COMPUTATION OF AN UPPER CONFIDENCE LIMIT (UCL) OF THE UNKNOWN POPULATION MEAN

USING SOFTWARE ProUCL, VERSION 3.0

PART I

Anita Singh

Lockheed Martin Environmental Services

Disclaimer

- The United States Environmental Protection Agency (EPA) through its Office of Research and Development funded and managed the research described here. Mention of trade names or commercial products does not constitute endorsement or recommendation by the EPA for use.
- ProUCL software was developed by Lockheed Martin under a contract with the EPA and is made available through the EPA Technical Support Center in Las Vegas, Nevada.
- Use of any portion of ProUCL that does not comply with the ProUCL User Guide is not recommended.
- ProUCL contains embedded licensed software. Any modification of the ProUCL source code may violate the embedded licensed software agreements and is expressly forbidden.
- ProUCL software provided by the EPA was scanned with McAfee VirusScan v4.5.1 SP1 and is certified free of viruses.
- With respect to ProUCL distributed software and documentation, neither the EPA nor any of their employees, assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed. Furthermore, software and documentation are supplied "as-is" without guarantee or warranty, expressed or implied, including without limitation, any warranty of merchantability or fitness for a specific purpose.

Acknowledgements

- The development of software ProUCL, Version 3.0 has been supported by the Office of Superfund Remediation and Technical Information, USEPA, Washington, D.C.

- ProUCL is managed by the Technical Support Center of USEPA in Las Vegas, NV.

Workshop Objectives - ProUCL

- The main objective is to provide working knowledge about ProUCL using real Superfund examples.

- Demonstrate why the default use of a lognormal distribution should be avoided to compute upper confidence limits (UCLs) of mean.

- Demonstrate why other skewed distribution such as a gamma distribution is better suited to compute UCLs.

- Discuss the availability of distribution - free UCL methods (e.g., bootstrap, Chebyshev inequality) in ProUCL, V 3.0.

Workshop Objectives - ProUCL

- Demonstrate the undue influence of outliers on the computation of UCLs, EPC terms.
- Discussion about the treatments of outliers.
- Discuss how to interpret the various output results produced by ProUCL.
- Discuss and show how to select the most appropriate UCL in ProUCL to estimate the EPC.
- Discuss how to use/interpret recommendations made by ProUCL.

Workshop Outline

- Introduction:What are UCLs and assumptions needed.

- Distributions and goodness-of-fit tests in ProUCL.

- Parametric UCL methods based upon:
  - Normal
  - Lognormal, and
  - Gamma

- Non-parametric (distribution-free) UCL methods:
  - Large sample UCLs, skewness adjusted UCLs
  - Chebyshev UCLs
  - Bootstrap UCLs

Workshop Outline

- Illustrations of parametric and non-parametric UCLs using real Superfund data sets.

- Outliers and their influence on UCLs.
  – Illustrations using real data sets.

- Summary of a recommended procedure to compute an appropriate UCL of the mean.

Why Do We Need to Compute a UCL?

- In Superfund exposure and risk assessment studies - a 95% UCL of unknown mean, μ_1 , of a contaminant of potential concern (COPC) is used to estimate the Exposure Point Concentration (EPC) Term.

- A UCL is computed using sampled data and a discernible (e.g., normal, lognormal, gamma) probability distribution (if any), or a non-parametric (distribution free) method.

Tools and Guidance to Compute UCLs / EPC Terms

- EPA, December 2002: Calculating UCLs Guidance Document for Hazardous Waste Sites - OSWER 9285.670

- EPA, April 2004: ProUCL, Version 3.0
  Available at:
  http://www.epa.gov/nerlesd1/tsc/software.htm

What is a UCL?

- A $(1-\alpha)100\%$ UCL of mean, μ_1 is a random value (based upon sampled data) given by the probability statement:

$$P(\mu_1 \leq UCL) = 1 - \alpha$$

- A UCL should represent a realistic value of practical merit providing approximately $(1-\alpha)100\%$ coverage to mean, μ_1.
  - This requires use of:
    - Appropriate probability distributions, and
    - UCL computation methods

Fundamental Assumptions Needed - for all UCL Computation Methods

- Make sure that there are no outliers and/or multiple populations (e.g, mixture of site and background data):
  - UCL is computed for the unknown mean of a single population.
  - If there are multiple populations, separate them before proceeding with UCL computations – seek statistician's help, use graphical displays such as quantile-quantile (Q-Q) plots.
  - Outliers when present distort all statistics, UCLs, etc.
  - If justified, remove all outliers before computing UCLs.

Distributions in ProUCL,Version 3.0

- Normal distribution (symmetric):
  - But environmental data are often positively skewed.
- Lognormal distribution (positively skewed):
  - Historically often used as a default model.
  - Lognormal model accommodates: outliers, impractically large mean values, and also multiple populations.
  - Its use often results in unstable, impractical, unreliable UCL values leading to:
    - Further investigation recommendation.
    - Use of the maximum value as an estimate of exposure point concentration (EPC) term.

Distributions in ProUCL,Version 3.0

- Conclusion: Several recent studies suggested that Lognormal model is not appropriate to compute a UCL, especially for small sample size, n.

- Gamma distribution (positively skewed):
  - It seems to work well on environmental data sets (EPA Issue Paper, 2002: Singh, Singh, and Iaci).

- Non-parametric (distribution free) UCL computation methods - several in ProUCL.

Lognormal Distribution

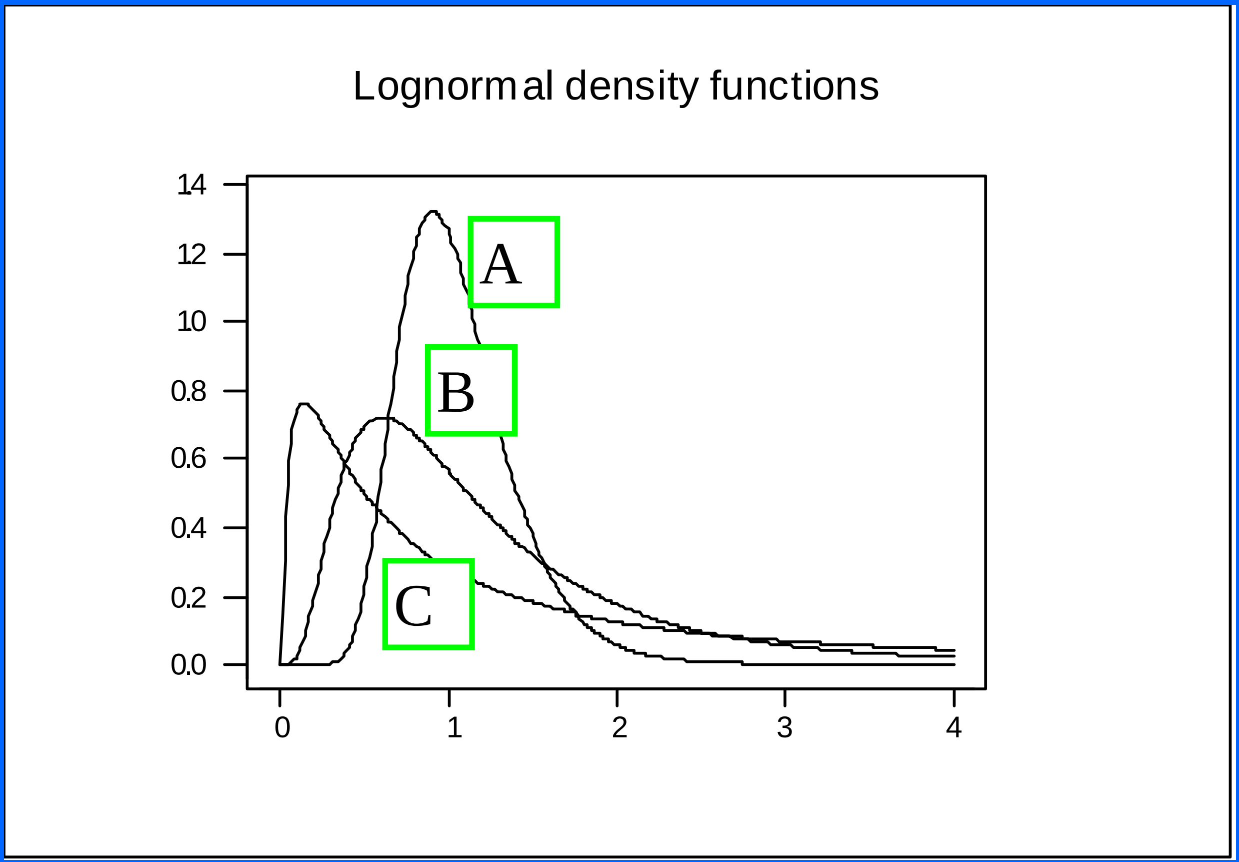
- X is lognormal, LN(μ , σ) if Y=ln(X) is normal with mean, μ and sd, σ
  - This is probably why it became so popular.
  - Its use however leads to unstable and impractical results.
  - Not a practical model to compute a UCL of the mean.

- Lognormal mean, $\mu_1 = \exp(\mu + \frac{\sigma^2}{2})$
  - Can be unduly large.

- Coefficient of Variation (CV) = $\sqrt{\exp(\sigma^2) - 1}$

- Skewness = $(CV)^3 + 3(CV)$, depends upon σ only, and often results in unrealistically large values.

Lognormal Distribution



A=LN(0, σ = 0.316), B=LN(0,0.707), C=LN(0,1.414)

Skewness as a function of σ Sd of Log-transformed Data

Skewness as a Function of σ (or its MLE, $s_y = \hat{\sigma}$)

Standard Deviation	Skewness
$\sigma < 0.5$	Symmetric to mild skewness
$0.5 \le \sigma < 1.0$	Mild Skewness to Moderate Skewness
$1.0 \le \sigma < 1.5$	Moderate Skewness to High Skewness
$1.5 \le \sigma < 2.0$	High skewness
$2.0 \le \sigma < 3.0$	Extremely high skewness
$\sigma \ge 3.0$	Not well-behaved data sets - require further investigation

Gamma Probability Model in ProUCL, Version 3.0

A two-parameter Gamma Model, $G(k,\theta)$ is:

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

$k =$ shape, and $\theta =$ scale parameters

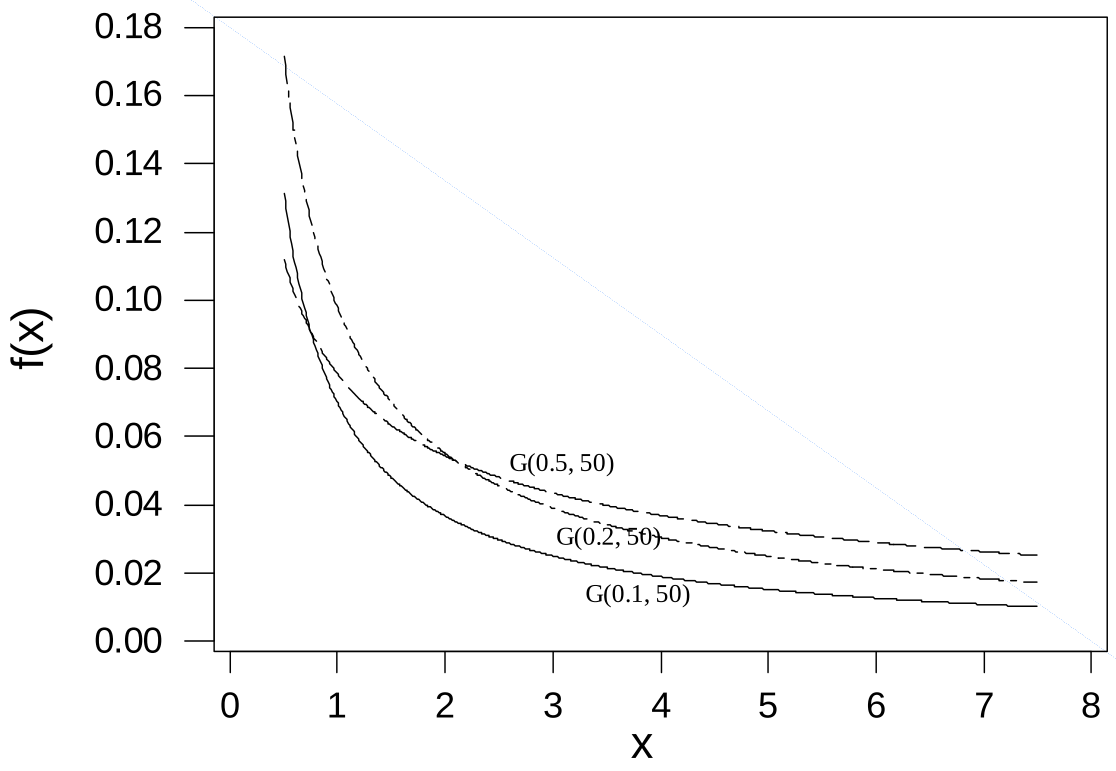
Mean $= \mu_1 = k\theta$

variance $= k\theta^2$

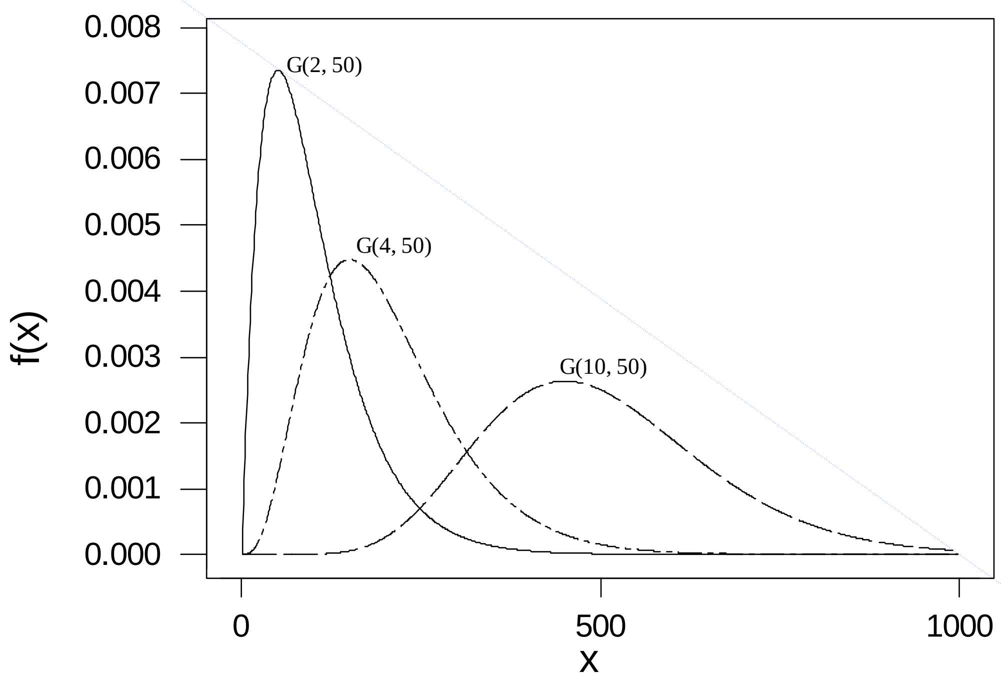
skewness $= 2/\sqrt{k}$

As k exceeds 6, it starts becoming symmetrical, and an approx. normal model may be used.

Gamma Densities



Gamma Densities



Random Sample of Size n

- Let $x_1, x_2, \ldots, x_n$ be a random sample (data) on X (e.g., X = conc. of Pb, PCB) from a population (site, EA) with unknown mean, μ_1 .
  - – Any potential outliers/multiple populations requiring separate treatment?
  - – Do data follow a discernible probability model?

- Raw sample mean and standard deviation (sd):

$$\bar{x} = \sum_{i=1}^{n} x_i / n$$

$$s_x = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)}$$

Normal/Lognormal Distributions

- Tests for normality and lognormality of data:
  - Graphical Q-Q plot and Histogram
  - Shapiro - Wilk test (sample size, n<=50)
  - Lilliefors test (n>50) – a generalization of K-S test
- Computes various statistics:
  - summary statistics
  - maximum likelihood estimates (MLEs) and minimum variance unbiased estimates (MVUEs) of mean, sd, quantiles, CV, skewness, SE of mean

Example 1. Consider a well-behaved data set (Grice.dat) of size n=20 from the literature.

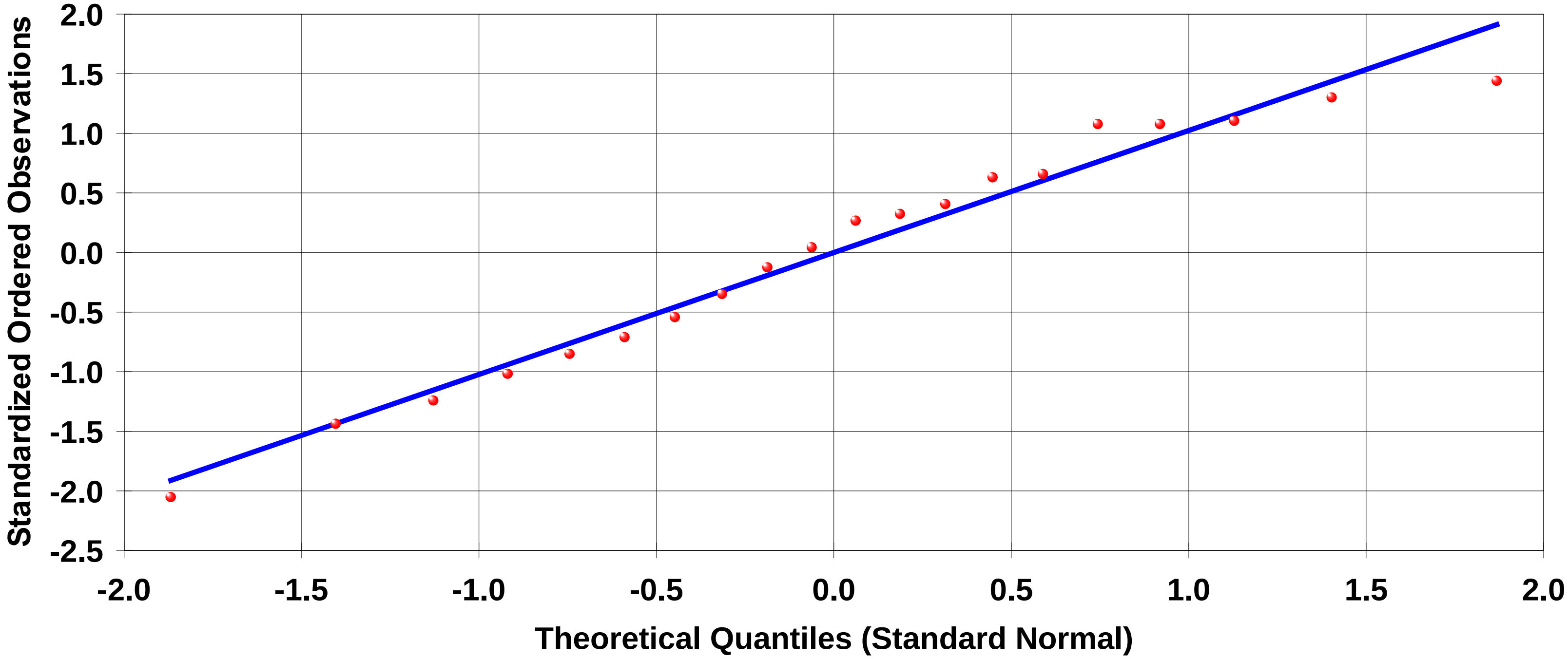
152, 152, 115, 109,137, 88, 94, 77, 160, 165, 125, 40, 128, 123, 136, 101, 62, 153, 83, and 69.

- Data are both normal and lognormal – see ProUCL output

- For this data, any of the two UCLs can be used:

  - normal 95% UCL = 127.29
  - lognormal 95% UCL = 134.73

Normal Q-Q Plot for Grice

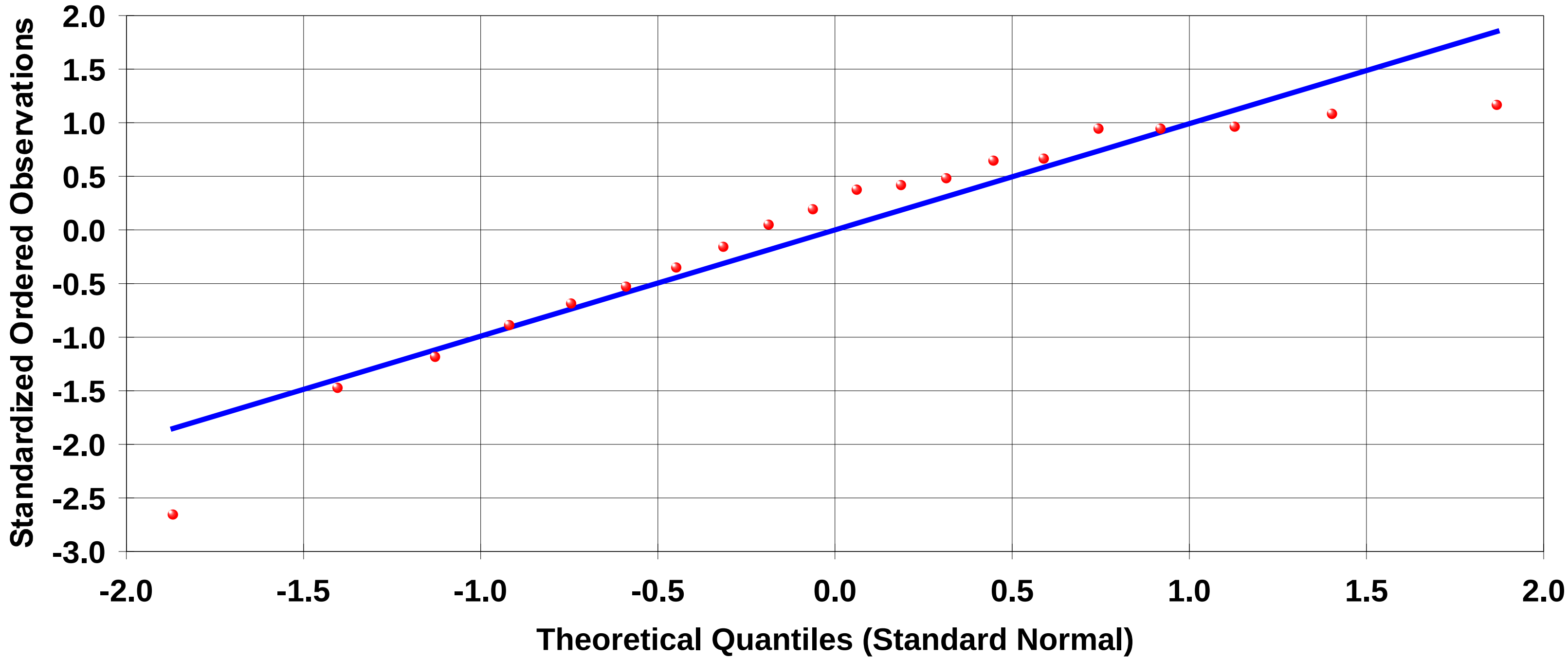


N = 20, Mean = 113.4500, Stdv = 35.7896

Slope = 1.0233, Intercept = 0.0000, Correlation, R = 0.98577759

Shapiro-Wilk Statistic = 0.961, Critical Value(0.05) = 0.905, Data are Normal

Lognormal Q-Q Plot for Grice



N = 20, Mean = 4.6735, Stdv = 0.3709

Slope = 0.9912, Intercept = 0.0000, Correlation, R = 0.95491839

Shapiro-Wilk Statistic = 0.912, Critical Value(0.05) = 0.905, Data are Lognormal

ProUCL Version 3.0 - [Lognormal UCL Statistics for Grice.dat]

File Edit View Options Summary Statistics Histogram Goodness-of-Fit Tests UCLs Window Help

