
Chapter 6: Distribution Fitting

Overview	129
Define Input Data	131
Sample Data	131
Density Data	132
Cumulative Data	132
Filtering Your Data	133
Bringing Data Into @RISK For Fitting	133
Select Distributions To Fit	135
Continuous vs. Discrete Distributions.....	135
Estimated Parameters vs. Predefined Distributions.....	135
Domain Limits	136
Run The Fit	139
Sample Data - Maximum Likelihood Estimators (MLEs).....	139
Curve Data - The Method of Least Squares.....	141
Interpret the Results	143
Ranking of Fits	143
Graphs	143
Statistics and Targets.....	146
Fit Statistics	146
P-Values and Critical Values	149
Using the Results of a Fit.....	151
Exporting Graphs and Reports	151
Specifying Distributions in Excel	152

Overview

@RISK allows you to fit probability distributions to your data (Professional and Industrial versions only). Fitting is done when you have a set of collected data that you want to use as the basis for an input distribution in your spreadsheet. For example, you may have collected historical data on a product price and you might want to create a distribution of possible future prices that is based on this data.

Fitting is done using an integrated copy of **BestFit**, Palisade Corporation's distribution fitting software package. This program may also be run without @RISK. BestFit running without @RISK looks very similar to the @RISK - Model window, except that there is no Model tab.

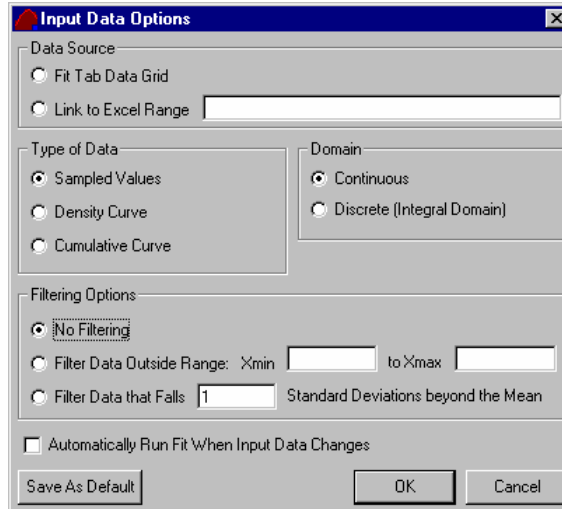
To fit distributions to data using @RISK, there are five steps that you should consider:

1. **Define Input Data**
2. **Specify Distributions to Fit**
3. **Run the Fit**
4. **Interpret the Results**
5. **Using the Results of a Fit**

Each of these steps is discussed in this chapter.

Define Input Data

@RISK allows you to analyze three kinds of data for distribution fitting: sample, density and cumulative. @RISK supports up to 100,000 data points for each of these types. The available data types are shown in the Input Data Options dialog box in the Model window.



Sample Data

Sample (or observation) data is a set of values drawn randomly from a large population. Distributions are fit to sample data to estimate the properties of that population.

Continuous vs. Discrete Samples

Sample data is either continuous or discrete. Continuous sample data can take on any value over a continuous range, while discrete data is limited to integer values. Discrete data can be entered in two formats. In the “standard” format, you enter each data point individually. In the “counted” format, the data is entered in pairs, where the first value is the sampled value and the second is the number of samples drawn with that value.

Data Requirements

Data requirements for sample data include:

- ◆ You must have at least five data values.
- ◆ Discrete data values must be integral.
- ◆ All sample values should fall in the range $-1E+37 \leq x \leq +1E+37$.

Density Data

Density data is a set of (x,y) points that describe the probability density function of a continuous distribution. Distributions are fit to density data to give the best representation of the curve points using a theoretical probability distribution.

Normalization of Density Data

Since all probability distribution functions must have unit area, @RISK automatically will scale your y-values so that the density curve described by your data has an area of one. Since the points you specify are isolated points on a continuum, linear interpolation between these points is used to calculate the normalization factor. In certain cases, such as fitting to data generated from a mathematical function already known to be normalized, it is undesirable to have @RISK apply its own normalization. In these cases, you may turn off this feature.

Data Requirements

Data requirements for density data include:

- ◆ You must have at least three (x,y) data pairs.
- ◆ All x-values must be in the range $-1E+37 \leq x \leq +1E+37$.
- ◆ All x-values should be distinct.
- ◆ All y-values must be in the range $0 \leq y \leq +1E+37$.
- ◆ At least one y-value must be non-zero.

