

## ***Distribution Fitting (Uncensored Data)***

### **Summary**

The **Distribution Fitting (Uncensored Data)** procedure fits any of 45 probability distributions to a column of numeric data. The data are assumed to be uncensored, i.e., the data represent random samples from the selected distribution. If the data have been censored due to a limit of detection or some other cause, use the *Distribution Fitting (Censored Data)* procedure instead.

**Sample StatFolio:** *distfit uncensored.sgp*

### **Sample Data:**

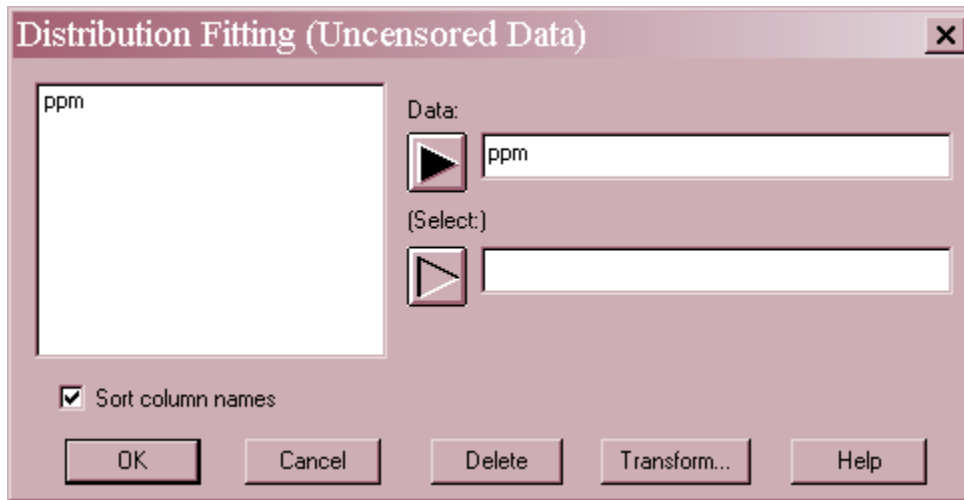
The file *groundwater.sgd* contains  $n = 47$  measurements of the concentration of uranium in groundwater samples taken from a location in northwest Texas. The table below shows a partial list of the data from that file:

<i>ppm</i>
8.25
2.82
4.16
18.66
12.72
8.75
2.29
7.22
9.76
7.72
27.38
5.14

It is desired to find a probability distribution that provides a suitable model for the variation amongst the samples in order to provide a background reference distribution against which to compare future samples.

## Data Input

The data to be analyzed consist of a single numeric column containing  $n = 2$  or more observations.



- **Data :** numeric column containing the data to be analyzed.
- **Select:** subset selection.

## Analysis Summary

The *Analysis Summary* shows the number of observations, the range of the data, and the values of the estimated parameters for each distribution that is fit to the data.

<u><a href="#">Uncensored Data - ppm</a></u>			
Data variable: ppm			
47 values ranging from 0.74 to 47.78			
Fitted Distributions			
<i>Gamma</i>	<i>Lognormal</i>	<i>Normal</i>	<i>Weibull</i>
shape = 1.56457	mean = 13.7033	mean = 12.8219	shape = 1.28496
scale = 0.122023	standard deviation = 15.6921	standard deviation = 10.445	scale = 13.8975
	Log scale: mean = 2.19873		
	Log scale: std. dev. = 0.915324		

The parameters displayed depend upon the distributions selected (see the documentation for the *Probability Distributions* procedure). Estimates are obtained using Maximum Likelihood Estimation (MLE). You can fit between 1 and 5 distributions at the same time using *Analysis Options*.

In the above table, 4 distributions have been fit to the groundwater data. The *gamma* and *Weibull* distributions are defined by their shape and scale parameters. The *lognormal* and *normal* distributions are defined by their means and standard deviations. In the case of the lognormal distribution, the mean and standard deviation of the natural logarithms of *ppm* are also shown.

## Analysis Options

- Distribution:** select between 1 and 5 distributions to fit to the data. Each distribution is described in detail in the *Probability Distributions* documentation. To help determine which distributions to fit, the *Comparison of Alternative Distributions* pane described below can be extremely helpful. The following tables may also be helpful.

### Discrete Distributions

<i>Distribution</i>	<i>Range of Data</i>	<i>Common Use</i>
Bernoulli	0 or 1	Model for event with 2 possible outcomes.
Binomial	0, 1, 2, ..., $m$	Number of successes in $m$ Bernoulli trials.
Discrete Uniform	$a, a+1, a+2, \dots, b$	Model for integers with fixed limits.
Geometric	0, 1, 2, ...	Waiting time until first Bernoulli success.
Hypergeometric	0, 1, 2, ..., $m$	Count when sampling from a finite population.
Negative Binomial	0, 1, 2, ...	Waiting time until $k$ -th Bernoulli success.
Poisson	0, 1, 2, ...	Number of events in fixed interval.

### Continuous Distributions

<i>Distribution</i>	<i>Range of Data</i>	<i>Common Use</i>
Beta	$0 \leq X \leq 1$	Distribution of a random proportion.
Beta (4-parameter)	$a \leq X \leq b$	Model for data with upper and lower thresholds.
Birnbaum-Saunders	$X > 0$	Failure times.
Cauchy	all real $X$	Measurement exhibiting long, flat tails.
Chi-Squared	$X \geq 0$	Reference distribution for sample variance.
Erlang	$X > 0$	Time between $k$ arrivals in a Poisson process.

Exponential	$X > 0$	Time between consecutive Poisson events.
Exponential (2-parms)	$X > a$	Lifetimes with fixed lower threshold.
Exponential power	all real $X$	Symmetric data with variable kurtosis.
F	$X \geq 0$	Ratio of 2 independent variance estimates.
Folded Normal	$X \geq 0$	Absolute values of data from normal distribution
Gamma	$X \geq 0$	Model for positively skewed measurements.
Gamma (3-parameter)	$X \geq a$	Positively skewed data with lower threshold.
Generalized Gamma	$X > 0$	Includes several distributions as special cases.
Generalized Logistic	All real $x$	Used for the analysis of extreme values.
Half Normal	$X \geq \mu$	Normal data folded about its mean.
Inverse Gaussian	$X > 0$	First passage time in Brownian motion.
Laplace	all real $X$	Data with pronounced peak and long tails.
Largest Extreme Value	all real $X$	Largest value in a sample.
Logistic	all real $X$	Growth model; common alternative to normal.
Loglogistic	$X > 0$	Logs of data from logistic distribution.
Loglogistic (3-parms)	$X > a$	Logs of data with fixed lower threshold.
Lognormal	$X > 0$	Positively skewed data.
Lognormal (3-parameter)	$X > a$	Positively skewed data with lower threshold.
Maxwell	$X > a$	Speed of a molecule in an ideal gas.
Noncentral Chi-squared	$X \geq 0$	Calculating power of chi-squared test.
Noncentral F	$X \geq 0$	Calculating power of F test.
Noncentral t	all real $X$	Calculating power of t test.
Normal	all real $X$	Data with many sources of variability.
Pareto	$X \geq 1$	Socio-economic quantities with long upper tails.
Pareto (2-parameter)	$X \geq a$	Socio-economic quantities with lower threshold.
Rayleigh	$X > a$	Distance between neighboring items.
Smallest Extreme Value	all real $X$	Smallest value in a sample.
Student's t	all real $X$	Reference distribution for sample mean.
Triangular	$a \leq X \leq b$	Rough model in the absence of data.
Uniform	$a \leq X \leq b$	Data with equal probability over an interval.
Weibull	$X \geq 0$	Product lifetimes.
Weibull (3-parameter)	$X \geq a$	Product lifetimes with lower threshold.