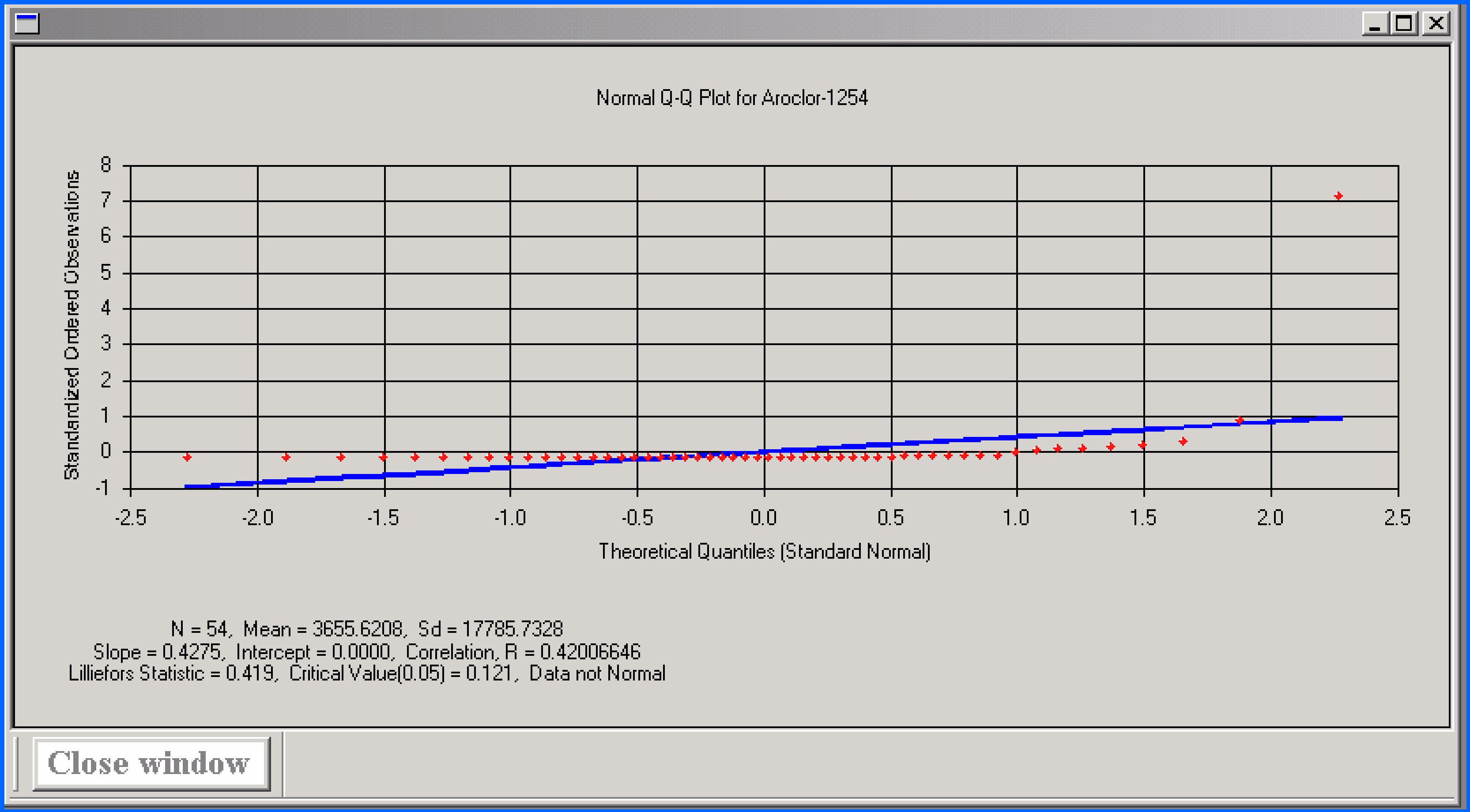
	A	B	C	D	E	F	G	H
1	Data File	D:\drive_c\sprfnd02\proucl2003\grice.dat				Variable:	Grice.dat	
2								
3	Number of Valid Samples				20			
4	Number of Distinct Samples				19			
5	Minimum of log data				3.6888795			
6	Maximum of log data				5.1059455			
7	Mean of log data				4.673464			
8	Standard Deviation of log data				0.3708584			
9	Variance of log data				0.1375359			
10								
11	Shapiro-Wilk Test Statisitic				0.9115046			
12	Shapiro-Wilk 5% Critical Value				0.905			
13	Data are lognormal at 5% significance level							
14								
15	95% UCL (Assuming Normal Distribution)							
16	Student's-t				127.28788			
17								
18	Estimates Assuming Lognormal Distribution							
19	MLE Mean				114.6899			
20	MLE Standard Deviation				44.03897			
21	MLE Coefficient of Variation				0.383983			
22	MLE Skewness				1.2085645			
23	MLE Median				107.06799			
24	MLE 80% Quantile				146.47274			
25	MLE 90% Quantile				172.43443			
26	MLE 95% Quantile				197.0635			
27	MLE 99% Quantile				253.68176			
28								
29	MVU Estimate of Median				106.70042			
30	MVU Estimate of Mean				114.27318			
31	MVU Estimate of Sd				43.305246			
32	MVU Estimate of SE of Mean				9.6740949			
33								
34	95% Non-parametric UCLs							
35	Adjusted-CLT UCL (Adjusted for Skewness)				125.93418			
36	Modified-t UCL (Adjusted for Skewness)				127.18193			
37	[illegible]				[illegible]			

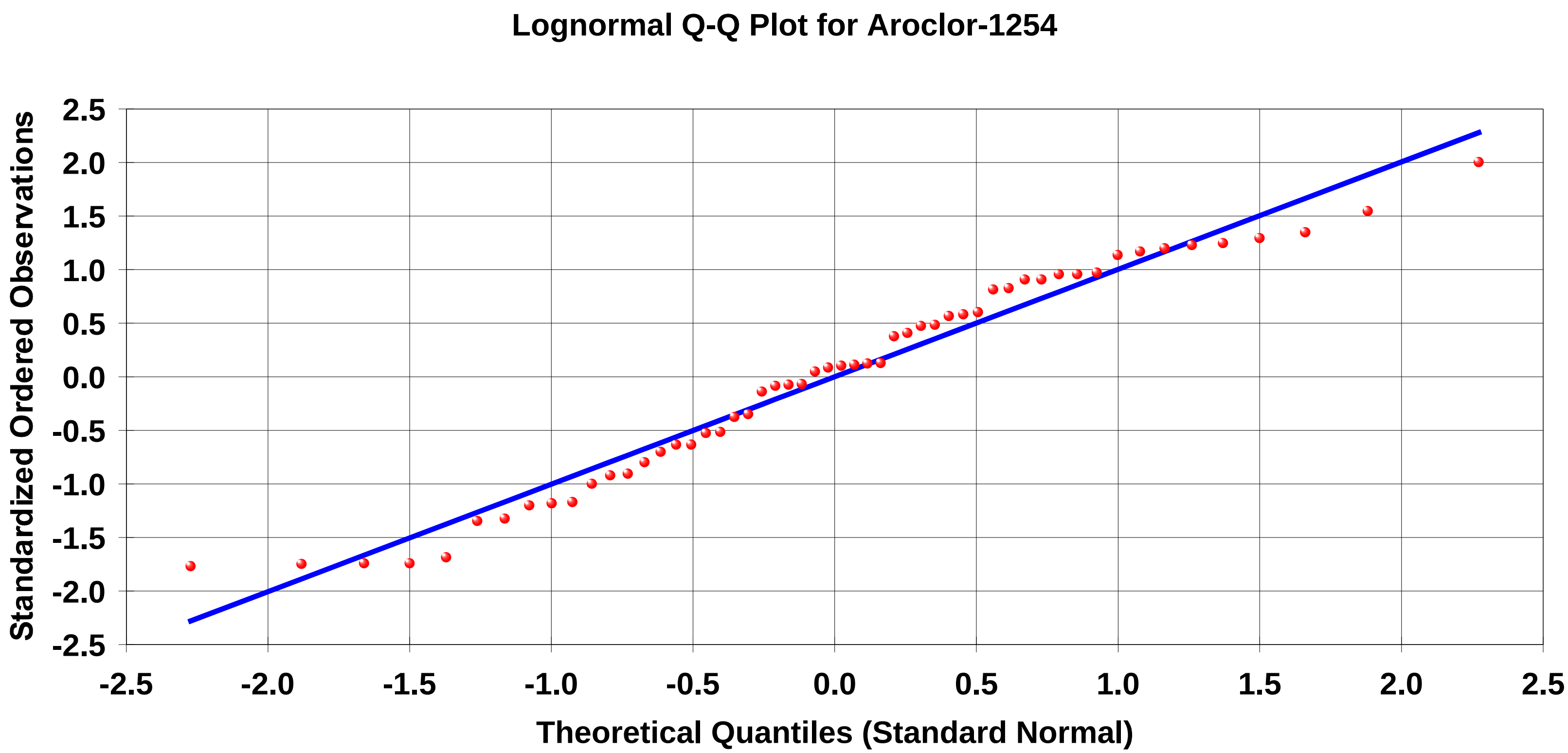
A Real Data Set
Cornell Dublier Site

Example 2. Consider a real data set of Aroclor –1254 concentrations of size 54 from a Superfund Site.

- At 5% significance level, data are lognormal (and not normal).
- Sd, $\hat{\sigma}$ of log data =4.20, very high.
- Data set has at least one outlier = 130,000.
- Lognormal distribution accommodates the outlier(s).

EXAMPLE 2: REAL DATA SET





N = 54, Mean = 3.3537, Stdv = 4.2015
Slope = 1.0028, Intercept = 0.0000, Correlation, R = 0.98534636
Lilliefors Statistic = 0.089, Critical Value(0.05) = 0.121, Data are Lognormal

Gamma Distribution is Missing From EPA Applications? Why?

- Statistical tools are not easily available:
  - Gamma goodness - of - fit tests not readily available.
  - Available in SAS, S-Plus, SPSS for limited values of n and k.
  - Estimation of gamma parameters, (k, θ) is computationally intensive – missing from text books.
  - Need complex numerical iterative methods (e.g., Newton-Raphson Method) which involve Digamma and Trigamma functions (Choi and Wette, 1969, Johnson, Kotz, Balakrishnan, 1994).

Available Goodness-of-Fit Tests for Gamma Model (e.g., in ExpertFit)

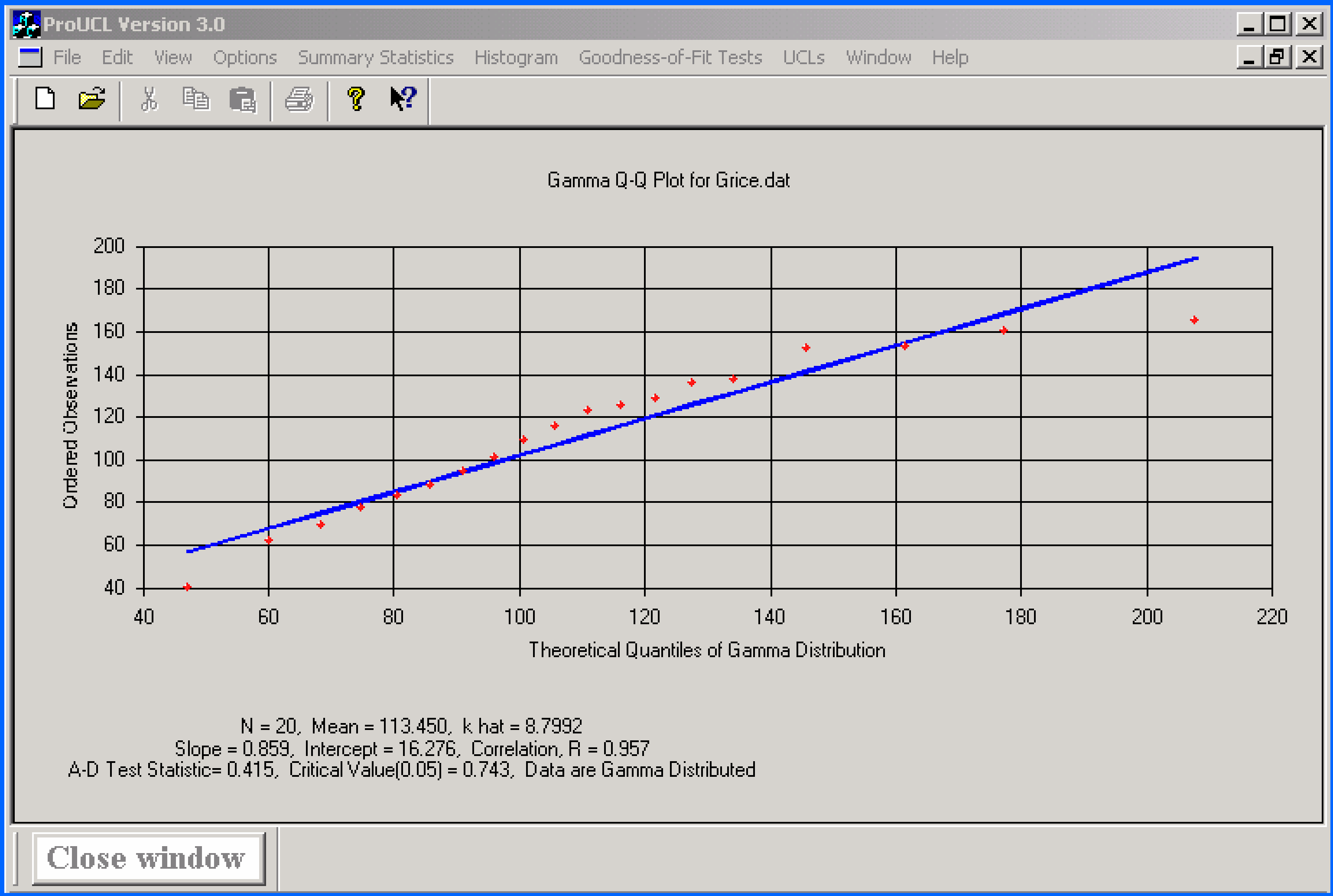
- Only limited critical values for selected values of k and sample size, n were available:
  - D'Agostino and Stephens (1986): A-D test
  - B. E. Schneider (1978): K-S test

- Software ExpertFit has A-D and K-S tests, uses generic critical values (for all distributions) of A-D and K-S tests, when all parameters are known (Stephens, 1970) – **not good enough.**

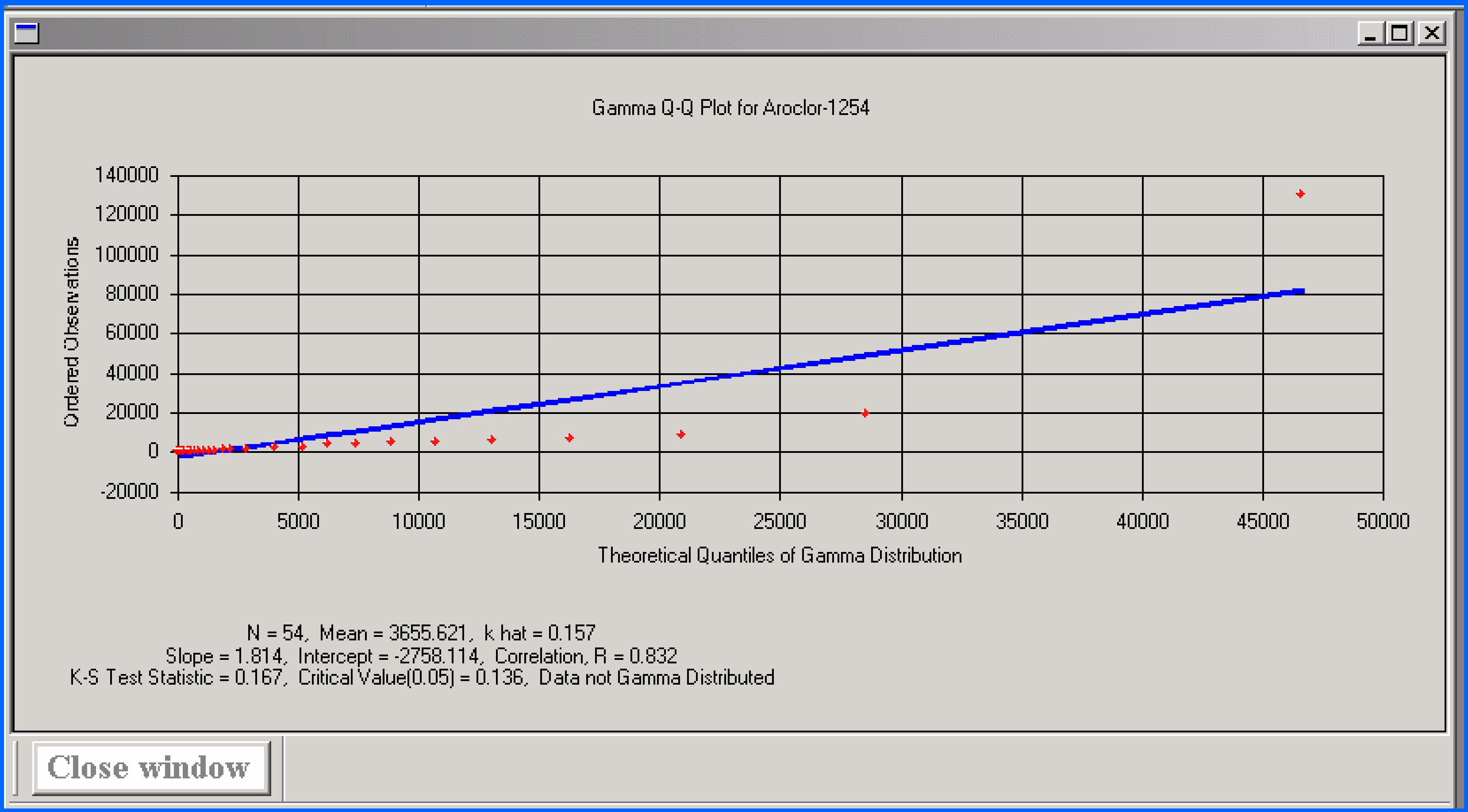
Goodness-of-Fit Tests for Gamma Distribution - in ProUCL, V 3.0

- We obtained simulated critical values for A-D and K-S tests for samples of size up to 2500 for various values of k – in ProUCL, V 3.0 .
  - ProUCL has graphical gamma Q-Q plot and histogram.
- For A-D (= A^2) and K-S (=D) tests, if calculated value is smaller than the respective critical value:
  - conclude data pass for a gamma model.
- ProUCL computes MLEs of k and θ

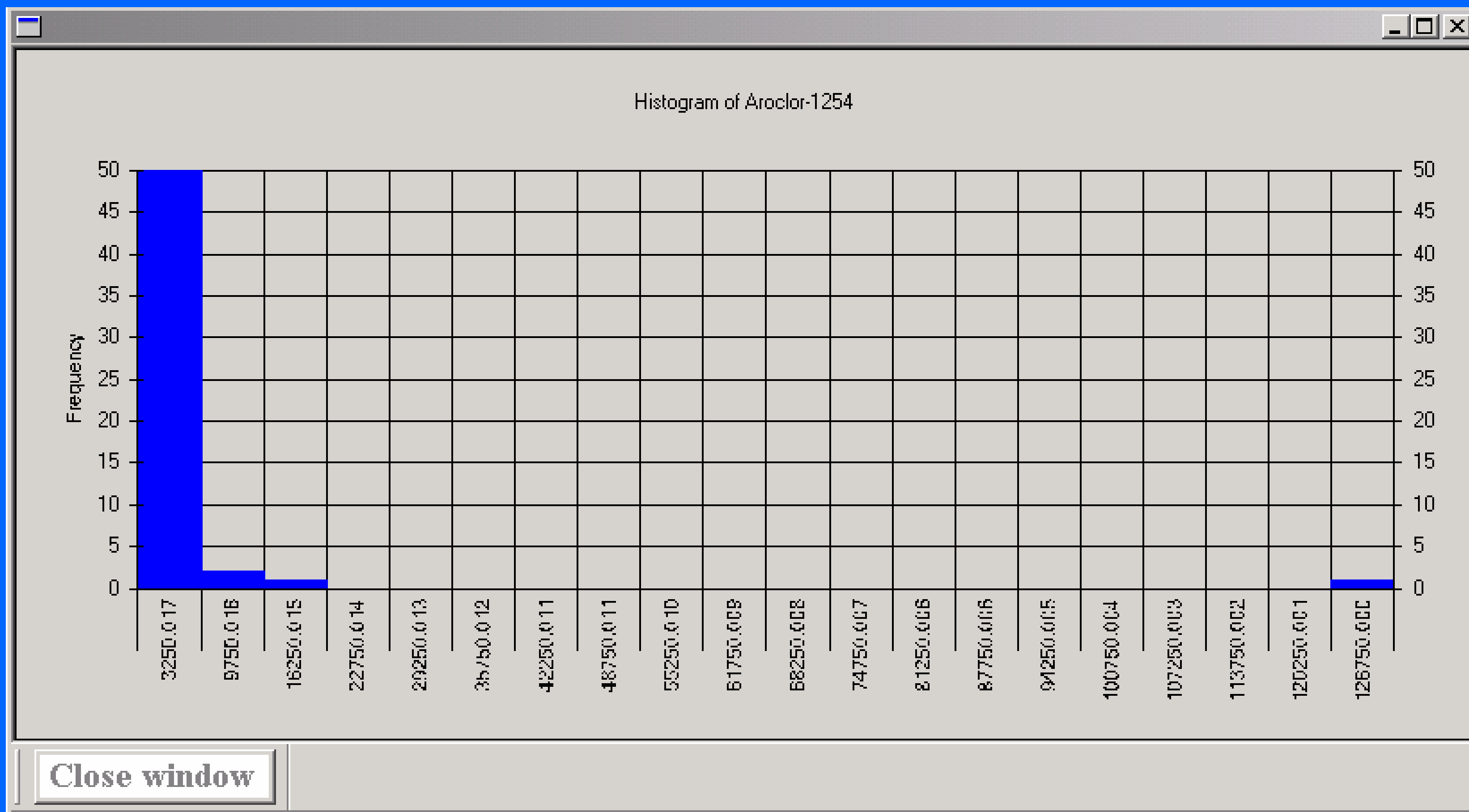
EXAMPLE 1: – WELL BEHAVED DATA SET - GRICE.DAT



EXAMPLE 2: - REAL DATA SET FOR AROCLOR – 1254



EXAMPLE 2: - REAL DATA SET FOR AROCLOR – 1254



Need For Accurate Critical Values (Rather Than using Generic Values) for Gamma Goodness –of-Fit- Tests.

Example 2 (continued) Aroclor - 1254: Test for Gamma Model using conservative generic critical values

	A-D test stat	K-S test stat
Calculated	2.18	0.167
5%Tabulated	2.49 (conservative)	~0.17

Calculated values < critical values, data pass for a gamma model at 0.05 level of significance – Incorrect conclusion based upon generic critical values.

