning: There are only 5 Values in this data
Note: It should be noted that even though bootstrap methods may be performed on this data set,
the resulting calculations may not be reliable enough to draw conclusions
The literature suggests to use bootstrap methods on data sets having more than 10-15 observations.

LOCKHEED MARTIN

Too Few Distinct Values in Data Set

- When data set has only 1 distinct value (datasets with or without nondetects), statistics are not computed

C15

Number of Valid Observations	5
Number of Distinct Observations	1
Minimum	3
Maximum	3

Warning: There is only one distinct observation value in this data set - resulting in '0' variance!
ProUCL (or any other software) should not be used on such a data set!
The data set for variable C15 was not processed!

It is suggested to collect at least 8 to 10 observations using these statistical methods!
If possible, compute and collect Data Quality Objectives (DQOs) based sample size and analytical results.
The Project Team may decide to use alternative site specific values to estimate environmental parameters (e.g., EPC, BTV).

LOCKHEED MARTIN

Negative values in the data

When there are negative values in a data set, UCLs and background statistics based on gamma and lognormal distributions can not be computed

Lognormal UCL Statistics for Full Data Sets	
User Selected Options	
From File	Sheet1.wst
Full Precision	OFF
Confidence Coefficient	95%
Number of Bootstrap Operations	2000

*

Data contains <= 0 Values. Unable to derive log transformed statistics

Gamma Background Statistics for Data Sets with Non-Detects	
User Selected Options	
From File	Sheet1.wst
Full Precision	OFF
Confidence Coefficient	95%
Coverage	90%

*

Dataset Contains Values <= 0 - Cannot Derive Gamma Statistics

LOCKHEED MARTIN

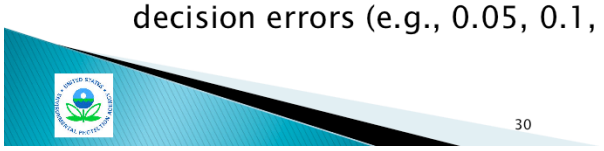
ProUCL 4.1.00

DQOs Based Sample Sizes to Address Project Objectives

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

Uncertainties in Statistics and Decisions Made Based on Those Statistics

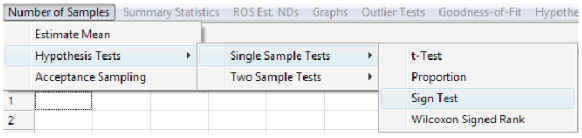
- All statistics: Upper confidence limit of mean (UCL), Upper prediction and Upper tolerance limits (UPL, UTL), T-test, Wilcoxon Rank Sum (WRS) test are computed using sampled data.
- Therefore:
  - Those statistics suffer from uncertainties (e.g., confidence coefficient – 0.90, 0.95), and
  - Conclusions based upon those statistics suffer from decision errors (e.g., 0.05, 0.1, 0.2)



LOCKHEED MARTIN

Data Quality Objectives (DQOs)

- DQOs are designed to manage uncertainties and control decision errors
- DQOs based sampling strategies can be used to collect adequate amount of representative data to address uncertainties and control decision errors
- DQOs based sample size strategies in ProUCL 4.1 are shown below:

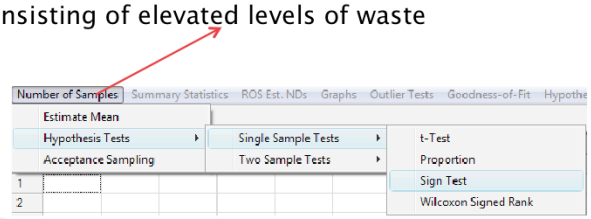


LOCKHEED MARTIN

Sample Size Determination in ProUCL 4.1

- ProUCL can compute DQOs based samples sizes for:

  - Estimation of site mean
  - Verification of the attainment of cleanup standard, C_s
  - Test for proportion of concentrations (e.g., in a MW, site AOC) exceeding an Action Level, A_0
  - Performing site versus background comparisons; upgradient versus downgradient wells comparisons
  - Acceptance sampling to accept or reject a lot: Determine number of drums that should be sampled from a batch of drums with p% of drums consisting of elevated levels of waste



LOCKHEED MARTIN

Data Quality Objectives (DQOs)

- Confidence Coefficient (CC): Specify desired CC and allowable error margin (width of gray region) in estimates of parameters (e.g., mean, proportion)

- Decision Errors: Specify allowable errors in decisions to be made using hypothesis testing approaches

  - Type 1 error= false positive error (e.g., 0.05, 0.1) = level of significance = Probability (Reject null hypothesis when in fact it is true – declare a clean area dirty) =α =false rejection rate

  - Type 2 error =false negative error (e.g., 0.1, 0.15) = Probability (Do not reject null hypothesis when in fact it is false – declare a dirty area clean) = (1–β) = false acceptance rate

LOCKHEED MARTIN

Data Quality Objectives (DQOs)

- Width of gray region is determined by alternative values of the parameters of interest: mean, proportion
  - In gray region consequences of committing errors are likely to be not very significant.
- Example:
  - Null Hypothesis H_0: Mean $\leq$ 100 (Baseline)
  - Alternative Hypothesis H_1: Mean $>$100 (right sided)
  - Width of gray region = 20
  - This implies that mean greater than 120 is considered significantly higher than baseline mean of 100.

LOCKHEED MARTIN

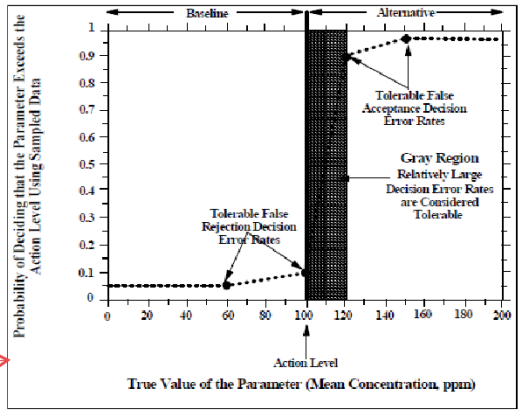
Gray Region for Right-Sided Alternative H_1

Null Hypothesis H_0: Mean ≤ Action Level =100 (Baseline)
Alternative Hypothesis H_1: Mean >100

Width of Gray Region =20

When mean exceeds 120= action level + gray region

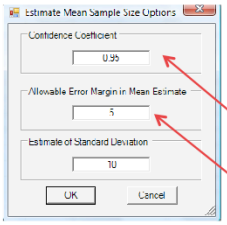
Consequence of false acceptance(of H_0) rate become serious

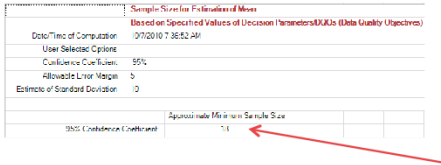


From EPA (2006)

DQOs Based Sample Size to Estimate Mean

- Objective: Estimation of site mean





How many samples to be collected to make sure that error (bias) in mean estimate will be within 5 units with 0.95 confidence coefficient?

DQOs:

Confidence coefficient=0.95

Width of gray region = Error Margin = 5

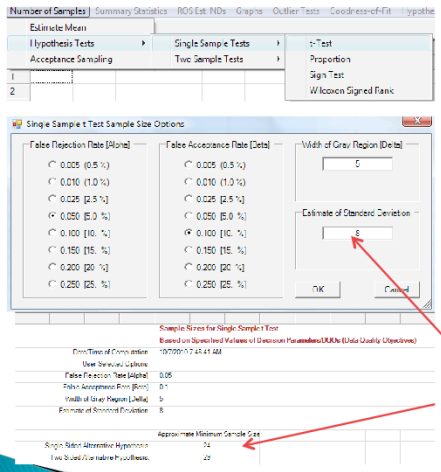
An estimate of variability (standard deviation) =10

Minimum sample size = 18

LOCKHEED MARTIN

DQOs Based Sample Size to Perform t-test

- Objective: Verify attainment of cleanup standard



Cleanup standard = 50 ppm

Null hypothesis, H_0: Mean ≤ 50 (standard met)
Alternative hypothesis, H_1: Mean> 50

DQOs:
False acceptance rate = 0.1
False rejection rate = 0.05 (significance level)

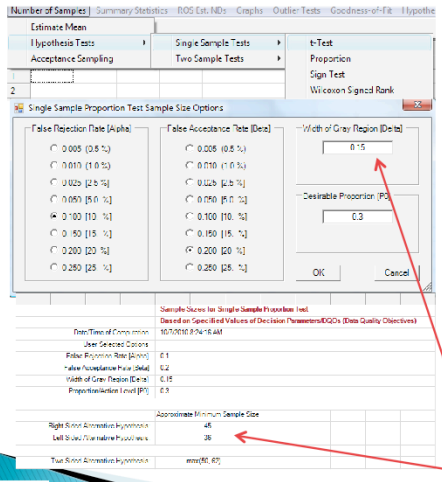
Width of gray region = 5 ppm
(mean > 55 considered significantly > 50)
An estimate of standard deviation = 8

Minimum sample size = 24 (one-sided H_1)

LOCKHEED MARTIN

DQO Based Sample Size for Proportion Test

- Determine if proportion of sites with elevated uranium (exceeding action level) at a uranium mine is < 0.3.



How many sites to sample to demonstrate that proportion of sites with elevated uranium levels is < 30%?