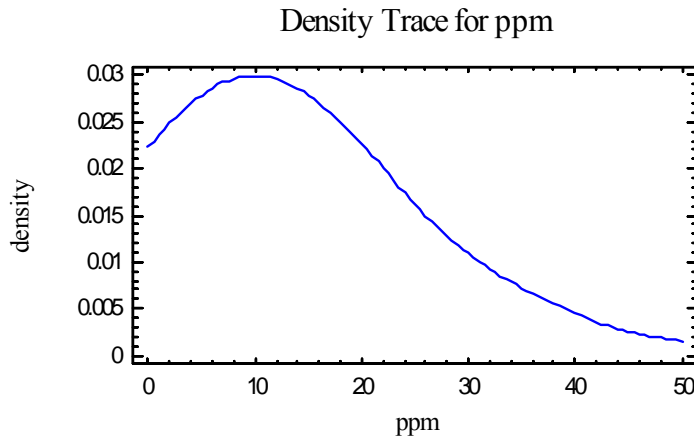
- **Binomial Trials** – when fitting the binomial distribution, you must specify the sample size  $n$ .
- **Hypergeometric Trials** – when fitting the hypergeometric distribution, you must specify the sample size  $n$ . You may either specify the population size parameter  $N$  or estimate it from the data.
- **Negative Binomial Trials** – when fitting the negative binomial distribution, you may either specify the parameter  $k$  or estimate it from the data.
- **Extended Threshold Parameters** – when fitting distributions that have one or more threshold parameters, you may specify those parameters or estimate them from the data. The relevant distributions are:

beta (4-parameter) – lower and upper  
 exponential (2-parameter) – lower only  
 half normal (2-parameter) – lower only  
 gamma (3-parameter) – lower only

- loglogistic (3-parameter) – lower only
- lognormal (3-parameter) – lower only
- Maxwell (2-parameter) – lower only
- Pareto (2-parameter) – lower only
- Rayleigh (2-parameter) – lower only
- Weibull (2-parameter) – lower only

## Density Trace

A good place to begin when selecting a distribution for a set of data is with the *Density Trace*. The *Density Trace* provides a nonparametric estimate of the probability density function of the population from which the data were sampled. It is created by counting the number of observations that fall within a window of fixed width moved across the range of the data.



The estimated density function is given by:

$$f(x) = \frac{1}{hn} \sum_{i=1}^n W\left(\frac{x - x_i}{h}\right) \tag{1}$$

where  $h$  is the width of the window in units of  $X$  and  $W(u)$  is a weighting function determined by the selection on the *Pane Options* dialog box. Two forms of weighting function are offered:

### Boxcar Function

$$W(u) = \begin{cases} 1 & \text{if } |u| \leq 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

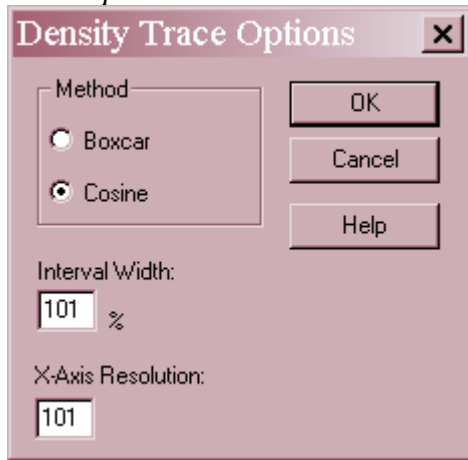
### Cosine Function

$$W(u) = \begin{cases} 1 + \cos(2\pi u) & \text{if } |u| < 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The latter selection usually gives a smoother result, with the desirable value of  $h$  depending on the size of the data sample.

In the case of the groundwater data, the density trace starts out relatively high at  $X = 0$ , increases to a peak, and then falls off rather slowly in the positive direction. A positively skewed distribution will clearly be necessary to model this data.

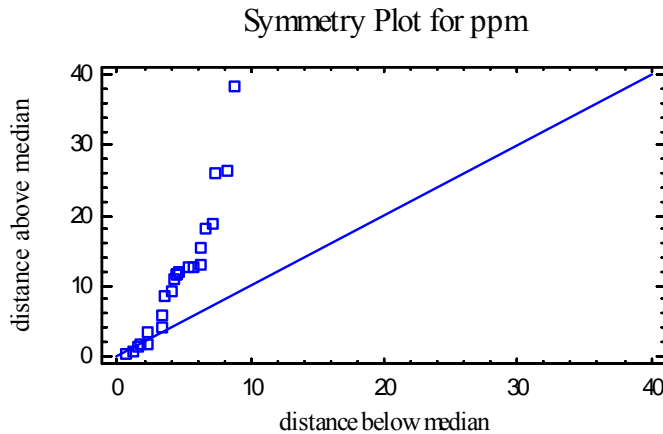
*Pane Options*



- **Method:** the desired weighting function. The boxcar function weights all values within the window equally. The cosine function gives decreasing weight to observations further from the center of the window. The default selection is determined by the setting on the *EDA* tab of the *Preferences* dialog box accessible from the *Edit* menu.
- **Interval Width:** the width of the window  $h$  within which observations affect the estimated density, as a percentage of the range covered by the x-axis.  $h = 60\%$  is not unreasonable for a small sample but may not give as much detail as a smaller value in larger samples.
- **X-Axis Resolution:** the number of points at which the density is estimated.

**Symmetry Plot**

The *Symmetry Plot* can also be used to help judge whether the data come from a symmetric distribution, i.e., a distribution that has a density function with the same shape on either side of the median.



To create this plot, the data values are sorted and then paired based on their location with respect to the median. For example, with 47 observations, the sorted points are paired as:

$$(X_{(23)}, X_{(25)}), (X_{(22)}, X_{(26)}), (X_{(21)}, X_{(27)}), \dots, (X_{(1)}, X_{(47)})$$

The distance of each pair below and above the median is plotted. If the data come from a symmetric distribution, the points should lie close to a 45-degree line. If not, the points will deviate from the line in a particular direction. The points in the plot above deviate well above the diagonal, indicating a longer upper tail than lower tail.

### Tests for Normality

The *Tests for Normality* pane performs up to 4 different tests designed to determine whether or not the data could reasonably have come from a normal distribution. For each test, the hypotheses of interest are:

- Null hypothesis: data are independent samples from a normal distribution
- Alt. hypothesis: data are not independent samples from a normal distribution

Tests for Normality for ppm		
Test	Statistic	P-Value
Chi-Squared	34.5745	0.00282602
Shapiro-Wilks W	0.871657	0.0000283121
Skewness Z-score	2.34972	0.0187876
Kurtosis Z-score	1.93069	0.0535207

The tests to be run are selected using *Pane Options*. Each test is displayed with its associated test statistic and *P-Value*. Small P-values lead to a rejection of the null hypothesis and thus to a rejection of the normal distribution. In the above table, the P-value for the Shapiro-Wilks test and the Chi-Squared test are both well below 0.01, leading to a rejection of the normal distribution at the 1% significance level.

