

---

# *ProUCL Version 3.0*

## *Statistical Software to Compute Upper Confidence Limits of the Unknown Population Mean*

---

### **Introduction**

The EPA Technical Support Center (TSC) in Las Vegas has developed ProUCL Version 3.0 software to support risk assessment and cleanup decisions at contaminated sites. Risk assessments, exposure evaluations, and cleanup decisions are often made based upon the mean concentrations of the contaminants of potential concern (COPCs). The true population mean concentrations of the COPCs at a contaminated site are often unknown, and are frequently estimated by the respective sample means based upon the data collected from the site under investigation. In order to address the uncertainties associated with the estimates of the true unknown mean concentrations of the COPCs, appropriate 95% upper confidence limits (UCLs) of the respective unknown means are frequently used in many environmental applications. The computation of an appropriate 95% UCL of practical merit depends upon the data distribution and the skewness associated with the data set under study. ProUCL can be used to compute an appropriate UCL of the unknown population mean using a discernible probability distribution (e.g., normal, lognormal, gamma) and/or a suitable non-parametric distribution-free method.

### **The Need for UCL Computational Software**

*A 95% UCL of the unknown population arithmetic mean,  $\mu$ , of a COPC is often used in environmental applications to:*

- Estimate the exposure point concentration (EPC) term,
- Support risk assessment applications,
- Determine the attainment of cleanup standards,

- Estimate background level mean contaminant concentrations, or
- Compare the soil mean concentrations with site-specific soil screening levels.

In December 2002, the EPA revised the Guidance Document to Calculate the Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites (OSWER 9285.670). ProUCL, Version 3.0 consists of all parametric and non-parametric UCL computation methods as described in this revised EPA UCL Guidance Document.

ProUCL also computes the UCLs of the unknown population mean based upon the positively skewed gamma distribution, which is often better suited to model environmental data sets than the lognormal distribution. For positively skewed data sets, the default use of a lognormal distribution often results in impractically large UCLs, especially when the data sets are small. This is also illustrated in the example included in this fact sheet. In order to obtain accurate and stable UCLs of practical merit, other distributions such as a gamma distribution should be used to model positively skewed data sets. ProUCL, Version 3.0 has procedures to perform the gamma goodness-of-fit test and to compute UCLs of the population mean based upon gamma distributed data sets.

### **ProUCL Version 3.0 Capabilities**

*Performs Goodness-of-Fit Tests to Assess Normality/Lognormality of a Data Set Using:*

- Informal graphical quantile-quantile (Q-Q) plot (normal probability plot) and histogram.
- Shapiro - Wilk test: to be used when the sample size is less than or equal to 50.

ADMIN RECORD

# ProUCL Version 3.0 Statistical Software to Compute Upper Confidence Limits of the Unknown Population Mean

---

- Lilliefors test: to be used when the sample size is large (e.g., greater than 50).

## *Performs Goodness-of-Fit Test for Gamma Distribution Using:*

- Informal graphical quantile-quantile (Q-Q) plot (gamma probability plot) and histogram.
- Kolmogorov-Smirnov test: available for sample sizes in the range 4-2500 (critical values computed using Monte Carlo simulations) and values of the estimated shape parameter,  $k$  in the interval [0.01, 100.0].
- Anderson-Darling test: available for sample sizes in the range 4-2500 (critical values computed using Monte Carlo simulations) and values of the estimated shape parameter,  $k$  in the interval [0.01, 100.0].

## *Computes the Estimates of Relevant Population Parameters:*

- Computes all relevant descriptive summary statistics for raw and log-transformed data.
- Computes the maximum likelihood (ML) and minimum variance unbiased (MVU) estimates of the various population parameters such as the mean, standard deviation, quantiles, coefficient of variation, skewness, and also the MLEs of the shape parameter  $k$  and scale parameter  $\theta$  of a gamma distribution.

## *Computes Five Parametric UCLs:*

A  $(1-\alpha)$  100% (for all values of the confidence coefficient,  $(1-\alpha)$  in the interval [0.5, 1.0] including 0.95 except for a couple of methods) UCL of the unknown population mean,  $\mu_1$ , using five (5) parametric methods for normal, lognormal, and gamma distributions. The five parametric UCL computation methods incorporated in ProUCL are:

1. Student's-t UCL: to be used for normally (or at least approximately normally) distributed data sets. Student's-t UCL is available for all confidence coefficients,  $(1-\alpha)$  in the interval [0.5, 1.0].