Null hypothesis, H_0: proportion ≤ 0.3
Alternative hypothesis, H_1: proportion > 0.3

DQOs:

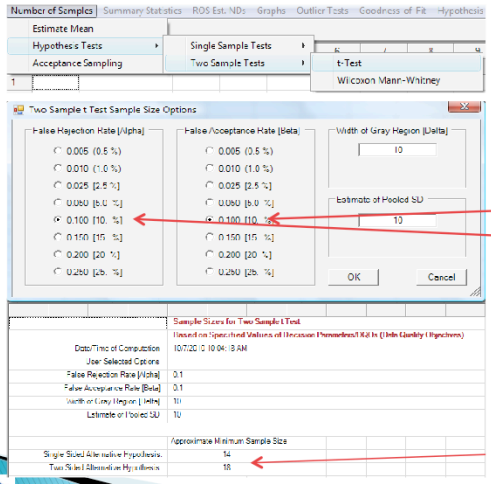
False acceptance rate = 0.2
False rejection rate = 0.1
Width of gray area = 0.15
(proportion > 0.45 considered significantly large)

Minimum # of sites to be sampled = 36

LOCKHEED MARTIN

DQOs Based Sample Sizes for Parametric Two Sample t-test

- Determine if site and background means are comparable



How many samples to collect to compare means of two populations

H_0: site mean = background mean
H_1: site mean ≠background mean

DQOs:
False acceptance rate = 0.1
False rejection rate = 0.1

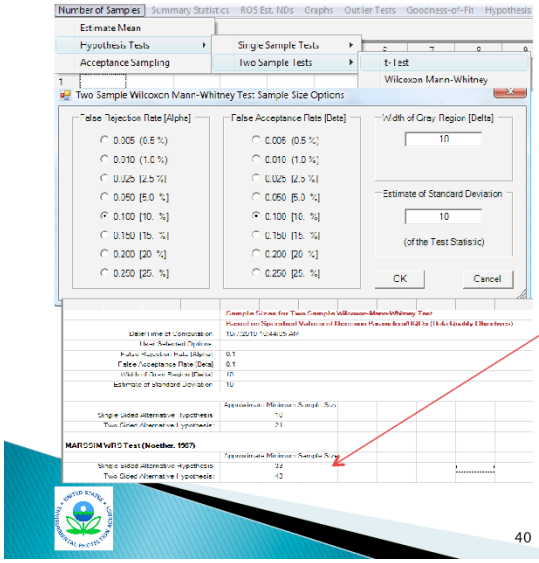
Width of gray area = 10
(Diff. between site and background mean > 10 considered significantly different)

Minimum # samples needed from each population = 18 (two-sided H_1)

LOCKHEED MARTIN

DQOs Based Sample Sizes –Nonparametric Two Sample Wilcoxon– Mann–Whitney Test

- Determine if site and background medians are comparable



H_0: site median ≥ background median
H_1: site median <background median

Specified DQOs:

False acceptance rate = 0.1
False rejection rate = 0.1
Width of gray area = 10

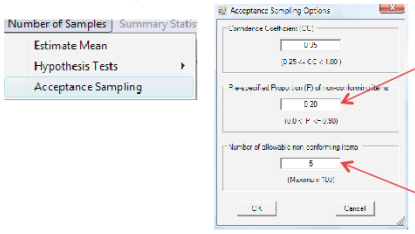
Minimum # samples needed from each population = 33 (one-sided H_1)

More samples needed for nonparametric tests

LOCKHEED MARTIN

DQOs Based Sample Size
Acceptance Sampling for Discrete Items

- Objective: Determine how many drums need to be sampled to reject or accept a batch of drums consisting of p% drums with unacceptable levels of hazardous waste.

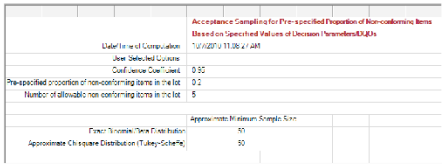


Proportion of non-conforming drums in lot= 20%

DQOs:

Confidence coefficient = 0.95
Allowable # of non-conforming drums in sampled drums = 5

Minimum # drums to be sampled = 50



Allowable # of non-conforming drums in sampled drums = 0

Minimum # drums to be sampled = 14

LOCKHEED MARTIN

DQOs and Sample Sizes

- Data collection also depends upon budget and resource constraints to minimize sampling and analyses costs
- When budget does not allow to collect DQOs based number of samples – minimum of 8–10 samples should be collected from each population under investigation

LOCKHEED MARTIN

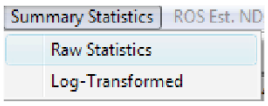
ProUCL 4.1.00

General Statistics and Graphical Capabilities: Boxplots, Q-Q plots, Time-Series Plots

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

General Summary Statistics

- ProUCL computes several general statistics for raw and log-transformed data sets:



- Consider a PCB concentrations data set of size 38: 1.92, 8.66, 4.58, 1.17, 2.48, 1.18, 5.62, 2.54, 25.15, 7.72, 1.02, 2.91, 3.23, 2.87, 2.49,1.71, 5.04, 6.01, 1.46, 1.68, 3.33, 2.89, 6.08, 4.45, 7.88, 22.22, 5.99, 1.16, 4.49, 5.53, 2.77, 2.97, 1.99, 2.06, 3.87, 3.22, 0.1, and 0.05.

LOCKHEED MARTIN

General Summary Statistics – PCB Data of Size 38

In addition to summary statistics, ProUCL computes lower and upper percentiles for raw as well as log-transformed data

Summary Statistics for Raw Full Dataset

Variable	NumObs	Minimum	Maximum	Mean	Median	Variance	SD	MAD/0.675	Skewness	Kurtosis	CV
PCB-Rosner	38	0.05	25.15	4.487	2.94	25.61	5.061	2.268	3.096	10.48	1.128

Percentiles for Raw Full Dataset

Variable	NumObs	5%ile	10%ile	20%ile	25%ile(Q1)	50%ile(Q2)	75%ile(Q3)	80%ile	90%ile	95%ile	99%ile
PCB-Rosner	38	0.882	1.167	1.692	1.938	2.94	5.41	5.842	7.768	10.69	24.07

Summary Statistics for Log-Transformed Full Dataset

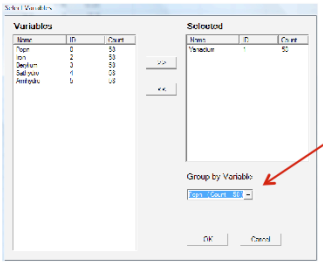
Variable	NumObs	Minimum	Maximum	Mean	Median	Variance	SD	MAD/0.675	Skewness	Kurtosis	CV
PCB-Rosner	38	-2.996	3.225	1.031	1.078	1.306	1.143	0.803	-1.538	4.875	1.108

Percentiles for Log-Transformed Full Dataset

Variable	NumObs	5%ile	10%ile	20%ile	25%ile(Q1)	50%ile(Q2)	75%ile(Q3)	80%ile	90%ile	95%ile	99%ile
PCB-Rosner	38	-0.329	0.154	0.526	0.661	1.078	1.687	1.765	2.05	2.3	3.179

LOCKHEED MARTIN

General Summary Statistics by Groups



Group ID variable should be present
Variable " Popn" represents 3 groups

Statistics for Vanadium computed for 3 groups

Summary Statistics for Raw Full Data Sets

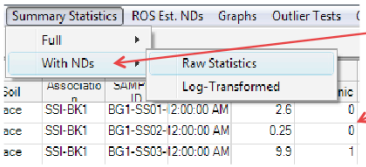
Variable	NumObs	Minimum	Maximum	Mean	Median	Variance	SD	MAD/0.675	Skewness	Kurtosis	CV
Vanadium (1)	9	2.7	3.9	3.167	3.1	0.238	0.487	0.593	0.727	-1.144	0.154
Vanadium (2)	11	1.2	8.4	4.445	4.2	3.853	1.963	1.186	0.753	1.116	0.442
Vanadium (3)	38	1.7	11	7.226	7.5	3.981	1.995	1.927	-0.65	0.39	0.276

Percentiles for Raw Full Data Sets

Variable	NumObs	5%ile	10%ile	20%ile	25%ile(Q1)	50%ile(Q2)	75%ile(Q3)	80%ile	90%ile	95%ile	99%ile
Vanadium (1)	9	2.7	2.7	2.76	2.8	3.1	3.5	3.66	3.9	3.9	3.9
Vanadium (2)	11	2.1	3	3.4	3.65	4.2	4.7	5	7.3	7.85	8.29
Vanadium (3)	38	3.94	4.38	5.84	6.2	7.5	8.4	9	9.5	9.575	10.63

LOCKHEED MARTIN

General Statistics – Data Set with NDs



Data file should have a ND column for "With NDs" options to get activated

Summary Statistics for Raw Data Sets with NDs using Detected Data Only

				Raw Statistics using Detected Observations							
Variable	Num Ds	NumNDs	% NDs	Minimum	Maximum	Mean	Median	SD	MAD/0.675	Skewness	CV
Arsenic (subsurface)	20	20	50.00%	2	18.6	6.335	4.2	5.038	2.298	1.365	0.795
Arsenic (surface)	23	17	42.50%	0.094	12.6	3.048	1.5	3.369	1.156	1.609	1.105

ProUCL can compute these statistics using various other methods such as DL/2, DL, ROS, and KM methods

To be discussed later

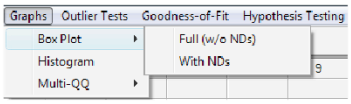
LOCKHEED MARTIN

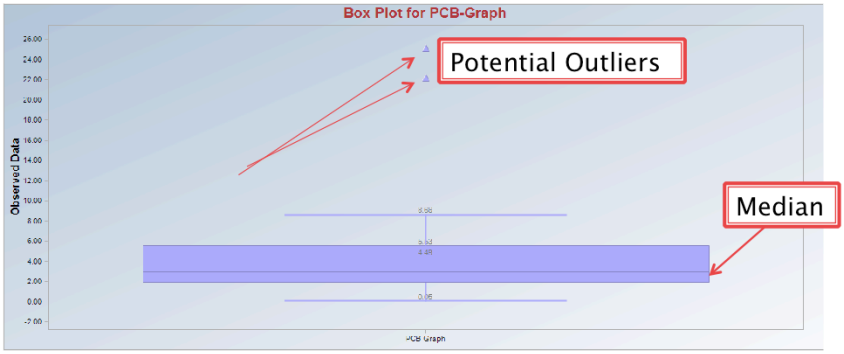
Graphical Displays

- Graphical displays help:
  - Determine data distribution – symmetric, skewed
  - Identify potential outliers
  - Compare data from two or more populations
  - Identify trends in concentrations over time
  - Confirm conclusions derived using test statistics
- No substitute for graphical displays of data