Cumulative Data

Cumulative data is a set of (x,p) points that describe a continuous cumulative distribution function. The p-value associated with a given x-value is the probability of obtaining a value less than or equal to x. Distributions are fit to cumulative data to give the best representation of the curve points using a theoretical probability distribution.

Endpoint Interpolation

In order to calculate statistics and display graphs of your cumulative data, @RISK needs to know where the input minimum and maximum are (that is, the points with p=0 and p=1). If you do not explicitly supply these points, @RISK will linearly interpolate them from your data. In general, it is recommended that you always include the p=0 and p=1 points in your data set, if possible.

Data Requirements

Data requirements for cumulative data include:

- ◆ You must have at least three (x,p) data pairs.
- ◆ All x-values must be in the range $-1E+37 \leq x \leq +1E+37$.
- ◆ All x-values must be distinct.
- ◆ All p-values must be in the range $0 \leq p \leq 1$.
- ◆ Increasing x-values must always correspond to increasing p-values.

Filtering Your Data

You can further refine your input data by applying an input filter. Filtering tells @RISK to ignore outliers, based on criteria you specify, without requiring you to explicitly remove them from your data set. For example, you may wish to only analyze x-values greater than zero. Or, you may wish to filter values that lie outside two standard deviations from the mean.

Bringing Data Into @RISK For Fitting

Numerous methods are available for bringing your data into @RISK. You may type it directly into the input data grid, paste it from another Windows application, fit directly from the @RISK-Results window, or even create a link between @RISK and an Excel spreadsheet.

The **Paste** command imports data to the input sheet from the Windows Clipboard. To do so, highlight the data you want to copy and paste into your @RISK input sheet. To import data from a non-spreadsheet application, be sure that paired data is tab-delimited and each set ends with a carriage return.

Data that resulted from an @RISK simulation can be fit very easily by choosing the **Fit** command from the Explorer list pop-up menu in the @RISK-Results window.

You can also link the data on a fit tab to a range in Microsoft Excel. To do this, highlight the range you want to link, and click the **Fit Distributions to Data** command from within Excel. Alternatively, you can set the link directly in the **Input Data Options** dialog in the @RISK-Model window.

Select Distributions To Fit

After you define your data set, you must specify the distributions you want @RISK to attempt to fit. There are three general questions you must answer to do this.

Continuous vs. Discrete Distributions

For sample data, you should first decide if your data is continuous or discrete. Discrete distributions always return integer values. For example, presume you have a set of data describing the number of failures in a series of 100 trial batches. You would only want to fit discrete distributions to this set because partial failures are not allowed. In contrast, continuous data can take on any value in a range. For example, presume you have a set of data describing the height, in inches, of 300 people. You would want to fit continuous distributions to this data, since heights are not restricted to integral values.

If you specify that your data is discrete, all your data values must be integers. Keep in mind, however, that the converse is not true. Just because you have all integral data values does not mean you have to fit only discrete distributions. In the previous example, the height data set may be rounded to the nearest inch, but fitting to continuous distributions is still appropriate.

@RISK does not support the fitting of discrete distributions to density and cumulative curve data.

You can specify whether your data set is continuous or discrete in the **Input Data Options** dialog.

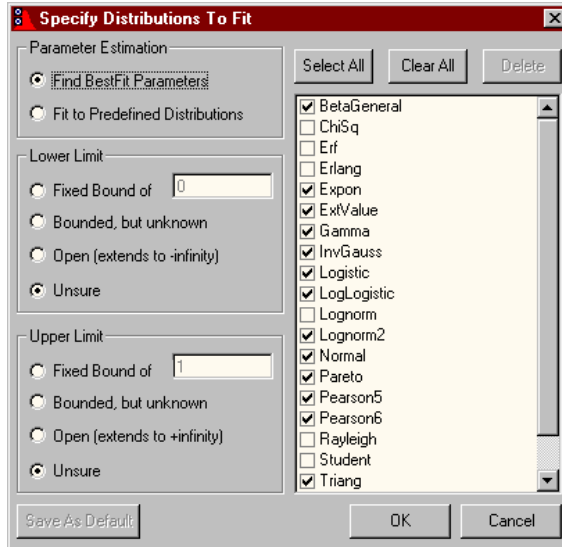
Estimated Parameters vs. Predefined Distributions

Generally, you will want @RISK to estimate the parameters of your distributions. However, in some cases, you may want to specify exactly what distributions to use. For example you may want to have @RISK compare two competing hypotheses and tell you which one is a better description of your data.