2. Approximate Gamma UCL: to be used for gamma distributed data and is typically used when  $k$  hat (ML estimate of the shape parameter,  $k$ ) is greater than or equal to 0.5. Approximate gamma UCL is available for all confidence coefficients  $(1-\alpha)$  in the interval [0.5, 1.0].
3. Adjusted Gamma UCL: to be used for gamma distributed data sets and should be used when  $k$  hat is greater than 0.1 and less than 0.5. Adjusted gamma UCL is available only for three confidence coefficients: 0.90, 0.95, and 0.99.
4. H-UCL based upon Land's H-statistic: to be used for lognormally distributed data sets. In ProUCL, H-UCL is available only for two confidence coefficients: 0.90 and 0.95. ProUCL can compute H-UCL for samples of size up to 2001.

**Caution: For highly skewed data sets, H-UCL should be avoided as the H-statistic often results in unrealistically large, impractical and unusable H-UCL values. ProUCL provides warning messages and recommends the use of alternative UCLs for such highly skewed lognormally distributed data sets.**

5. Chebyshev (MVUE) UCL: to be used for lognormally distributed data sets. This UCL computation method uses the MVU estimates of the standard deviation of the mean and of other parameters of a lognormal distribution. Chebyshev (MVUE) UCL is available for all confidence coefficients,  $(1-\alpha)$  in the interval [0.5, 1.0].

## *Computes Ten Non-parametric UCLs Based Upon Bootstrap Procedures and Chebyshev Inequality:*

These UCLs can be computed for all confidence coefficients,  $(1-\alpha)$  in the interval [0.5, 1.0].

1. Central Limit Theorem (CLT) based UCL: to be used when the sample size is large.
2. Adjusted-CLT (Adjusted for skewness) UCL: to be used for skewed data sets of large sizes.

# ProUCL Version 3.0 Statistical Software to Compute Upper Confidence Limits of the Unknown Population Mean

---

3. Modified-t statistic (Adjusted for skewness) based UCL: may be used for mildly skewed data.
4. Chebyshev (Mean, Sd) UCL: based upon the sample mean and standard deviation, Sd.
5. Jackknife UCL for mean (same as Student's-t UCL).
6. Standard Bootstrap UCL.
7. Bootstrap-t UCL.
8. Hall's Bootstrap UCL.
9. Percentile Bootstrap UCL.
10. Bias-corrected accelerated (BCA) Bootstrap UCL.

As mentioned before, for most of the UCL computation methods, ProUCL can compute the UCLs for all confidence coefficients,  $(1-\alpha)$  in the interval  $[0.5, 1.0]$ . However, since in most environmental applications (e.g., estimation of EPC), a 95% UCL of mean is used, therefore, ProUCL makes recommendations for the most appropriate 95% UCL(s) which may be used to estimate the unknown population mean concentration. The basis and theoretical justification for these recommendations can be found in the references listed in this fact sheet.