LOCKHEED MARTIN

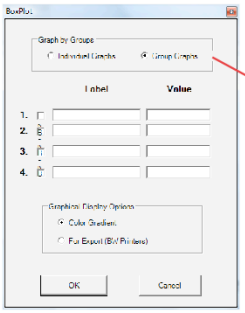
Box Plot of PCB Data Set of Size 38

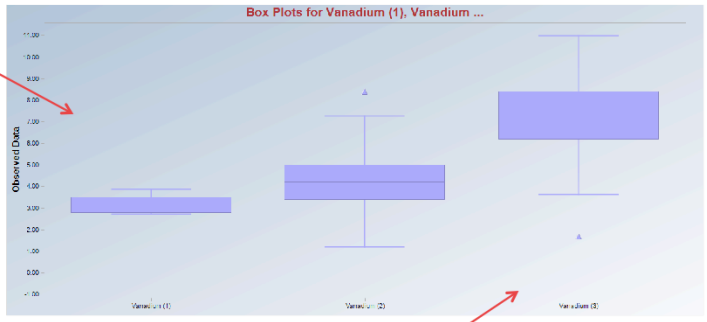




25th percentile =1.94, Median=2.94, Mean=4.49, 75th percentile= 5.41
Box Plot identifies potential outliers, Mean>Median, data positively skewed
Length of upper whisker > length of lower whisker implies data positively skewed

Side-by-Side Box Plots: Comparing Vanadium of Three Groups

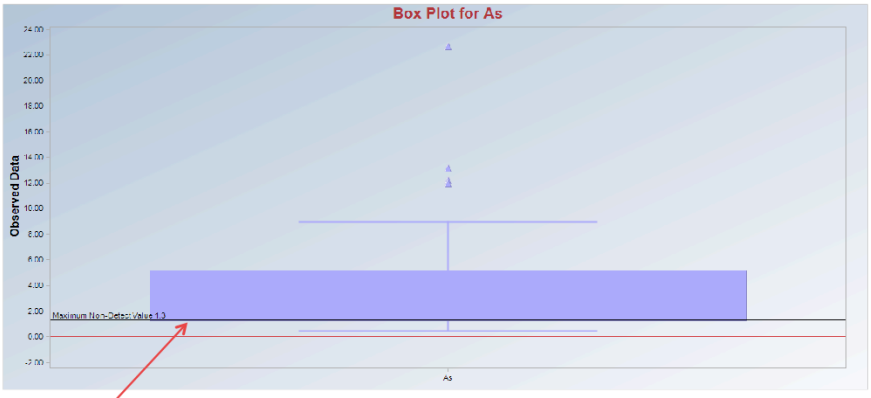




Group 3 exhibits highest level of Vanadium concentrations

LOCKHEED MARTIN

Box Plot of Arsenic (with NDs) From a Large Federal Facility– Depths Combined

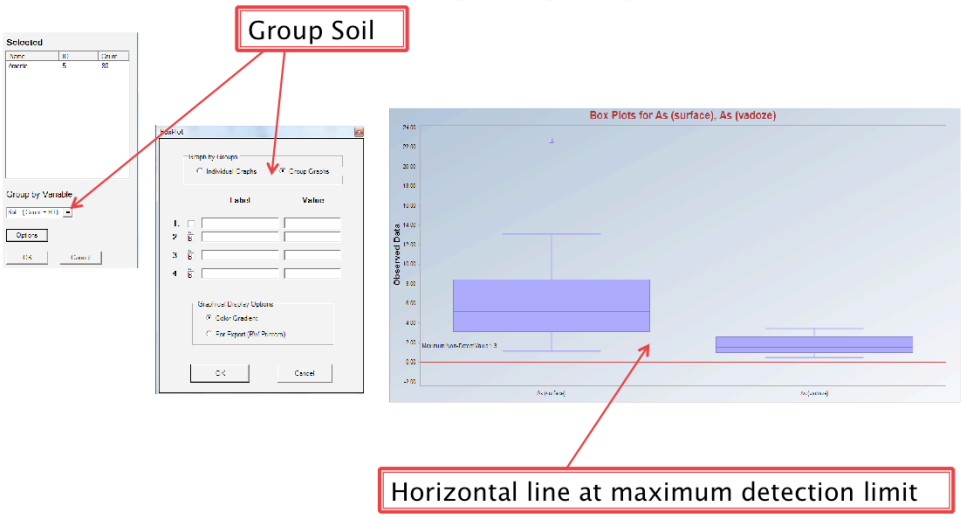


Box Plot with Multiple Detection Limits (DLs)

Max ND

LOCKHEED MARTIN

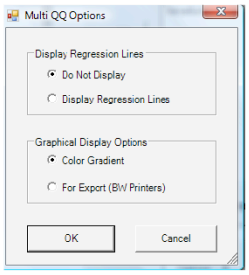
Side-by-Side Box Plots of Arsenic with NDs From a Federal Facility – by depth



LOCKHEED MARTIN

Normal Quantile-Quantile (Q-Q) Plot

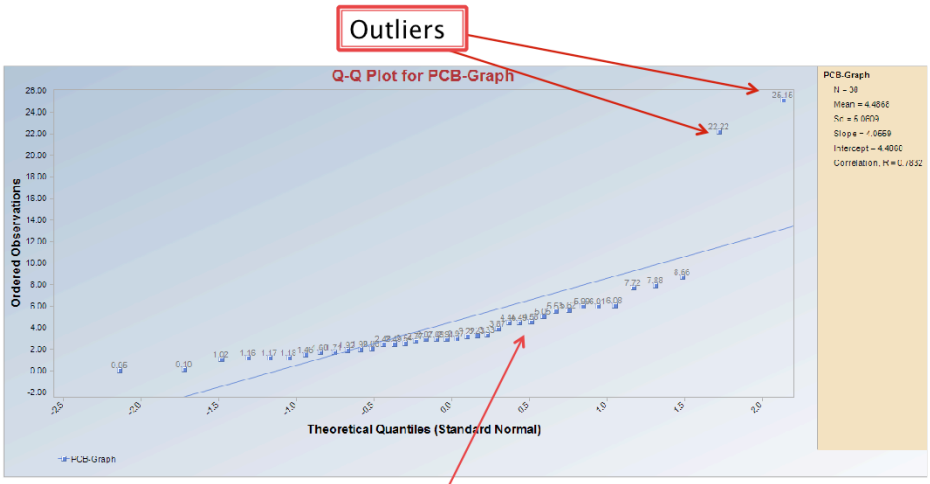
- Normal Q-Q plot represents an informal graphical method to test for approximate normality:
  - Linear pattern displayed by bulk data suggests approximate normality or lognormality (when performed on log-transformed data)
- On Q-Q plot, values well separated from bulk data represent potential outliers
  - Obvious jumps and breaks in Q-Q plot suggest presence of multiple populations/groups



LOCKHEED MARTIN

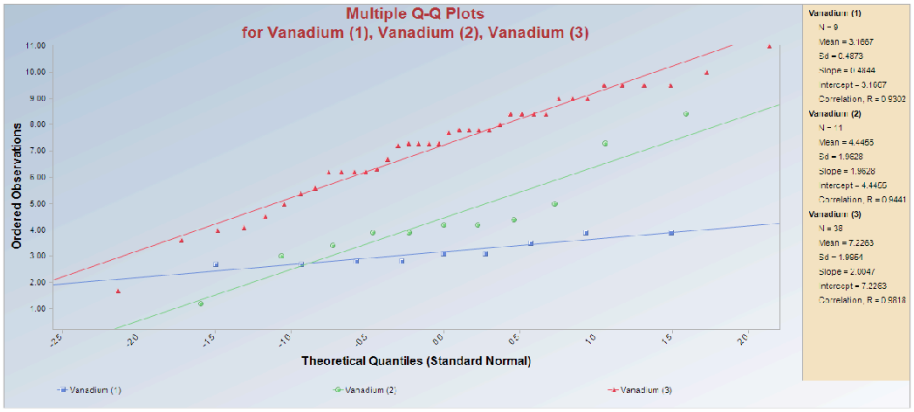
Normal Quantile-Quantile (Q-Q) Plot – PCB Data

- Q-Q plot suggests that data set has two potential outliers



Data not normal as linear pattern not displayed

Q-Q Plots – Comparing Vanadium of Three Groups

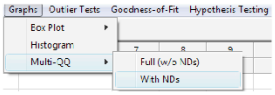


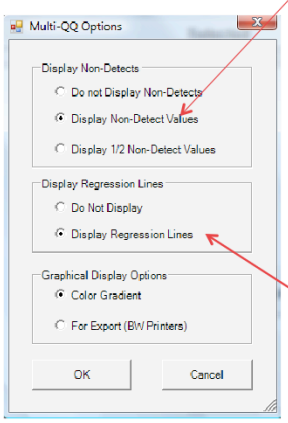
Group 3 exhibits highest Vanadium concentrations
Q-Q plot for Group 3 displays linear pattern
suggesting approximate normality for Group 3 data