Predefined distributions can be set in the **Specify Distributions to Fit** dialog.

Domain Limits

For continuous data sets (sample or curve data) you can specify how you want @RISK to treat the upper and lower limits of the distributions. For both limits there are four choices: fixed bound, unknown bound, open bound, and unsure. Domain limits can be set in the **Specify Distributions to Fit** dialog.



Fixed Bound

If you specify a fixed bound, you are telling @RISK that the limit of the distribution must be the value you specify. For example, if you have a data set of the times between arrivals of customers in a queue, you might want to fit distributions which have a fixed lower bound of zero, since it is impossible to have a negative time between events.

Unknown Bound

If you specify an unknown bound, you are telling @RISK that the limit of the distribution has a finite bound (that is, it does not extend to plus or minus infinity). Unlike a fixed bound, however, you do not know what the actual value of the limit is. You want @RISK to choose the value for you as it performs its fit.

Open Bound

If you specify an open bound, you are telling @RISK that the limit of the distribution must extend to minus infinity (for a lower bound) or plus infinity (for an upper bound).

Unsure

This is the default option. It is the combination of an unknown bound and an open bound. The limits of distributions that are non-asymptotic are treated as in the unknown bound case, while asymptotic distributions are still included as in the open bound case.

Note, not all distributions functions are compatible with all the possible choices. For example, you can not specify a fixed or unknown lower bound for the Normal distribution, since it asymptotically extends to minus infinity.

Run The Fit

To start the fitting process, click the **Run Fit** icon on the Fitting toolbar.

For each of the distributions specified in the previous step, @RISK will try to find the set of parameters that make the closest match between the distribution function and your data set. Keep in mind, @RISK does not produce an absolute answer, but rather identifies a distribution that *most likely* produced your data. Always evaluate your @RISK results quantitatively and qualitatively, examining both the comparison graphs and statistics before using a result.

@RISK uses two methods to calculate the best distributions for your data set. For sample data, distribution parameters are estimated using Maximum Likelihood Estimators (MLEs). For density and cumulative data (collectively known as curve data), the method of least squares is used to minimize the root-mean square error between the curve points and the theoretical function.

Sample Data – Maximum Likelihood Estimators (MLEs)

The MLEs of a distribution are the parameters of that function that maximize the probability of obtaining the given data set.

Definition

For any density distribution $f(x)$ with one parameter α , and a corresponding set of n sampled values X_i , an expression called the likelihood may be defined:

$$L = \prod_{i=1}^n f(X_i, \alpha)$$

To find the MLE, simply maximize L with respect to α :

$$\frac{dL}{d\alpha} = 0$$

and solve for α . The method described above can be easily generalized to distributions with more than one parameter.

A Simple Example

An exponential function with a fixed lower bound of zero has only one adjustable parameter, and its MLE is easily calculated. The distribution's density function is:

$$f(x) = \frac{1}{\beta} e^{-x/\beta}$$

and the likelihood function is:

$$L(\beta) = \prod_{i=1}^n \frac{1}{\beta} e^{-X_i/\beta} = \beta^{-n} \exp\left(-\frac{1}{\beta} \sum_{i=1}^n X_i\right)$$

To simplify matters, we can use the natural log of the likelihood function:

$$l(\beta) = \ln L(\beta) = -n \ln(\beta) - \frac{1}{\beta} \sum_{i=1}^n X_i$$

To maximize the log of the likelihood, simply set its derivative with respect to β to zero:

$$\frac{dl}{d\beta} = \frac{-n}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^n X_i$$

which equals zero when:

$$\beta = \sum_{i=1}^n \frac{X_i}{n}$$

Therefore, when @RISK tries to fit your data to the best Exponential function with a fixed lower bound of zero, it first finds the mean of the input data and uses it as the MLE for β .

Modifications to the MLE Method

For some distributions, the MLE method described above does not work. For example, a 3-parameter Gamma distribution (a Gamma distribution whose lower bound is allowed to vary) can not always be fit using MLEs. In these cases @RISK will resort to a hybrid algorithm, which combines the standard MLE approach with a moment matching procedure.