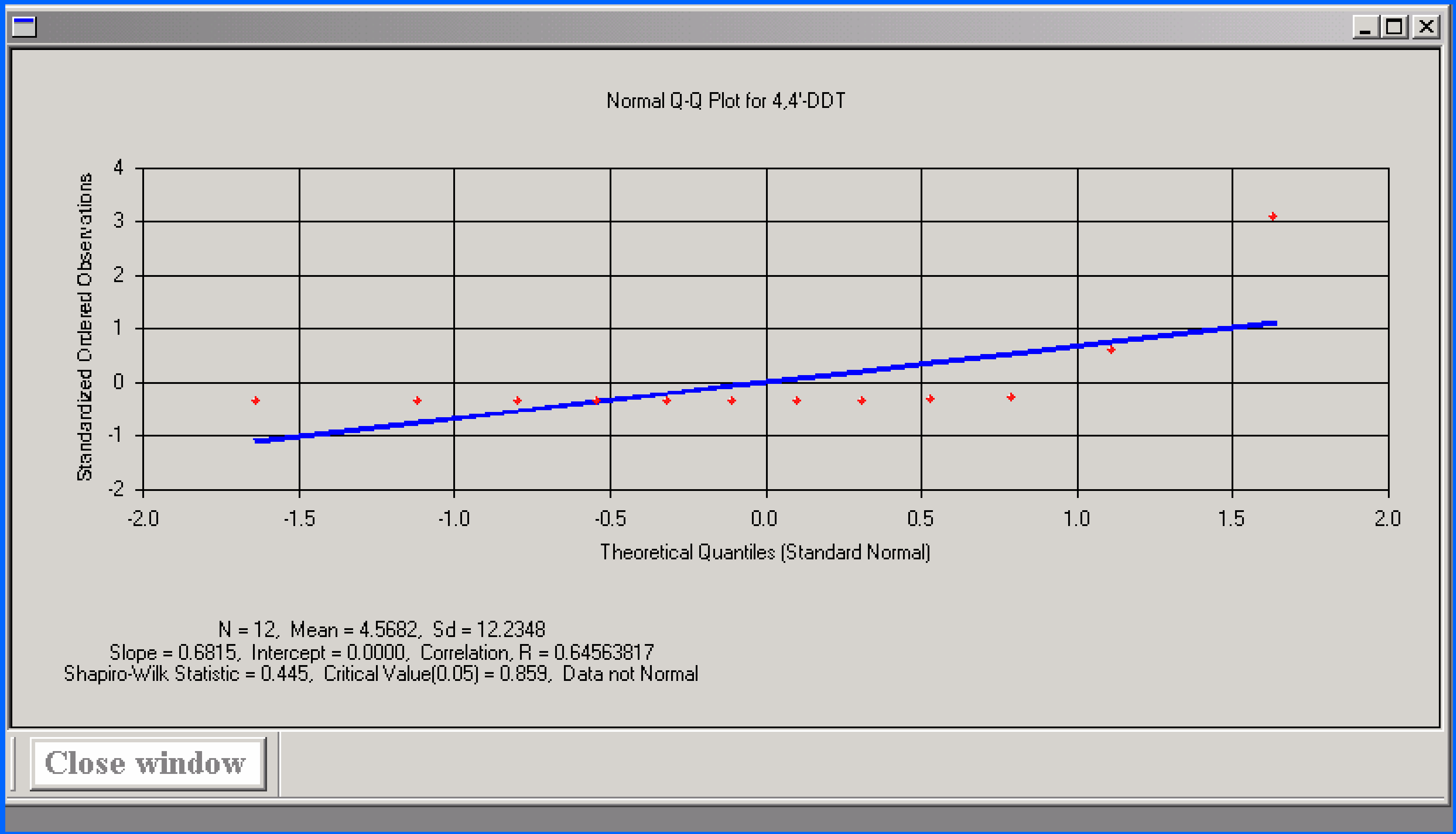
EXAMPLE 3: REAL DATA SET 4,4' – DDT CONCENTRATIONS

Test for Gamma Model using conservative generic critical values.

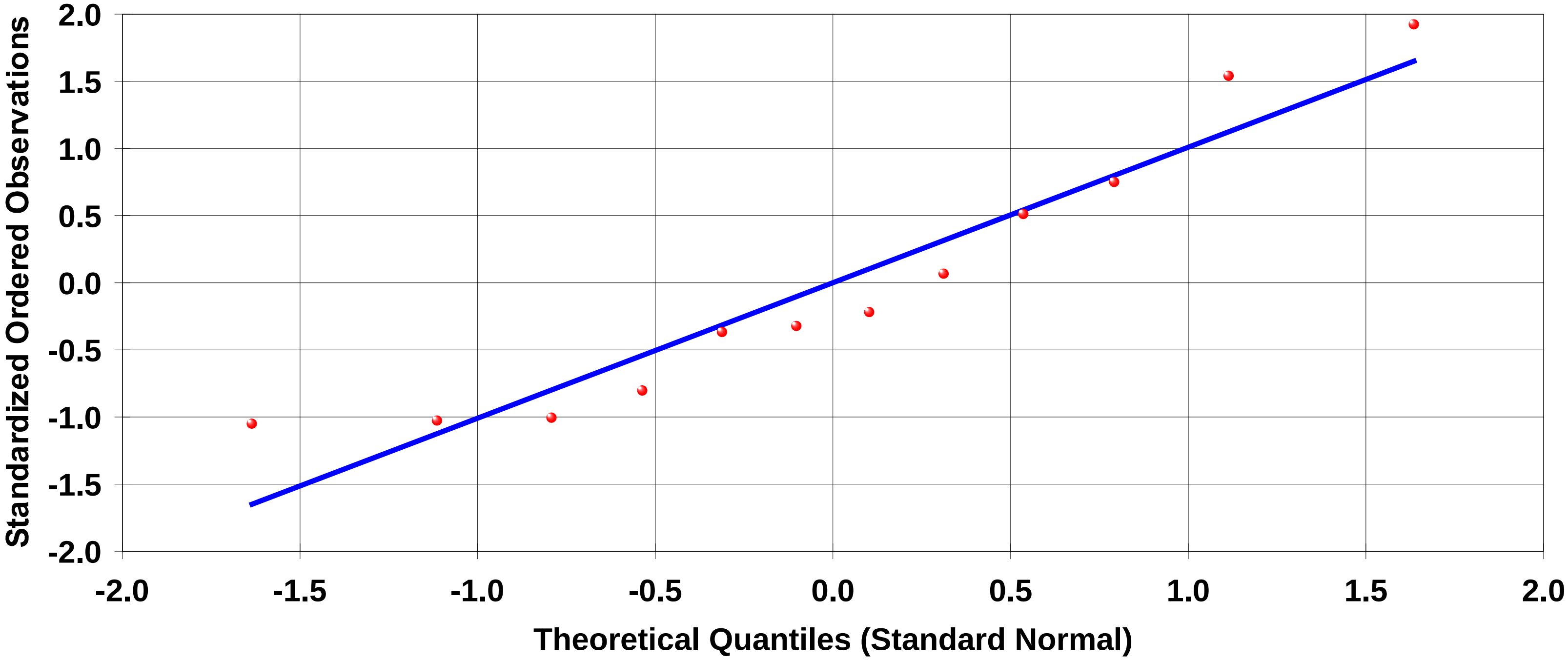
	A-D test stat	K-S test stat
Calculated	1.26	0.275
5%Tabulated	2.49 (conservative)	~0.29

Conclusion: Data pass for a gamma model - this may be incorrect.

EXAMPLE 3: REAL 4,4' DDT DATA SET

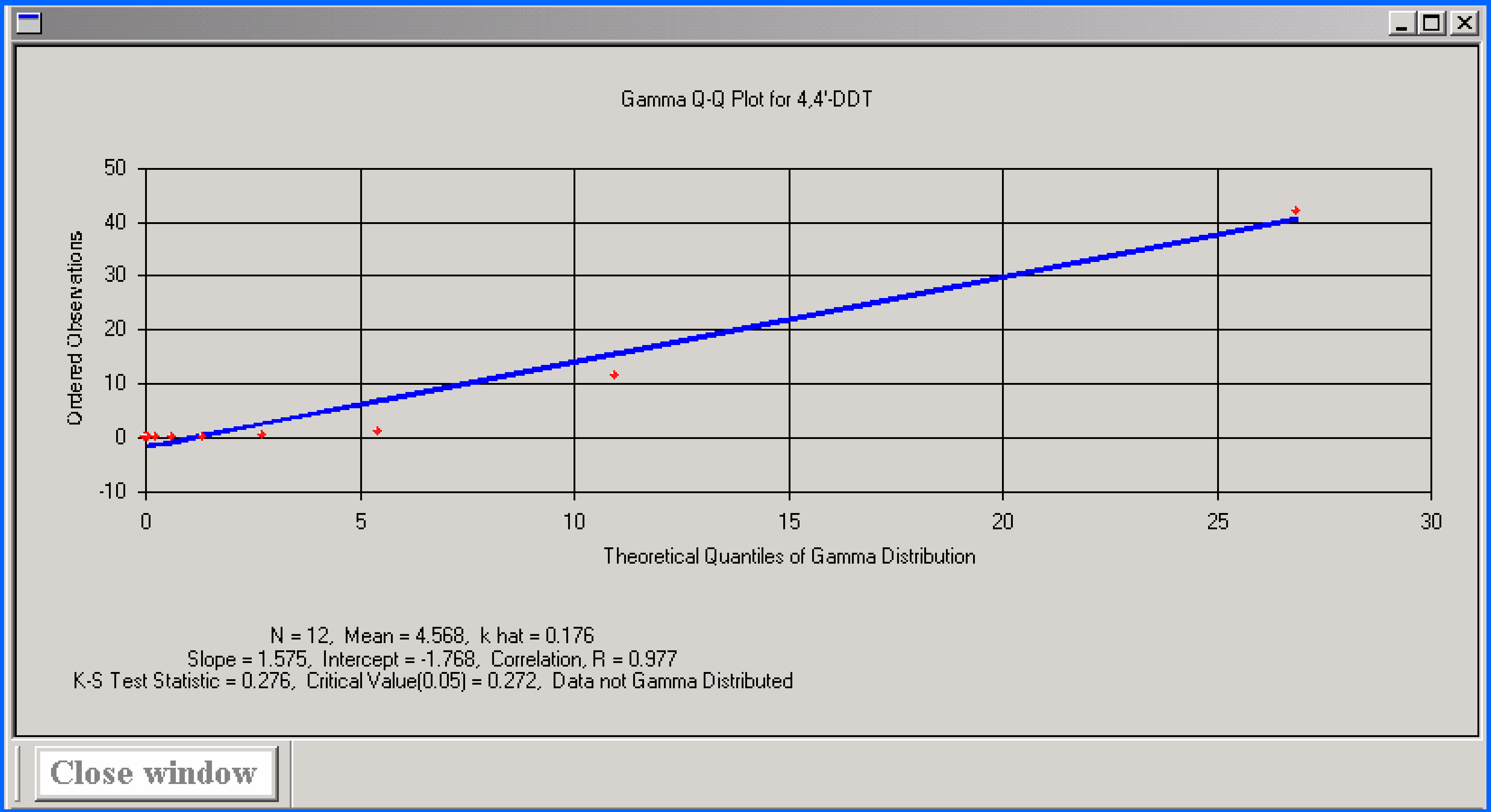


Lognormal Q-Q Plot for 4,4'-DDT

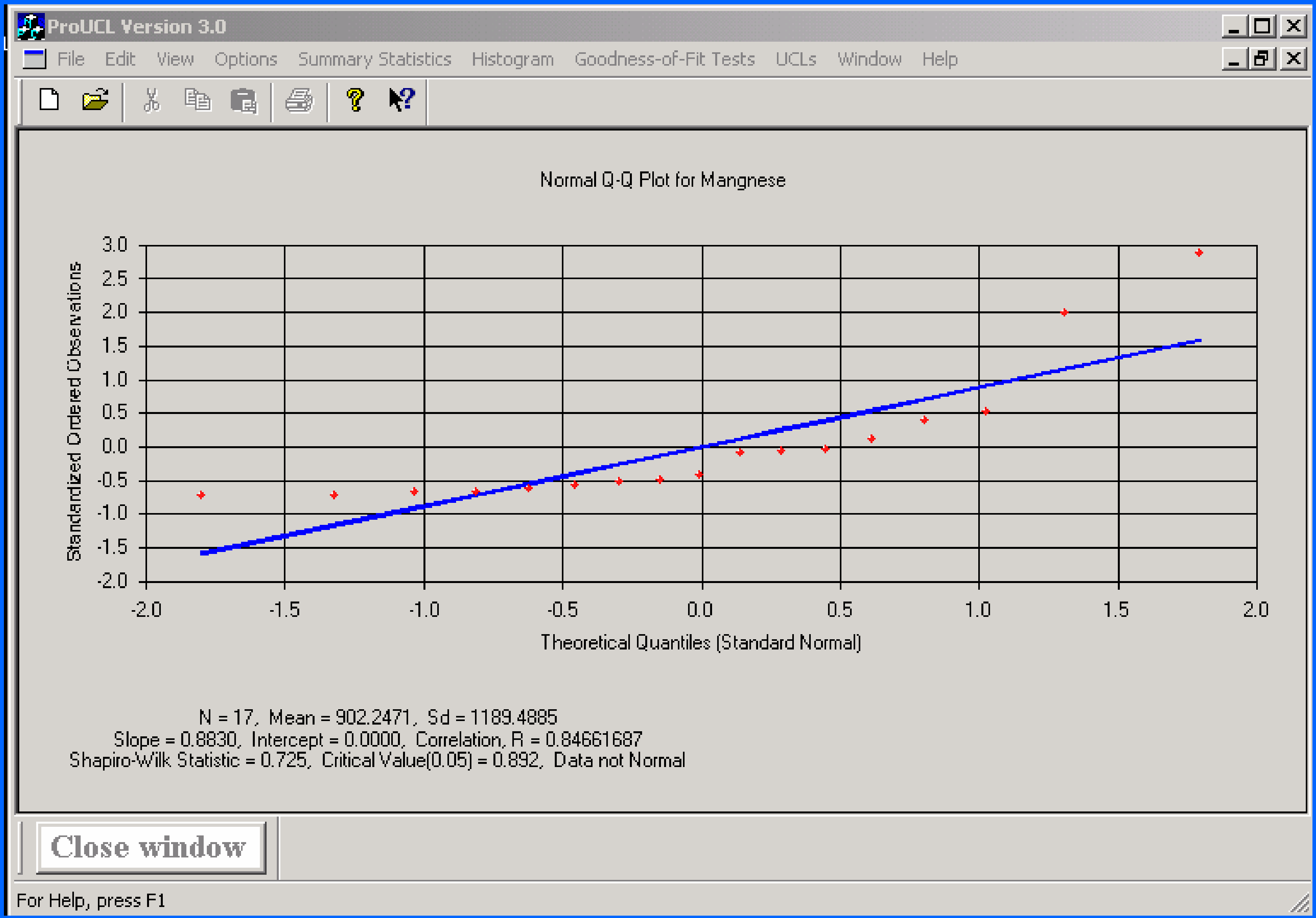


N = 12, Mean = -2.7522, Stdv = 3.3728
Slope = 1.0090, Intercept = 0.0000, Correlation, R = 0.95591841
Shapiro-Wilk Statistic = 0.899, Critical Value(0.05) = 0.859, Data are Lognormal

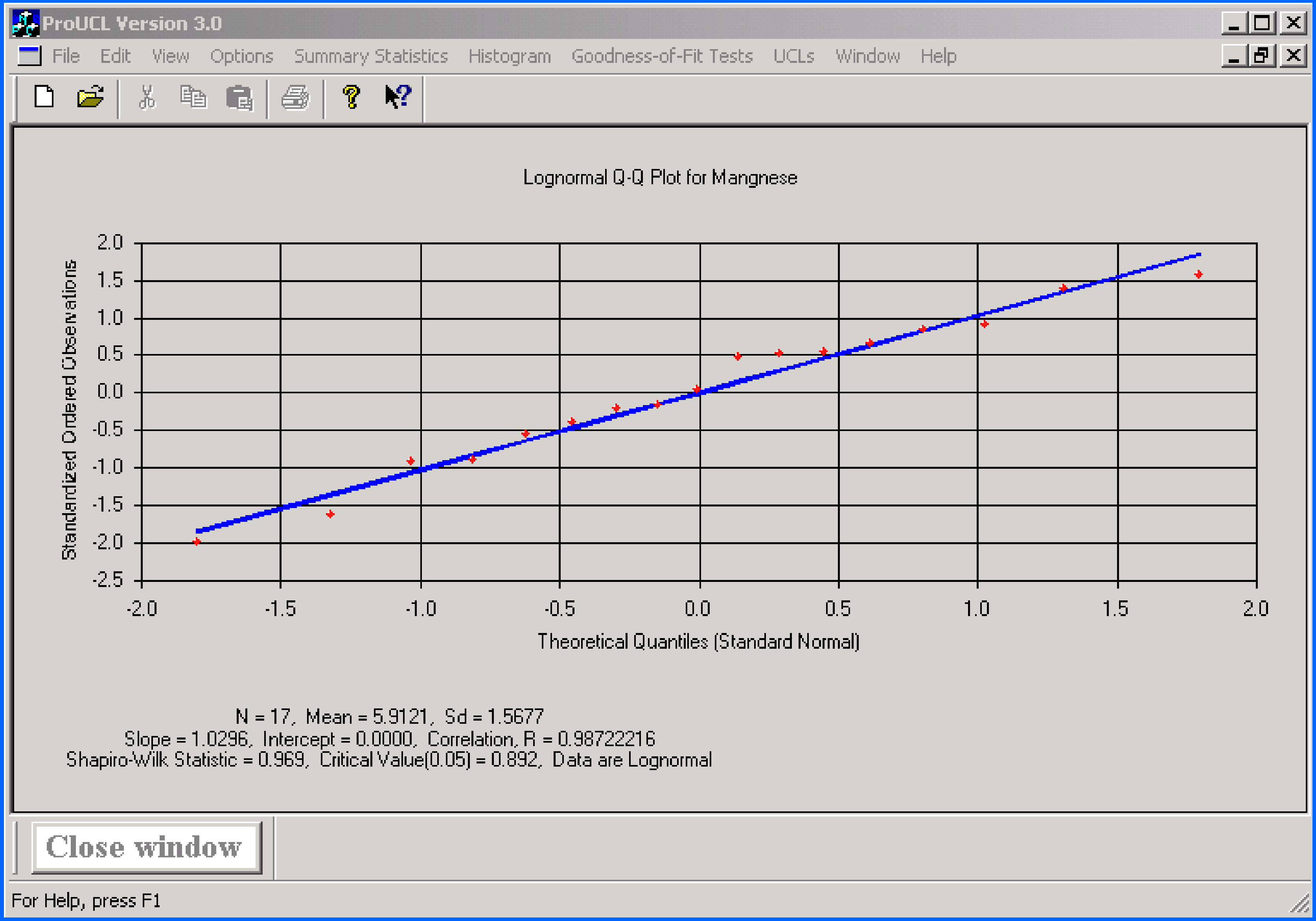
EXAMPLE 3: REAL 4,4' DDT DATA SET



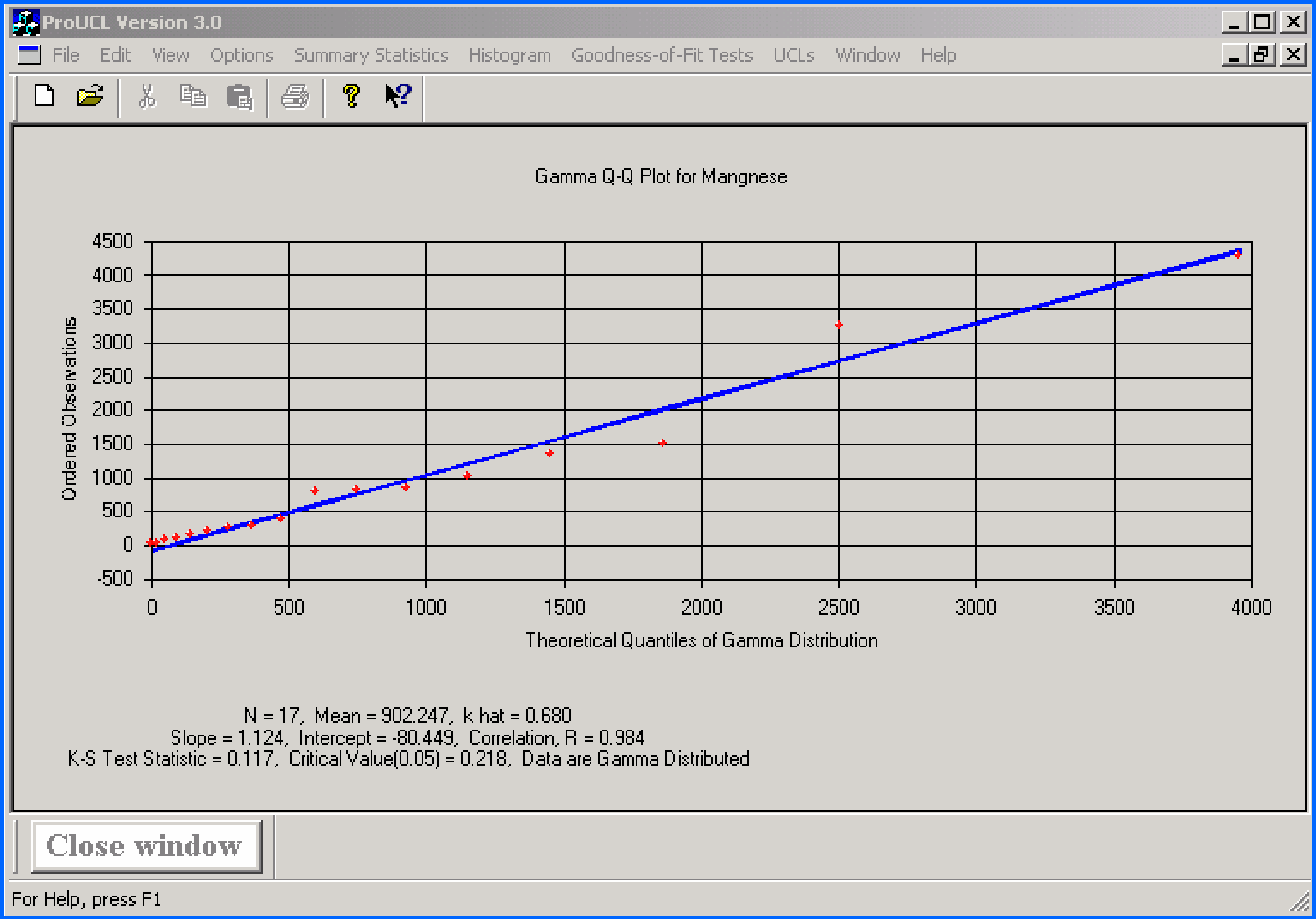
Example 4: Real Mn Data - NCBC Site



Example 4: Real Mn Data - NCBC Site



Example 4: Real Mn Data - NCBC Site



UCLs in ProUCL, Version 3.0

- Computes UCLs for normal, lognormal, Gamma models and also for non-parametric data sets.

- 5 Parametric UCL Methods:
  - Student's - t (normal)
  - H-UCL, Chebyshev (MVUE) UCL (lognormal)
  - Approximate and Adjusted gamma UCLs

- Several Non-parametric UCL Methods:
  - Modified - t, CLT, Adjusted - CLT, Chebyshev (Mean, Sd)

UCLs in ProUCL, Version 3.0

- Non-parametric Resampling UCL Methods: Jackknife and Bootstrap (Efron, 1981, 1982, Hall, 1992).
  - Jackknife UCL of mean – same as Student’s-t UCL
  - Standard Bootstrap
  - Bootstrap – t: **use with caution**
  - Hall’s Bootstrap procedure: **use with caution**
  - BCA Bootstrap Method
  - Percentile Bootstrap Method
- Caution: Bootstrap-t and Hall’s methods can be easily influenced by outliers, and can result in erratic inflated UCL values (Efron & Tibshirani,1993).

UCL Based Upon a Normal Distribution

- Normal distribution is symmetric, skewness ~ 0.

- Student's – t statistic is used to compute a UCL of the mean (in EPA 1992 guidance document):

$$UCL = \overline{x} + t_{n-1,1-\alpha} \frac{s_x}{\sqrt{n}}$$

- Here $t_{n-1,1-\alpha}$ is the upper α th percentile of student's t-distribution.
  - Sensitive to outliers, which may need to be removed.
  - UCL does not provide adequate coverage for mean of skewed probability models.

UCL of Mean Based Upon a Lognormal Distribution

- Land's H-Statistic based $(1-\alpha)100\%$ UCL of mean (in 1992 EPA guidance):

$$H-UCL = \exp(\bar{y} + 0.5s_y^2 + s_y \frac{H_{1-\alpha}}{\sqrt{n-1}})$$

$\bar{y}$ = mean of $y_i = \ln(x_i)$, s_y = sd of y_i

$H_{1-\alpha}$ = value from H-tables (Land, 1975)

$H_{1-\alpha}$ increases with s_y

  - Even a small increase in sd, s_y can cause an unjustifiable increase in mean and its H-UCL.
  - Often results in unstable and impractically large H-UCL.
  - H-UCL is very sensitive to outliers (small and large).

Lognormal Chebyshev UCL

- A $(1-\alpha)100\%$ lognormal UCL of mean (Singh et al., 1997, 1999, 2000) is given by:

$$UCL = \text{MVUE of mean} +$$

$$\sqrt{((1/\alpha)-1)}\ \text{MVUE of SE of mean}$$

  - Tends to provide a conservative estimate, Chebyshev (MVUE) UCL, especially when sample size is large (e.g., >20).
  - Sensitive to outliers, which may need to be removed.