The 4 available tests are defined as follows:

**Chi-Square Test** – This test divides the range of the data into a set of  $k$  equiprobable classes, where

$$k = \min\{100, \text{ceiling}(3.7653(n-1)^{0.4})\} \tag{4}$$

It then calculates the number of observations  $O_i$  falling into each class and the expected frequencies  $E_i$  based on the fitted distribution. A chi-squared statistic is computed according to

$$X^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \tag{5}$$

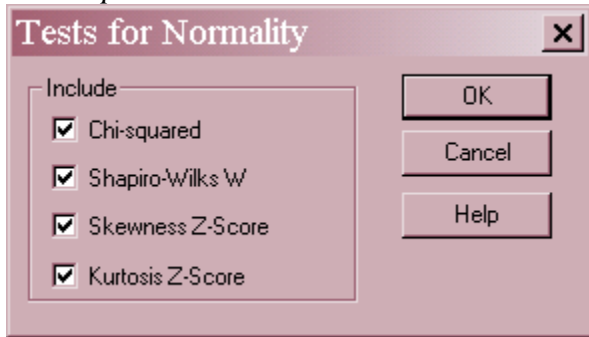
and compared to a chi-squared distribution with  $(k-3)$  degrees of freedom.

**Shapiro-Wilks Test** - This test, available when  $2 \leq n \leq 2000$ , uses a statistic derived by calculating how well the data fall along a straight line when plotted on a normal probability plot. In computing the statistic and its P-value, STATGRAPHICS uses Roysten’s method as outlined in Section 1.2 of Madansky (1988).

**Z-score for skewness** - calculates the sample skewness and determines whether it is significantly different from zero. The Z score is calculated according to the  $S_U$  approximation described on p.377 of D’Agostino and Stephens (1986) and is available only if  $n \geq 8$ .

**Z-score for kurtosis** - calculates the sample kurtosis and determines whether it is significantly different from zero. The Z score is calculated according to the Anscombe and Glynn approximation described on p.388 of D’Agostino and Stephens (1986) and is available only if  $n \geq 20$ .

*Pane Options*



- **Include** – select the tests to include in the output. The default tests are defined on the *Dist. Fit* tab of the *Preferences* dialog box accessed from the *Edit* menu.

**Goodness-of-Fit Tests**

The *Goodness-of-Fit Tests* pane performs up to 7 different tests to determine whether or not the data could reasonably have come from each fitted distribution. For all tests, the hypotheses of interest are:

- Null hypothesis: data are independent samples from the specified distribution
- Alt. hypothesis: data are not independent samples from the specified distribution

The tests to be run are selected using *Pane Options*.

The first 2 tests are general purpose tests that can be applied to any set of data:

Goodness-of-Fit Tests for ppm				
Chi-Squared Test				
	<i>Gamma</i>	<i>Lognormal</i>	<i>Normal</i>	<i>Weibull</i>
Chi-Squared	8.53204	10.0639	34.575	8.53191
D.f.	15	15	15	15
P-Value	0.900688	0.815699	0.00282556	0.900694



<b>Goodness-of-Fit Tests for ppm</b>				
Kolmogorov-Smirnov Test				
	<i>Gamma</i>	<i>Lognormal</i>	<i>Normal</i>	<i>Weibull</i>
DPLUS	0.077953	0.0441855	0.181741	0.0889679
DMINUS	0.0905744	0.0953022	0.123694	0.0833416
DN	0.0905744	0.0953022	0.181741	0.0889679
P-Value	0.835393	0.786792	0.0896715	0.850863

**Chi-Squared Test** - This test divides the range of  $X$  into  $k$  intervals and compares the observed counts

$$O_j = \text{number of data values observed in interval } j$$

to the number expected given the fitted distribution

$$E_j = \text{number of data values expected in interval } j.$$

The test statistic is given by

$$\chi^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j} \tag{6}$$

which is compared to a chi-squared distribution with  $k-p-1$  degrees of freedom, where  $p$  is the number of parameters estimated when fitting the selected distribution. For a discrete distribution, the intervals are formed by taking each unique value of  $X$  and grouping values together from each end, forming intervals with expected values  $E_j \geq 2$ . For a continuous distribution, equiprobable intervals are formed (intervals with equal  $E_j$ ) and  $k$  is selected so as to achieve the largest number of intervals with all  $E_j \geq 2$ . Small P-values lead to a rejection of the hypothesized distribution. In the above table, the test rejects the hypothesis of a *normal distribution* at the 1% significance level since the P-value is less than 0.01. However, the other distributions are all reasonable candidates.

**Kolmogorov-Smirnov Test** - This test compares the cumulative distribution of the data to the fitted cumulative distribution (as illustrated in the *Quantile Plot* below). It first evaluates the fitted cumulative distribution at each of the data values:

$$z_{(i)} = \hat{F}(x_{(i)}) \tag{7}$$

It then computes and displays the maximum distance of the empirical c.d.f. above the fitted c.d.f.

$$D^+ = \max_i \left\{ \frac{i}{n} - z_{(i)} \right\} \tag{8}$$

and the maximum distance of the empirical c.d.f. below the fitted c.d.f.

$$D^- = \max_i \left\{ z_{(i)} - \frac{i-1}{n} \right\} \tag{9}$$

The Kolmogorov statistic is the greater of the two distances

$$D = \max(D^+, D^-) \tag{10}$$

An approximate P-value is then computed. In the above table, none of the distributions is rejected by this test at the 5% significance level.

The other 5 tests, 2 of which are shown below, have both a standard form and a modified form:

<b>Goodness-of-Fit Tests for ppm</b>				
Modified Kolmogorov-Smirnov D				
	<i>Gamma</i>	<i>Lognormal</i>	<i>Normal</i>	<i>Weibull</i>
D	0.0905744	0.0953022	0.181741	0.0889679
Modified Form	0.633269	0.666324	1.26667	0.609933
P-Value	>=0.10*	>=0.10*	<0.01*	>=0.10*

Anderson-Darling A^2				
	<i>Gamma</i>	<i>Lognormal</i>	<i>Normal</i>	<i>Weibull</i>
A^2	0.33168	0.322124	1.87405	0.372536
Modified Form		0.322124	1.90586	0.383404
P-Value	*	>=0.10*	0.0000734208*	>=0.10*

\*Indicates that the P-Value has been compared to tables of critical values specially constructed for fitting the selected distribution. Other P-values are based on general tables and may be very conservative (except for the Chi-Squared Test).

The “modified form” of the statistic is specific to the distribution being fit. To determine whether to reject a specified distribution, one of two approaches is then taken:

1. In most cases, the modified statistic is compared to a table of critical values that have been obtained through Monte Carlo studies. In such a case, the output will display one of the following:

“>=0.10” if the statistic is less than or equal to the tabulated value for  $\alpha=0.10$ .