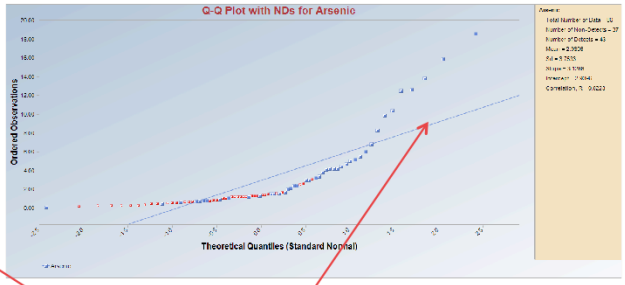
LOCKHEED MARTIN

Normal Q-Q Plot for Arsenic with NDs

- Data set has a ND column
- Nondetects at DLs are in red



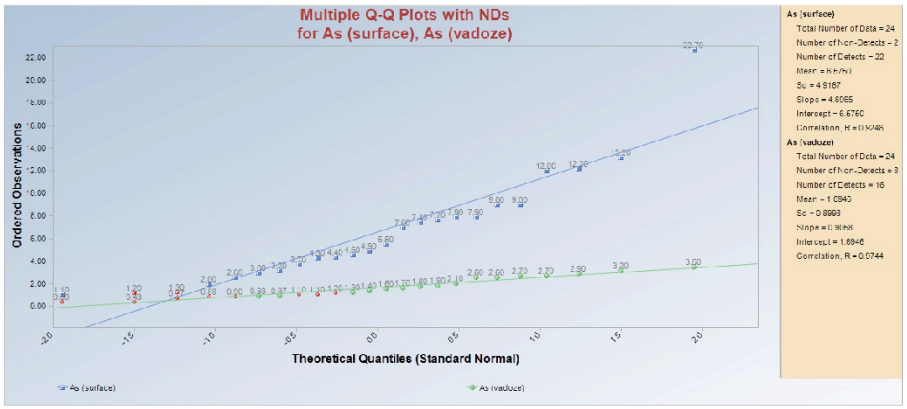




Display Regression Line Option Chosen

LOCKHEED MARTIN

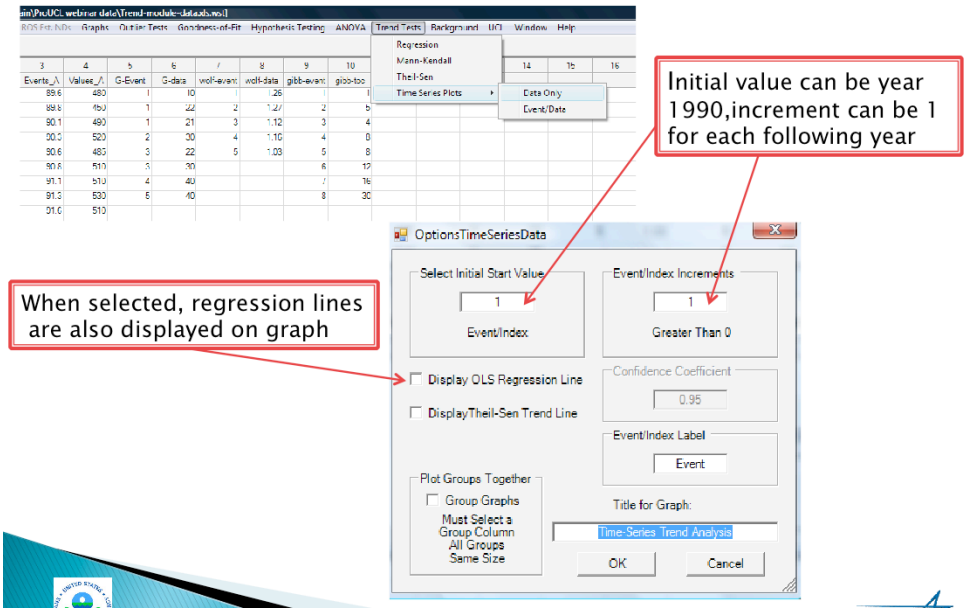
Q-Q Plots – Comparing Arsenic (with NDs) in Soils Surface vs. vadose Zone (Federal Facility)



Potential outlier (=22.7) in surface soils
NDs at detection limits shown in red
Surface soils exhibit higher arsenic than subsurface soils

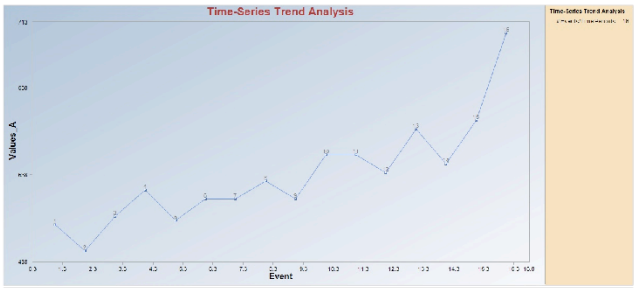
LOCKHEED MARTIN

Time Series Plot with Data Only Option

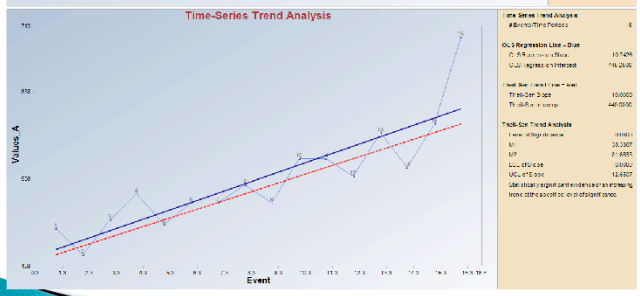


LOCKHEED MARTIN

Time Series Plots with Data Only Option



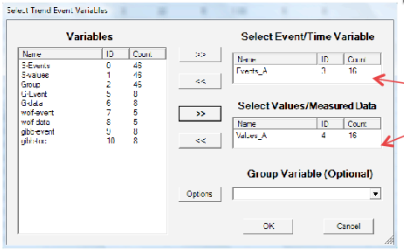
Connected index plot vs. chosen event increments



Plot as a function of chosen event increments with trend lines

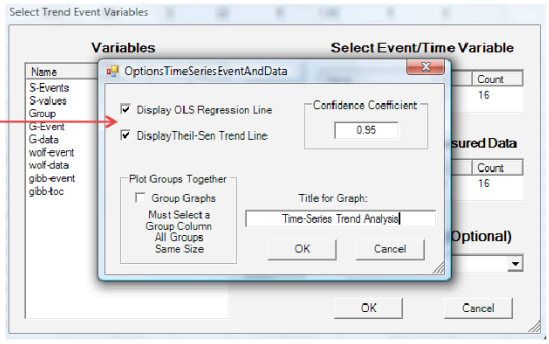
LOCKHEED MARTIN

Time Series Plots for Data vs. Sampling Events



Same no. of values

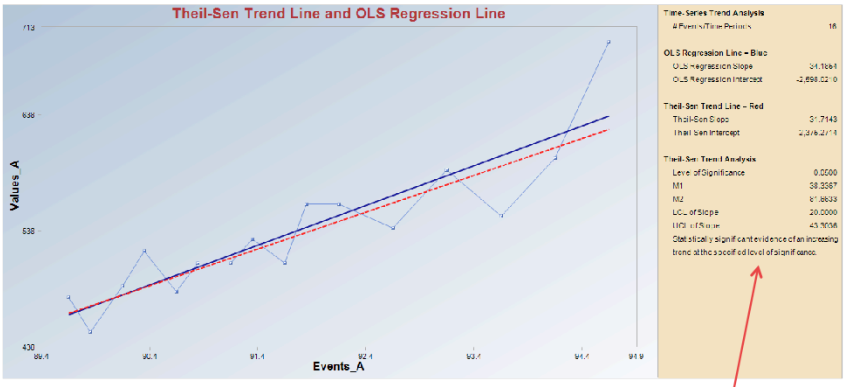
When selected, regression lines are also displayed on time series plot



LOCKHEED MARTIN

Time Series Plot – Identifying Trend in Contaminant Concentration Over Time

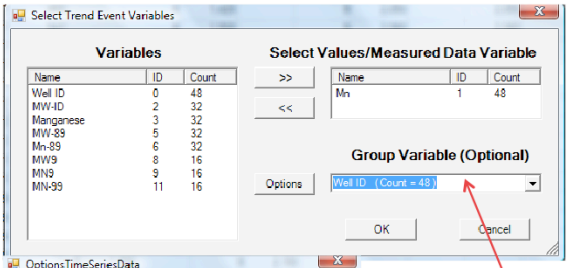
- Time Series plot identifying trend as a function of time

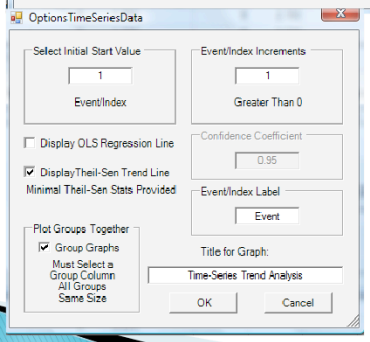


- Graph suggests there is an upward trend – and confirmed by trend test statistics

LOCKHEED MARTIN

Time Series Plots – Comparing Concentrations of Multiple Groups (Wells) versus Sampling Events





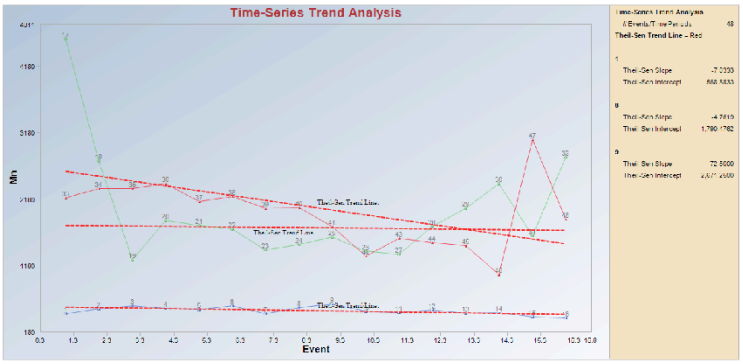
Data file should have a group ID variable

There should be same number of observations/sampling events for each group (MW)

