Proceedings of the 1996 Winter Simulation Conference
ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain

BESTFIT, DISTRIBUTION FITTING SOFTWARE BY PALISADE CORPORATION

Linda Jankauskas
Sam McLafferty

Palisade Corporation
31 Decker Road
Newfield, NY 14867, U.S.A.

ABSTRACT

BestFit is distribution-fitting software for Microsoft Windows that finds the statistical distribution function that best fits a data set. BestFit provides a flexible, easy-to-use environment for analysis that allows users to focus on their data, not their software.

Users of simulation software can use distributions produced by BestFit to define uncertainty in their simulation models. BestFit helps you find the best representation of randomness in your model, making your simulation results more accurate. Users of @RISK, ProModel or any other simulation software will find BestFit an invaluable tool for defining uncertainty.

1 OVERVIEW OF THE SOFTWARE

BestFit's goal is to find the distribution that best fits your input data. BestFit does not produce an absolute answer, it identifies a distribution that most likely produced your data. For a given distribution, BestFit looks for the parameters of the function that optimize the goodness-of-fit, a measurement of the probability that the input data was produced by the given distribution.

Since its release in 1993, BestFit has undergone several upgrades. Each upgrade incorporated new features and capabilities directly requested by BestFit users. Palisade Corporation is dedicated to responding to its customers needs by offering new features and compatibility in a timely manner.

1.1 Entering a Data Set

BestFit allows three kinds of data: sample, density and cumulative. Methods for bringing your data into BestFit include: typing it directly into the input sheet, copying it from another Windows application, opening a text file containing the data or creating a link between BestFit and an Excel or Lotus 1-2-3 spreadsheet.

1.2 The BestFit Calculation

For each distribution type, BestFit makes a first guess of the best distribution parameters using maximum-likelihood estimators (MLEs). The MLEs of a distribution are the parameters of that function that maximize the likelihood of the distribution given a set of observational data.

BestFit optimizes the parameters calculated from the MLEs using the Levenberg-Marquardt method, an algorithm that maximizes the goodness-of-fit between a data set and a distribution function. This method takes a first guess of the parameters (the MLEs) and varies each parameter slightly until a good fit is found.

1.3 Graphs in BestFit

A good way of interpreting the results of a BestFit calculation is to visually assess how well a distribution agrees with the input data. Four graphs are available for this purpose: comparison, difference, P-P and Q-Q.

Comparison Graphs: The comparison graph superimposes the input and result distributions on the same graph, allowing you to visually compare them either as density or cumulative curves. This graph allows you to determine if the best fit distribution is well matched to 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: The difference graph displays the absolute error between the input and result distributions. This error is defined as the difference between the input and result probability for each input value of X. Comparing the magnitude of the error to the magnitude of the result gives you an idea of the extent to which the result deviates from the input.

Probability-Probability Graphs: Probability-Probability (or P-P) graphs plot the distribution of the input data (P_i) against the distribution of the result ($F(x_i)$). If the fit is good the plot will be nearly linear.

Quantile-Quantile Graphs: Quantile-Quantile (or Q-Q) graphs plot the plot percentile values of the input distribution (x_i) against percentile values of the result ($F^{1}(P_i)$). If the fit is good the plot will be nearly linear.

1.4 The BestFit Statistical Report

Basic statistics (mean, variance, mode, etc.) for each distribution are reported and can be compared to the statistics of the input. For each result, goodness-of-fit values and the corresponding confidence intervals are reported. 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.

The targeting option in BestFit compares percentile values and probabilities between distributions and the input data. For example, if the 5th and 95th percentile values are important to you, you can easily compare them in the BestFit statistical report.

Statistical results, including distribution functions, can be transferred to other Windows applications using the clipboard. In addition, results can be saved to a text file for export to other operating systems.

1.5 System Requirements and Compatibility

BestFit is available in 16-bit and 32-bit versions. The 32-bit version requires Microsoft Windows 95 or Windows NT. The 16-bit version requires Windows 3.1, Windows NT or Windows 95. Both versions will run under the minimum system requirements for the operating system. A multi-user, network version of BestFit can be used on any server-based PC network and is available in 16-bit and 32-bit versions.

2 HOW BESTFIT WORKS

For a given distribution, BestFit looks for the parameters of the function that optimize the goodness-of-fit. BestFit follows these steps when finding the best fit for your input data:

1. For each distribution type selected, parameters values are estimated using MLEs (maximum-likelihood estimators).

2. The fit is optimized using the Levenberg-Marquardt method (optional).

3. The goodness-of-fit is measured for the optimized function using chi-square, Kolmogorov-Smirnov and/or Anderson-Darling tests.

4. All functions are compared and the one with the lowest goodness-of-fit values is considered the best fit.

5. Detailed statistics (mean, variance, skewness, etc.) are reported for the input data and all results.

6. Goodness-of-fit statistics and their confidence intervals are reported for each distribution function.

2.1 Probability Distributions

BestFit offers 26 distribution functions. The density and distribution (when available) used for each function is defined in Evans, et al. (1993) and Law and Kelton (1982).

Table 1: Distributions Available in BestFit

Beta	Lognormal,
Binomial	Lognormal2
Chi-Square	Negative Binomial
Error Function	Normal
Erlang	Pareto
Exponential	Pearson Type V
Extreme Value	Pearson Type VI
Gamma	Poisson
Geometric	Rayleigh
Hypergeometric	Student's t
Inverse Gaussian	Triangular
Logistic	Uniform
Log-Logistic	Weibull

2.1 Maximum-Likelihood Estimators

As defined by Law and Kelton (1982), the MLEs of a distribution are the parameters of that function that maximize the likelihood of the distribution given a set of observational data. For any density distribution $f(x)$ with one parameter α and a corresponding set of observational data X_i, an expression called the likelihood may be defined:

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

To find the MLE, use (1) to maximize L with respect to α:

$$\frac{dL}{d\alpha} = 0 \tag{2}$$

and solve (2) for α.