In certain distributions, a strict MLE method produces parameters which are heavily biased for small sample sizes. For example, the MLE of the “shift” parameter of an exponential distribution and the minimum and maximum parameters of the uniform distribution are heavily biased for small sample sizes. Where possible, @RISK will correct for the bias.

Curve Data – The Method of Least Squares

The root-mean square error (RMSErr) between set of n curve points (X_i , Y_i) and a theoretical distribution function $f(x)$ with one parameter α is:

$$RMSErr = \sqrt{\frac{1}{n} \sum_{i=1}^n (f(x_i, \alpha) - y_i)^2}$$

The value of α that minimizes this value is called the least squares fit. In a sense, this value minimizes the “distance” between the theoretical curve and the data. The formula above is easily generalized to more than one parameter.

This method is used to calculate the best distribution for both density and cumulative curve data.

Interpret the Results

Once @RISK has completed the fitting process, you should review its results. @RISK provides a powerful array of graphs, statistics, and reports to help you evaluate fits and select the best choice for your models.

Ranking of Fits

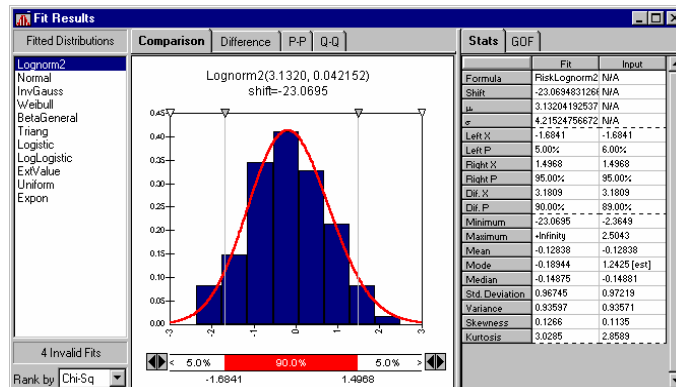
@RISK ranks all the fitted distributions using one or more fit statistics. For continuous sample data, you can choose to rank fits by their chi-squared statistic, Anderson-Darling statistic, or Kolmogorov-Smirnov statistic. Each of these statistics is discussed in more detail, later in this section. For discrete sample data, only the chi-squared statistic can be used. For density and cumulative curve data, the fits are ranked by their RMS Error value.

Graphs

@RISK provides four types of graphs to help you visually assess the quality of your fits.

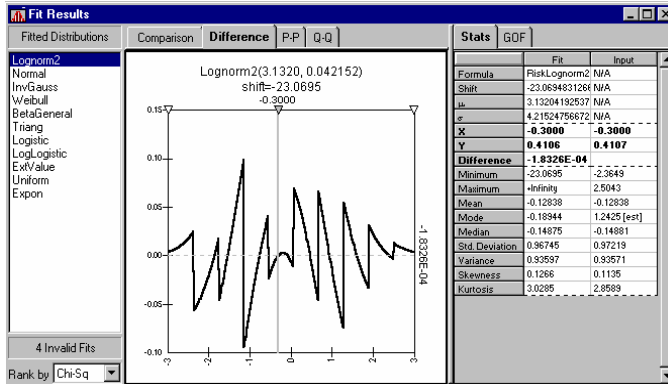
Comparison Graphs

A comparison graph superimposes the input data and fitted distribution on the same graph, allowing you to visually compare them either as density or cumulative curves. This graph allows you to determine if the fitted distribution matches the input data in specific areas. For example, it may be important to have a good match around the mean or in the tails.



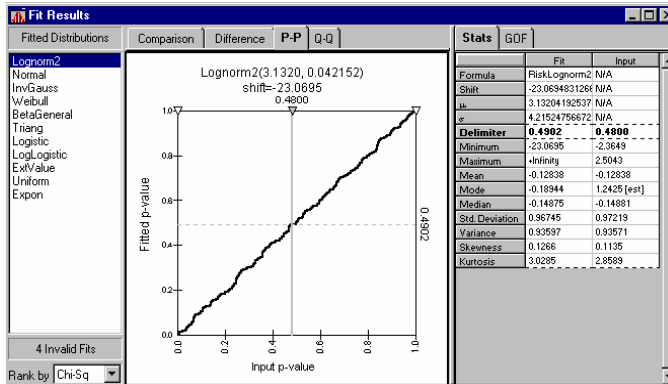
Difference Graphs

A difference graph displays the absolute error between the fitted distribution and the input data.