	A–C	D	E–H	I
1	Data File D:\drive_c\sprfnd02\proucl2003\grice.dat		Variable: Grice.dat	
2				
3	Raw Statistics		Normal Distribution Test	
4	Number of Valid Samples	20	Shapiro-Wilk Test Statisitic	0.9613402
5	Number of Unique Samples	19	Shapiro-Wilk 5% Critical Value	0.905
6	Minimum	40	Data are normal at 5% significance level	
7	Maximum	165		
8	Mean	113.45	95% UCL (Assuming Normal Distribution)	
9	Median	119	Student's-t UCL	127.28788
10	Standard Deviation	35.789553		
11	Variance	1280.8921	Gamma Distribution Test	
12	Coefficient of Variation	0.3154654	A-D Test Statistic	0.414965
13	Skewness	-0.355233	A-D 5% Critical Value	0.7426541
14			K-S Test Statistic	0.1386766
15	Gamma Statistics		K-S 5% Critical Value	0.1939989
16	k hat	8.7992147	Data follow gamma distribution	
17	k star (bias corrected)	7.5126658	at 5% significance level	
18	Theta hat	12.893196		
19	Theta star	15.101164	95% UCLs (Assuming Gamma Distribution)	
20	nu hat	351.96859	Approximate Gamma UCL	130.45122
21	nu star	300.50663	Adjusted Gamma UCL	131.90595
22	Approx.Chi Square Value (.05)	261.34273		
23	Adjusted Level of Significance	0.038	Lognormal Distribution Test	
24	Adjusted Chi Square Value	258.46051	Shapiro-Wilk Test Statisitic	0.9115046
25			Shapiro-Wilk 5% Critical Value	0.905
26	Log-transformed Statistics		Data are lognormal at 5% significance level	
27	Minimum of log data	3.6888795		
28	Maximum of log data	5.1059455	95% UCLs (Assuming Lognormal Distribution)	
29	Mean of log data	4.673464	95% H-UCL	134.72948
30	Standard Deviation of log data	0.3708584	95% Chebyshev (MVUE) UCL	156.44158
31	Variance of log data	0.1375359	97.5% Chebyshev (MVUE) UCL	174.68788
32			99% Chebyshev (MVUE) UCL	210.52921
33				
34			95% Non-parametric UCLs	
35			CLT UCL	126.61341
36			Adj-CLT UCL (Adjusted for skewness)	125.93418

Example 1 (Cont.). Two (2) below detection limit values = 0.05 are added to Grice.dat resulting in a sample of 22.

- Normal Student's-t UCL= 120.63
- Lognormal H-UCL=7144.51
  - Note earlier in Example 1:

    Normal UCL = 127.29, and H-UCL = 134.73
- Not a reasonable behavior of H-UCL

Grice1.dat with 2 ND observations = 0.05

	A	B	C	D	E	F	G	H	I	J
3	Raw Statistics				Normal Distribution Test					
4	Number of Valid Samples			22	Shapiro-Wilk Test Statisitic				0.9253104	
5	Number of Unique Samples			20	Shapiro-Wilk 5% Critical Value				0.911	
6	Minimum			0.05	Data are normal at 5% significance level					
7	Maximum			165						
8	Mean			103.14091	95% UCL (Assuming Normal Distribution)					
9	Median			112	Student's-t UCL				120.62874	
10	Standard Deviation			47.668482						
11	Variance			2272.2842	Gamma Distribution Test					
12	Coefficient of Variation			0.4621685	A-D Test Statistic				3.8974314	
13	Skewness			-0.819751	A-D 5% Critical Value				0.7762745	
14					K-S Test Statistic				0.3346237	
15	Gamma Statistics				K-S 5% Critical Value				0.1916182	
16	k hat			0.8875401	Data do not follow gamma distribution					
17	k star (bias corrected)			0.796815	at 5% significance level					
18	Theta hat			116.20985						
19	Theta star			129.44148	95% UCLs (Assuming Gamma Distribution)					
20	nu hat			39.051766	Approximate Gamma UCL				160.63787	
21	nu star			35.059858	Adjusted Gamma UCL				166.17908	
22	Approx.Chi Square Value (.05)			22.510917						
23	Adjusted Level of Significance			0.0386	Lognormal Distribution Test					
24	Adjusted Chi Square Value			21.760294	Shapiro-Wilk Test Statisitic				0.4652824	
25					Shapiro-Wilk 5% Critical Value				0.911	
26	Log-transformed Statistics				Data not lognormal at 5% significance level					
27	Minimum of log data			-2.995732						
28	Maximum of log data			5.1059455	95% UCLs (Assuming Lognormal Distribution)					
29	Mean of log data			3.9762644	95% H-UCL				7144.5048	
30	Standard Deviation of log data			2.2840274	95% Chebyshev (MVUE) UCL				1911.3177	
31	Variance of log data			5.2167812	97.5% Chebyshev (MVUE) UCL				2514.1904	
32					99% Chebyshev (MVUE) UCL				3698.417	
33										
34					95% Non-parametric UCLs					
35					CLT UCL				119.85748	
36					Adj-CLT UCL (Adjusted for skewness)				117.95959	
37					Mod-t UCL (Adjusted for skewness)				120.33271	
38					Jackknife UCL				120.62874	

Modified – t Statistic UCL

- A $(1-\alpha)100\%$ UCL of mean (Johnson (1978)) is given by:

$$UCL = \overline{x} + \frac{\hat{\mu}_3}{6ns_x^2} + t_{n-1,1-\alpha}\frac{s_x}{\sqrt{n}}$$

$$\hat{\mu}_3 = \frac{n\sum(x_i - \overline{x})^3}{(n-1)(n-2)}$$

  – A Non-parametric method for skewed data sets.
  – This UCL does not provide adequate coverage for moderately to highly skewed data sets (e.g., gamma).
  – Sensitive to outliers, which may need to be removed.

UCL Based Upon Adjusted-CLT

- A $(1-\alpha)100\%$ UCL of mean (to be used when sample size is large) is given by (Chen, JASA 1995):

$$UCL = \overline{x} + [z_\alpha + \hat{k}_3 (\frac{1 + 2z_\alpha^2}{6\sqrt{n}})] \frac{s_x}{\sqrt{n}}$$

$$z_\alpha = \text{upper } \alpha \text{th percentile of SND}$$

$$\hat{k}_3 = \frac{\hat{\mu}_3}{s_x^3}, \text{ skewness coef. (raw data)}$$

  – A Non-parametric method for skewed data sets.
  – UCL does not provide adequate coverage for moderately skewed to highly skewed data (e.g., gamma).
  – Sensitive to outliers, which may need to be removed.

Non-parametric Chebyshev UCL

- A $(1-\alpha)100\%$ non-parametric UCL of mean (Singh et al., 1997, 1999, 2000) is given by:

$$UCL = \overline{x} + \sqrt{((1/\alpha)-1)}\,\frac{s_x}{\sqrt{n}}$$

  - This method tends to provide a conservative but reasonable estimate (providing at least $(1-\alpha)100\%$ coverage) of UCL (= Chebyshev (mean,std)), especially when sample size is large (e.g., >20).
  - Sensitive to outliers, which may need to be removed.

UCL Based Upon Standard Bootstrap Method

- A $(1-\alpha)100\%$ UCL of mean is given by:

$$UCL = \bar{x} + z_{\alpha}\hat{\sigma}_B, \text{ where}$$

$$\hat{\sigma}_{\mathrm{B}} = \sqrt{\frac{\sum(\bar{x}_i - \bar{x}_B)^2}{(N-1)}},$$

N = number of bootstrap samples (e.g., 2000)

$$\bar{x}_B = \sum_{i=1}^{N} \bar{x}_i / N,$$

$\bar{x}_i$ = mean of the ith bootstrap sample

- – This UCL does not provide adequate coverage for skewed models (e.g., gamma).

Bootstrap-t UCL Method

- A $(1-\alpha)100\%$ UCL of mean is given by:

$$UCL = \bar{x} - t_{(\alpha N)} s_x / \sqrt{n}, \text{ where}$$

$$N = \text{ number of bootstrap samples}$$

$$t_i = \sqrt{n}(\bar{x}_i - \bar{x}) / s_{x,i}, i = 1, 2, ..., N$$

$$t_{(\alpha N)} = \text{lower } \alpha \text{th quantile of } t_i, i = 1, 2, ..., N$$

  - Outliers can influence this UCL substantially.
  - For Gamma distribution – provides better coverage than Adj-CLT, Modified-t, standard, BCA bootstrap methods.

UCL Based Upon Hall's Bootstrap

- A $(1-\alpha)100\%$ UCL (Manly, 1997) is:

$$UCL = \bar{x} - W(q_\alpha)s_x$$

$$W(q_\alpha) = 3[(1+\hat{k}_3(q_\alpha - \hat{k}_3/(6n)))^{1/3} - 1]/\hat{k}_3, \text{ where}$$

$$q_\alpha = (\alpha N)th \text{ ordered value of } Q_i(W_i), i = 1, 2, ..., N$$

$$Q_i(W_i) = W_i + \hat{k}_{3i}W_i^2/3 + \hat{k}_{3i}^2W_i^3/27 + \hat{k}_{3i}/(6n)$$

$$W_i = (\bar{x}_i - \bar{x})/s_{x,i}$$

  - Influenced by outliers, which may need to be removed.

UCL Based Upon a Gamma Model

- Chi-square distribution can be used.
- A $(1-\alpha)100\%$ uniformly most accurate UCL of mean is given by (Grice and Bain, 1980):

$$P(\mu_1 \leq 2nk\bar{x} / \chi^2_{2nk}(\alpha)) = 1-\alpha, \text{ where}$$

$\chi^2_{2nk}(\alpha) =$ lower αth percentile of χ^2_{2nk} distribution

- k needs to be estimated from data, therefore coverage is not guaranteed - needs an adjustment.

UCL Based Upon a Gamma Model

- An approximate $(1-\alpha)100\%$ UCL of mean:

$$UCL = 2n\hat{k}\overline{x} / \chi^2_{2n\hat{k}}(\alpha)$$

- An Adjusted $(1-\alpha)100\%$ UCL of mean:

$$UCL = 2n\hat{k}\overline{x} / \chi^2_{2n\hat{k}}(\beta)$$

- Adjusted α are given by β in Table 1.

- Note Chi-square critical values increase with df.

UCL Based Upon Gamma Model

Table 1

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

UCL Based Upon Gamma Model

- Note gamma UCL does not depend upon scale parameter, θ or its MLE.
- Also note that Gamma UCL is the only UCL which does not depend on the sd of data, therefore, outliers have reduced influence on estimation of gamma UCL.

Recommendations in ProUCL

- ProUCL makes recommendations based upon the most appropriate data distribution, associated skewness, and coverage probabilities.
  - ProUCL prints out message(s) about data distribution(s) - (normal, gamma, lognormal, or non-parametric).
  - ProUCL also recommends which 95% UCL(s) to use.
  - It is the user's responsibility to select the most appropriate UCL.
    - » This may require testing and removing of outliers – especially when bootstrap UCLs are recommended.

Table 1
Summary Table for the Computation of a 95% *UCL* of the Unknown Mean, μ_1 of a Gamma Distribution

$\hat{k}$	***Sample Size, n***	***Recommendation***
$\hat{k} \geq 0.5$	For all n	Approximate Gamma 95% *UCL*
$0.1 \leq \hat{k} < 0.5$	For all n	Adjusted Gamma 95% *UCL*
$\hat{k} < 0.1$	n < 15	95% *UCL* Based Upon Bootstrap-t or Hall's Bootstrap Method *
	n ≥ 15	Adjusted Gamma 95% *UCL* if available, otherwise use Approximate Gamma 95% *UCL*

* If bootstrap-t or Hall's bootstrap methods yield erratic, inflated, and unstable *UCL* values (which often happens when outliers are present), the *UCL* of the mean should be computed using adjusted gamma *UCL*.

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

$\hat{\sigma}$	*Sample Size, n*	*Recommendation*
$\hat{\sigma} < 0.5$	For all n	Student's-t, modified-t, or *H-UCL*
$0.5 \le \hat{\sigma} < 1.0$	For all n	*H-UCL*
$1.0 \le \hat{\sigma} < 1.5$	$n < 25$	95% Chebyshev *(MVUE) UCL*
	$n \ge 25$	*H-UCL*
$1.5 \le \hat{\sigma} < 2.0$	$n < 20$	99% Chebyshev *(MVUE) UCL*
	$20 \le n < 50$	95% Chebyshev *(MVUE) UCL*
	$n \ge 50$	*H-UCL*
$2.0 \le \hat{\sigma} < 2.5$	$n < 20$	99% Chebyshev *(MVUE) UCL*
	$20 \le n < 50$	97.5% Chebyshev *(MVUE) UCL*
	$50 \le n < 70$	95% Chebyshev *(MVUE) UCL*
	$n \ge 70$	*H-UCL*
$2.5 \le \hat{\sigma} < 3.0$	$n < 30$	Larger of (99% Chebyshev *(MVUE) UCL*, 99% Chebyshev(Mean, Sd))
	$30 \le n < 70$	97.5% Chebyshev *(MVUE) UCL*
	$70 \le n < 100$	95% Chebyshev *(MVUE) UCL*
	$n \ge 100$	*H-UCL*

TABLE 2 - CONTINUED

$1.5 \le \hat{\sigma} < 2.0$	$n < 20$	99% Chebyshev *(MVUE) UCL*
	$20 \le n < 50$	95% Chebyshev *(MVUE) UCL*
	$n \ge 50$	*H-UCL*
$2.0 \le \hat{\sigma} < 2.5$	$n < 20$	99% Chebyshev *(MVUE) UCL*
	$20 \le n < 50$	97.5% Chebyshev *(MVUE) UCL*
	$50 \le n < 70$	95% Chebyshev *(MVUE) UCL*
	$n \ge 70$	*H-UCL*
$2.5 \le \hat{\sigma} < 3.0$	$n < 30$	Larger of (99% Chebyshev *(MVUE) UCL*, 99% Chebyshev(Mean, Sd))
	$30 \le n < 70$	97.5% Chebyshev *(MVUE) UCL*
	$70 \le n < 100$	95% Chebyshev *(MVUE) UCL*
	$n \ge 100$	*H-UCL*
$3.0 \le \hat{\sigma} \le 3.5$	$n < 15$	Hall's bootstrap method *
	$15 \le n < 50$	Larger of (99% Chebyshev *(MVUE) UCL*, 99% Chebyshev(Mean, Sd))
	$50 \le n < 100$	97.5% Chebyshev *(MVUE) UCL*
	$100 \le n < 150$	95% Chebyshev *(MVUE) UCL*
	$n \ge 150$	*H-UCL*
$\hat{\sigma} > 3.5$	For all n	Use non-parametric methods *

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

Table 3
Summary Table for the Computation of a 95% *UCL* of the Unknown Mean, μ_1 of a Skewed Non-parametric Distribution with all Positive Values, Where $\hat{\sigma}$ is the Sd of Log-transformed Data

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

* If the Hall's bootstrap method yields an erratic and unstable *UCL* value (e.g., this tends to happen when outliers are present), the EPC term may be computed using the 99% Chebyshev (Mean, Sd) UCL.

Maximum Value Should Not be Used to Estimate EPC Term

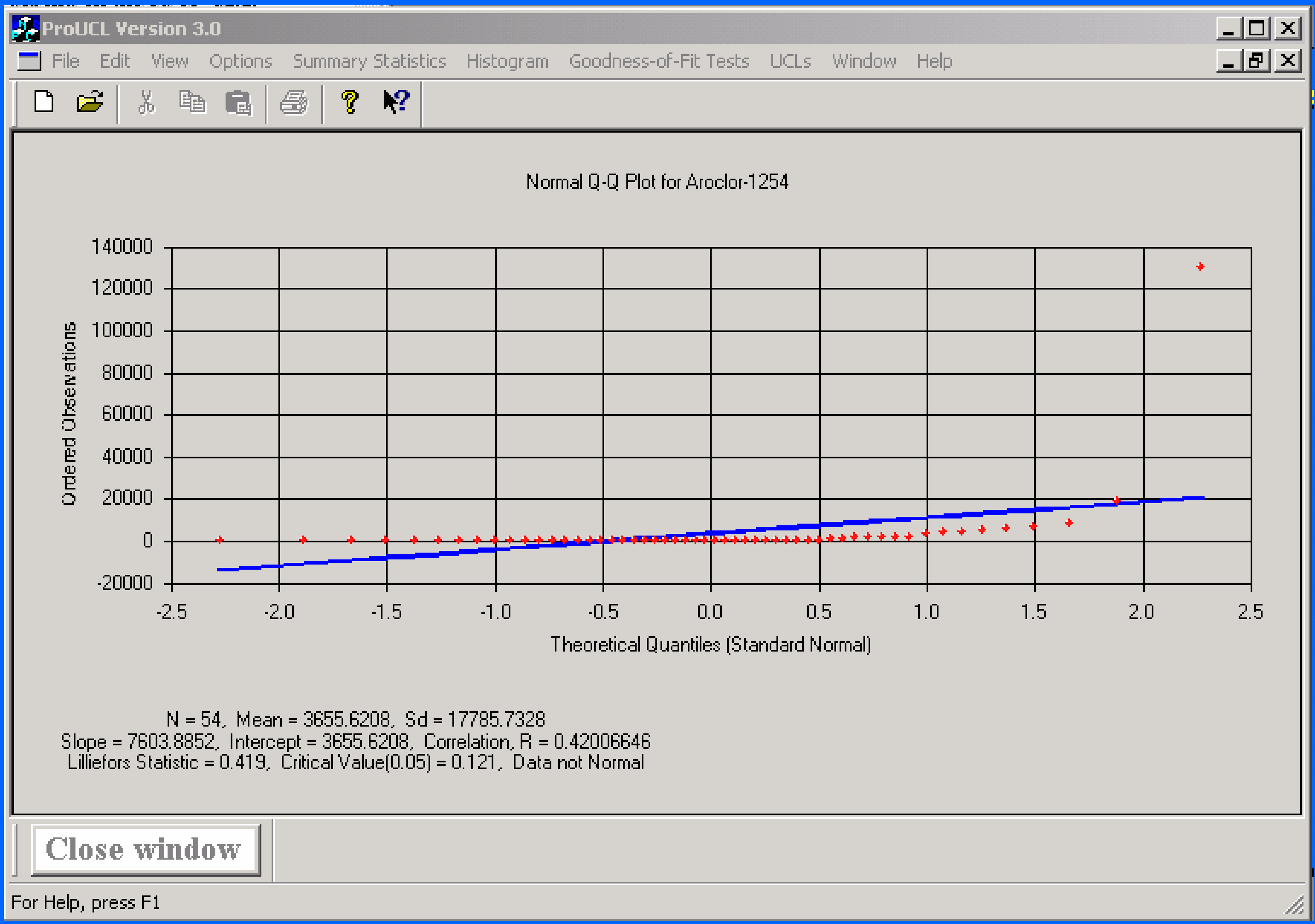
- The EPC term represents average exposure over an exposure area (EA) during a long period of time.

- Therefore, the Max value should not be used as an estimate of EPC term – **ignores most info in data set**.

- ProUCL displays a warning message when the recommended 95% UCL (e.g., H-UCL, Hall's bootstrap UCL etc.) exceeds the Max value.

- For such cases, alternative UCL computation method such as the Chebyshev (mean, Sd) should be used.

Example 2: Aroclor 1254 – with Outlier = 130,000



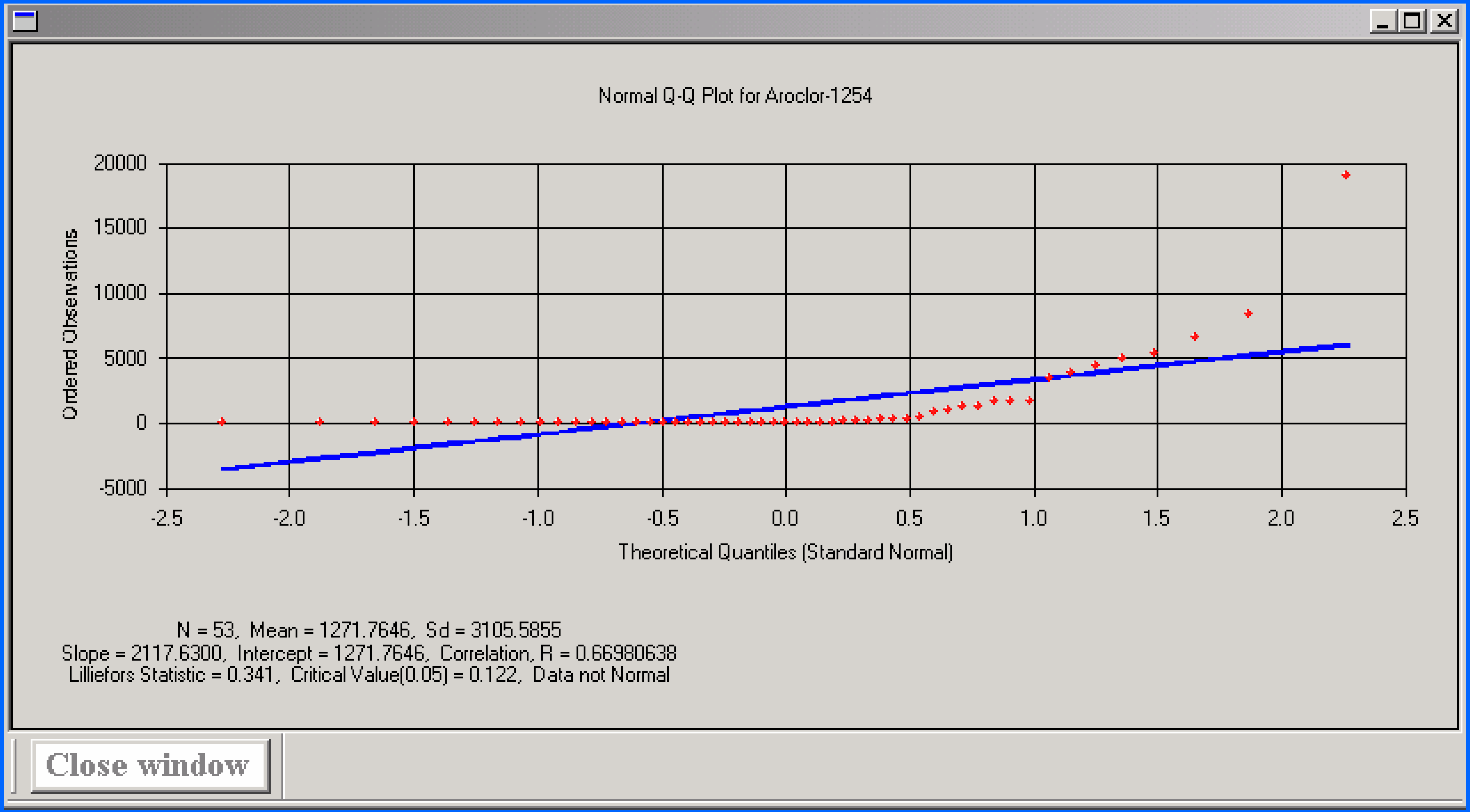
Example 2: Aroclor 1254 – with Outlier = 130,000

	A	B	C	D	E	F	G	H	I	J
1	Data File	D:\drive_c\SPRFND95\cornell\A-54.dat				Variable:	Aroclor-1254			
2										
3	Raw Statistics				Normal Distribution Test					
4	Number of Valid Samples			54	Lilliefors Test Statisitic				0.4185768	
5	Number of Unique Samples			50	Lilliefors 5% Critical Value				0.1205693	
6	Minimum			0.017	Data not normal at 5% significance level					
7	Maximum			130000						
8	Mean			3655.6208	95% UCL (Assuming Normal Distribution)					
9	Median			42.5	Student's-t UCL				7707.5365	
10	Standard Deviation			17785.733						
11	Variance			3.2E+008	Gamma Distribution Test					
12	Coefficient of Variation			4.8653112	A-D Test Statistic				2.183231	
13	Skewness			7.0302526	A-D 5% Critical Value				0.9380814	
14					K-S Test Statistic				0.1666673	
15	Gamma Statistics				K-S 5% Critical Value				0.1357658	
16	k hat			0.1573544	Data do not follow gamma distribution					
17	k star (bias corrected)			0.1609582	at 5% significance level					
18	Theta hat			23231.765						
19	Theta star			22711.618	95% UCLs (Assuming Gamma Distribution)					
20	nu hat			16.994277	Approximate Gamma UCL				7103.8453	
21	nu star			17.383484	Adjusted Gamma UCL				7238.6552	
22	Approx.Chi Square Value (.05)			8.9454967						
23	Adjusted Level of Significance			0.0455556	Lognormal Distribution Test					
24	Adjusted Chi Square Value			8.7788993	Lilliefors Test Statisitic				0.0888924	
25					Lilliefors 5% Critical Value				0.1205693	
26	Log-transformed Statistics				Data are lognormal at 5% significance level					
27	Minimum of log data			-4.074542						
28	Maximum of log data			11.77529	95% UCLs (Assuming Lognormal Distribution)					
29	Mean of log data			3.3537198	95% H-UCL				9176544.8	
30	Standard Deviation of log data			4.2015474	95% Chebyshev (MVUE) UCL				293727.75	
31	Variance of log data			17.653001	97.5% Chebyshev (MVUE) UCL				394494.73	
32					99% Chebyshev (MVUE) UCL				592431.96	