“<0.10” if the statistic is greater than the tabulated value for  $\alpha=0.10$  and less than or equal to the tabulated value for  $\alpha=0.05$ .

“<0.05” if the statistic is greater than the tabulated value for  $\alpha=0.05$  and less than or equal to the tabulated value for  $\alpha=0.01$ .

“<0.01” if the statistic is greater than the tabulated value for  $\alpha=0.01$ .

2. In a few cases, approximate P-values are computed.

Details and tables of critical values may be found in D’Agostino and Stephens (1988).

The available statistics are:

**Kolmogorov-Smirnov D** – This statistic calculates the maximum distance between the empirical c.d.f. and the fitted c.d.f., as discussed earlier.

**Kuiper V** - This statistic, calculated from the Kolmogorov statistics according to

$$V = D^+ + D^- \tag{11}$$

is often used for measurements of points which are distributed on a circle.

**Cramer-Von Mises  $W^2$**  - This statistic is related to the area between the empirical and fitted c.d.f.'s. It is calculated according to:

$$W^2 = \sum_{i=1}^n \left( z_{(i)} - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n} \tag{12}$$

**Watson  $U^2$**  - This statistic is a modified version of  $W^2$  designed for data recorded on a circle. It is calculated according to:

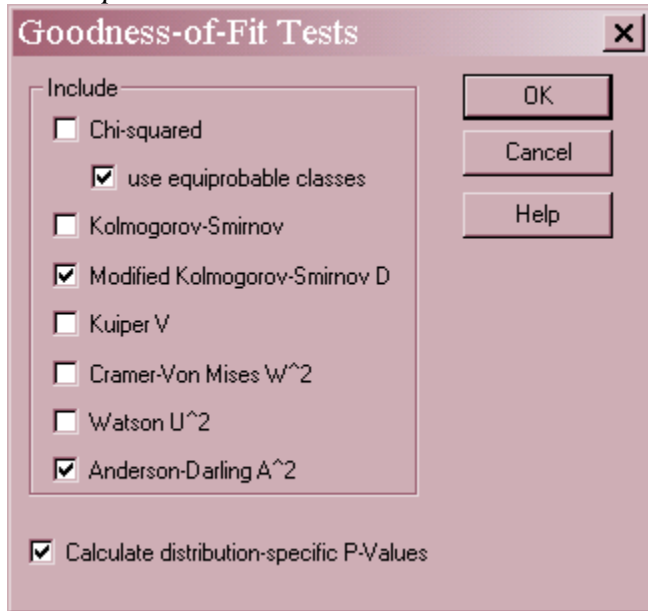
$$U^2 = W^2 - n(\bar{z} - 0.5)^2 \tag{13}$$

**Anderson-Darling  $A^2$**  - This statistic is a weighted measure of the area between the empirical and fitted c.d.f.'s. It is calculated according to:

$$A^2 = -n - \frac{\sum_{i=1}^n ((2i-1)\ln(z_{(i)}) + (2n+1-2i)\ln(1-z_{(i)}))}{n} \tag{14}$$

According to the tests displayed in the table above, any of the 3 distributions other than the normal should provide a reasonable model for the data.

*Pane Options*

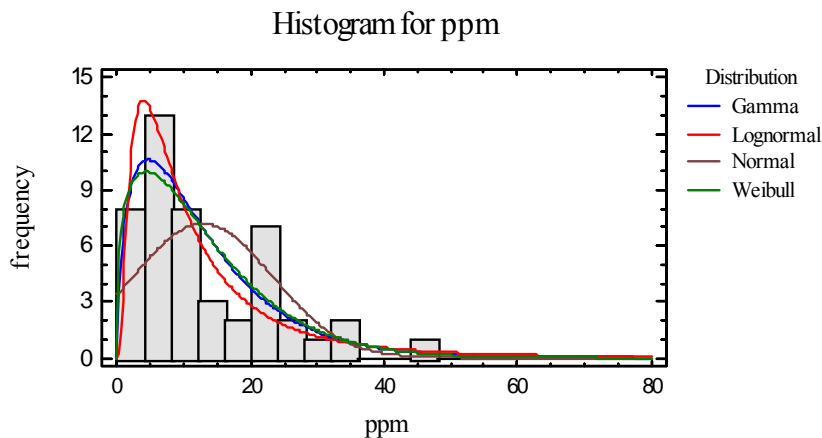


- **Include:** select one or more tests to perform. For the chi-squared test, select *use equiprobable classes* to group data into classes with equal expected frequencies. If this option is not checked, classes will be created that match the *Frequency Histogram*.

- **Calculate distribution-specific P-Values** – if checked, the P-Values will be based on tables or formulas specifically developed for the distribution being tested. Otherwise, the P-Values will be based on a general table or formula that applies to all distributions. The general approach is more conservative (will not reject a distribution as easily) but may be preferred when comparing P-Values amongst different distributions.

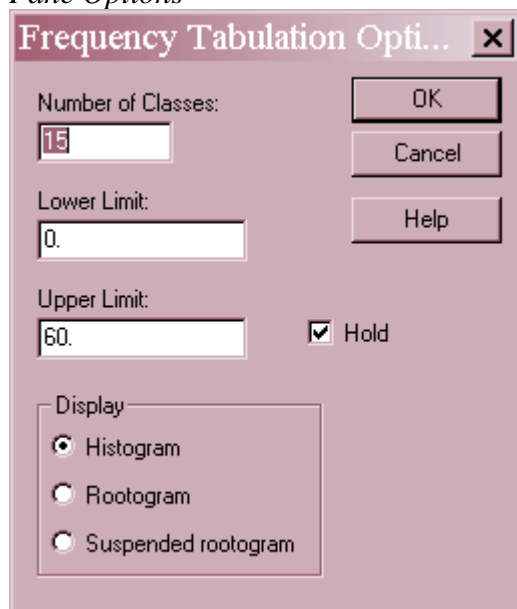
## Frequency Histogram

The best way to view the fitted distributions is through the *Frequency Histogram*. This pane shows a histogram of the data as a set of vertical bars, together with the estimated probability density or mass functions.



Notice that the 3 non-normal distributions are all positively skewed. The gamma and Weibull distributions are nearly identical, with the lognormal distribution having a somewhat higher peak.

### Pane Options

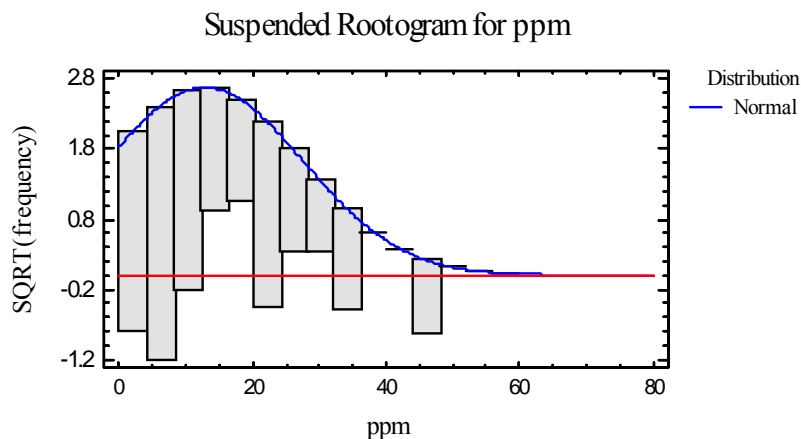


- **Number of classes:** the number of intervals into which the data will be divided. Intervals are adjacent to each other and of equal width. The number of intervals into which the data is

grouped by default is set by the rule specified on the *EDA* tab of the *Preferences* dialog box on the *Edit* menu.