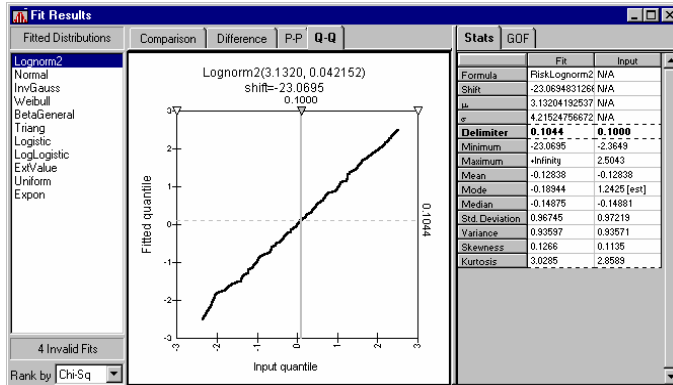
P-P Graphs

Probability-Probability (P-P) graphs plot the distribution of the input data (P_i) vs. the distribution of the result ($F(x_i)$). If the fit is "good," the plot will be nearly linear. P-P graphs are only available for fits to sample data.



Q-Q Graphs

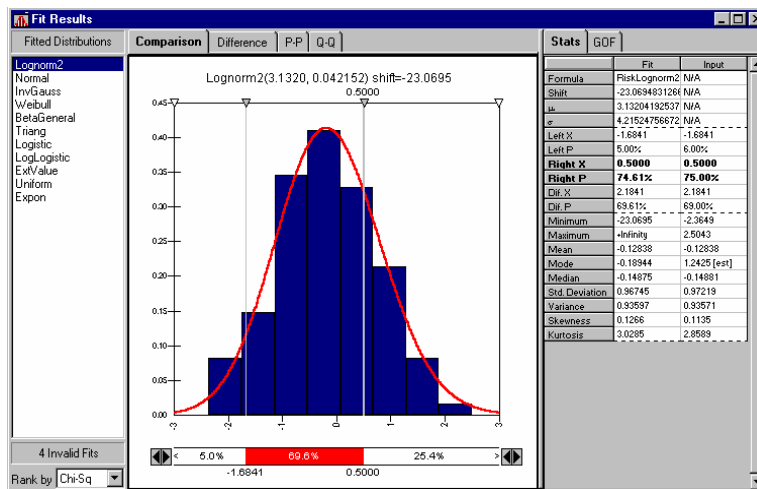
Quantile-Quantile (Q-Q) graphs plots percentile values of the input distribution (x_i) vs. percentile values of the result ($F^{-1}(P_i)$). If the fit is "good," the plot will be nearly linear. Q-Q graphs are only available for fits to continuous sample data.



Statistics and Targets

@RISK reports basic statistics (mean, variance, mode, etc.) for each fitted distribution, which can easily be compared to the same statistics for the input data.

@RISK allows you to compare percentiles and target values between distributions and the input data. For example, perhaps the 5th and 95th percentiles values are important to you. This can be done in two ways. First, all @RISK graphs have a set of “delimiters” which allow you to visually set two different targets or percentiles. Second, the @RISK Summary Report has a data entry area for specifying up to ten different targets or percentiles.



Fit Statistics

For each fit, @RISK reports one or more fit statistics. These statistics measure how good the distribution fits the input data and how confident you can be that the data was produced by the distribution function. For each of these statistics, the smaller the value, the better the fit. @RISK makes use of four different fit statistics: chi-squared, Kolmogorov-Smirnov, Anderson-Darling, and Root-Mean Squared Error.

When more than one fit statistic is available, there is no hard rule to decide which test will give you the "best" result. Each test has its strengths and weaknesses. You must decide which information is most important to you when considering which test to use.

Chi-Squared Statistic

The chi-squared statistic is the best known goodness-of-fit statistic. It can be used with both continuous and discrete sample data. To calculate the chi-squared statistic, you first must break up the x-axis domain into several “bins”. The chi-squared statistic is then defined as:

$$\chi^2 = \sum_{i=1}^K \frac{(N_i - E_i)^2}{E_i}$$

where

K = the number of bins

N_i = the observed number of samples in the i^{th} bin

E_i = the expected number of samples in the i^{th} bin.