Example 2: Aroclor 1254 – with Outlier = 130,000

19	Theta star	22711.618	95% UCLs (Assuming Gamma Distribution)	
20	nu hat	16.994277	Approximate Gamma UCL	7103.8453
21	nu star	17.383484	Adjusted Gamma UCL	7238.6552
22	Approx.Chi Square Value (.05)	8.9454967		
23	Adjusted Level of Significance	0.0455556	Lognormal Distribution Test	
24	Adjusted Chi Square Value	8.7788993	Lilliefors Test Statisitic	0.0888924
25			Lilliefors 5% Critical Value	0.1205693
26	Log-transformed Statistics		Data are lognormal at 5% significance level	
27	Minimum of log data	-4.074542		
28	Maximum of log data	11.77529	95% UCLs (Assuming Lognormal Distribution)	
29	Mean of log data	3.3537198	95% H-UCL	9176544.8
30	Standard Deviation of log data	4.2015474	95% Chebyshev (MVUE) UCL	293727.75
31	Variance of log data	17.653001	97.5% Chebyshev (MVUE) UCL	394494.73
32			99% Chebyshev (MVUE) UCL	592431.96
33				
34			95% Non-parametric UCLs	
35			CLT UCL	7636.7121
36			Adj-CLT UCL (Adjusted for skewness)	10110.881
37			Mod-t UCL (Adjusted for skewness)	8093.4568
38			Jackknife UCL	7707.5365
39			Standard Bootstrap UCL	7522.7743
40			Bootstrap-t UCL	31956.2
41	RECOMMENDATION		Hall's Bootstrap UCL	21720.324
42	Data are lognormal (0.05)		Percentile Bootstrap UCL	8348.5959
43			BCA Bootstrap UCL	8131.3965
44	Use Hall's Bootstrap UCL		95% Chebyshev (Mean, Sd) UCL	14205.602
45			97.5% Chebyshev (Mean, Sd) UCL	18770.587
46	In case Hall's Bootstrap method yields		99% Chebyshev (Mean, Sd) UCL	27737.617
47	an erratic, unreasonably large UCL value,			
48	use 99% Chebyshev (Mean, Sd) UCL			
49				

Example 2: Aroclor 1254 – Without Outlier = 130,000



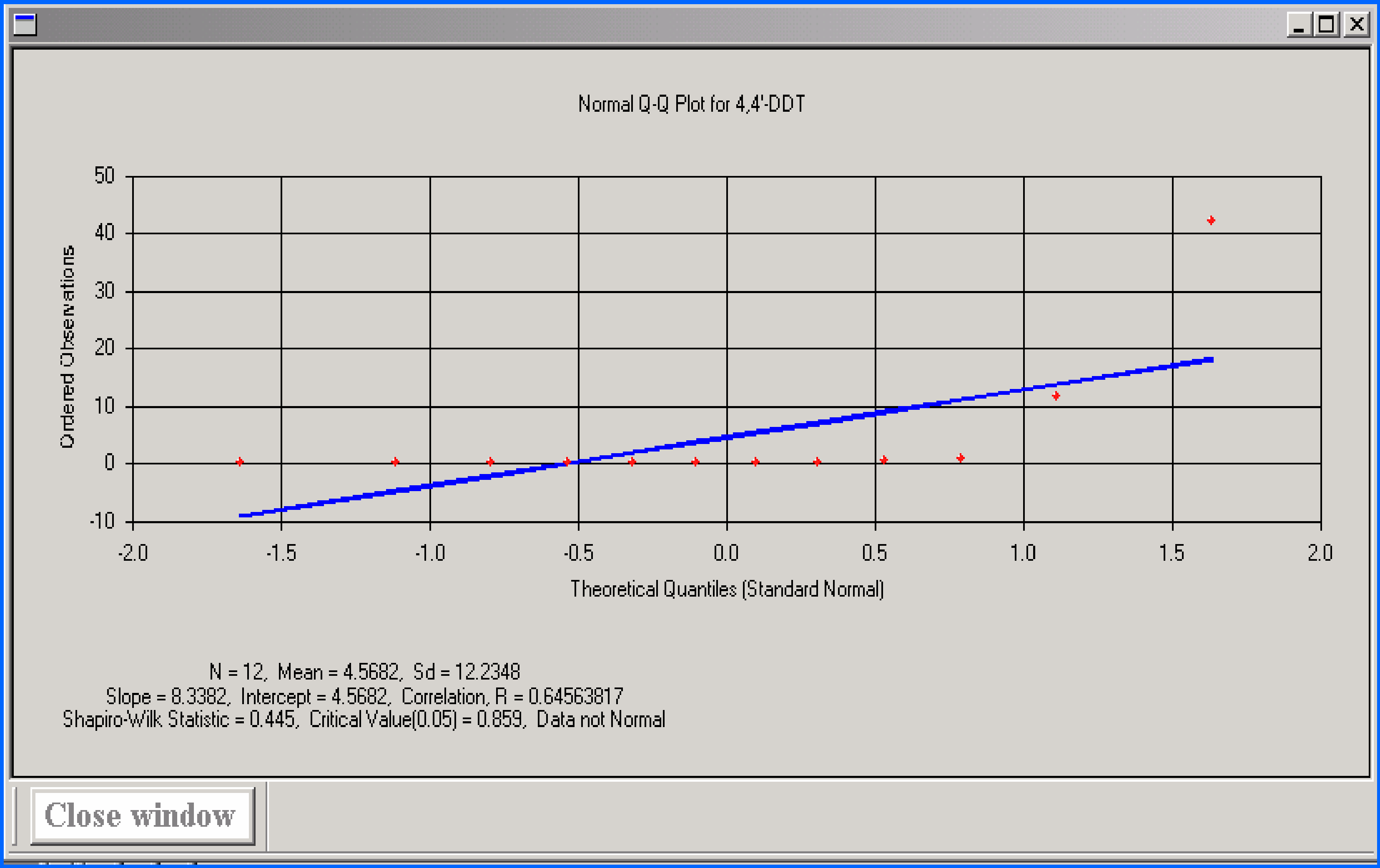
Example 2: Aroclor 1254 – Without Outlier = 130,000

	A	B	C	D	E	F	G	H	I
1	Data File	D:\drive_c\sprfnd02\proucl2004\A-53.dat				Variable:	Aroclor-1254		
2									
3	Raw Statistics				Normal Distribution Test				
4	Number of Valid Samples			53	Lilliefors Test Statisitic				0.3410852
5	Number of Unique Samples			49	Lilliefors 5% Critical Value				0.1217015
6	Minimum			0.017	Data not normal at 5% significance level				
7	Maximum			19000					
8	Mean			1271.7646	95% UCL (Assuming Normal Distribution)				
9	Median			41	Student's-t UCL				1986.1617
10	Standard Deviation			3105.5855					
11	Variance			9644661.5	Gamma Distribution Test				
12	Coefficient of Variation			2.44195	A-D Test Statistic				1.132286
13	Skewness			4.1242405	A-D 5% Critical Value				0.9164426
14					K-S Test Statistic				0.1555879
15	Gamma Statistics				K-S 5% Critical Value				0.1358339
16	k hat			0.1879527	Data do not follow gamma distribution				
17	k star (bias corrected)			0.1898925	at 5% significance level				
18	Theta hat			6766.4064					
19	Theta star			6697.2864	95% UCLs (Assuming Gamma Distribution)				
20	nu hat			19.92299	Approximate Gamma UCL				2338.9686
21	nu star			20.128607	Adjusted Gamma UCL				2380.2993
22	Approx.Chi Square Value (.05)			10.944503					
23	Adjusted Level of Significance			0.0454717	Lognormal Distribution Test				
24	Adjusted Chi Square Value			10.754467	Lilliefors Test Statisitic				0.0935493
25					Lilliefors 5% Critical Value				0.1217015
26	Log-transformed Statistics				Data are lognormal at 5% significance level				

Example 2: Aroclor 1254 – Without Outlier = 130,000

17	k star (bias corrected)	0.1898925	at 5% significance level	
18	Theta hat	6766.4064		
19	Theta star	6697.2864	95% UCLs (Assuming Gamma Distribution)	
20	nu hat	19.92299	Approximate Gamma UCL	2338.9686
21	nu star	20.128607	Adjusted Gamma UCL	2380.2993
22	Approx.Chi Square Value (.05)	10.944503		
23	Adjusted Level of Significance	0.0454717	Lognormal Distribution Test	
24	Adjusted Chi Square Value	10.754467	Lilliefors Test Statisitic	0.0935493
25			Lilliefors 5% Critical Value	0.1217015
26	Log-transformed Statistics		Data are lognormal at 5% significance level	
27	Minimum of log data	-4.074542		
28	Maximum of log data	9.8521943	95% UCLs (Assuming Lognormal Distribution)	
29	Mean of log data	3.1948223	95% H-UCL	3876782.2
30	Standard Deviation of log data	4.0746591	95% Chebyshev (MVUE) UCL	161169.26
31	Variance of log data	16.602847	97.5% Chebyshev (MVUE) UCL	216268.24
32			99% Chebyshev (MVUE) UCL	324499.54
33				
34			95% Non-parametric UCLs	
35			CLT UCL	1973.4344
36			Adj-CLT UCL (Adjusted for skewness)	2231.6557
37			Mod-t UCL (Adjusted for skewness)	2026.439
38			Jackknife UCL	1986.1617
39			Standard Bootstrap UCL	1985.9029
40			Bootstrap-t UCL	2727.3543
41	RECOMMENDATION		Hall's Bootstrap UCL	4682.1017
42	Data are lognormal (0.05)		Percentile Bootstrap UCL	2024.076
43			BCA Bootstrap UCL	1858.5446
44	Use Hall's Bootstrap UCL		95% Chebyshev (Mean, Sd) UCL	3131.2054
45			97.5% Chebyshev (Mean, Sd) UCL	3935.7869
46	In case Hall's Bootstrap method yields		99% Chebyshev (Mean, Sd) UCL	5516.2315
47	an erratic, unreasonably large UCL value,			
48	use 99% Chebyshev (Mean, Sd) UCL			

Example 3: 4'4 DDT Data, n=12



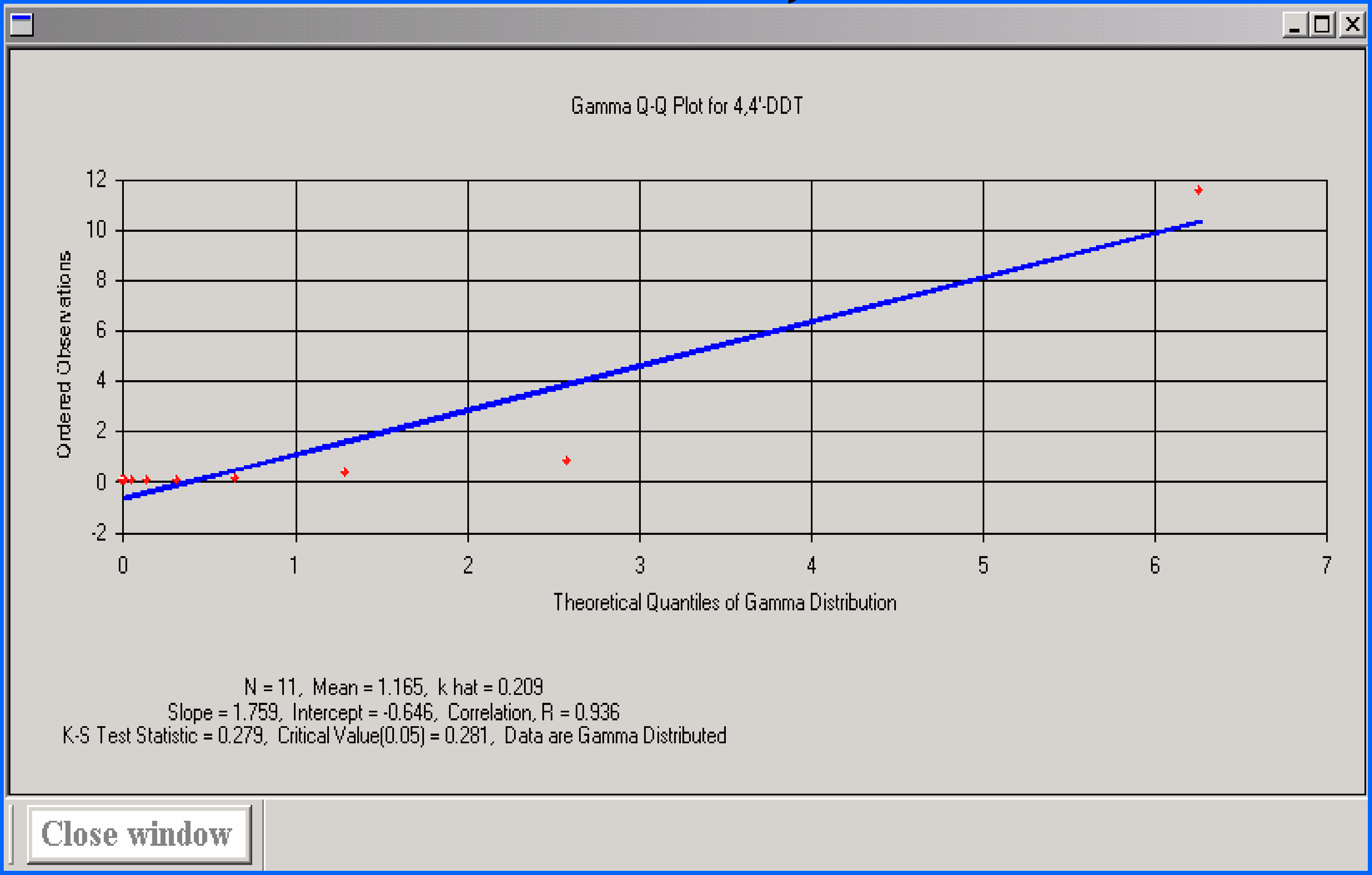
Example 3: 4'4 DDT Data, n=12

	A	B	C	D	E	F	G	H	I
1	Data File	D:\drive_c\sprfnd02\proucl2004\DDT-12.dat				Variable:	4,4'-DDT		
2									
3	Raw Statistics				Normal Distribution Test				
4	Number of Valid Samples			12	Shapiro-Wilk Test Statisitic				0.4449076
5	Number of Unique Samples			12	Shapiro-Wilk 5% Critical Value				0.859
6	Minimum			0.00185	Data not normal at 5% significance level				
7	Maximum			42					
8	Mean			4.5682083	95% UCL (Assuming Normal Distribution)				
9	Median			0.026	Student's-t UCL				10.911082
10	Standard Deviation			12.234838					
11	Variance			149.69126	Gamma Distribution Test				
12	Coefficient of Variation			2.6782574	A-D Test Statistic				1.2597627
13	Skewness			3.0914053	A-D 5% Critical Value				0.8736293
14					K-S Test Statistic				0.2756245
15	Gamma Statistics				K-S 5% Critical Value				0.2720702
16	k hat			0.1757525	Data do not follow gamma distribution				
17	k star (bias corrected)			0.1873699	at 5% significance level				
18	Theta hat			25.99228					
19	Theta star			24.38069	95% UCLs (Assuming Gamma Distribution)				
20	nu hat			4.2180601	Approximate Gamma UCL				22.162437
21	nu star			4.4968784	Adjusted Gamma UCL				28.962989
22	Approx.Chi Square Value (.05)			0.9269142					
23	Adjusted Level of Significance			0.02896	Lognormal Distribution Test				
24	Adjusted Chi Square Value			0.7092734	Shapiro-Wilk Test Statisitic				0.8992571
25					Shapiro-Wilk 5% Critical Value				0.859
26	Log-transformed Statistics				Data are lognormal at 5% significance level				

Example 3: 4'4 DDT Data, n=12

26	Log-transformed Statistics		Data are lognormal at 5% significance level	
27	Minimum of log data	-6.29257		
28	Maximum of log data	3.7376696	95% UCLs (Assuming Lognormal Distribution)	
29	Mean of log data	-2.752233	95% H-UCL	66699.737
30	Standard Deviation of log data	3.3727967	95% Chebyshev (MVUE) UCL	17.795819
31	Variance of log data	11.375758	97.5% Chebyshev (MVUE) UCL	23.915811
32			99% Chebyshev (MVUE) UCL	35.937349
33				
34			95% Non-parametric UCLs	
35			CLT UCL	10.377656
36			Adj-CLT UCL (Adjusted for skewness)	13.745511
37			Mod-t UCL (Adjusted for skewness)	11.436399
38			Jackknife UCL	10.911082
39			Standard Bootstrap UCL	10.282371
40			Bootstrap-t UCL	232.03236
41	RECOMMENDATION		Hall's Bootstrap UCL	189.66219
42	Data are lognormal (0.05)		Percentile Bootstrap UCL	10.543208
43			BCA Bootstrap UCL	10.582163
44	Use Hall's Bootstrap UCL		95% Chebyshev (Mean, Sd) UCL	19.963375
45			97.5% Chebyshev (Mean, Sd) UCL	26.624876
46			99% Chebyshev (Mean, Sd) UCL	39.710104
47	Recommended UCL exceeds the maximum observation			
48				
49	In case Hall's Bootstrap method yields			
50	an erratic, unreasonably large UCL value,			
51	use 99% Chebyshev (Mean, Sd) UCL			

Example 3: 4'4 DDT Data, without outlier = 42, n=11



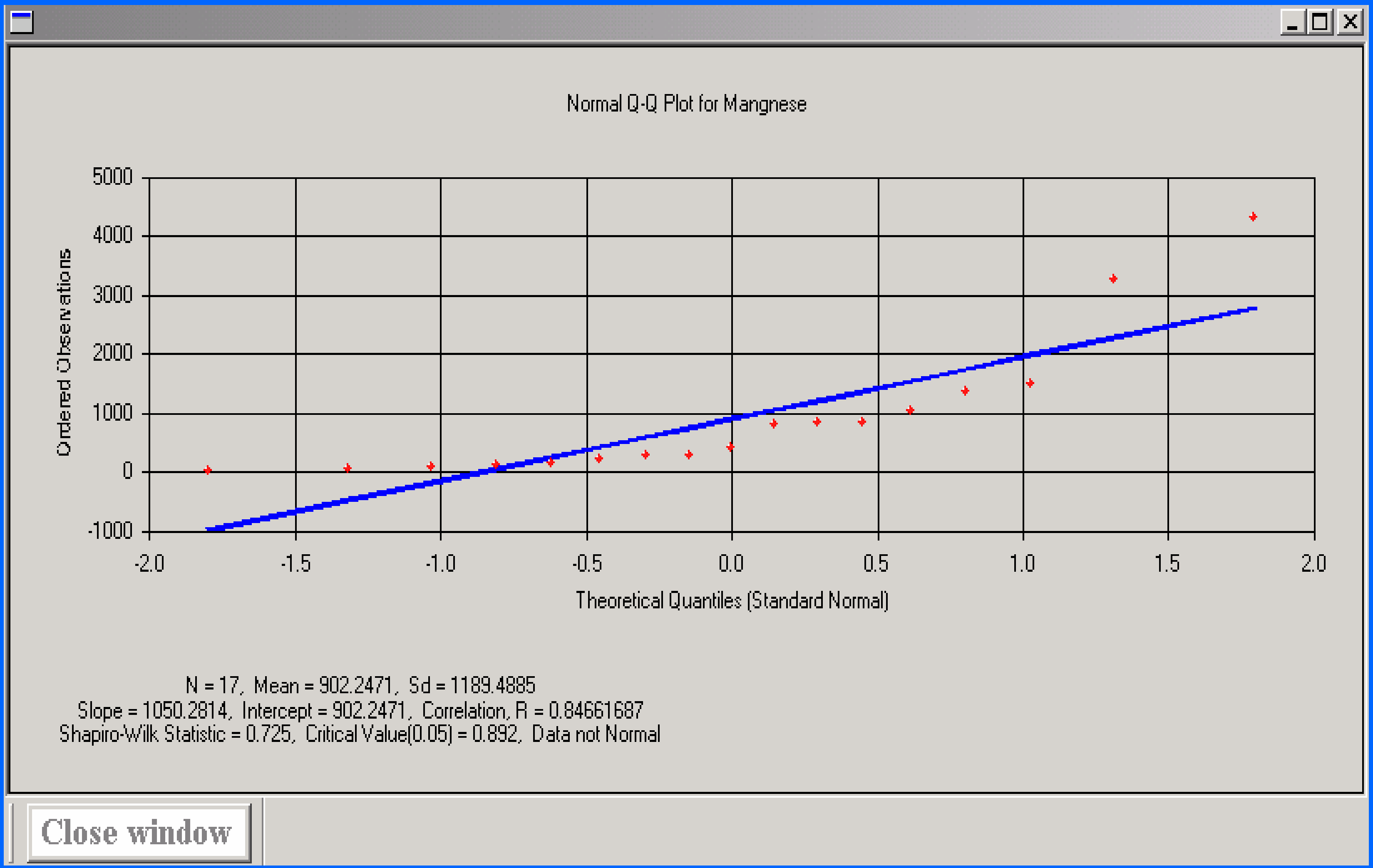
Example 3: 4’4 DDT Data, without outlier = 42, n=11

	A	B	C	D	E	F	G	H	I	J
1	Data File	D:\drive_c\sprfnd02\proucl2004\DDT-11.dat			Variable:	4,4'-DDT				
2										
3	Raw Statistics				Normal Distribution Test					
4	Number of Valid Samples			11	Shapiro-Wilk Test Statisitic				0.3911878	
5	Number of Unique Samples			11	Shapiro-Wilk 5% Critical Value				0.85	
6	Minimum			0.00185	Data not normal at 5% significance level					
7	Maximum			11.5						
8	Mean			1.1653182	95% UCL (Assuming Normal Distribution)					
9	Median			0.0215	Student's-t UCL				3.0432343	
10	Standard Deviation			3.436401						
11	Variance			11.808852	Gamma Distribution Test					
12	Coefficient of Variation			2.948895	A-D Test Statistic				1.2125145	
13	Skewness			3.2865534	A-D 5% Critical Value				0.8564558	
14					K-S Test Statistic				0.2787718	
15	Gamma Statistics				K-S 5% Critical Value				0.2813882	
16	k hat			0.2090901	Data follow approximate gamma distibution					
17	k star (bias corrected)			0.2126716	at 5% significance level					
18	Theta hat			5.5732837						
19	Theta star			5.4794267	95% UCLs (Assuming Gamma Distribution)					
20	nu hat			4.5999812	Approximate Gamma UCL				5.4175272	
21	nu star			4.6787742	Adjusted Gamma UCL				7.1607382	
22	Approx.Chi Square Value (.05)			1.0064113						
23	Adjusted Level of Significance			0.02783	Lognormal Distribution Test					
24	Adjusted Chi Square Value			0.7614104	Shapiro-Wilk Test Statisitic				0.9099724	
25					Shapiro-Wilk 5% Critical Value				0.85	
26	Log-transformed Statistics				Data are lognormal at 5% significance level					
27	Minimum of log data			-6.29257						