- **Lower Limit:** lower limit of the first interval.
- **Upper Limit:** upper limit of the last interval.
- **Hold:** maintains the selected number of intervals and limits even if the source data changes. By default, the number of classes and the limits are recalculated whenever the data changes. This is necessary so that all observations are displayed even if some of the updated data fall beyond the original limits.
- **Display:** the manner in which to display the frequencies. A *Histogram* scales the bars according to the number of observations in each class. A *Rootogram* scales the bars according to the square root of the number of observations. A *Suspended Rootogram* scales by the square roots and suspends the bars from the curve.

#### Example – Suspended Rootogram for a Normal Distribution



The idea of using square roots is to equalize the variance of the deviations between the bars and the curve, which otherwise would increase with increasing frequency. The idea of suspending the bars from the curve is to allow an easier visual comparison with the horizontal line drawn at 0, since visual comparison with a curved line may be deceiving. Statistically, there are large discrepancies between the histogram and the fitted normal distribution in the above plot throughout the range of X.

## Comparison of Alternative Distributions

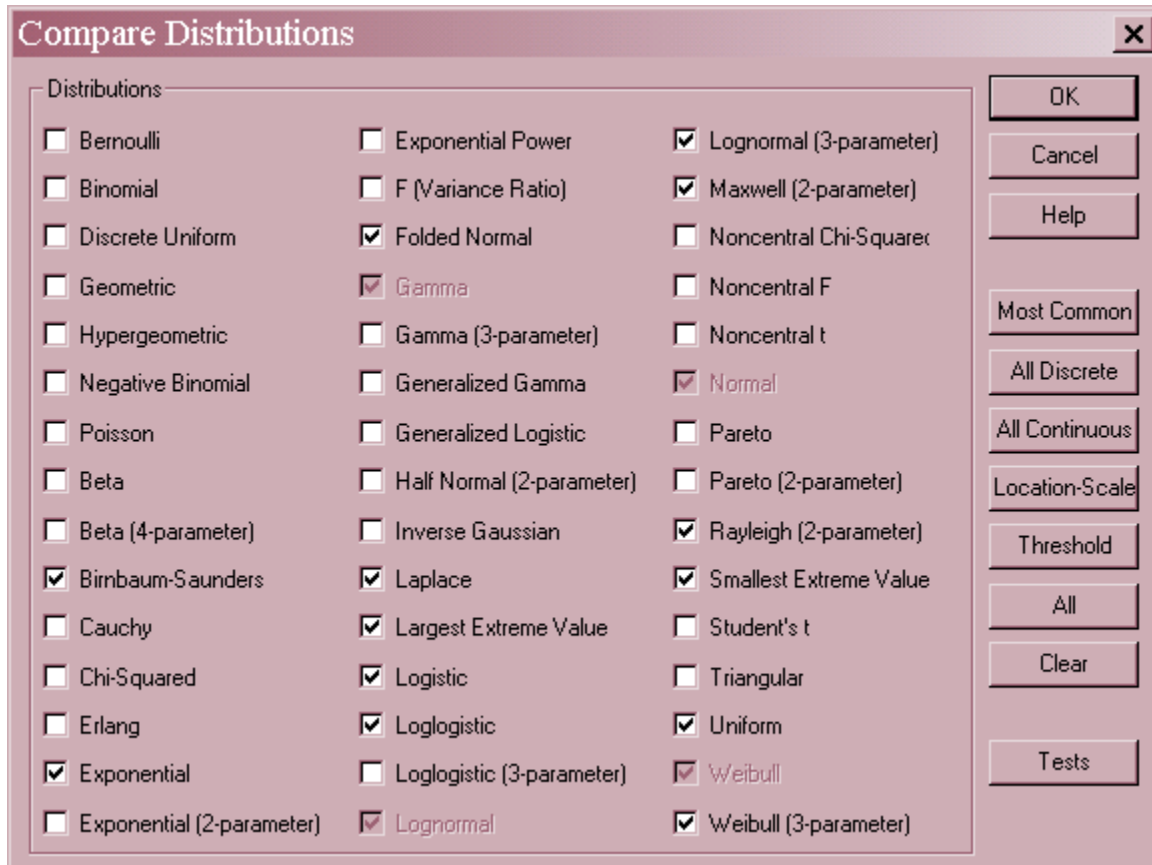
This pane automatically fits a collection of different distributions and displays them from top to bottom according to how well they fit the data.

Comparison of Alternative Distributions				
Distribution	Est. Parameters	Log Likelihood	KS D	A <sup>2</sup>
Weibull (3-Parameter)	3	-163.681	0.0858491	0.278173
Lognormal (3-Parameter)	3	-164.894	0.100446	0.295287
Lognormal	2	-165.372	0.0953022	0.322124
Gamma	2	-164.4	0.0905804	0.331721
Loglogistic	2	-165.949	0.0988569	0.342383
Weibull	2	-164.705	0.0889679	0.372536
Birnbaum-Saunders	2	-165.678	0.100989	0.559682
Folded Normal	2	-165.778	0.116524	0.572653
Exponential	1	-166.904	0.123457	0.975354
Largest Extreme Value	2	-168.64	0.112849	1.00414
Logistic	2	-175.464	0.130926	1.58494
Maxwell	2	-171.962	0.178343	1.71863
Rayleigh	2	-170.128	0.181461	1.82623
Normal	2	-176.458	0.181741	1.87405
Laplace	2	-175.979	0.16334	2.15572
Smallest Extreme Value	2	-188.009	0.223254	3.30569
Uniform	2	-180.997	0.414645	

The table shows:

- **Distribution** – the name of the distribution fit. You may select additional distributions using *Pane Options*.
- **Est. Parameters** – the number of estimated parameters for that distribution.
- **Log Likelihood** – the natural logarithm of the likelihood function. Larger values tend to indicate better fitting distributions.
- **KS D, A<sup>2</sup>, and other statistics** – values of various goodness-of-fit statistics, selected using the *Tests* button on the *Pane Options* dialog box. Smaller values tend to indicate better fitting distributions.

The distributions are sorted from best to worst according to one of the goodness-of-fit columns. That column is selected using the *Tests* button on the *Pane Options* dialog box. The above table shows the distributions sorted according to the value of the Anderson-Darling A<sup>2</sup> statistic. According to that statistic, the 3-parameter Weibull fits best.