A weakness of the chi-squared statistic is that there are no clear guidelines for selecting the number and location of the bins. In some situations, you can reach different conclusions from the same data depending on how you specified the bins.

Some of the arbitrariness of the bin selection can be removed by telling @RISK to use equiprobable bins. In this mode, @RISK will adjust the bin sizes based on the fitted distribution, trying to make each bin contain an equal amount of probability. For continuous distributions this is straightforward. For discrete distributions, however, @RISK will only be able to make the bins approximately equal.

@RISK allows you full control of how bins are defined for the chi-squared test. These settings are in the **Define Chi-Squared Binning** dialog.

Kolmogorov- Smirnov Statistic (K-S)

Another fit statistic that can be used for continuous sample data is the Kolmogorov-Smirnov statistic, which is defined as

$$D_n = \sup \left[\left| F_n(x) - \bar{F}(x) \right| \right]$$

where

n = total number of data points

$\bar{F}(\hat{x})$ = the fitted cumulative distribution function

$$F_n(x) = \frac{N_x}{n}$$

N_x = the number of X_i 's less than x .

The K-S statistic does not require binning, which makes it less arbitrary than the chi-squared statistic. A weakness of the K-S statistic is that it does not detect tail discrepancies very well.

Anderson-Darling Statistic (A-D)

The final fit statistic that can be used with continuous sample data is the Anderson-Darling Statistic, which is defined as:

$$A_n^2 = n \int_{-\infty}^{+\infty} [F_n(x) - \bar{F}(x)]^2 \Psi(x) \bar{f}(x) dx$$

where

n = total number of data points

$$\psi^2 = \frac{1}{\bar{f}(\hat{x})[1 - \bar{F}(\hat{x})]}$$

$\bar{f}(\hat{x})$ = the hypothesized density function

$\bar{F}(\hat{x})$ = the hypothesized cumulative distribution function

$$F_n(x) = \frac{N_x}{n}$$

N_x = the number of X_i 's less than x .

Like the K-S statistic, the A-D statistic does not require binning. But unlike the K-S statistic, which focuses in the middle of the distribution, the A-D statistic highlights differences between the tails of the fitted distribution and input data.

Root-Mean Squared Error (RMSErr)

For density and cumulative curve data, the only fit statistic used is the Root-Mean Squared Error. This is the same quantity that @RISK minimized to determine the distribution parameters during its fitting process. It is a measure of the "average" squared error between the input and fitted curve.

P-Values and Critical Values

The goodness-of-fit statistic reports a measure of the deviation of the fitted distribution from the input data. As mentioned earlier, the smaller the fit statistic is, the better the fit. But how small a value is needed for a “good” fit? For fits to sample data, this section explains how P-values and critical values can be used to analyze the “goodness” of a fit.

For the discussion below, suppose we have a distribution fitted to a set of N sampled values, and a corresponding fit statistic, s .

P-Values

How likely is it that a new set N samples drawn from the fitted distribution would generate a fit statistic greater than or equal to s ? This probability is referred to as the P-value and is sometimes called the “observed significance level” of the test. As the P-value decreases to zero, we are less and less confident that the fitted distribution could possibly have generated our original data set. Conversely, as the P-value approaches one, we have no basis to reject the hypothesis that the fitted distribution actually generated our data set.

Critical Values

Often we want to turn the same question around and specify a particular level of significance to use, usually denoted by α . This value is the probability that we will incorrectly reject a distribution because it generated, due to statistical fluctuations, a value of s that was very large. Now we want to know, given this significance level, what the largest value of s is that we would accept as a valid fit. This value of s is called the “critical value” of the fit statistic at the α level of significance. Any fit that has a value of s above the critical value is rejected, while fits with values of s below the critical value are accepted. Typically, critical values depend on the type of distribution fit, the particular fit statistic being used, the number of data points, and the significance level.