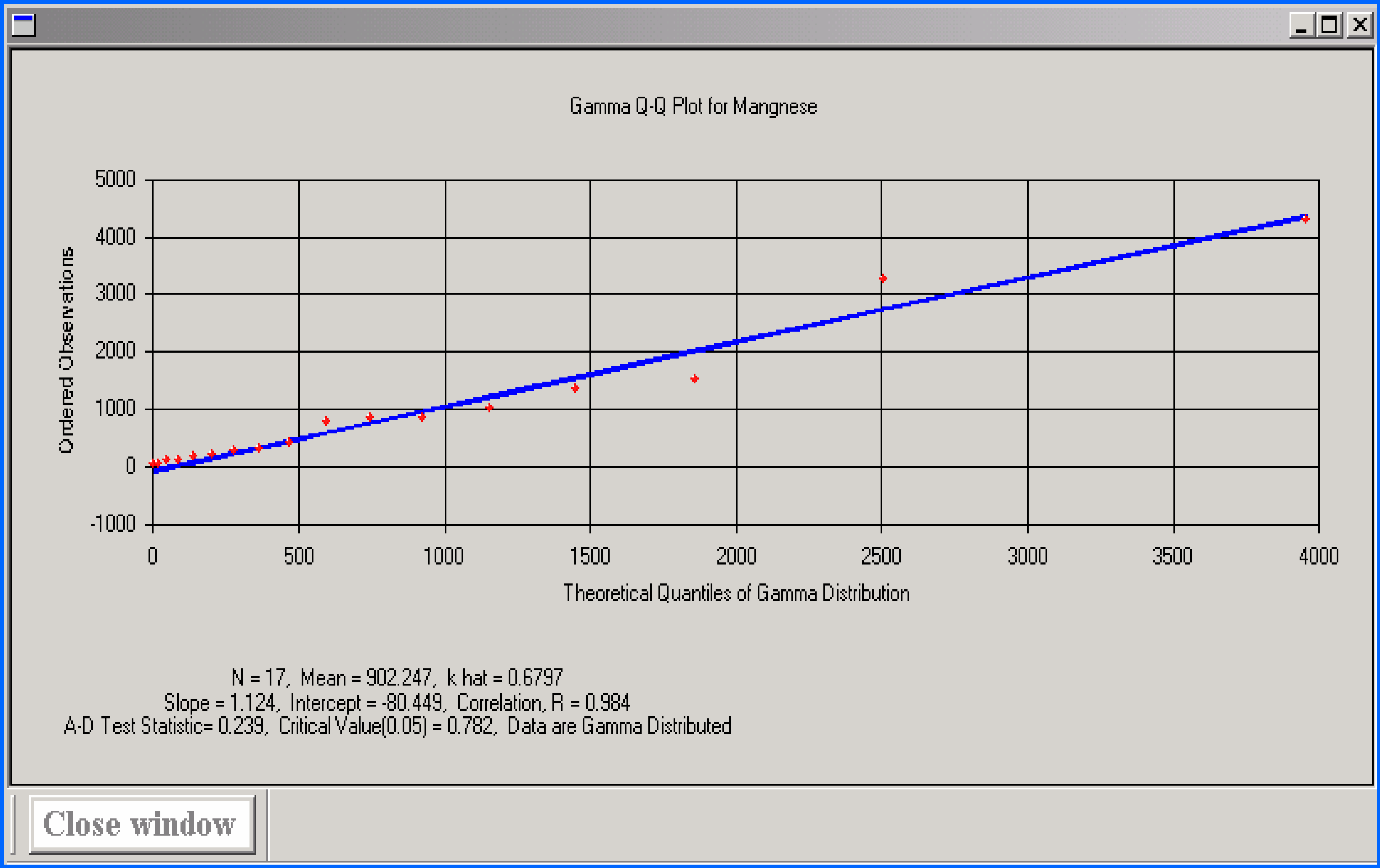
Example 3: 4'4 DDT Data, without outlier = 42, n=11

18	Theta hat	5.5732837		
19	Theta star	5.4794267	95% UCLs (Assuming Gamma Distribution)	
20	nu hat	4.5999812	Approximate Gamma UCL	5.4175272
21	nu star	4.6787742	Adjusted Gamma UCL	7.1607382
22	Approx.Chi Square Value (.05)	1.0064113		
23	Adjusted Level of Significance	0.02783	Lognormal Distribution Test	
24	Adjusted Chi Square Value	0.7614104	Shapiro-Wilk Test Statisitic	0.9099724
25			Shapiro-Wilk 5% Critical Value	0.85
26	Log-transformed Statistics		Data are lognormal at 5% significance level	
27	Minimum of log data	-6.29257		
28	Maximum of log data	2.442347	95% UCLs (Assuming Lognormal Distribution)	
29	Mean of log data	-3.342224	95% H-UCL	1014.1227
30	Standard Deviation of log data	2.8139921	95% Chebyshev (MVUE) UCL	3.0095789
31	Variance of log data	7.9185516	97.5% Chebyshev (MVUE) UCL	4.026235
32			99% Chebyshev (MVUE) UCL	6.023259
33				
34			95% Non-parametric UCLs	
35			CLT UCL	2.8695739
36			Adj-CLT UCL (Adjusted for skewness)	3.9666386
37			Mod-t UCL (Adjusted for skewness)	3.2143542
38			Jackknife UCL	3.0432343
39			Standard Bootstrap UCL	2.7967694
40			Bootstrap-t UCL	37.870738
41	RECOMMENDATION		Hall's Bootstrap UCL	35.520978
42	Assuming gamma distribution (0.05)		Percentile Bootstrap UCL	3.2183591
43			BCA Bootstrap UCL	3.1806727
44	Use Adjusted Gamma UCL		95% Chebyshev (Mean, Sd) UCL	5.6816339
45			97.5% Chebyshev (Mean, Sd) UCL	7.6358473
46			99% Chebyshev (Mean, Sd) UCL	11.474521

Example 4. Mn at NCBC Site



Example 4. Mn at NCBC Site



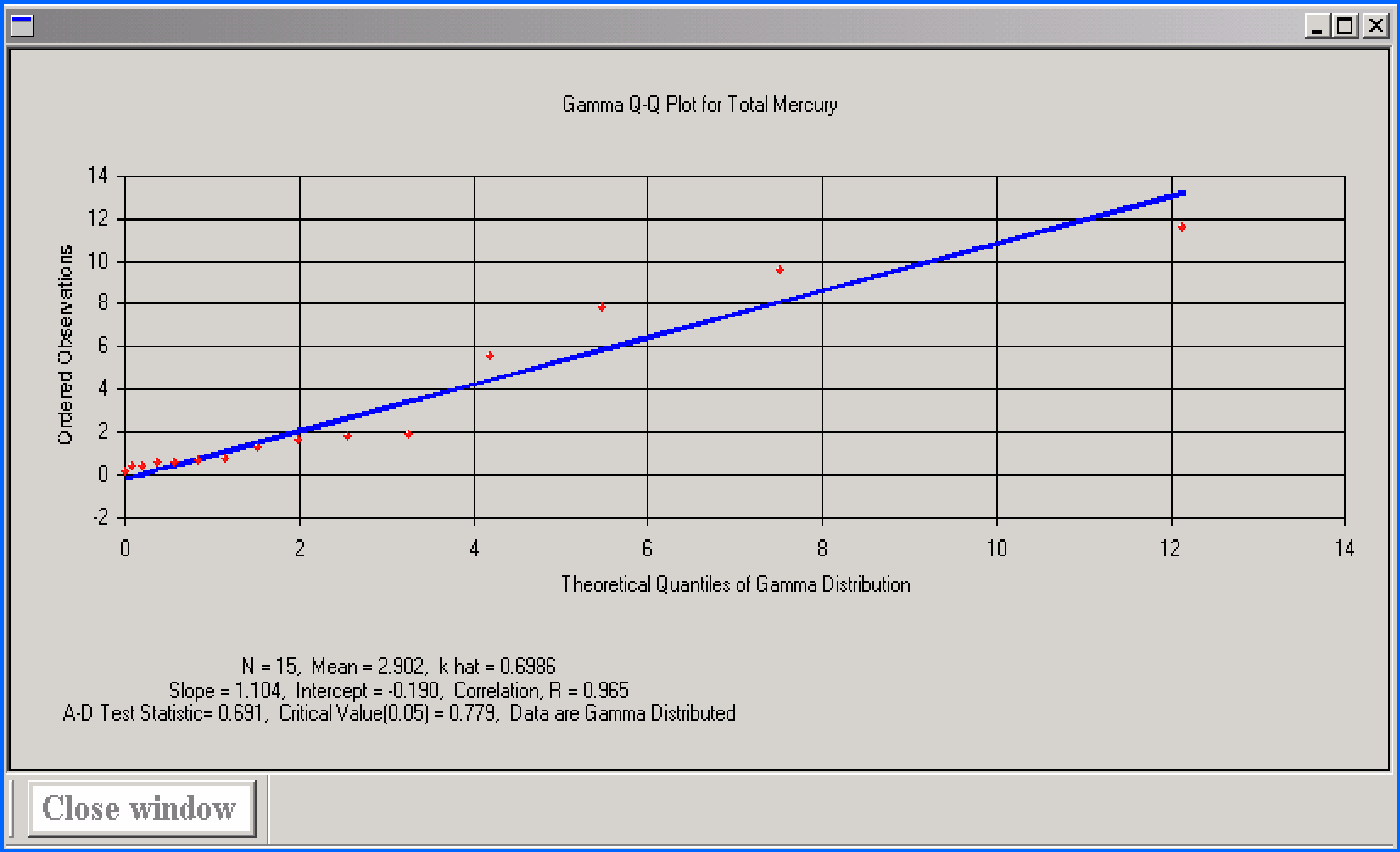
Example 4. Mn at NCBC Site

	A	B	C	D	E	F	G	H	I
1	Data File	D:\drive_c\sprfnd02\proucl2004\NSBC-MAN(				Variable:	Mangnese		
2									
3	Raw Statistics				Normal Distribution Test				
4	Number of Valid Samples			17	Shapiro-Wilk Test Statisitic				0.7247918
5	Number of Unique Samples			17	Shapiro-Wilk 5% Critical Value				0.892
6	Minimum			15.8	Data not normal at 5% significance level				
7	Maximum			4300					
8	Mean			902.24706	95% UCL (Assuming Normal Distribution)				
9	Median			390	Student's-t UCL				1405.9228
10	Standard Deviation			1189.4885					
11	Variance			1414882.9	Gamma Distribution Test				
12	Coefficient of Variation			1.3183623	A-D Test Statistic				0.2390536
13	Skewness			2.0455011	A-D 5% Critical Value				0.7824581
14					K-S Test Statistic				0.1168327
15	Gamma Statistics				K-S 5% Critical Value				0.2182914
16	k hat			0.6797386	Data follow gamma distribution				
17	k star (bias corrected)			0.5990004	at 5% significance level				
18	Theta hat			1327.3442					
19	Theta star			1506.2545	95% UCLs (Assuming Gamma Distribution)				
20	nu hat			23.111112	Approximate Gamma UCL				1652.4595
21	nu star			20.366014	Adjusted Gamma UCL				1765.4423
22	Approx.Chi Square Value (.05)			11.119895					
23	Adjusted Level of Significance			0.03461	Lognormal Distribution Test				
24	Adjusted Chi Square Value			10.408256	Shapiro-Wilk Test Statisitic				0.9688235
25					Shapiro-Wilk 5% Critical Value				0.892
26	Log-transformed Statistics				Data are lognormal at 5% significance level				

Example 4. Mn at NCBC Site

19	Theta star	1506.2545	95% UCLs (Assuming Gamma Distribution)	
20	nu hat	23.111112	Approximate Gamma UCL	1652.4595
21	nu star	20.366014	Adjusted Gamma UCL	1765.4423
22	Approx.Chi Square Value (.05)	11.119895		
23	Adjusted Level of Significance	0.03461	Lognormal Distribution Test	
24	Adjusted Chi Square Value	10.408256	Shapiro-Wilk Test Statisitic	0.9688235
25			Shapiro-Wilk 5% Critical Value	0.892
26	Log-transformed Statistics		Data are lognormal at 5% significance level	
27	Minimum of log data	2.7600099		
28	Maximum of log data	8.3663703	95% UCLs (Assuming Lognormal Distribution)	
29	Mean of log data	5.9121322	95% H-UCL	5239.7026
30	Standard Deviation of log data	1.5676658	95% Chebyshev (MVUE) UCL	3237.4892
31	Variance of log data	2.4575762	97.5% Chebyshev (MVUE) UCL	4161.9824
32			99% Chebyshev (MVUE) UCL	5977.9704
33				
34			95% Non-parametric UCLs	
35			CLT UCL	1376.7764
36			Adj-CLT UCL (Adjusted for skewness)	1529.7059
37			Mod-t UCL (Adjusted for skewness)	1429.7768
38			Jackknife UCL	1405.9228
39			Standard Bootstrap UCL	1366.0647
40			Bootstrap-t UCL	1980.3313
41	RECOMMENDATION		Hall's Bootstrap UCL	3827.6703
42	Data follow gamma distribution (0.05)		Percentile Bootstrap UCL	1372.7059
43			BCA Bootstrap UCL	1717.4824
44	Use Approximate Gamma UCL		95% Chebyshev (Mean, Sd) UCL	2159.7604
45			97.5% Chebyshev (Mean, Sd) UCL	2703.8874
46			99% Chebyshev (Mean, Sd) UCL	3772.7195

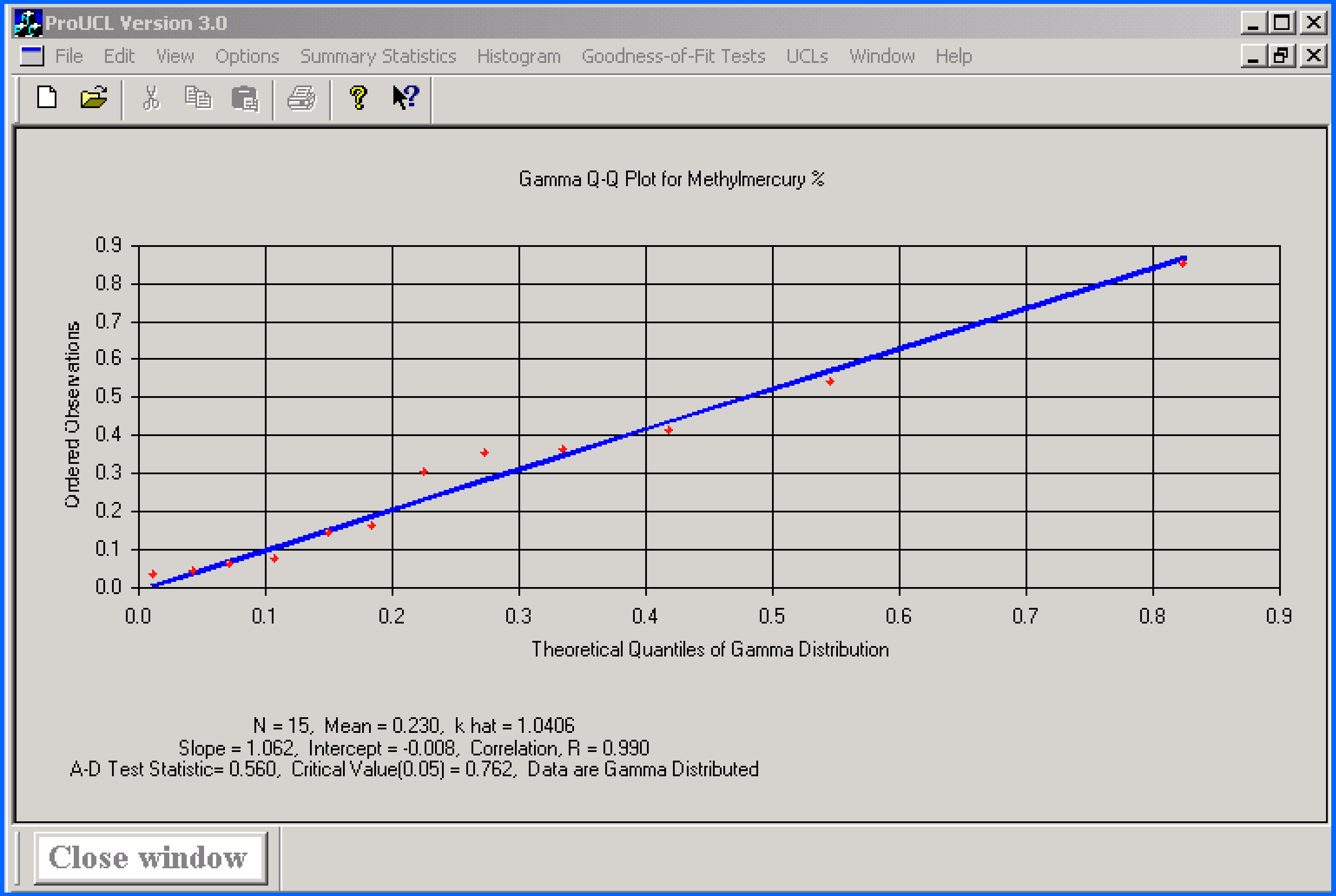
Mercury Data From Nepera Site



Mercury Data From Nepera Site

1	Data File	D:\drive_c\sprfnd02\nepera\Table2.xls	Variable:	Total Mercury
2				
3	Raw Statistics		Normal Distribution Test	
4	Number of Valid Samples	15	Shapiro-Wilk Test Statisitic	0.7288956
5	Number of Unique Samples	15	Shapiro-Wilk 5% Critical Value	0.881
6	Minimum	0.074	Data not normal at 5% significance level	
7	Maximum	11.548		
8	Mean	2.9022	95% UCL (Assuming Normal Distribution)	
9	Median	1.179	Student's-t UCL	4.6180086
10	Standard Deviation	3.7729294		
11	Variance	14.234996	Gamma Distribution Test	
12	Coefficient of Variation	1.3000239	A-D Test Statistic	0.6905374
13	Skewness	1.453894	A-D 5% Critical Value	0.7787848
14			K-S Test Statistic	0.2173563
15	Gamma Statistics		K-S 5% Critical Value	0.2309578
16	k hat	0.6986189	Data follow gamma distribution	
17	k star (bias corrected)	0.6033396	at 5% significance level	
18	Theta hat	4.1541959		
19	Theta star	4.8102262	95% UCLs (Assuming Gamma Distribution)	
20	nu hat	20.958568	Approximate Gamma UCL	5.5517622
21	nu star	18.100188	Adjusted Gamma UCL	6.0371894
22	Approx.Chi Square Value (.05)	9.4619264		
23	Adjusted Level of Significance	0.03235	Lognormal Distribution Test	
24	Adjusted Chi Square Value	8.7011293	Shapiro-Wilk Test Statisitic	0.951452
25			Shapiro-Wilk 5% Critical Value	0.881
26	Log-transformed Statistics		Data are lognormal at 5% significance level	
27	Minimum of log data	-2.60369		
28	Maximum of log data	2.4465123	95% UCLs (Assuming Lognormal Distribution)	
29	Mean of log data	0.2001752	95% H-UCL	13.354141
30	Standard Deviation of log data	1.4423611	95% Chebyshev (MVUE) UCL	8.7170048
31	Variance of log data	2.0804054	97.5% Chebyshev (MVUE) UCL	11.163643
32			99% Chebyshev (MVUE) UCL	15.96959
33				