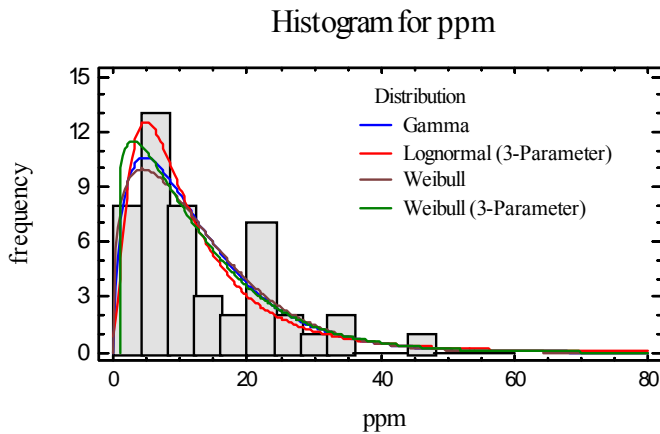
*Pane Options*

- **Distribution:** select the distributions to be fit to the data. The currently selected distributions are grayed out since they will be always included.
- **Most Common** – push this button to select the most commonly used distributions for variable (continuous) data.
- **All Discrete** – push this button to select all discrete distributions.
- **All Continuous** – push this button to select all continuous distributions.
- **Location-Scale** – push this button to select all distributions that are parameterized by a location parameter (such as a mean) and a scale parameter (such as a standard deviation).
- **Threshold** – push this button to select all distributions that contain a lower threshold parameter.
- **All** – push this button to select all distributions.
- **Clear** – push this button to deselect all distributions.
- **Tests** – push this button to display the dialog box used to specify the desired goodness-of-fit statistics:



- **Include** – select the goodness-of-fit statistics to be included in the table. The list includes the likelihood function and various statistics displayed on the *Goodness-of-Fit* pane.
- **Sort By** – select one of the included statistics to use to sort the distributions from best to worst.

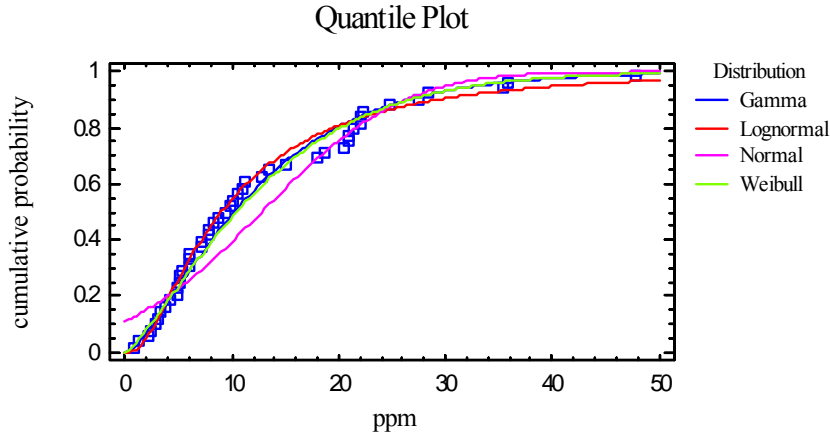
A plot of the 4 best-fitting distributions is shown below:





### Quantile Plot

The *Quantile Plot* shows the fraction of observations at or below X, together with the cumulative distribution function of the fitted distributions.



To create the plot, the data are sorted from smallest to largest and plotted at the coordinates

$$\left( x_{(j)}, \frac{j-0.5}{n} \right) \tag{15}$$

Ideally, the points will lie close to the line for the fitted distribution, as is the case in the plot above for each distribution other than the normal.

### Tail Areas

This pane shows the value of the cumulative distribution at up to 5 values of X.

Tail Areas for ppm				
Lower Tail Area (<=)				
X	Gamma	Lognormal	Normal	Weibull
10.0	0.490226	0.545171	0.393514	0.48063
15.0	0.679937	0.711046	0.582594	0.66815
20.0	0.805086	0.808051	0.75403	0.797378
25.0	0.883615	0.867472	0.878177	0.88075
30.0	0.931462	0.905528	0.949976	0.931978
Upper Tail Area (>)				
X	Gamma	Lognormal	Normal	Weibull
10.0	0.509774	0.454829	0.606486	0.51937
15.0	0.320063	0.288954	0.417406	0.33185
20.0	0.194914	0.191949	0.24597	0.202622
25.0	0.116385	0.132528	0.121823	0.11925
30.0	0.0685384	0.0944719	0.0500244	0.0680221

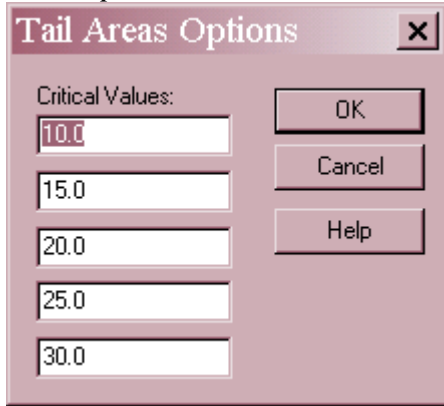
The table displays:

- **Lower Tail Area** – the probability that the random variable is less than or equal to X.

- **Upper Tail Area** – the probability that the random variable is greater than X.

For example, the probability of being less than or equal to  $X = 10$  for the gamma distribution is approximately 0.4902.

*Pane Options*



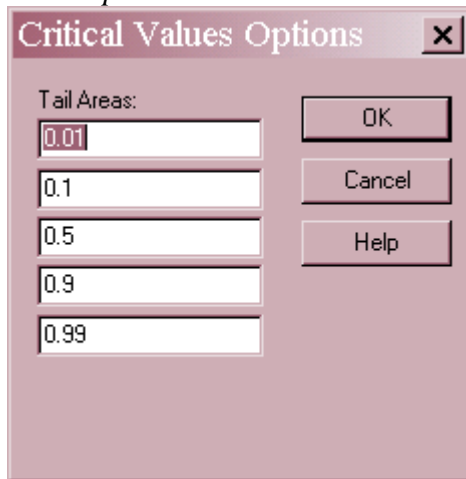
- **Critical Values:** values of X at which the cumulative probability is to be calculated.

**Critical Values**

This pane calculates the value of the random variable X below which lies a specified probability.

Critical Values for ppm				
<i>Lower Tail Area (&lt;=)</i>	<i>Gamma</i>	<i>Lognormal</i>	<i>Normal</i>	<i>Weibull</i>
0.01	0.547556	1.07182	-11.4769	0.387408
0.1	2.62818	2.78902	-0.563978	2.41186
0.5	10.2174	9.01355	12.8219	10.4487
0.9	26.4455	29.1299	26.2078	26.5964
0.99	47.5454	75.7997	37.1208	45.6136

The table displays the smallest value of X such that the probability of being less than or equal to X is at least the tail area desired. The table above shows that the c.d.f. of the fitted gamma distribution equals 0.01 at  $X = 0.548$ .