LOCKHEED MARTIN

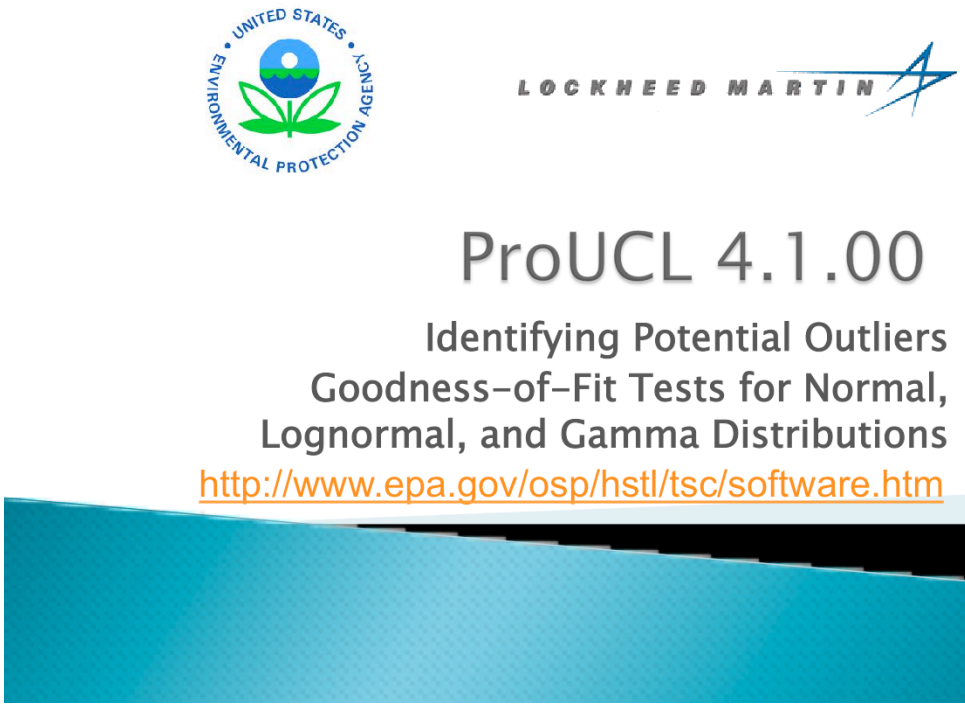
Time Series Plots – Comparing Arsenic in Upgradient and Monitoring Wells

- Groundwater data from 3 MW wells: Well 1 is upgradient well, and wells 8 and 9 are MW wells



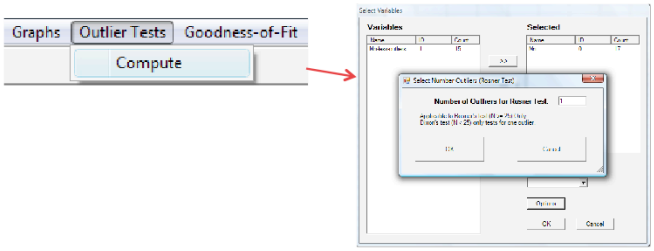
- Graph suggests that As in MW 8 and MW 9 are much higher than upgradient well 1

LOCKHEED MARTIN


UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
LOCKHEED MARTIN
ProUCL 4.1.00
Identifying Potential Outliers
Goodness-of-Fit Tests for Normal,
Lognormal, and Gamma Distributions
http://www.epa.gov/osp/hstl/tsc/software.htm
64

Outlier Tests

- Outliers are values that do not belong to the main population represented by majority of data



- Project Team should determine reasons associated with identified statistical outliers; and

- Project Team should make decisions about disposition (use or not use) of identified outliers

LOCKHEED MARTIN

Identifying Potential Outliers

- Dixon Test for data sets of size ≤ 25
  - Tests for one low and one high outlier at a time
    - Null Hypothesis, H_0: There is no outlier in data set
    - Alternative Hypothesis, H_1: Largest and smallest values are outliers
- Rosner Test for data sets of size ≥ 25
  - Tests for up to 10 outliers
    - Null Hypothesis, H_0: There are no outliers in data
    - Alternative Hypothesis, H_1: There are m (≤ 10) outliers
    - This test identifies outliers as a group
- Outlier tests should be supplemented with graphical displays

LOCKHEED MARTIN

Dixon Test on PCB Data

- PCB levels measured from 16 surface soil samples from a dirt road sprayed with waste oil are: 1.92, 8.66, 4.58, 1.17, 2.48, 1.18, 5.62, 2.54, 25.15, 7.72, 1.02, 2.91, 3.23, 2.87, 2.49, and 1.71.

Dixon's Outlier Test for PCB-out1
Number of data = 16
10% critical value: 0.454
5% critical value: 0.507
1% critical value: 0.595
1. Data Value 25.15 is a Potential Outlier (Upper Tail)?
Test Statistic: 0.727
For 10% significance level, 25.15 is an outlier.
For 5% significance level, 25.15 is an outlier.
For 1% significance level, 25.15 is an outlier.

Test Statistic > Critical Value
0.727> 0.507 (0.05 level)

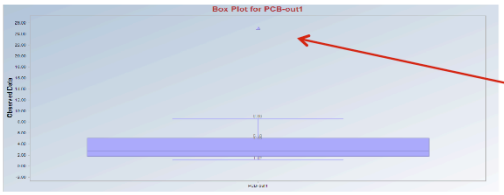
Reject null H_0 of no outlier; and

Conclude 25.15 is an outlier

Note: Extremeness of outliers can be determined only by graphical displays

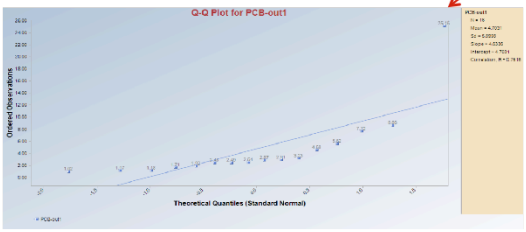
LOCKHEED MARTIN

Graphical Methods– PCB Data with 1 Outlier



Outlier

Figure1. Box Plot – Five Point Summary



These figures suggest that 25.15 is an outlier and is well separated from rest of data

Figure 2. Quantile- Quantile Plot – 25.15 is an outlier

LOCKHEED MARTIN

Dixon Test – PCB Data Set with 3 Outliers

Two additional outliers : 28.89 and 30.23 added to PCB data

Dixon's Outlier Test for PCB-3out
Number of data = 18
10% critical value: 0.424
5% critical value: 0.475
1% critical value: 0.561
1. Data Value 30.23 is a Potential Outlier (Upper Tail)?
Test Statistic: 0.175
For 10% significance level, 30.23 is not an outlier.
For 5% significance level, 30.23 is not an outlier.
For 1% significance level, 30.23 is not an outlier.

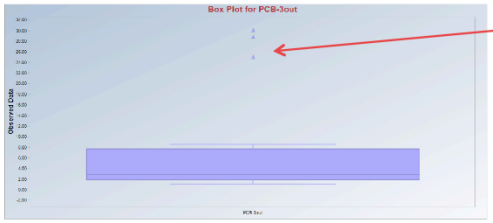
Due to masking – Dixon test could not identify any of 3 outliers

Graphical displays suggest presence of outliers as shown next

Data set is of small size < 25, Rosner test can not be used

LOCKHEED MARTIN

Graphical Methods–Identified 3 Outliers

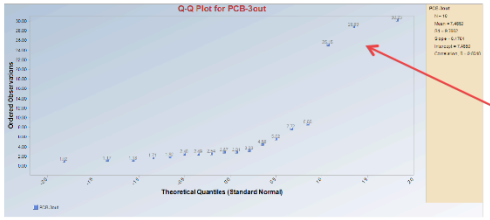


Outliers

Fig.3 Box Plot – PCB data with 3 outliers

Figs. 3,4 suggest 25.15, 28.89, 30.23 are outlying

They are well separated from bulk data



Outliers

Fig. 4 Q-Q Plot- with 3 outliers -25.15, 28.89, 30.23

LOCKHEED MARTIN

Rosner Test –Manganese Data of Size 25

Mn data of size 25: 5 12.1 16.9 21.6 2 5 7.7 53.6 9.5 45.9 5 5.3 12.6 106.3 34.5 6.3 11.9 10 2 77.2 17.9 3.3 8.4 2 22.7

In practice, we do not know no. of outliers in a data set

Need to try Rosner test several times to identify all outliers

Graphical displays help to determine number of outliers

Rosner Test –Manganese Data of Size 25

ProUCL generated outputs for Rosner Test: 2 and 4 outliers

Rosner's Outlier Test for Mn

Mean	**20.19**
Standard Deviation	**25.63**
Number of data	**25**
Number of suspected outliers	**2**

#	Mean	sd	Potential outlier	Obs. Number	Test value	Critical value (5%)	Critical value (1%)
1	20.19	25.11	106.3	14	3.43	2.82	3.14
2	16.6	18.69	77.2	20	3.242	2.8	3.11

For 5% significance level, there are 2 Potential Outliers

Therefore, Potential Statistical Outliers are

106.3, 77.2

Rosner's Outlier Test for Mn

Mean	**20.19**
Standard Deviation	**25.63**
Number of data	**25**
Number of suspected outliers	**4**

#	Mean	sd	Potential outlier	Obs. Number	Test value	Critical value (5%)	Critical value (1%)
1	20.19	25.11	106.3	14	3.43	2.82	3.14
2	16.6	18.69	77.2	20	3.242	2.8	3.11
3	13.97	13.82	53.6	8	2.867	2.78	3.09
4	12.16	11.05	45.9	10	3.054	2.76	3.06

For 5% significance level, there are 4 Potential Outliers

Therefore, Potential Statistical Outliers are

106.3, 77.2, 53.6, 45.9

For 1% Significance Level, there are 2 Potential Outliers

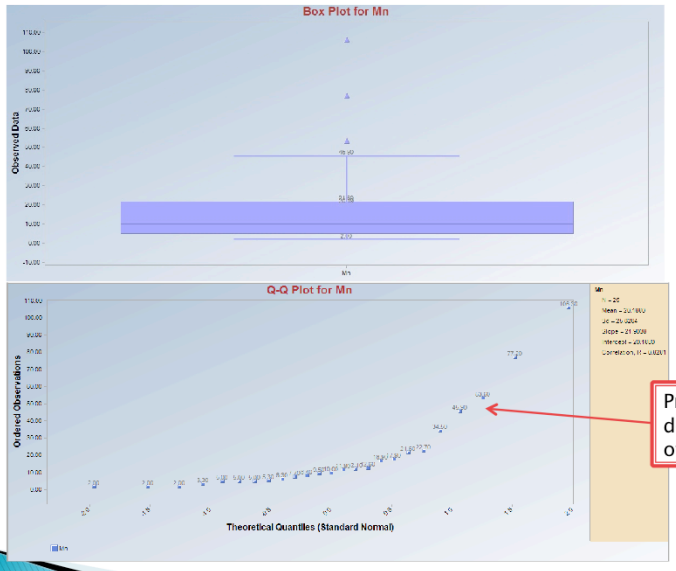
Therefore, Potential Statistical Outliers are