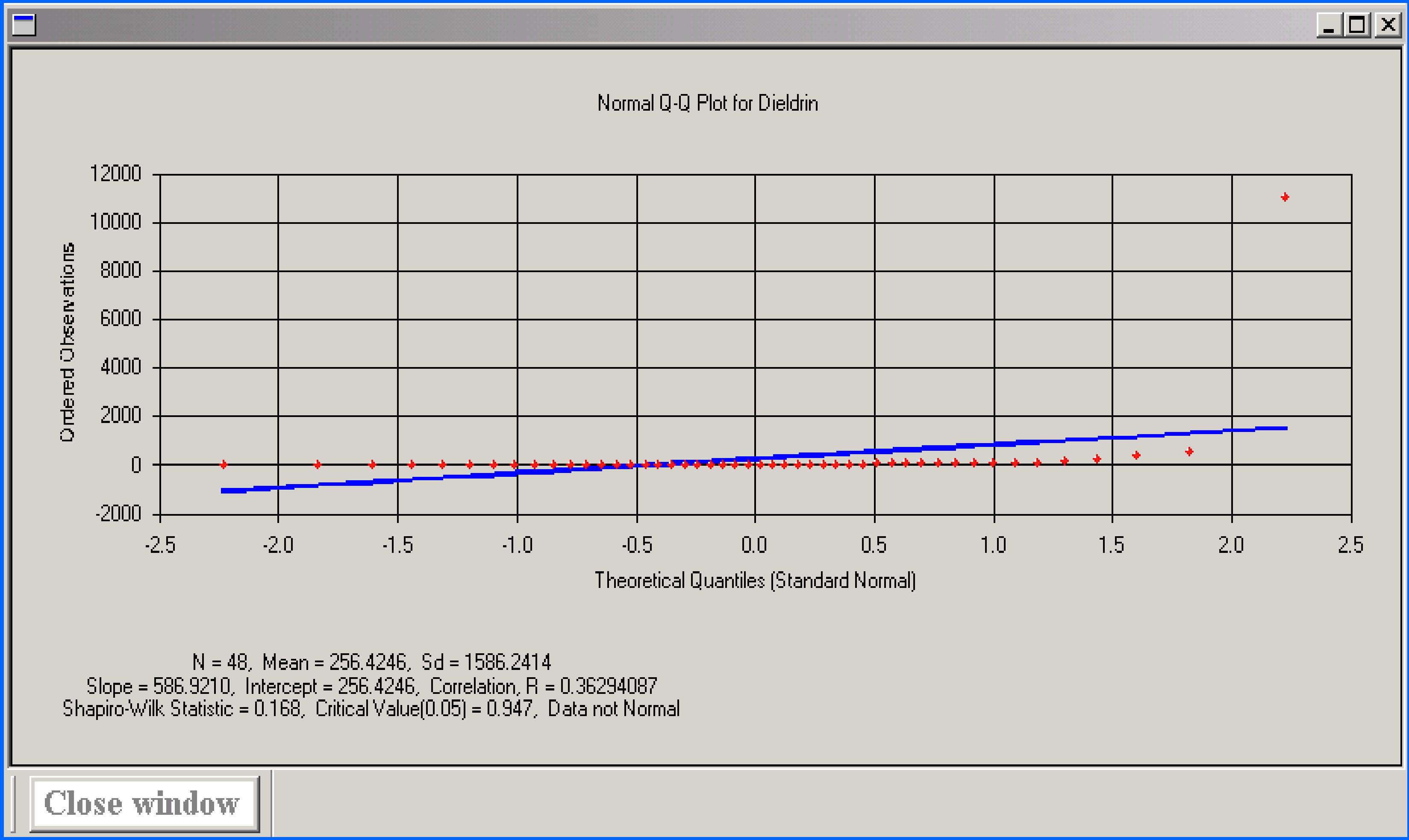
Mercury Data From Nepera Site



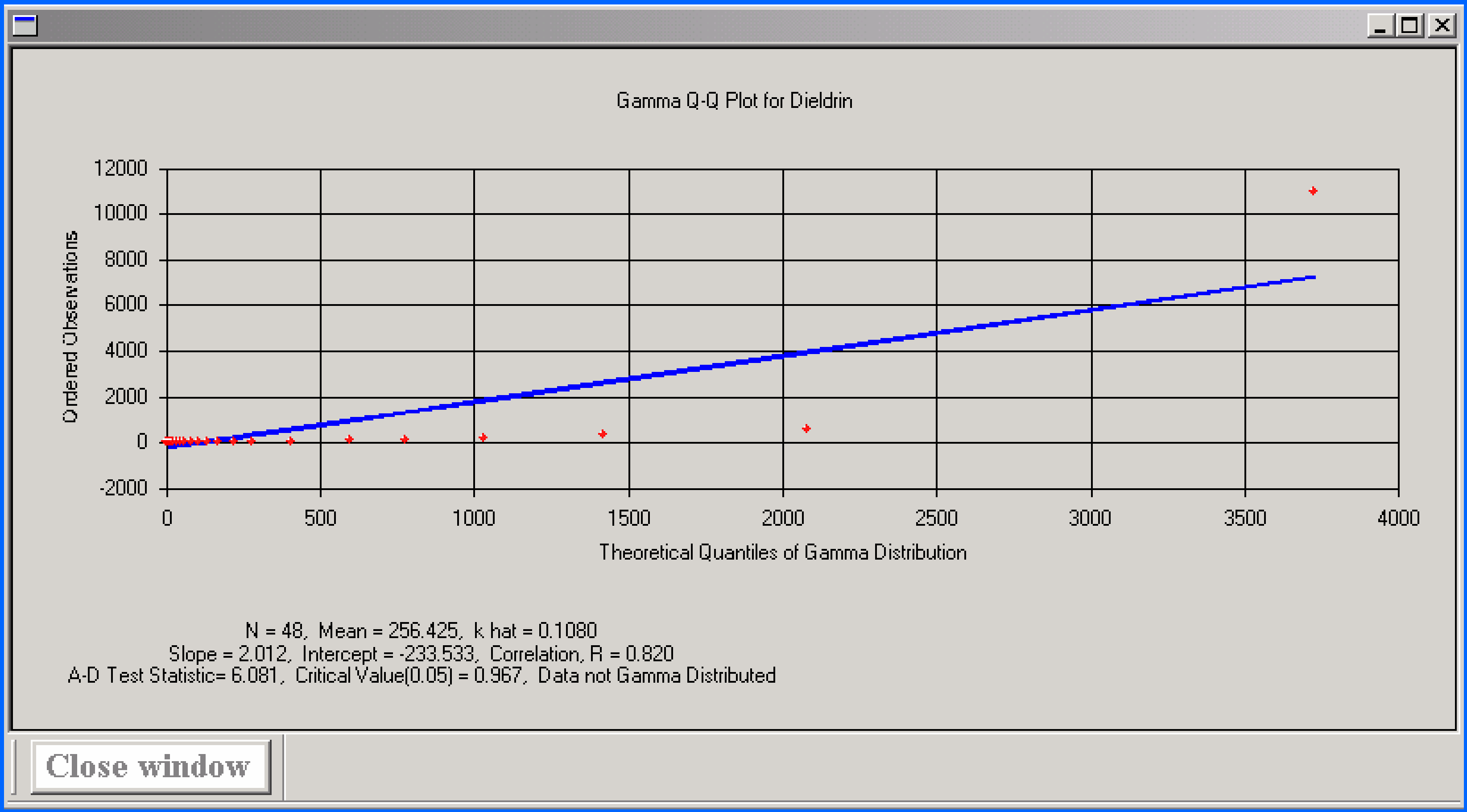
Mercury Data From Nepera Site

	A	B	C	D	E	F	G	H	I
1	Data File	D:\drive_c\sprfnd02\nepera\Table2.xls				Variable:	Methylmercury %		
2									
3	Raw Statistics				Normal Distribution Test				
4	Number of Valid Samples			15	Shapiro-Wilk Test Statisitic				0.8191722
5	Number of Unique Samples			12	Shapiro-Wilk 5% Critical Value				0.881
6	Minimum			0.03	Data not normal at 5% significance level				
7	Maximum			0.85					
8	Mean			0.23	95% UCL (Assuming Normal Distribution)				
9	Median			0.14	Student's-t UCL				0.3383293
10	Standard Deviation			0.2382076					
11	Variance			0.0567429	Gamma Distribution Test				
12	Coefficient of Variation			1.0356852	A-D Test Statistic				0.5595622
13	Skewness			1.4287947	A-D 5% Critical Value				0.7622797
14					K-S Test Statistic				0.2129509
15	Gamma Statistics				K-S 5% Critical Value				0.2277243
16	k hat			1.0406006	Data follow gamma distribution				
17	k star (bias corrected)			0.8769249	at 5% significance level				
18	Theta hat			0.2210262					
19	Theta star			0.2622801	95% UCLs (Assuming Gamma Distribution)				
20	nu hat			31.218018	Approximate Gamma UCL				0.3875053
21	nu star			26.307748	Adjusted Gamma UCL				0.4142477
22	Approx.Chi Square Value (.05)			15.614706					
23	Adjusted Level of Significance			0.03235	Lognormal Distribution Test				
24	Adjusted Chi Square Value			14.606676	Shapiro-Wilk Test Statisitic				0.9179229
25					Shapiro-Wilk 5% Critical Value				0.881
26	Log-transformed Statistics				Data are lognormal at 5% significance level				
27	Minimum of log data			-3.506558					
28	Maximum of log data			-0.162519	95% UCLs (Assuming Lognormal Distribution)				
29	Mean of log data			-2.021816	95% H-UCL				0.6217778
30	Standard Deviation of log data			1.1363214	95% Chebyshev (MVUE) UCL				0.5728518
31	Variance of log data			1.2912263	97.5% Chebyshev (MVUE) UCL				0.7182186
32					99% Chebyshev (MVUE) UCL				1.0037635
33									

Example 5 – Real Data of 48 Dieldrin Concentrations



Example 5 – Real Data of 48 Dieldrin Concentrations



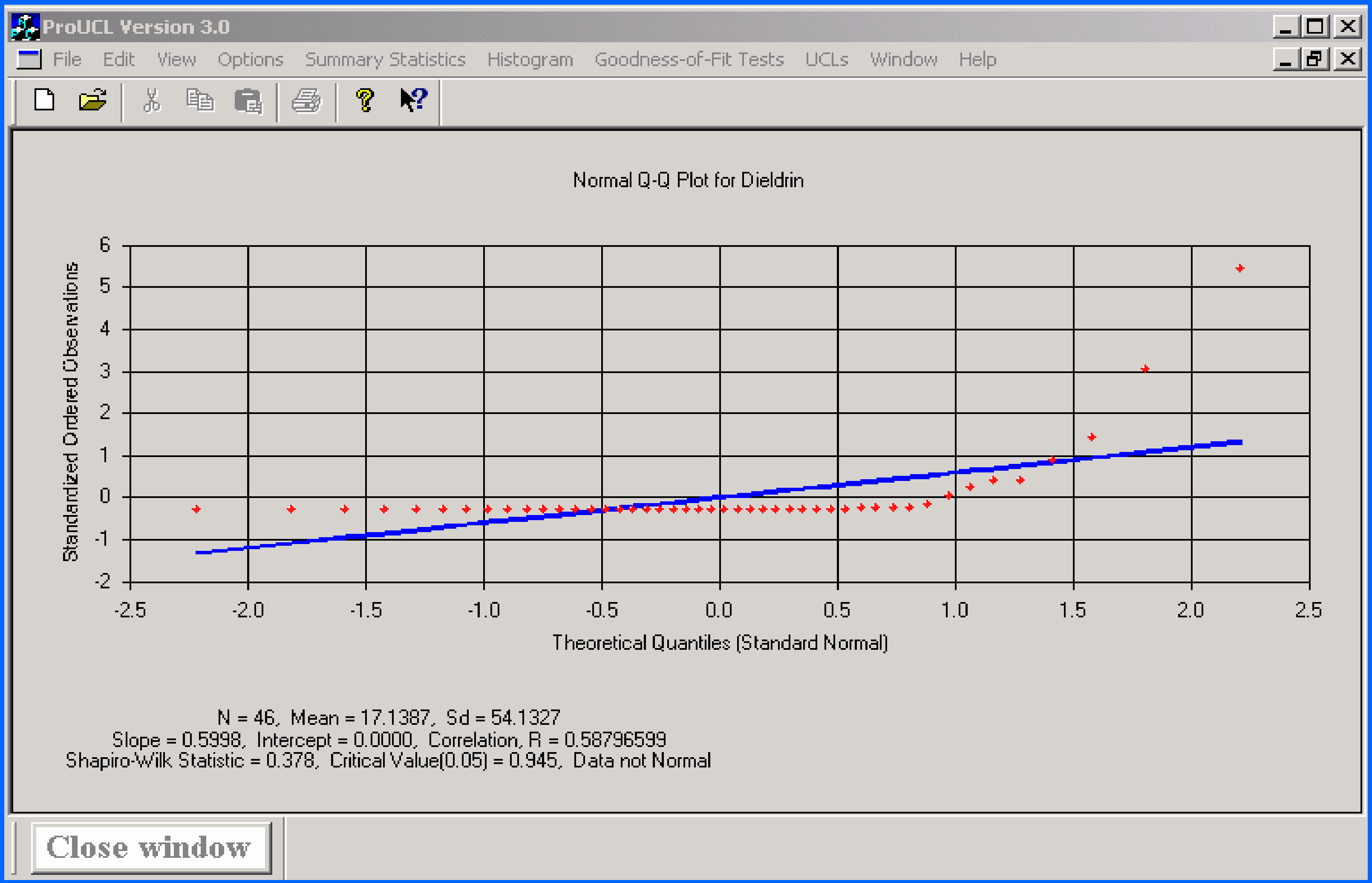
Example 5 – Real Data of 48 Dieldrin Concentrations

	A	B	C	D	E	F	G	H	I	J
1	Data File	D:\drive_c\sprfnd02\proucl2004\D-48.DAT				Variable:	Dieldrin			
2										
3	Raw Statistics				Normal Distribution Test					
4	Number of Valid Samples			48	Shapiro-Wilk Test Statisitic				0.1680365	
5	Number of Unique Samples			38	Shapiro-Wilk 5% Critical Value				0.947	
6	Minimum			0.00032	Data not normal at 5% significance level					
7	Maximum			11000						
8	Mean			256.42459	95% UCL (Assuming Normal Distribution)					
9	Median			0.125	Student's-t UCL				640.59292	
10	Standard Deviation			1586.2414						
11	Variance			2516161.8	Gamma Distribution Test					
12	Coefficient of Variation			6.1859958	A-D Test Statistic				6.0814679	
13	Skewness			6.8942515	A-D 5% Critical Value				0.9665944	
14					K-S Test Statistic				0.2799734	
15	Gamma Statistics				K-S 5% Critical Value				0.1450752	
16	k hat			0.1080292	Data do not follow gamma distribution					
17	k star (bias corrected)			0.1151663	at 5% significance level					
18	Theta hat			2373.6599						
19	Theta star			2226.5599	95% UCLs (Assuming Gamma Distribution)					
20	nu hat			10.370803	Approximate Gamma UCL				614.82655	
21	nu star			11.055962	Adjusted Gamma UCL				632.51946	
22	Approx.Chi Square Value (.05)			4.6110897						
23	Adjusted Level of Significance			0.045	Lognormal Distribution Test					
24	Adjusted Chi Square Value			4.4821078	Shapiro-Wilk Test Statisitic				0.9088836	
25					Shapiro-Wilk 5% Critical Value				0.947	
26	Log-transformed Statistics				Data not lognormal at 5% significance level					
27	Minimum of log data			-8.04719						

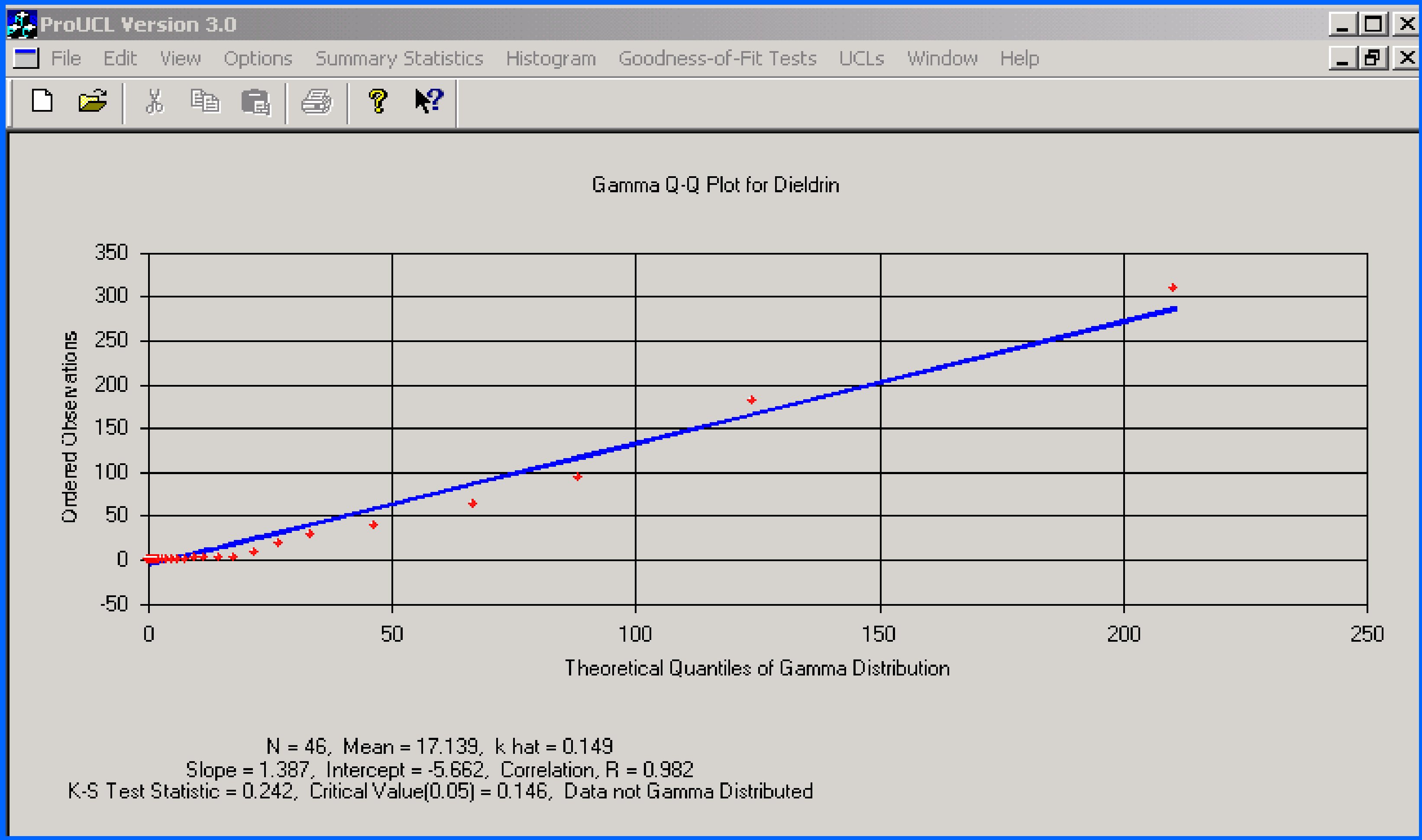
Example 5 – Real Data of 48 Dieldrin Concentrations

22	Approx.Chi Square Value (.05)	4.6110897		
23	Adjusted Level of Significance	0.045	Lognormal Distribution Test	
24	Adjusted Chi Square Value	4.4821078	Shapiro-Wilk Test Statisitic	0.9088836
25			Shapiro-Wilk 5% Critical Value	0.947
26	Log-transformed Statistics		Data not lognormal at 5% significance level	
27	Minimum of log data	-8.04719		
28	Maximum of log data	9.3056506	95% UCLs (Assuming Lognormal Distribution)	
29	Mean of log data	-1.896874	95% H-UCL	166544.37
30	Standard Deviation of log data	4.348021	95% Chebyshev (MVUE) UCL	2281.4309
31	Variance of log data	18.905286	97.5% Chebyshev (MVUE) UCL	3068.8723
32			99% Chebyshev (MVUE) UCL	4615.6484
33				
34			95% Non-parametric UCLs	
35			CLT UCL	633.02078
36			Adj-CLT UCL (Adjusted for skewness)	876.46282
37			Mod-t UCL (Adjusted for skewness)	678.56496
38			Jackknife UCL	640.59292
39			Standard Bootstrap UCL	612.73957
40			Bootstrap-t UCL	8583.6274
41	RECOMMENDATION		Hall's Bootstrap UCL	5755.5691
42	Data are Non-parametric (0.05)		Percentile Bootstrap UCL	717.06843
43			BCA Bootstrap UCL	704.22294
44	Use Hall's Bootstrap UCL		95% Chebyshev (Mean, Sd) UCL	1254.4129
45			97.5% Chebyshev (Mean, Sd) UCL	1686.2433
46	In case Hall's Bootstrap method yields		99% Chebyshev (Mean, Sd) UCL	2534.4904
47	an erratic, unreasonably large UCL value,			
48	use 99% Chebyshev (Mean, Sd) UCL			
49				

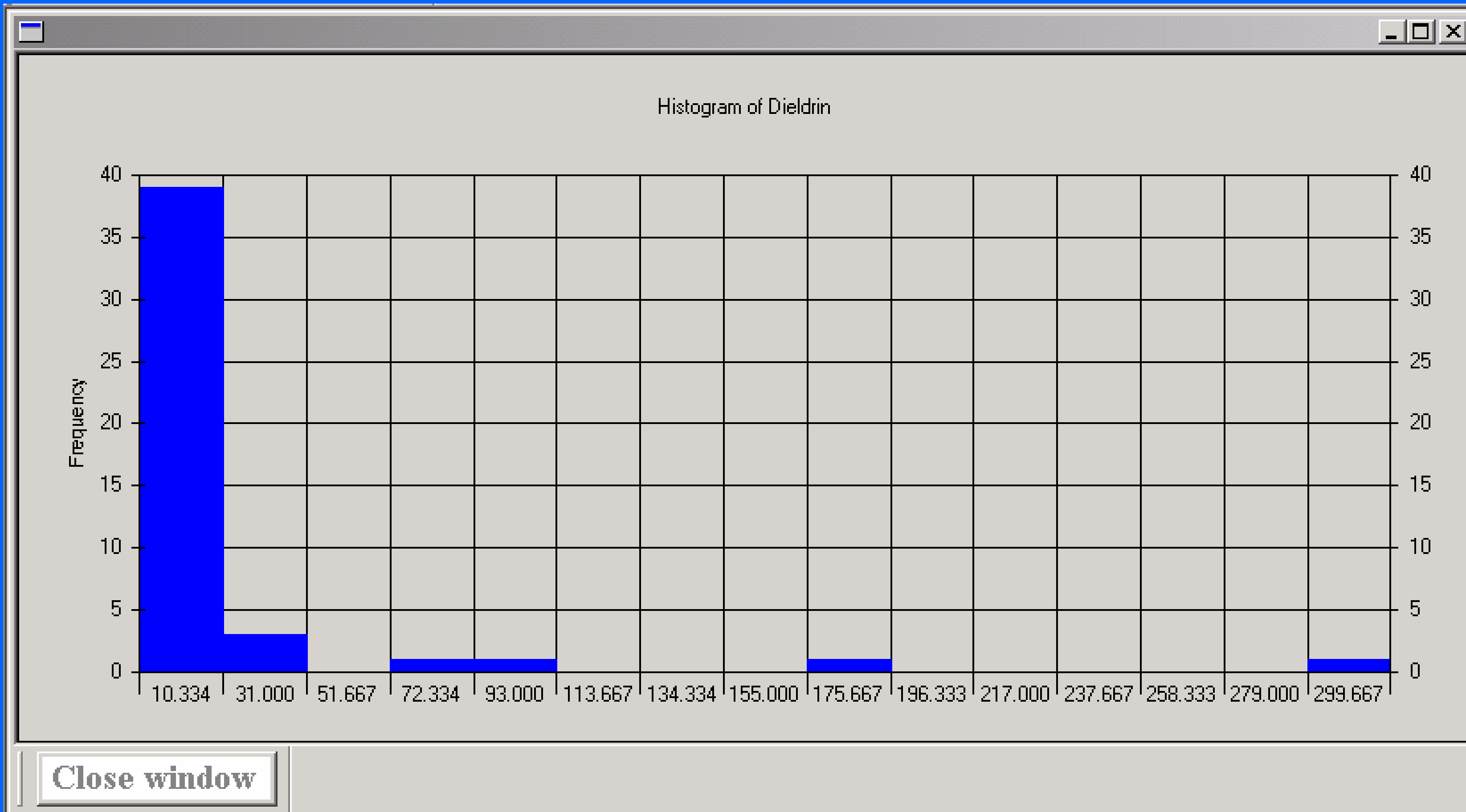
Example 5: Dieldrin Data without 2 Outlying Concentrations



Example 5: Dieldrin Data without 2 Outlying Concentrations



Example 5: Dieldrin Data without 2 Outlying Concentrations



Example 5: Dieldrin Data without 2 Outlying Concentrations

	A	B	C	D	E	F	G	H	I
1	Data File	D:\drive_c\sprfnd02\proucl2004\D-46.DAT				Variable:	Dieldrin		
2									
3	Raw Statistics				Normal Distribution Test				
4	Number of Valid Samples			46	Shapiro-Wilk Test Statisitic				0.3779341
5	Number of Unique Samples			36	Shapiro-Wilk 5% Critical Value				0.945
6	Minimum			0.00032	Data not normal at 5% significance level				
7	Maximum			310					
8	Mean			17.138701	95% UCL (Assuming Normal Distribution)				
9	Median			0.05075	Student's-t UCL				30.542925
10	Standard Deviation			54.13267					
11	Variance			2930.3459	Gamma Distribution Test				
12	Coefficient of Variation			3.1585049	A-D Test Statistic				3.9018848
13	Skewness			4.3671579	A-D 5% Critical Value				0.9384433
14					K-S Test Statistic				0.2420304
15	Gamma Statistics				K-S 5% Critical Value				0.1464637
16	k hat			0.1491094	Data do not follow gamma distribution				
17	k star (bias corrected)			0.1538776	at 5% significance level				
18	Theta hat			114.94044					
19	Theta star			111.37876	95% UCLs (Assuming Gamma Distribution)				
20	nu hat			13.718065	Approximate Gamma UCL				36.334358
21	nu star			14.156742	Adjusted Gamma UCL				37.264868
22	Approx.Chi Square Value (.05)			6.6776513					
23	Adjusted Level of Significance			0.0447826	Lognormal Distribution Test				
24	Adjusted Chi Square Value			6.5109091	Shapiro-Wilk Test Statisitic				0.8996916
25					Shapiro-Wilk 5% Critical Value				0.945
26	Log-transformed Statistics				Data not lognormal at 5% significance level				

Example 5: Dieldrin Data without 2 Outlying Concentrations

22	Approx.Chi Square Value (.05)	6.6776513		
23	Adjusted Level of Significance	0.0447826	Lognormal Distribution Test	
24	Adjusted Chi Square Value	6.5109091	Shapiro-Wilk Test Statisitic	0.8996916
25			Shapiro-Wilk 5% Critical Value	0.945
26	Log-transformed Statistics		Data not lognormal at 5% significance level	
27	Minimum of log data	-8.04719		
28	Maximum of log data	5.7365723	95% UCLs (Assuming Lognormal Distribution)	
29	Mean of log data	-2.317597	95% H-UCL	8832.1207
30	Standard Deviation of log data	3.9114521	95% Chebyshev (MVUE) UCL	346.15005
31	Variance of log data	15.299458	97.5% Chebyshev (MVUE) UCL	464.2486
32			99% Chebyshev (MVUE) UCL	696.23034
33				
34			95% Non-parametric UCLs	
35			CLT UCL	30.266979
36			Adj-CLT UCL (Adjusted for skewness)	35.758352
37			Mod-t UCL (Adjusted for skewness)	31.399468
38			Jackknife UCL	30.542925
39			Standard Bootstrap UCL	29.959657
40			Bootstrap-t UCL	56.771826
41	RECOMMENDATION		Hall's Bootstrap UCL	78.489944
42	Data are Non-parametric (0.05)		Percentile Bootstrap UCL	30.967498
43			BCA Bootstrap UCL	28.532952
44	Use Hall's Bootstrap UCL		95% Chebyshev (Mean, Sd) UCL	51.928933
45			97.5% Chebyshev (Mean, Sd) UCL	66.982695
46	In case Hall's Bootstrap method yields		99% Chebyshev (Mean, Sd) UCL	96.552894
47	an erratic, unreasonably large UCL value,			
48	use 99% Chebyshev (Mean, Sd) UCL			
49				

Procedure to Compute a 95% UCL

- Identify potential outliers/multiple populations.
  - If justified, study them separately.
- Perform goodness-of-fit tests, look at data graphically using histogram, Q-Q plots - **Never skip this step**.
  - **In order to automate the EPC computation process for multiple variables, some users want to skip this step which is not recommended, as it may lead to incorrect conclusions by accommodating outliers/multiple populations.**
- If data follow a normal model (or approximate normal) – use Student's - t 95% UCL.
- If data follow a gamma model – use adjusted or approximate gamma 95% UCL as described in Table 2.

Procedure to Compute 95% UCL

- Avoid the use of a lognormal model, as:
  - It accommodates outliers and multiple populations.
  - It often yields unstable/ impractical UCLs, especially for highly skewed small data sets (e.g., n <10-20 etc.).
  - Use the procedure described in Table 1 with **caution**.
- For nonparametric data sets, compute 95% UCLs using the procedure as summarized in Table 3.