*Pane Options*


**Critical Values Options** [X]

Tail Areas:

0.01

0.1

0.5

0.9

0.99

OK

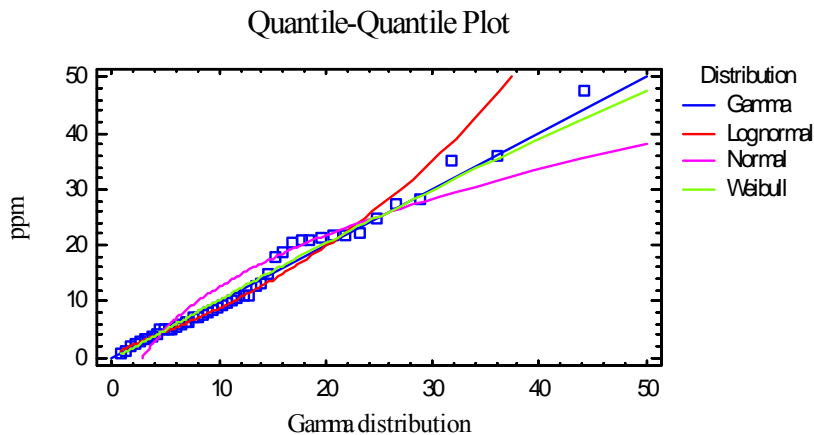
Cancel

Help

- **Tail Areas:** values of the c.d.f. at least to determine percentiles of the fitted distributions.

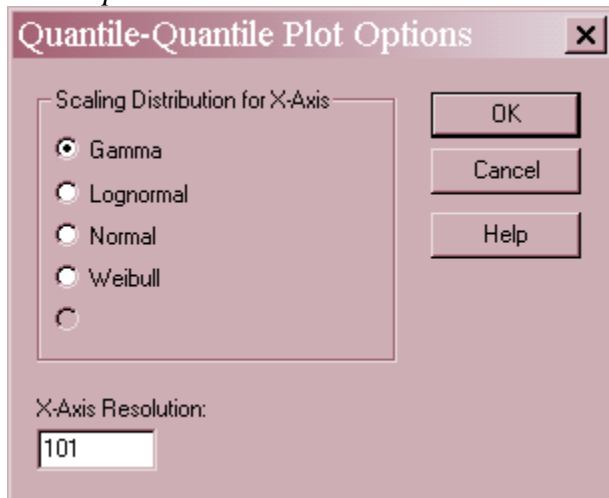
**Quantile-Quantile Plot**

The *Quantile-Quantile Plot* shows the fraction of observations at or below  $X$  plotted versus the equivalent percentiles of the fitted distributions.



One distribution, selected using *Pane Options*, is used to define the X-axis and is represented by the diagonal line. The others are represented by curves.

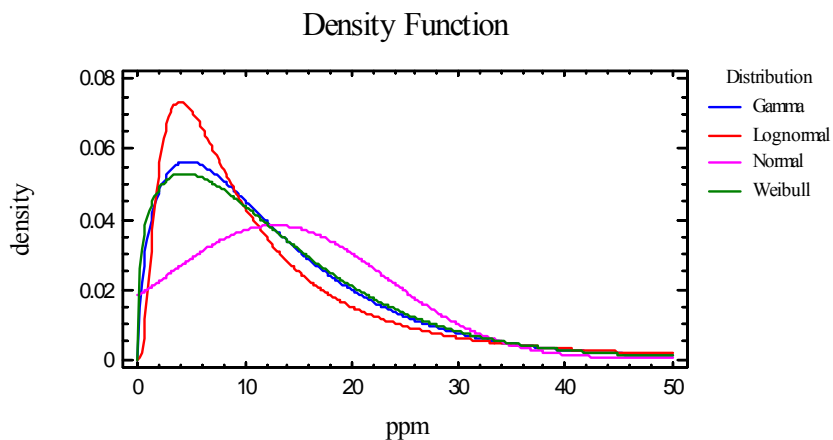
In the above plot, the fitted gamma distribution has been used to define the X-axis. The fact that the points lie close to the diagonal line confirms the fact that the gamma distribution provides a good model for the data, as does the Weibull distribution. The lognormal line is close at the lower end, but deviates away from the data at higher values of  $X$ . Evidently, the tail of the lognormal distribution is too fat. The line for the normal distribution completely misses the data.

*Pane Options*

- **Scaling Distribution for X-Axis:** the distribution used to scale the horizontal axis, corresponding to the diagonal line.
- **X-Axis Resolution** – the number of X locations at which the functions are plotted. Increase this value if the lines are not smooth enough.

**Distribution Functions 1 and 2**

These two panes plot various functions for the fitted distributions.

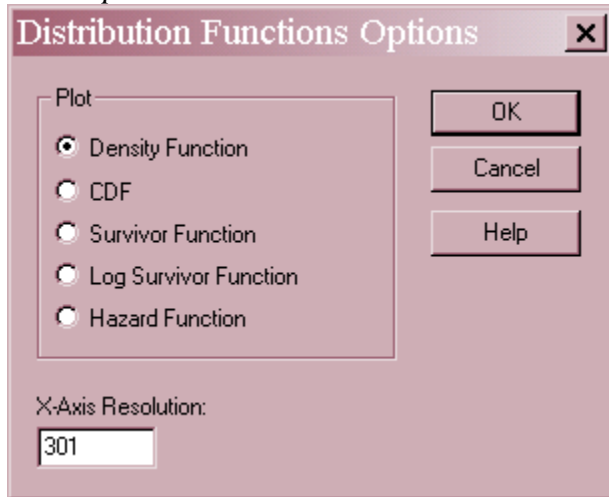


Using *Pane Options*, you may plot any of the following:

1. Probability density or mass function
2. Cumulative distribution function
3. Survivor function
4. Log survivor function
5. Hazard function

For definitions of these functions, see the documentation for *Probability Distributions*.

*Pane Options*



- **Plot:** the function to plot.
- **X-Axis Resolution** – the number of X locations at which the function is plotted. Increase this value if the lines are not smooth enough.

**Normal Tolerance Limits**

Statistical tolerance limits give a range of values for X such that one may be 100(1-α)% confident that P percent of the population from which a data sample comes falls within that range. Assuming that the data come from a normal distribution, a two-sided tolerance limit may be calculated by taking the sample mean plus and minus a multiple of the standard deviation, according to

$$\bar{x} \pm Ks \tag{16}$$

The factor K depends upon the sample size *n*, the level of confidence (1-α), and the specified percentage *P*.

<p><b>Normal Tolerance Limits for ppm</b></p> <p>Normal distribution                  Sample size = 47                  Mean = 12.8219                  Sigma = 10.445</p> <p>95.0% tolerance interval for 99.73% of the population                  Xbar +/- 3.66641 sigma                  Upper: 51.1177                  Lower: -25.4739</p>
--

For example, the above table states that one may be 95% confident that 99.73% of groundwater samples taken from the selected location in northwest Texas would have uranium concentrations between -25.5 and 51.1. This result is obviously bogus, however, since the data have been shown to NOT come from a normal distribution.

It is important to note that the above interval is not simply the interval under the fitted normal curve containing an area of 99.73%, which would correspond to  $\pm 3$  sigma. It is wider than that since it allows for sampling variability in both the mean and standard deviation.

You may select values of  $\alpha$  and  $P$  using *Pane Options*.