Calculation Methods in @RISK

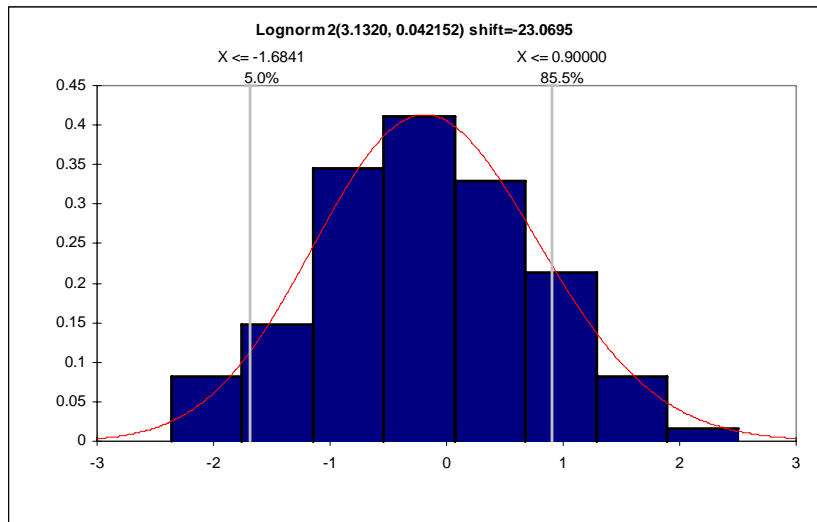
For the chi-squared test, the P-values and critical values can be calculated by finding the appropriate points on a chi-square distribution with $k-1$ degrees of freedom (where k is the number of bins). While this method is exactly correct when predefined distributions are used, it turns out to be only an approximation for distributions where @RISK estimated one or more distribution parameters. Conveniently, however, the approximation is always a conservative one. That is, the reported values for both your critical values and P-value will be slightly higher than the exact values. More information about this can be found in *Appendix D: Recommended Readings* in this manual.

Most critical values and P-values for the A-D and K-S fit statistics have been found by very detailed Monte-Carlo studies (see *Appendix D: Recommended Readings* for references). Unfortunately, not all distributions have been analyzed in enough detail for @RISK to be able to report them. Where possible @RISK will report the appropriate P-values and critical values. Often, where an exact P-value calculation is not possible, a range is returned for the P-value, indicating that the true P-value lies between the specified upper and lower limit.

Using the Results of a Fit

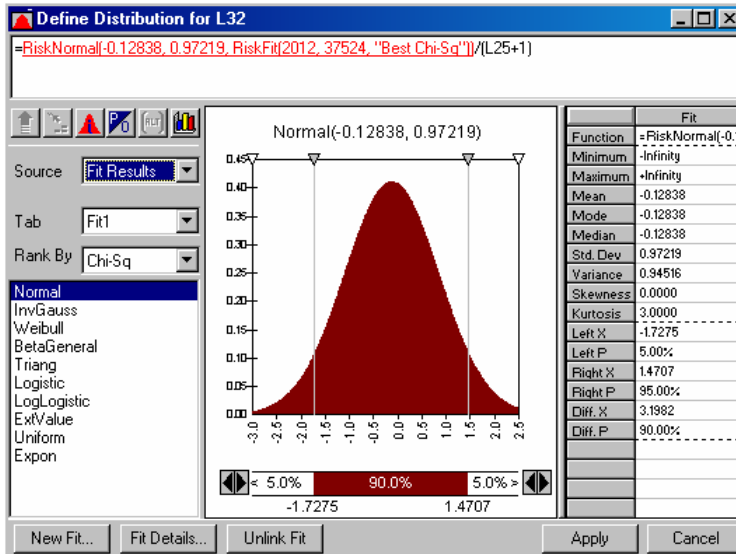
Exporting Graphs and Reports

Once you have analyzed the results of your calculation, you may wish to export the results to another program. Of course, you can always copy and paste any @RISK graph or report into Excel or another Windows application via the clipboard. In addition, using the **Graph In Excel** command, @RISK allows you to create a copy of the current @RISK graph in Excel's native chart format.



Specifying Distributions in Excel

Often you will want to use the result of a fit in an @RISK model. If you are using @RISK Professional or Industrial with the built-in fitting capabilities of BestFit, this is easy. In the **Define Distribution** window, simply change the source list box to say “Fit Result” and select your distribution.



A quick method for using a fit to define a distribution in an @RISK model using data that resides in an Excel spreadsheet is to click the **New Fit** button from the **Define Distribution** window. Just pop-up the Define Distribution window when selecting the cell with the uncertain value. Then, click New Fit and select the data in Excel to fit. Your fit will automatically run and return the fit results to the window. Then you can select your distribution from the Fit Results.