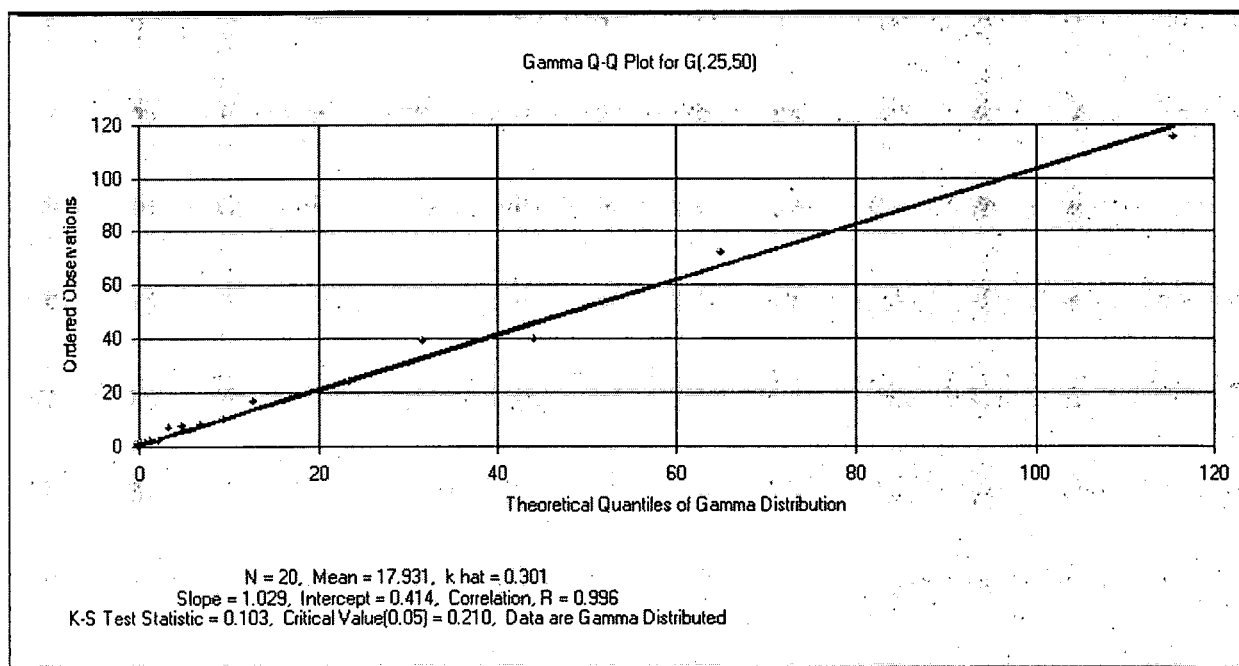
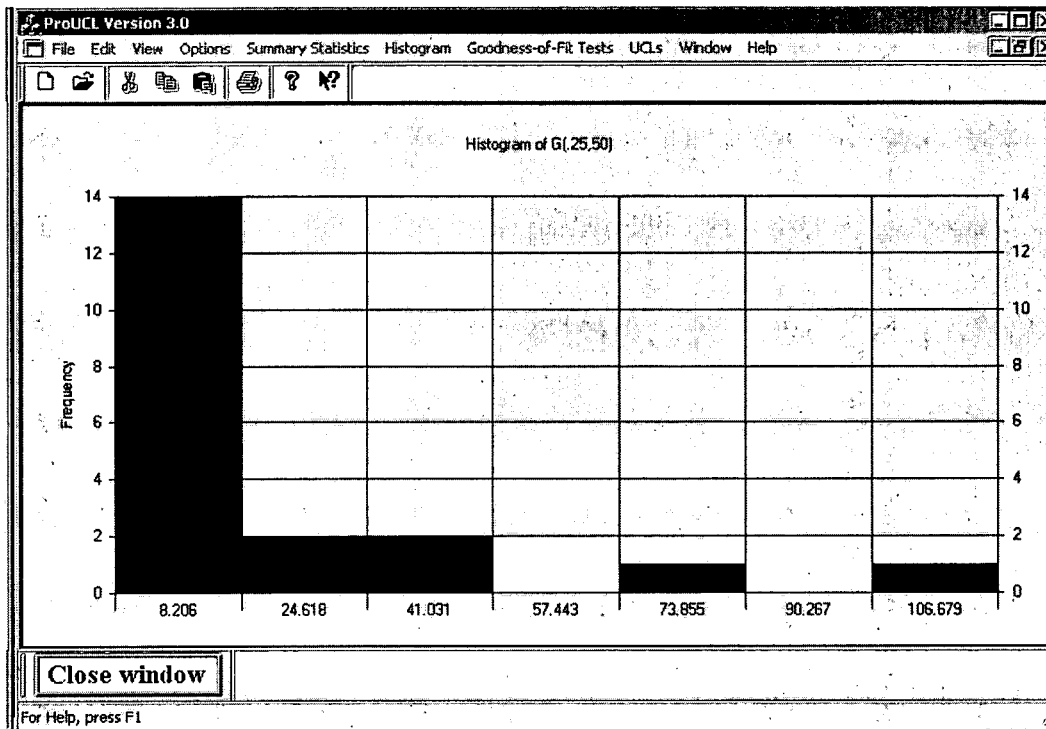
**Example:** An example illustrating the importance of the use of gamma distribution based UCL computation methods is discussed next. Consider a simulated positively skewed data set of size 20: 0.0086284103, 16.18078972, 7.334853523, 6.12856E-005, 1.756500498, 1.394359005, 23.41857632, 7.516539628, 0.8594623274, 39.06134332, 17.97357103, 114.885481, 9.251610362, 39.44123801, 71.64271025, 6.271065467, 0.9742964478, 0.1558884758, 0.4817911951, 0.0065875373. This data set was