106.3, 77.2

Outliers when
test values>critical values

LOCKHEED MARTIN

Graphical Displays- Manganese Data



Project Team should decide about disposition of these outliers

LOCKHEED MARTIN

Goodness-of-Fit (GOF) Tests

- A GOF test starts with hypotheses statements:
  - Null Hypothesis H_0:data set follows a normal distribution
    vs.
  - Alternative Hypothesis H_1:data set does not follow a normal distribution

  - H_0 :data are gamma distributed
    vs.
  - H_1:data are not gamma distributed

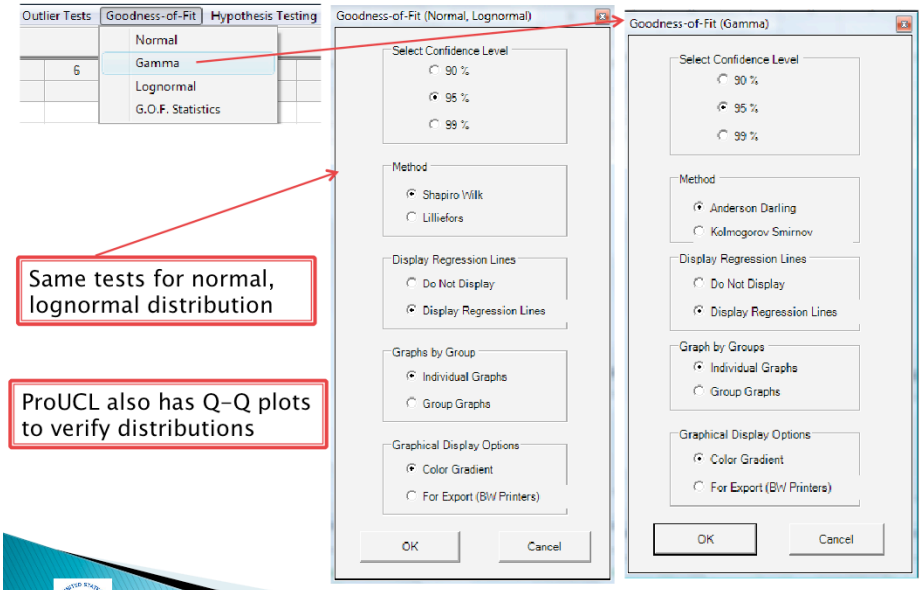
LOCKHEED MARTIN

What is a P-value?

- P- value is associated with a test statistic such as Shapiro-Wilk test statistic
- Smaller the p-value, the more strongly the test statistic (e.g., S-W statistic) rejects null hypothesis
- 1%, 5%, and 10% are common significance levels to which p-values are compared
- A p-value $< .05$ rejects the null hypothesis at " 5% level"

LOCKHEED MARTIN

Goodness-of-Fit (GOF) Tests



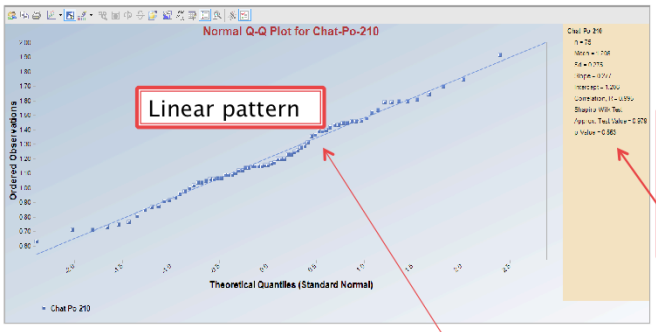
Same tests for normal, lognormal distribution

ProUCL also has Q-Q plots to verify distributions

LOCKHEED MARTIN

Normal Shapiro-Wilk (S-W) GOF Test

- S-W test on Plonium-210 data with 75 values:



H_0: Data normal
H_1: Data not normal

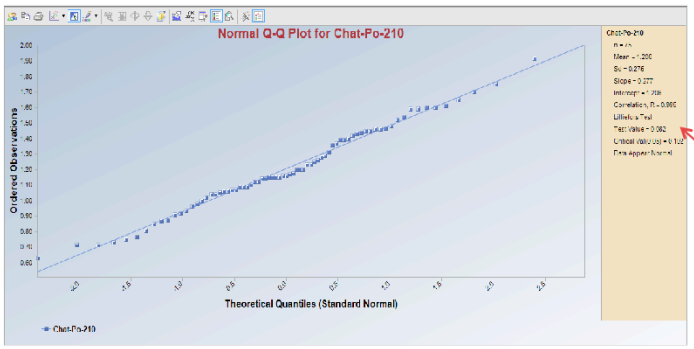
P-value=0.553>0.05
Do not reject H_0

- S-W test and normal Q-Q plot suggest that data are normally distributed with p-value = 0.553 (>0.05, 0.1)

LOCKHEED MARTIN

Normal Lilliefors GOF Test

- Lilliefors GOF test results on Po-210 data



H_0: Data normal
H_1: Data not normal

Critical value 0.102
< Test value 0.082

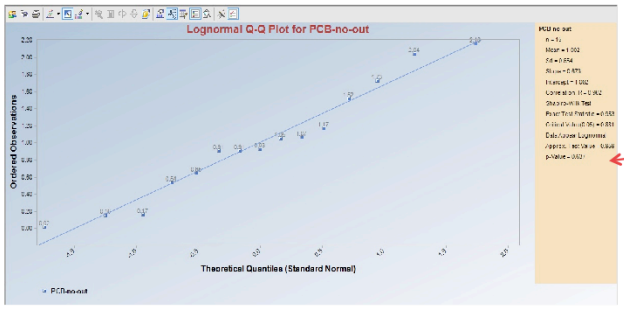
Do not reject H_0

- Based upon Lilliefors test and Q-Q plot, conclude that data follow a normal distribution with critical value of 0.102 < test value 0.082

LOCKHEED MARTIN

Lognormal S-W GOF Test

- PCB concentrations from 15 surface soil samples from a dirt road sprayed with waste oil are: 1.92, 8.66, 4.58, 1.17, 2.48, 1.18, 5.62, 2.54, 7.72, 1.02, 2.91, 3.23, 2.87, 2.49, and 1.71ppm



Null Hypothesis H_0: Data lognormal

p-value=0.637

Do not reject H_0

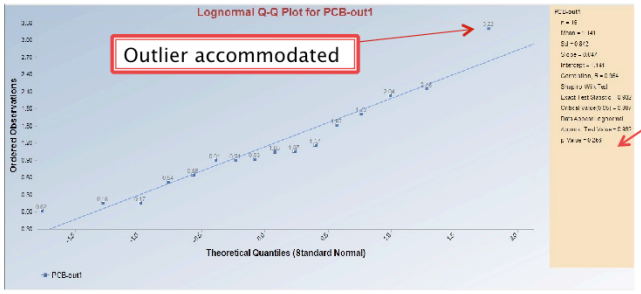
- Based upon S-W test statistic, conclude that data follow a lognormal distribution with p-value=0.637 (>0.05, 0.1)

Data from Drs. J. Warren and B. Nussbaum's 2010 NARPM Workshop

LOCKHEED MARTIN

Lognormal GOF Test on Data with Outlier

- An outlying PCB value=25. 15ppm (log (25.15) = 3.22) is added
- Lognormal GOF test results on data with outlier:



H_0: Data lognormal
H_1: Data not lognormal

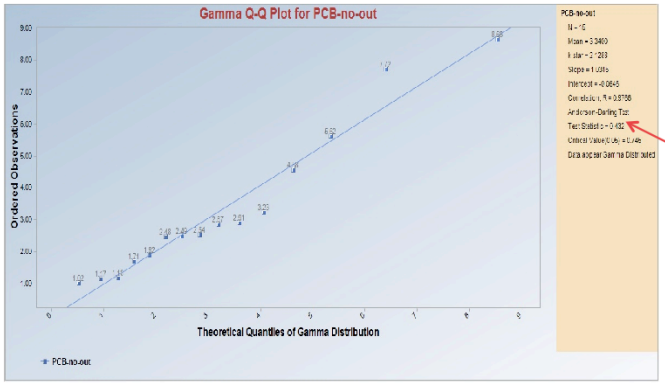
p-value=0.258>0.05
Do not reject H_0

- Using S-W test, conclude data are lognormally distributed
- Lognormal distribution accommodated outlier 25.15

LOCKHEED MARTIN

Gamma Anderson-Darling (A-D) GOF Test

- Based upon A-D GOF test, PCB data set of size 15 appears to follow a gamma distribution



H_0: Data are gamma distributed

H_1: Data are not gamma distributed

A-D Test Stat=0.432
5% Critical value=0.745

Test Stat< Critical Value
0.432<0.745

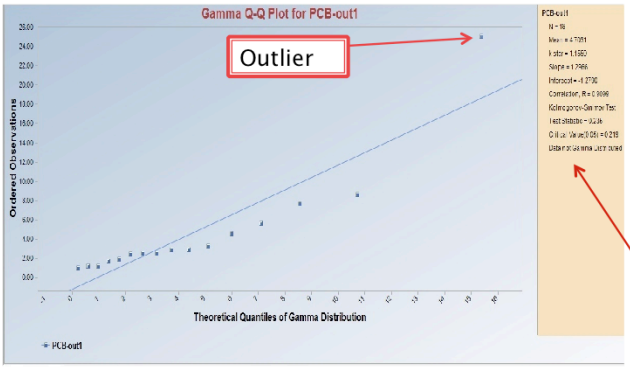
Do not reject H_0 at 5% level
Data follow gamma