#### *Pane Options*

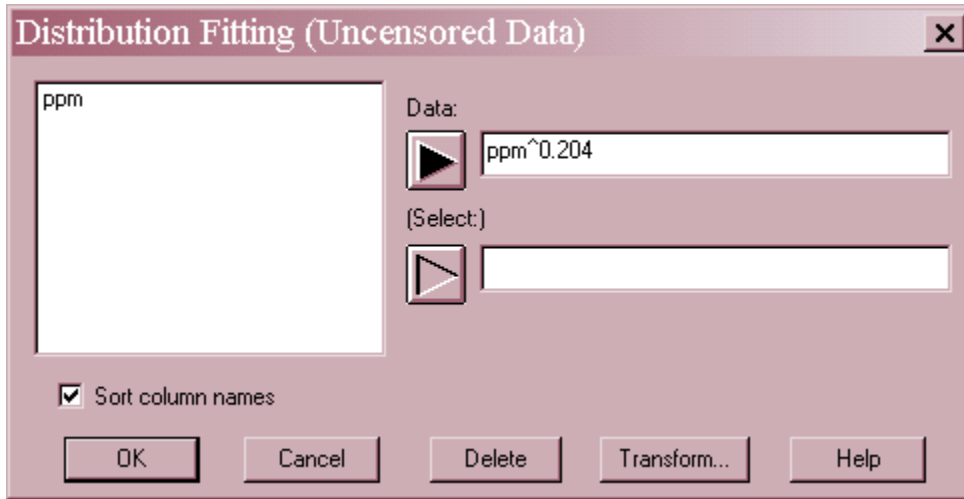
The image shows a dialog box titled "Normal Tolerance Limits Op...". It has three input fields: "Confidence Level" with a value of 95%, "Population Proportion" with a value of 99.73%, and a "Limits" section with three radio buttons: "Two-Sided" (selected), "Upper Only", and "Lower Only". There are three buttons on the right: "OK", "Cancel", and "Help".

- **Confidence Level** – specify the level of confidence for the tolerance limits, i.e.,  $100(1-\alpha)\%$ .
- **Population Proportion** – specify the percentage of the population  $P$  that the tolerance limits bound.
- **Limits** – select either two-sided tolerance limits or a one-sided tolerance bound.

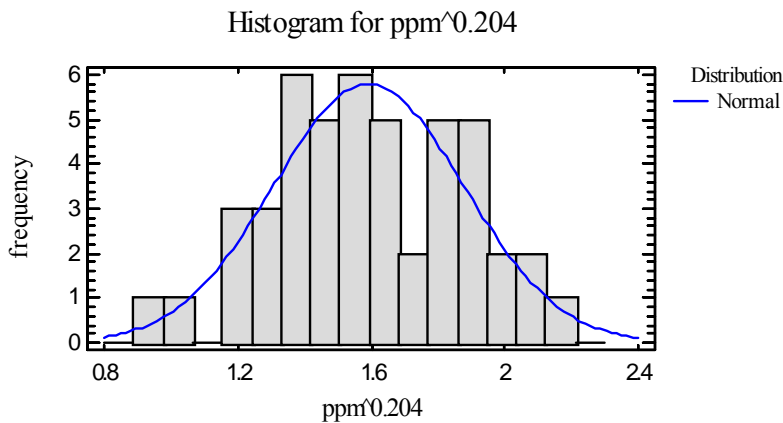
#### Example – One-Sided Tolerance Bound for Uranium Concentrations

One of the primary reasons for studying the uranium concentrations was to determine a reference distribution against which to compare future samples. For this purpose, it would be useful to derive an upper bound on the concentration beyond which a measurement might indicate an unusual event. A 99.9% upper tolerance limit would give a value that would be exceeded by chance only 1 time out of 1000.

In the documentation on *Power Transformations*, it was discovered that  $ppm^{0.204}$  was approximately normally distributed. In order to use the normal tolerance limits, the data input dialog box will be edited as shown below:



The *Frequency Histogram* verifies that the transformed values are approximately normal:



Requesting an upper 99.9% normal tolerance bound results in:

<p><b>Normal Tolerance Limits for ppm<sup>0.204</sup></b>                  Normal distribution                  Sample size = 47                  Mean = 1.59219                  Sigma = 0.28628</p> <p>95.0% tolerance bound for 99.9% of the population                  Xbar + 3.79169 sigma                  Upper: 2.67767</p>
--

The bound states that we can be 95% certain that 99.9% of all groundwater samples taken from the selected location will have values of ppm<sup>0.204</sup> no greater than 2.67767. Inverting the transformation yields the following upper bound in the original metric:

$$2.67767^{1/0.204} = 124.98$$

Concentrations of 125 or greater could therefore be considered as a signal of an unusual event.

## Distribution-Free Limits

The  $k$ -th smallest and  $k$ -th largest values in a data sample may be used to construct a tolerance limit for the population from which the data come without assuming any specific distribution. The resulting tolerance limits give a range of values for  $X$  such that one may be  $100(1-\alpha)\%$  confident that **at least**  $P$  percent of the population from which a data sample comes falls within that range. The interval can be quite conservative, with the actual percentage being much larger than that stated.

Distribution-Free Tolerance Limits for ppm	
Data summary	
Count =	47
Maximum =	47.78
Median =	9.44
Minimum =	0.74
95.0% tolerance interval for 90.2933% of the population	
Upper:	47.78
Lower:	0.74
(Based on an interval depth = 1)	

For example, the above table takes the most extreme values of *ppm* and states that one can be 95% confident that at least 90.2933% of all samples would have concentrations between 0.74 and 47.78.

In this procedure, you can select *Pane Options* to choose either the level of confidence  $100(1-\alpha)$  or the population percentage  $P$ , but not both.

### Pane Options

The screenshot shows a dialog box titled "Distribution-Free Tolerance Li...". It has three main sections: "Input", "Limits", and control buttons. In the "Input" section, "Confidence Level" is selected with a value of 95%, and "Population Proportion" is also set to 95%. The "Interval Depth" is set to 1. In the "Limits" section, "Two-Sided" is selected. There are "OK", "Cancel", and "Help" buttons on the right side.

- **Input** - specify either the level of confidence for the interval  $100(1-\alpha)$  or the population percentage  $P$ .
- **Interval Depth** – specify the value of  $k$  used to select the order statistics upon which the limits are based. In creating the interval, the procedure uses the  $k$ -th smallest and  $k$ -th largest data values.
- **Limits** – select either two-sided tolerance limits or a one-sided tolerance bound.



## Save Results

The following results can be saved to the datasheet:

1. *X* – up to 5 values at which tail areas were calculated.
2. *Tail Areas* – the calculated tail areas.
3. *P* – up to 5 lower tail areas at which critical values were calculated.
4. *Critical Values* – the calculated critical values.

Calculations**Kolmogorov-Smirnov P-Value**

Let  $d = \sqrt{n}D$ . Then:

$$P = 1 \quad \text{if } d < 0.22 \quad (17)$$

$$P = 1 - \frac{\sqrt{2\pi}}{d} \exp\left(\frac{-\pi^2}{8d^2}\right) \quad \text{if } 0.22 \leq d \leq 0.80 \quad (18)$$

$$P = 2e^{-2d^2} + e^{-8d^2} - e^{-18d^2} \quad \text{if } 0.80 < d \leq 3.15 \quad (19)$$

$$P = 0 \quad \text{if } d > 3.15 \quad (20)$$