generated from a gamma distribution with the shape parameter,  $k = 0.25$  and the scale parameter,  $\theta = 50$ . This can be seen from the two enclosed figures, a histogram and the associated gamma probability plot generated by ProUCL. This data set does not follow a normal or a lognormal distribution. However, using the default lognormal (as some times done in practice for positively skewed environmental data sets) distribution, one will get an unrealistically large 95% H-UCL (=498390.58) as given in the enclosed output generated by ProUCL. This in turn will lead to the use of the maximum value (=114.885) as an estimate of the EPC term - a practice often used in environmental applications. However, since the data set does follow a gamma distribution, therefore, an appropriate and more accurate estimate of the EPC term (representing the average exposure over a long period of time over an exposure area) will be obtained by using the 95% UCL (=45.067) based upon the gamma distribution as recommended in the enclosed output table generated by ProUCL, Version 3.0.

## Summary

ProUCL computes parametric UCLs based upon a normal, lognormal, and a gamma distribution. ProUCL also computes UCLs using several non-parametric methods. The computation of an appropriate UCL of the unknown population mean depends upon the data distribution, therefore goodness-of-fit tests need to be performed to assess the data distribution before using one of the UCL computation methods available in ProUCL. Based upon an appropriate data distribution and the associated skewness, ProUCL provides recommendations about one or more 95% UCL computation methods that may be used to estimate the unknown mean concentration of a COPC.

# ProUCL Version 3.0 Statistical Software to Compute Upper Confidence Limits of the Unknown Population Mean



# ProUCL Version 3.0 Statistical Software to Compute Upper Confidence Limits of the Unknown Population Mean

ProUCL Version 3.0										
File Edit View Options Summary Statistics Histogram Goodness-of-Fit Tests UCLs Window Help										
UCL Statistics for G(,25,50)										
	A	B	C	D	E	F	G	H	I	
1	Data File	D:\PData\Gamma-Test-File2.txt			Variable:		G(,25,50)			
2										
3	Raw Statistics				Normal Distribution Test					
4	Number of Valid Samples			20	Shapiro-Wilk Test Statistic			0.6627653		
5	Number of Unique Samples			20	Shapiro-Wilk 5% Critical Value			0.905		
6	Minimum			6.13E-005	Data not normal at 5% significance level					
7	Maximum			114.88548						
8	Mean			17.930768	95% UCL (Assuming Normal Distribution)					
9	Median			6.8029595	Student's-t UCL			29.275688		
10	Standard Deviation			29.341896						
11	Variance			860.94688	Gamma Distribution Test					
12	Coefficient of Variation			1.6363993	A-D Test Statistic			0.1492209		
13	Skewness			2.3948728	A-D 5% Critical Value			0.8445354		
14					K-S Test Statistic			0.1026697		
15	Gamma Statistics				K-S 5% Critical Value			0.2101084		
16	k hat			0.3008281	Data follow gamma distribution					
17	k star (bias corrected)			0.2890372	at 5% significance level					
18	Theta hat			59.604702						
19	Theta star			62.036194	95% UCLs (Assuming Gamma Distribution)					
20	nu hat			12.033123	Approximate Gamma UCL			41.976248		
21	nu star			11.561488	Adjusted Gamma UCL			45.066538		
22	Approx. Chi Square Value (.05)			4.9386585						
23	Adjusted Level of Significance			0.038	Lognormal Distribution Test					
24	Adjusted Chi Square Value			4.6000062	Shapiro-Wilk Test Statistic			0.8641123		
25					Shapiro-Wilk 5% Critical Value			0.905		
26	Log-transformed Statistics				Data not lognormal at 5% significance level					
27	Minimum of log data			-9.699966						
28	Maximum of log data			4.7439358	95% UCLs (Assuming Lognormal Distribution)					
29	Mean of log data			0.5953246	95% H-UCL			498390.58		
30	Standard Deviation of log data			3.6226212	95% Chebyshev (MVUE) UCL			1438.355		
31	Variance of log data			13.123384	97.5% Chebyshev (MVUE) UCL			1932.73		
32					99% Chebyshev (MVUE) UCL			2903.834		
33										
34					95% Non-parametric UCLs					
35					CLT UCL			28.72273		
36					Adj-CLT UCL (Adjusted for skewness)			32.476962		
37					Mod-t UCL (Adjusted for skewness)			29.861273		
38					Jackknife UCL			29.275688		
39					Standard Bootstrap UCL			28.161388		
40					Bootstrap-t UCL			39.768558		
41	RECOMMENDATION				Hall's Bootstrap UCL			71.935053		
42	Data follow gamma distribution (0.05)				Percentile Bootstrap UCL			29.194241		
43					BCA Bootstrap UCL			34.604005		
44	Use Adjusted Gamma UCL				95% Chebyshev (Mean, Sd) UCL			46.529711		
45					97.5% Chebyshev (Mean, Sd) UCL			58.904496		
46					99% Chebyshev (Mean, Sd) UCL			83.212366		
47										
General Statistics										
For Help, press F1										