Data from Drs. J. Warren and B. Nussbaum's 2010 NARPM Workshop

LOCKHEED MARTIN

Gamma GOF Test on Data with Outlier

- Kolmogorov-Smirnov (K-S) GOF test results on data with outlier = 25.15 are shown below:



H_0: Data are gamma distributed

H_1: Data are not gamma distributed

K-S Test Stat=0.235
5% Critical value=0.219

Test Stat> Critical Value
0.235>0.219

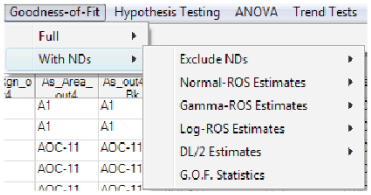
Reject H_0 at 5% level
Data not gamma distributed

- Gamma model did not accommodate outlier, 25.15

LOCKHEED MARTIN

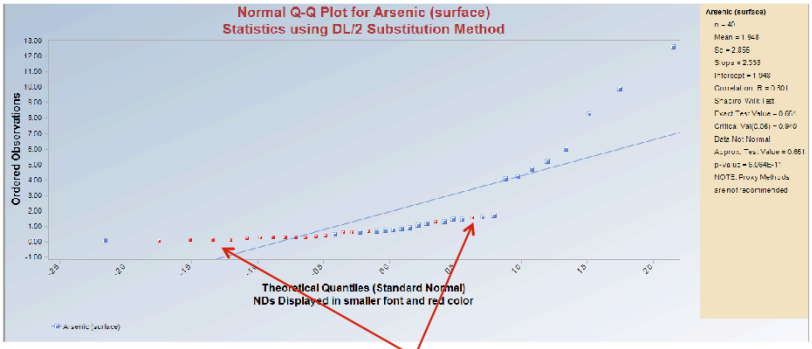
GOF Tests on Data with Nondetects (NDs)

- For data sets with NDs, GOF tests may be used on data sets obtained:
  - Replacing NDs by ½ (detection limit)= ½ (DL/2);
  - Using Regression on Order Statistics (ROS) methods: normal ROS (NROS), lognormal ROS (LROS), gamma ROS (GROS);
  - Excluding nondetect observations



GOF Tests on Data with Nondetects (NDs)

- Consider arsenic data with NDs collected from surface soils at a Federal facility
- Use DL/2 Substitution method
- P-value ~ 0.0; Reject null hypothesis and conclude data are not normally distributed

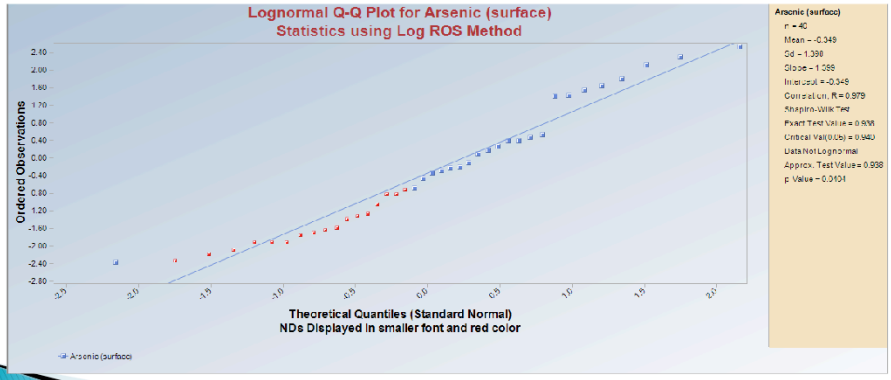


NDs shown in red

LOCKHEED MARTIN

GOF Tests on Data with Nondetects (NDs)

- Use LROS method (ROS method on logged data)
- For S-W test, p-value ~ 0.04, reject null hypothesis at 0.05 and conclude data are not lognormally distributed
- Since p-value = 0.04> 0.01, null hypothesis is not rejected at 0.01 level of significance



LOCKHEED MARTIN

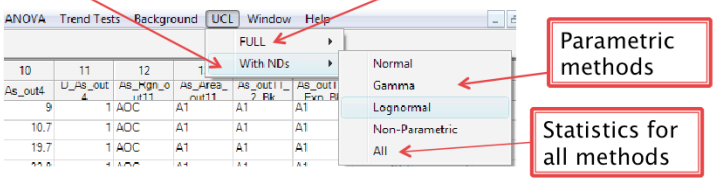
ProUCL 4.1.00

Computing 95% Upper Confidence Limit (UCL95) of Mean

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

Computing UCL95 of Mean

- Exposure Point Concentration (EPC) term represents "average" exposure contracted by a receptor over an exposure area during a long period of time
  - To address uncertainties associated with average (mean) exposure, a UCL95 is used to estimate the EPC term
- UCL module computes Parametric and Non-Parametric UCL95 for data " With NDs" and uncensored "Full" data without NDs



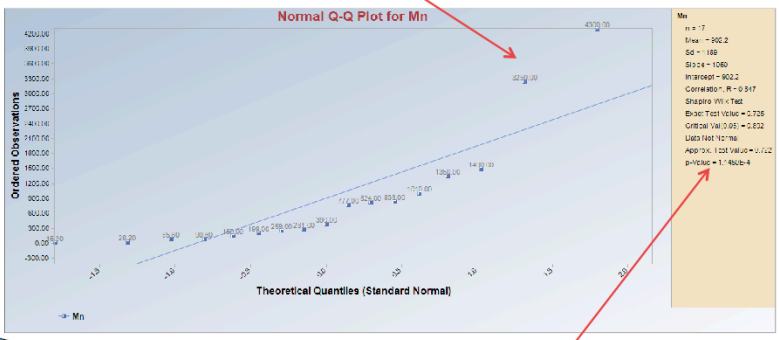
UCL95 for "Full" Data without NDs

- Parametric UCL methods:
  - Student's – t UCL: assumes approximate normality
  - Land's H-UCL: assumes lognormal distribution
  - Gamma distribution based UCLs
- Non-Parametric UCL Methods:
  - Modified – t, CLT, Adjusted – CLT, Chebyshev (Mean, Std)
  - Jackknife, standard bootstrap, bootstrap-t re-sampling methods

LOCKHEED MARTIN

UCL95 of Mean–Manganese Data

- Mn data from a Navy Site: 15.8 28.2 90.6 1490 85.6 281 4300 199 838 777 824 1010 1350 390 150 3250 259
  - Any outliers?
  - Data not normal
  - Data follow lognormal and gamma distributions

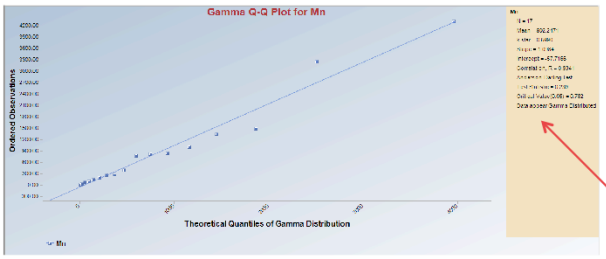


Data not normal, p-value = 0.001

LOCKHEED MARTIN

UCL95 of Mean- Manganese Data

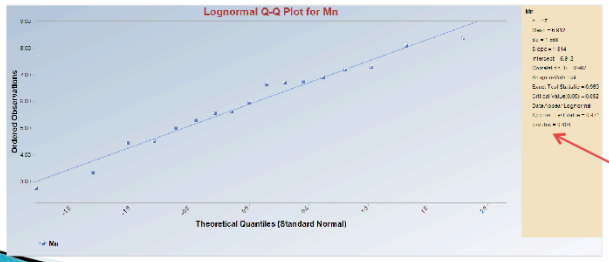
- Manganese data from Navy Site continued:



Data gamma distributed

A-D Test statistic=0.239

Critical value (.05) =0.782 > 0.239



Data are lognormal p-value = 0.808

LOCKHEED MARTIN

UCL95 of Mean-Mn Data (All Option)

Summary Statistics for Raw Full Dataset

Variable	NumObs	Minimum	Maximum	Mean	Median	Variance	SD	MAD/0.675	Skewness	Kurtosis	CV
Mn	17	15.8	4300	902.2	390	1414883	1189	554.8	2.046	3.934	1.318

Assuming Normal Distribution		Assuming Lognormal Distribution	
95% Student's-t UCL	1406	95% H-UCL	5182
95% UCLs (Adjusted for Skewness)		95% Chebyshev (MVUE) UCL	3237
95% Adjusted-CLT UCL (Chen-1995)	1530	97.5% Chebyshev (MVUE) UCL	4162
95% Modified-t UCL (Johnson-1978)	1430	99% Chebyshev (MVUE) UCL	5978
Gamma Distribution Test		**Data Distribution**	
k star (bias corrected)	0.599	Data appear Gamma Distributed at 5% Significance Level	
Theta Star	1506		
MLE of Mean	902.2		
MLE of Standard Deviation	1166		
nu star	20.37		
Approximate Chi Square Value (.05)	11.12	**Nonparametric Statistics**	
Adjusted Level of Significance	0.0346	95% CLT UCL	1377
Adjusted Chi Square Value	10.41	95% Jackknife UCL	1406
		95% Standard Bootstrap UCL	1352
Anderson-Darling Test Statistic	0.239	95% Bootstrap-t UCL	1923
Anderson-Darling 5% Critical Value	0.782	95% Hall's Bootstrap UCL	3851
Kolmogorov-Smirnov Test Statistic	0.117	95% Percentile Bootstrap UCL	1425
Kolmogorov-Smirnov 5% Critical Value	0.218	95% BCA Bootstrap UCL	1512
Data appear Gamma Distributed at 5% Significance Level		95% Chebyshev(Mean, Sd) UCL	2160
		97.5% Chebyshev(Mean, Sd) UCL	2704
Assuming Gamma Distribution		99% Chebyshev(Mean, Sd) UCL	3773
95% Approximate Gamma UCL	1652		
95% Adjusted Gamma UCL	1765		