Procedure to Compute 95% UCL

- **Caution:** When Hall's or bootstrap – t UCLs are recommended, make sure there are no outliers.
- Do not use the Max value to estimate the EPC Term. ProUCL recommends alternative UCL methods for EPC.
- Decision Tables 1-3 are programmed in ProUCL.
- ProUCL recommends the most suitable method(s) which may be used to compute an appropriate 95% UCL of the mean (EPC Term).

ProUCL Questions/Comments

- Contact Gareth Pearson at the USEPA Technical Support Center in Las Vegas
  - 702-798-2101 or 702-798-2270
  - pearson.gareth@epa.gov

- ProUCL, Version 3.0 is available for download at:

http://www.epa.gov/nerlesd1/tsc/software.htm

BACKGROUND VS SITE COMPARISONS & COMPUTATION OF BACKGROUND THRESHOLD VALUES (BTV)

PART II

BACKGROUND VS SITE COMPARISONS

- Why compare background contamination levels (mean, UPL, UTL, Percentiles) with contamination levels introduced by some potentially responsible party (PRP) at an industrial site?

- This is a high interest topic in the area of:
  - making cleanup decisions such as - Where to clean? How much (to what concentration level) to clean?
  - verification of the attainment of cleanup levels (such as represented by BTVs) at a polluted site - perhaps after performing some remediation actions.

Workshop Objectives - Background

- Discuss comparison of site vs background data.

1. Based upon two sample comparisons:

   – to be used when enough site and background data are available.

2. Based upon background threshold values:

   – when individual site values (typically used when enough site data are not available) are to be compared with a BTV.

Workshop Objectives - Background

- To estimate BTVs using:
  - 95% upper prediction limits (UPLs)
  - 95% upper percentiles – parametric and non-parametric
- To identify site outliers – perhaps representing contaminated areas, site hotspots in comparison with the site background:
  - Discussion on how to interpret site values exceeding BTVs.

Workshop Outline

- Comparison of background versus site data:
  – Two sample comparisons with enough data
  – Comparison of each site value with BTV
- Computation of BTVs
  – 95% UPLs
  – 95% upper percentiles (parameteric/nonparametric)
- Illustrations using real data sets
- Recommended procedure to compute BTVs
- How to interpret site values exceeding BTVs?

Evaluation of Background and Site Data

- Two Sample comparisons:
  - Site versus Background – used when enough data from the two populations are available.
  - Discussed in detail in EPA 2002 CERCLA Background Guidance Document.

- Computation of BTVs:
  - Compare individual site observations with some BTV.
  - If most site data (e.g., > 95%) fall below BTV, then site data can be considered as coming from the background population.
  - Site observations > BTV may require further investigation.

Evaluation of Background vs Site Data- Attainment of Cleanup Standard?

- Is site contaminated? Do Site and Background data come from the same statistical population?

- Two Ways to address these questions.

  1. If enough site and background data (e.g., ≥10 points):

  – Perform two sample comparisons: t-test, Wilcoxon Rank Sum Test / Mann - Whitney Test – discussed in background CERCLA document.

  2. When individual (and not mean) site values are to be compared with some background value:

  – Compute BTV – procedures not provided in CERCLA background document.

Two Sample Comparisons

- 1. Null (baseline condition) hypothesis is $H_0 : \mu_s \leq \mu_b$, vs alternative hypothesis, $H_1 : \mu_s > \mu_b$.
  - H_0 called - Background Test Form 1.
  - This hypothesis is useful to verify the attainment of cleanup standards after some remediation actions have been performed at a typical contaminated site.

It is assumed that the null hypothesis is true, that is the site has been cleaned enough to attain cleanup standards (such as background mean value).

  - Using the available data, the burden of proof is to prove it otherwise (e.g., reject the null hypothesis and conclude that the site mean still exceeds the background mean).

Two Sample Comparisons

- 2. Null (baseline condition) hypothesis is $H_0 : \mu_s \geq \mu_b$, vs alternative hypothesis, $H_1 : \mu_s < \mu_b$.
  - Called Background Test Form 2.

- This Null hypothesis is protective of the environment and human health.

- It is assumed that the null hypothesis is true (that is the site is dirty and may be impacted by the site activities).

- Some times, a factor, S>0 (e.g., $=s_b$, background sd) is added to the right hand side of the Form 2 hypotheses stated above.

Two Sample Comparisons

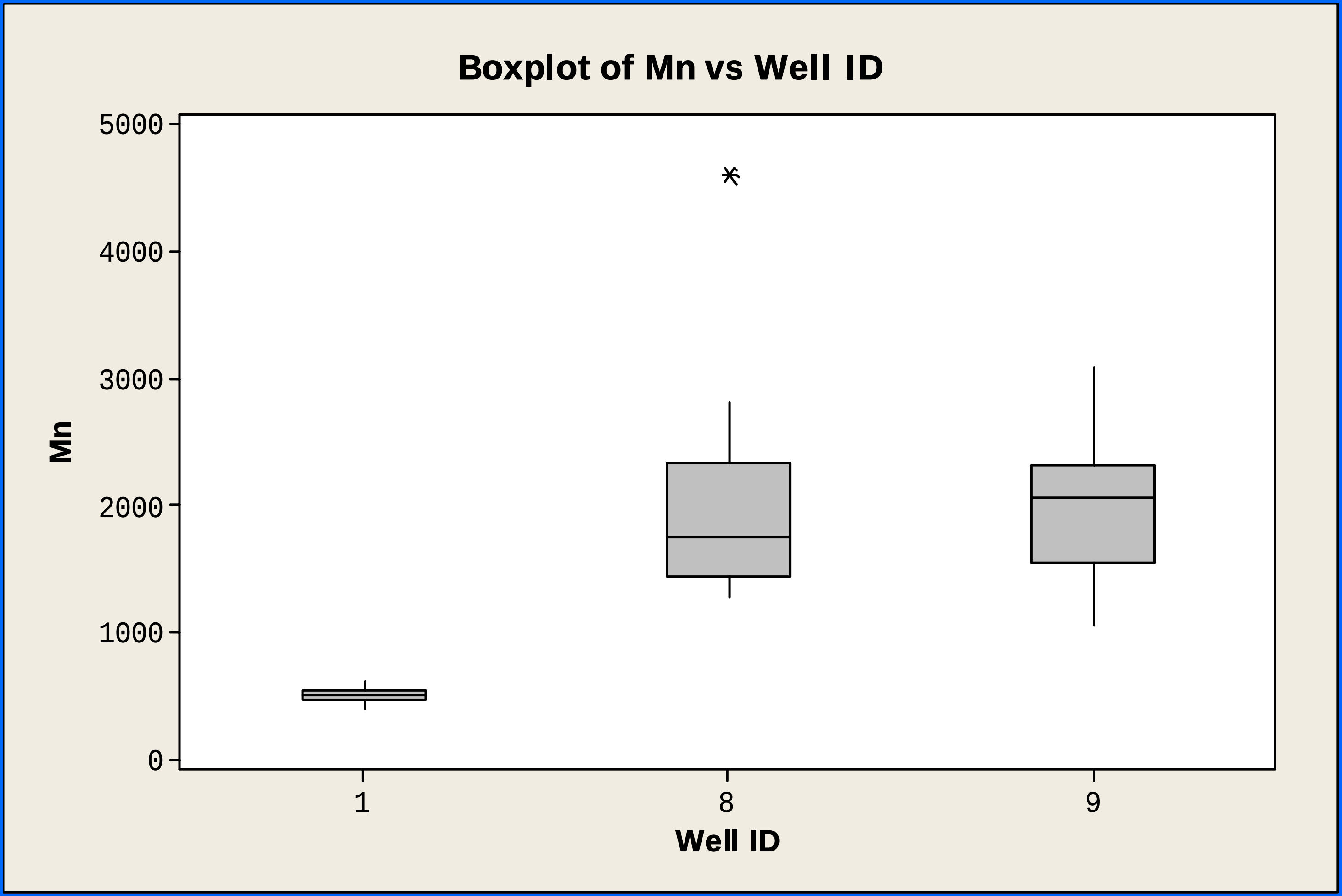
- Using the available data, the burden of proof is to prove it otherwise (e.g., reject the null hypothesis and conclude the site is clean).

- This Form 2 is often used when not much is known about an area of concern – such as prior to remediation actions etc.

Real Example:Two Sample Comparisons

- Real data set (C&R Battery Site) – COPCs are inorganics. Consider Mn
- MW 1= Upgradient background well
- MW6, MW7, MW8, MW9 = Downgradient wells
- Should perform ANOVA for more than 2 populations – beyond the scope of Workshop.
- Graphical Comparison First

Real Mn Data: Box Plots For Three Monitoring Wells



Assumptions Needed to Perform Two Sample Comparisons

- Samples should be normally distributed for t-test.
- Two populations should be independent.
- No distributional assumptions needed for WRS test or for Mann - Whitney two sample tests.

Caution:

- **Do not use t-test on log-transformed data to compare means of two populations – a common mistake.**
- **Use nonparametric tests instead.**

Real Mn Data:Two Sample Comparison

Two-Sample T-Test for Mn: Background (MW1) vs MW8

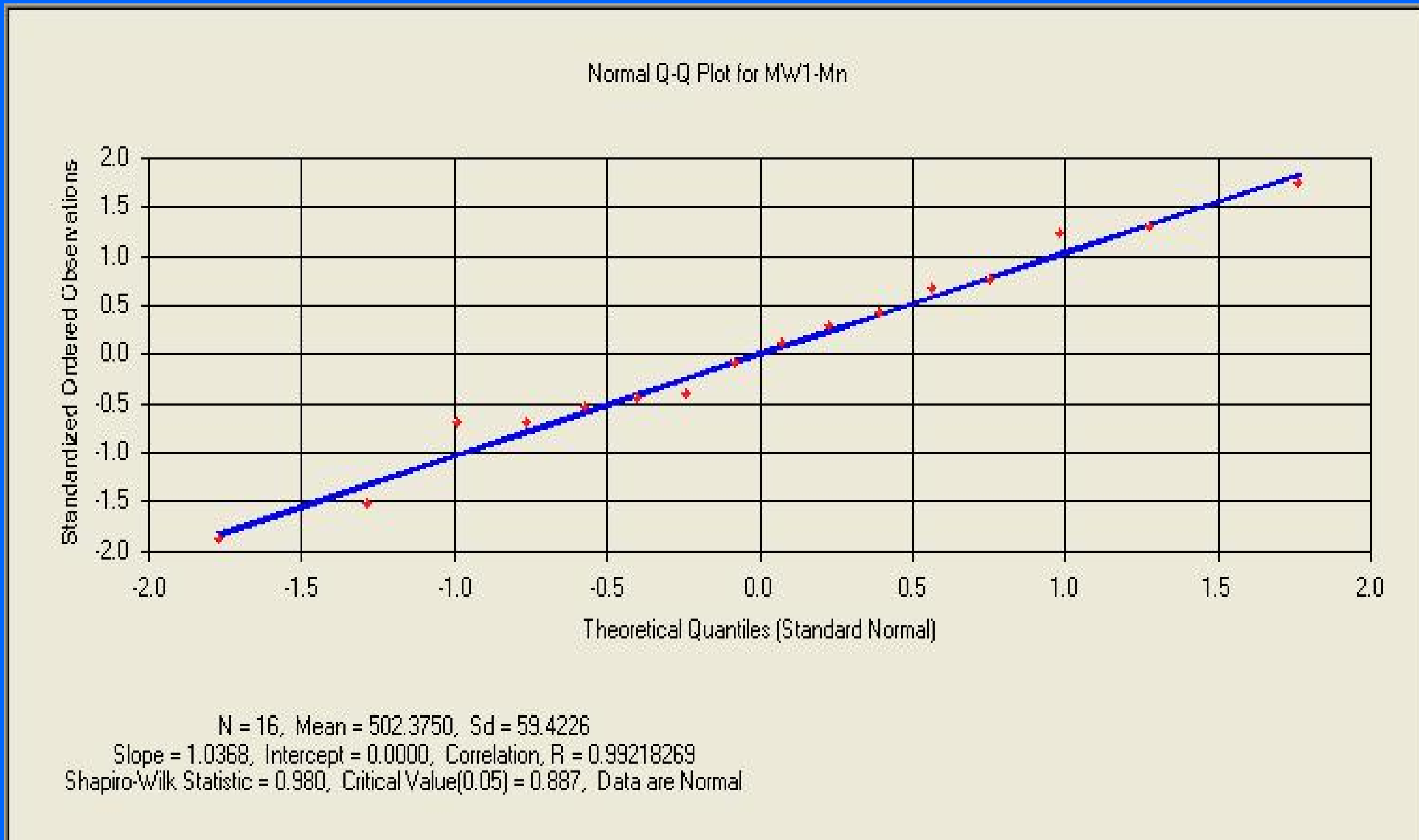
ID	N	Mean	StDev	SE Mean
1	16	502.4	59.4	15
8	16	1998	839	210

Difference = mu (1) - mu (8)
Estimate for difference: -1495.75
95% upper bound for difference: -1127.23

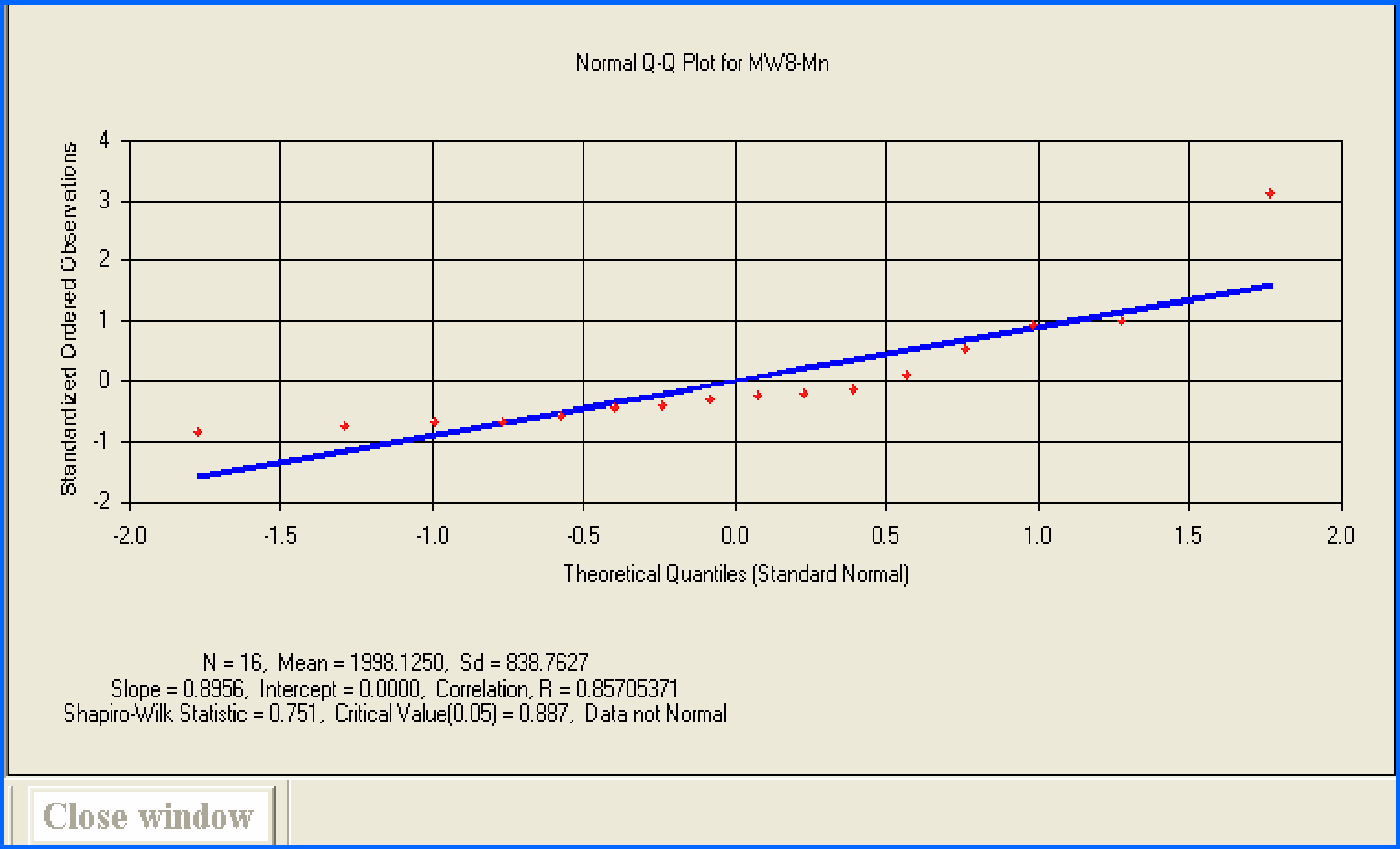
T-Test of difference = 0 (vs <): T-Value = -7.12 P-Value = 0.000
DF = 15, highly significant.

Conclusion: Reject H_0 and conclude MW8 has much higher mean Mn than that of MW1 (upgradient well).

Real Mn Data:Normality Test for MW1



Real Mn Data:Normality Test for MW8



Real Mn Data:Two Sample Comparison

Mann-Whitney Test and CI: MW1-Mn, MW8-Mn

	N	Median
MW1-Mn	16	502.0
MW8-Mn	16	1750.0

Point estimate for ETA1-ETA2 is -1237.0
95.2 Percent CI for ETA1-ETA2 is (-1509.1,-1025.2)

W = 136.0
Test of ETA1 = ETA2 vs ETA1 < ETA2 is significant at 0.0000
The test is significant at 0.0000 (adjusted for ties)

Conclusion: Reject H_0, and conclude MW8 has higher Mn than that of MW1

Real Mn Data:Two Sample Comparison

Two-Sample T-Test for Mn: Background (MW1) vs MW =9

ID	N	Mean	StDev	SE Mean
1	16	502.4	59.4	15
9	16	1968	500	125

Difference = mu (1) - mu (9)
Estimate for difference: -1465.75
95% upper bound for difference: -1244.99

T-Test of difference = 0 (vs <): T-Value = -11.64 P-Value = 0.000
DF = 15

Conclusion: Reject H_0 and conclude MW 9 has higher mean Mn than that of MW1 (Upgradient well).

How to Compute Background Threshold Values (BTV)?

- Often need to compute BTVs for Superfund Site Evaluations - such as used for:
  - Andrews Air Force Base
  - South Weymouth Naval Air Station
  - AMTL Charles River Site
  - Wallops Air Facility
  - Fort Benning Site, Georgia

- No clear guidelines provided in Background CERCLA Document (2002) for Soils (EPA 540-R-01-003, OSWER 9285.7-41) or in any other Navy or EPA document on how to compute BTVs.

How to Compute BTVs?

- Parametric Background Values:
  - 95% upper prediction limit (UPL) – also given in EPA 1992 RCRA document addendum.
  - 95th upper percentile based upon background data distribution – mentioned in some Navy documents.
  - However, formulae to compute the percentiles are missing.
- Make sure no significant outliers and/or multiple populations are present – treat them separately.

95% Parametric UPL

- (1-α)100% UPL for normal background data sets (EPA 1992 RCRA document).

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

- UPL for lognormal data - obtained using log-transformed data and then back transformation.

Normal 95% Upper Percentiles

- Determine background data distribution first.
- Make sure no significant outliers and/or multiple populations are present.
- For Normal distribution, the p100% percentile is:

$$\hat{x}_p = x + sz_p$$

- z_p = upper p100th (e.g., =95%) percentile of N(0,1).
- If distributions of site and background data are really the same (meaning no contamination due to site activities), then site data should lie below the background 95% upper percentile with 0.95 confidence coefficient.

Lognormal 95% Upper Percentiles

- For lognormal Distribution, the p100% percentile is:

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

- z_p = upper p100th percentile of N(0,1).
- If distributions of site and background data are really the same (meaning no contamination due to site), then site data should lie below the background 95% upper percentile with 0.95 confidence coefficient.
- Similarly, one can obtain upper percentiles for gamma distribution – using inverse gamma distribution.

Nonparametric Background Threshold Values

- Non-parametric Background Values:
  - Current practice: The largest or second largest value based upon professional judgment/site specific conditions is used to estimate BTV.
  - However, avoid its use, as it has no theoretical justification.
  - Use simple upper 95th percentile, $x_{0.95}$ of background data.
  - Where, X_p is that value such that p100% of background data lies at or below X_p.
- Real data from Fort Benning Site - discussed next:

Computing BTV Using 95% UPLs Real Data Set

	A	B	C	D	E	F	G	H
1		Shapiro-Wilks Statistics				95% Upper Prediction Limits		
2		Critical Value	Normal	Lognormal	Distribution	Normal	Lognormal	Non-Parametric
3	Barium	0.842	0.765	0.892	Lognormal	96.906	**125.927**	88.000
4	Chromium	0.842	0.672	0.953	Lognormal	18.850	**23.245**	7.200
5	Iron	0.940	0.574	0.963	Lognormal	59749.101	**63044.750**	52000.000
6	Iron w/o Outlier	0.939	0.892	0.939	Non-Parametric	35077.163	51503.240	**36900.000**
7	Manganese	0.842	0.661	0.991	Lognormal	498.213	**959.105**	230.000
8	Vandium	0.905	0.647	0.973	Lognormal	29.745	**31.094**	18.000
9								

Computing BTV - 95% Upper Percentiles
Real Data Set

	A	B	C	D	E	F	G	H
1		Shapiro-Wilks Statistics				95% Percentiles		
2		Critical Value	Normal	Lognormal	Distribution	Normal	Lognormal	Non-Parametric
3	Barium	0.842	0.765	0.892	Lognormal	87.94	99.53	92.95
4	Chromium	0.842	0.672	0.953	Lognormal	17.00	18.06	16.44
5	Iron	0.940	0.574	0.963	Lognormal	58264.84	59191.34	37655.00
6	Iron w/o Outlier	0.939	0.892	0.939	Non-Parametric	34333.12	48513.53	32490.00
7	Manganese	0.842	0.661	0.991	Lognormal	444.78	631.00	455.50
8	Vandium	0.905	0.647	0.973	Lognormal	28.37	28.08	19.70

Computing BTV - 95% Upper Percentiles and 95% UPLs - Real Data Set

	A	B	C	D	E	F	G	H
1		Data	95% Upper Prediction Limits			95% Percentiles		
2	COPC	Distribution	Normal	Lognormal	Non-Parametric	Normal	Lognormal	Non-Parametric
3	Barium	Lognormal	96.906	**125.927**	88.000	87.94	**99.53**	92.95
4	Chromium	Lognormal	18.850	**23.245**	7.200	17.00	**18.06**	16.44
5	Iron	Lognormal	59749.101	**63044.750**	52000.000	58264.84	**59191.34**	37655.00
6	Iron w/o Outlier	Non-Parametric	35077.163	51503.240	**36900.000**	34333.12	48513.53	**32490.00**
7	Manganese	Lognormal	498.213	**959.105**	230.000	444.78	**631.00**	455.50
8	Vandium	Lognormal	29.745	**31.094**	18.000	28.37	**28.08**	19.70

Recommended Procedure to Compute BTVs

- Make sure no significant outliers or multiple populations are present in the background data set.

- Background statistics should be computed based upon a single sample (from the background population) without outliers.

- Use graphical displays to visualize data – these provide useful info about outliers, multiple populations etc.

- Determine background data distribution.

Recommended Procedure to Compute BTVs

- Use 95% UPL or 95% upper percentile for normal, lognormal, or gamma distribution.

  – Note lognormal distribution yields higher BTVs.

- For nonparametric data sets, use the 95% upper percentiles instead of arbitrarily chosen largest or 2nd largest value.

Determining Outlying Site Values and Hot Spots- Per EPA 2002

- 2002 EPA CERCLA background document suggests the identification of site outliers based upon background 95% UTLs.

- Background 95% UPL or 95% upper percentile can also be used to identify contaminated site values.

- Site values exceeding a BTV may be considered as coming from a population different from the site background suggesting contamination due to site activities (perhaps representing a hot spot).

Determining Outlying Site Values and Hot Spots

- According to CERLA document, site outliers exceeding BTV can be interpreted as hot spots (contaminated parts of the site) – needing further investigation.
  - In practice, individual site values can exceed the BTV even when the site mean and the background mean appear to be the same.
  - It is desirable to use the two sample tests (provided enough data are available) as well as the BTVs to perform site and background comparisons.