# ProUCL Version 3.0 Statistical Software to Compute Upper Confidence Limits of the Unknown Population Mean

---

## Computer Requirements to Operate ProUCL

Installation of ProUCL Version 3.0 requires a microprocessor speed of at least 200MHz, 12 MB of hard drive space, 48 MB of memory (RAM), and Windows 98 (or newer) operating system. ProUCL is compatible with Windows NT-4, Windows 2000, Windows XP, and Windows ME.

## Installation

ProUCL can be downloaded from the TSC website at [www.epa.gov/nerlesd1/tsc/tsc.htm](http://www.epa.gov/nerlesd1/tsc/tsc.htm). The website contains download and usage instructions.

## Find More Information About ProUCL

The TSC website at [www.epa.gov/nerlesd1/tsc/tsc.htm](http://www.epa.gov/nerlesd1/tsc/tsc.htm) provides additional information. EPA technical issue papers used in the development of ProUCL are also available at the TSC website. For additional information, contact:

Gareth Pearson, TSC Director  
E-mail: [pearson.gareth@epa.gov](mailto:pearson.gareth@epa.gov)  
Phone: 702-798-2270

The website containing information on the 2002 EPA guidance for calculating the 95% UCLs is: [www.epa.gov/superfund/programs/risk/ragsa/ucl.pdf](http://www.epa.gov/superfund/programs/risk/ragsa/ucl.pdf).

## References

- EPA (2002), Calculating Upper Confidence Limits for Exposure Point Concentrations at Hazardous Waste Sites, OSWER 9285.6-10, December 2002.
- Schulz, T. W., and Griffin, S. (1999), Estimating Risk Assessment Exposure Point Concentrations When Data are Not Normal or Lognormal. Risk Analysis, Vol. 19, No. 4, 1999.
- Singh, A. K., Singh, A., and Engelhardt, M. (1997), "The Lognormal Distribution in Environmental Applications," EPA/600/R-97/006, December 1997.
- Singh, A., Singh, A. K., and Iaci, R. J. (2002). "Estimation of the Exposure Point Concentration Term Using a Gamma Distribution." EPA/600/R-02/084.
- Singh, A., Singh, A. K., Engelhardt, M., and Nocerino, J.M. (2003), "On the Computation of the Upper Confidence Limit of the Mean of Contaminant Data Distributions." Under EPA Review.
- Singh, A. and Singh, A.K. (2003). Estimation of the Exposure Point Concentration Term (95% UCL) Using Bias-Corrected Accelerated (BCA) Bootstrap Method and Several Other Methods for Normal, Lognormal, and Gamma Distributions. Draft EPA Internal Report.

## Notice

The U. S. Environmental Protection Agency through its Office of Research and Development funded and managed the research described here under Interagency Agreement DW47939416 to the Government Services Administration (GSA). It has been subject to the Agency's peer and administrative review and has been approved for publication as an EPA document.