Elevated lognormal UCL

Use Gamma UCL

LOCKHEED MARTIN

Influence of Outliers on UCL95 –Mn Data

- Manganese data from Navy Site continued:
  - Are 3250 and 4300 potential outliers?
  - Dixon outlier test used as sample size <25

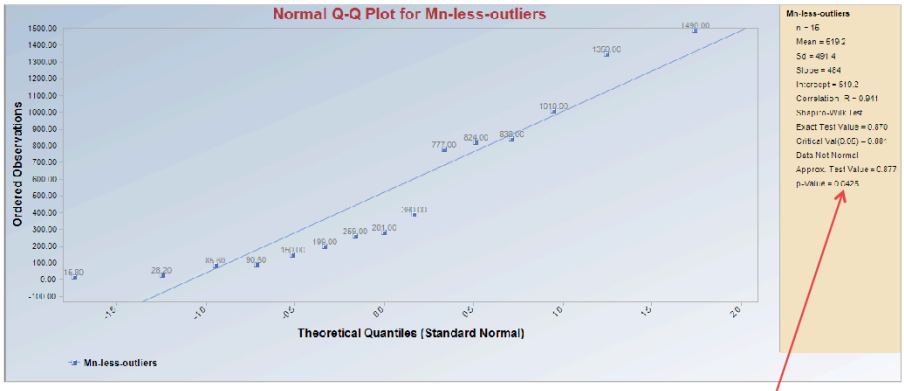
Dixon's Outlier Test for Mn
Number of data = 17
10% critical value: 0.438
5% critical value: 0.49
1% critical value: 0.577
1. Data Value 4300 is a Potential Outlier (Upper Tail)?
Test Statistic: 0.667
For 10% significance level, 4300 is an outlier.
For 5% significance level, 4300 is an outlier.
For 1% significance level, 4300 is an outlier.

Dixon's Outlier Test for Mn
Number of data = 16
10% critical value: 0.454
5% critical value: 0.507
1% critical value: 0.595
1. Data Value 3250 is a Potential Outlier (Upper Tail)?
Test Statistic: 0.600
For 10% significance level, 3250 is an outlier.
For 5% significance level, 3250 is an outlier.
For 1% significance level, 3250 is an outlier.

Both 4300 and 3250 are potential outliers
Project Team should decide about their disposition:– include or not include in statistical analysis

LOCKHEED MARTIN

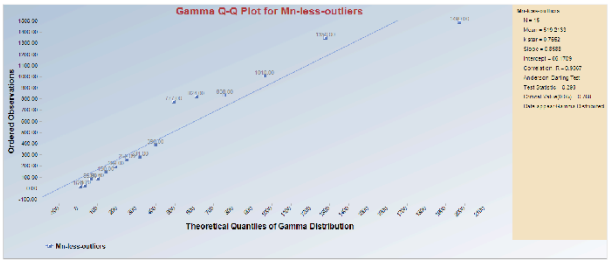
GOF Tests without 2 Outliers–Mn Data



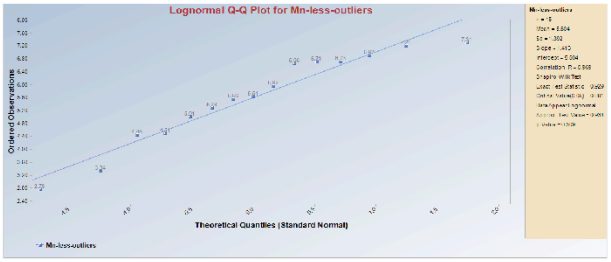
Data not normal (at 0.05), p–value of S–W test = 0.0425

LOCKHEED MARTIN

GOF Tests without 2 Outliers–Mn Data



Data without 2 outliers Gamma distributed



Data without 2 outliers are lognormal

LOCKHEED MARTIN

UCL95 less 2 Outliers–Mn Data (All Option)

Summary Statistics for Raw Full Dataset

Variable	NumObs	Minimum	Maximum	Mean	Median	Variance	SD	MAD/0.675	Skewness	Kurtosis	CV
Mn-less-outliers	15	15.8	1490	519.2	281	241481	491.4	374.8	0.807	-0.614	0.946

Assuming Normal Distribution	
95% Student's-t UCL	742.7
95% UCLs (Adjusted for Skewness)	
95% Adjusted-CLT UCL (Chen-1995)	756.2
95% Modified-t UCL (Johnson-1978)	747.1

Gamma Distribution Test	
k star (bias corrected)	0.765
Theta Star	678.5
MLE of Mean	519.2
MLE of Standard Deviation	593.5
nu star	22.96
Approximate Chi Square Value (.05)	13.06
Adjusted Level of Significance	0.0324
Adjusted Chi Square Value	12.15
Anderson-Darling Test Statistic	0.298
Anderson-Darling 5% Critical Value	0.768
Kolmogorov-Smirnov Test Statistic	0.175
Kolmogorov-Smirnov 5% Critical Value	0.229
Data appear Gamma Distributed at 5% Significance Level	
Assuming Gamma Distribution	
95% Approximate Gamma UCL	912.8
95% Adjusted Gamma UCL	981.3

Assuming Lognormal Distribution	
95% H-UCL	2550
95% Chebyshev (MVUE) UCL	1780
97.5% Chebyshev (MVUE) UCL	2272
99% Chebyshev (MVUE) UCL	3240

Data Distribution
Data appear Gamma Distributed at 5% Significance Level

Nonparametric Statistics	
95% CLT UCL	727.9
95% Jackknife UCL	742.7
95% Standard Bootstrap UCL	715.6
95% Bootstrap-t UCL	781.3
95% Hall's Bootstrap UCL	734.1
95% Percentile Bootstrap UCL	734
95% BCA Bootstrap UCL	725.1
95% Chebyshev(Mean, Sd) UCL	1072
97.5% Chebyshev(Mean, Sd) UCL	1312
99% Chebyshev(Mean, Sd) UCL	1782

LOCKHEED MARTIN

UCL95 Less 2 Outliers-Mn Data (Gamma)

- Statistics computed using Gamma distribution Option

Gamma Distribution Test		Data Distribution	
k star (bias corrected)	0.765	Data appear Gamma Distributed at 5% Significance Level	
Theta Star	678.5		
MLE of Mean	519.2		
MLE of Standard Deviation	593.5		
nu star	22.96		
Approximate Chi Square Value (.05)	13.06	Nonparametric Statistics	
Adjusted Level of Significance	0.0324	95% CLT UCL	727.9
Adjusted Chi Square Value	12.15	95% Jackknife UCL	742.7
		95% Standard Bootstrap UCL	724.4
Anderson-Darling Test Statistic	0.298	95% Bootstrap-t UCL	798.7
Anderson-Darling 5% Critical Value	0.768	95% Hall's Bootstrap UCL	738.9
Kolmogorov-Smirnov Test Statistic	0.175	95% Percentile Bootstrap UCL	727
Kolmogorov-Smirnov 5% Critical Value	0.229	95% BCA Bootstrap UCL	754.1
Data appear Gamma Distributed at 5% Significance Level		95% Chebyshev(Mean, Sd) UCL	1072
		97.5% Chebyshev(Mean, Sd) UCL	1312
Assuming Gamma Distribution		99% Chebyshev(Mean, Sd) UCL	1782
95% Approximate Gamma UCL	912.8		
95% Adjusted Gamma UCL	981.3		
Potential UCL to Use		Use 95% Approximate Gamma UCL	912.8

Note: Suggestions regarding the selection of a 95% UCL are provided to help the user to select the most appropriate 95% UCL.
These recommendations are based upon the results of the simulation studies summarized in Singh, Singh, and Iaci (2002)
and Singh and Singh (2003). For additional insight, the user may want to consult a statistician.

LOCKHEED MARTIN

UCL95 with and without 2 Outliers–Mn Data

- Two outliers distorted all statistics including UCL95

UCL95 Method	With outliers, n=17	Without outliers, n=15
Student's t UCL	1406	742.7
Gamma UCL	1652	912.8
Lognormal UCL	5182	2550
Bootstrap-t	1923	781.3
BCA Bootstrap	1512	725.1
Maximum	4300	1490

  - Project Team should make a decision about disposition of 2 outliers
  - Lognormal distribution resulted in unrealistic UCL95 > maximum value
  - Data are gamma distributed
    - Use of UCL95 based upon gamma distribution is recommended

LOCKHEED MARTIN

Steps to Compute UCL95 Using ProUCL

- Identify potential outliers/multiple populations
  - If justified, study them separately
  - Project Team should decide about disposition of outliers
- Perform GOF tests, look at data graphically using box plots and Q-Q plots to gain additional insight
- Use UCL95 as recommended by ProUCL
- Gamma distribution is better suited than lognormal distribution to model positively skewed uncensored environmental data sets without nondetects

LOCKHEED MARTIN

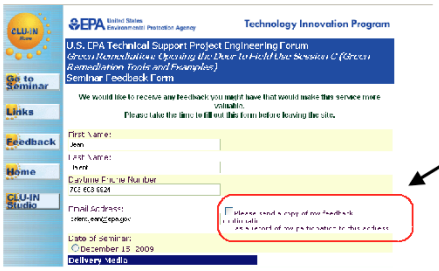
Avoid Lognormal Distribution and H-UCL

- Avoid use of a lognormal model as:
  - It accommodates outliers and multiple populations
  - It tends to yield impractically large UCL95, especially for highly skewed data sets of small sizes (e.g., <20)
- H-UCL95 often exceeds Max value
  - This results in use of Max value as an estimate of EPC term
  - EPC term represents average value:
    - Use of a measure of central tendency (and not extremes) should be used to estimate EPC
- NOTE: H-statistic yields unrealistically small H-UCL for large data sets of moderate skewness

LOCKHEED MARTIN

Resources & Feedback

- To view a complete list of resources for this seminar, please visit the **Additional Resources**
- Please complete the **Feedback Form** to help ensure events like this are offered in the future



Need confirmation of your participation today?

Fill out the feedback form and check box for confirmation email.

100