This paper was presented at the 1995 Winter Simulation Conference in Crystal City, VA.

BestFit calculates the maximum-likelihood estimators for each distribution using formulas defined in Evans et al. (1993) and Law and Kelton (1982).

2.2 Levenberg-Marquardt Method

BestFit can determine the best fit of a data set simply from its MLEs or it can optimize the parameters using the Levenberg-Marquardt method. The Levenberg-Marquardt Method is an algorithm used to maximize the goodness-of-fit between a data set and a distribution function. As explained by Press, et al. (1988), this method takes a first guess of the parameters of the distribution function, and then varies each parameter slightly until it finds a good fit.

The Levenberg-Marquardt method, which is a non-linear least-squares routine, takes an iterative approach to try to minimize the goodness-of-fit statistic (chi-square). The Levenberg-Marquardt method does not find the absolute minimum for chi-square; rather, it finds a local minima. The success of this method depends on the initial parameters used. Therefore, a good first guess (i.e., one produced from a well-defined input) will produce a good result, while a poor first guess might not provide a useful result.

The goodness-of-fit test used by BestFit for optimizing a distribution is the chi-square test. This test was chosen because it is the most frequently used and the fastest to calculate.

2.3 Goodness-Of-Fit Statistics

The process of calculating MLEs and optimizing the chi-square value gives a best guess for each distribution. BestFit then measures the quality of each fit using goodness-of-fit measurements.

2.3.1 What is Goodness-of-Fit?

Formally, goodness-of-fit is defined as the probability of the data given the parameters. In other words, the goodness-of-fit statistic tells you how probable it is that a given distribution function produced your data set. The goodness-of-fit statistic is usually used in a relative sense by comparing the values to the goodness-of-fit of other distribution functions.

BestFit offers three goodness-of-fit tests: chi-square, Kolmogorov-Smirnov and Anderson-Darling. The chi-square is the most common, but the others may supply more detailed information about your distribution.

2.3.2 Chi-Square Test

The chi-square test is the most common goodness-of-fit test. It can be used with any type of input data (sample, density or cumulative) and any type of distribution function (discrete or continuous).

A weakness of the chi-square test is that there are no clear guidelines for selecting intervals. In some situations, you can reach different conclusions from the same data depending on how you specified the intervals (number of classes).

The chi-square statistic used by BestFit, as defined in Law and Kelton (1982), is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(P_i - p_i)^2}{p_i} \qquad (3)$$

In (3) we define the probability values as

P_i = the observed probability value for a given histogram bar

p_i = the theoretical probability that a value will fall with the X range of the histogram bar

2.3.3 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test does not depend on the number of intervals, which makes it more powerful than the chi-square test. A weakness of the Kolmogorov-Smirnov test is that it does not detect tail discrepancies very well.

The Kolmogorov-Smirnov statistic used by BestFit, as defined in Law and Kelton (1982), is:

$$D_n = \sup\left[\left|F_n(x) - \hat{F}(x)\right|\right] \qquad (4)$$

In (4) we define the parameter values as

n = total number of data points

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

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

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

2.3.4 Anderson-Darling Test

The Anderson-Darling test is very similar to the Kolmogorov-Smirnov test, but it places more emphasis on tail values. It does not depend on the number of intervals.

This paper was presented at the 1995 Winter Simulation Conference in Crystal City, VA.

The Anderson-Darling Statistic used by BestFit, as defined in Anderson and Darling (1954), is:

$$A_n^2 = n \int_{-\infty}^{+\infty} \left[F_n(x) - \hat{F}(x)\right]^2 \Psi(x) \hat{f}(x) dx \qquad (5)$$

In (5) we define the parameter values as

$$\Psi = \frac{1}{\hat{F}(x)\left[1 - \hat{F}(x)\right]}$$

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

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

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

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

2.4 Confidence Levels and Critical Values

For the goodness of fit tests, the critical value determines whether you should accept or reject a fitted distribution at a given confidence level (α). The method of statistical hypothesis testing allows you to use a structured analytical method to make a decision regarding your BestFit results. This method allows you to control or measure the uncertainty involved in the decision.

Typically, the critical value depends on the type of distribution fit, the number of data points, and the confidence interval. Most critical values for the Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit statistics have been found by Monte Carlo studies (see Chandra, et al. 1981, Stephens 1974 and Stephens 1977). BestFit reports both the smallest confidence value that passes the null hypothesis test (has a critical value greater than the goodness-of-fit statistic) and the critical values for selected confidence intervals.

2.4.1 Critical Values for the Chi-Square Test

For the chi-square test, BestFit calculates the critical value by finding the target point for the 1-α percentile of a chi-square distribution with *N*-1 degrees of freedom (*N* is the number of classes).

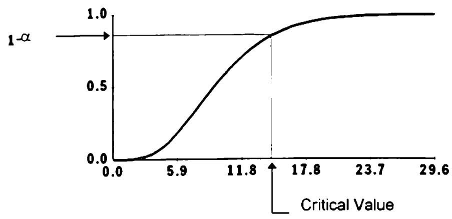

Figure 1: A Chi-Square Distribution With 10 Degrees of Freedom

When the calculated chi-square statistic is larger than the critical value, the null hypothesis should be rejected (the distribution is not a good fit). This particular measurement of the critical value is a conservative test, it does not take into account that the parameters were estimated from the data. The actual confidence is greater than or equal to α.

2.4.2 Critical Values for the Kolmogorov-Smirnov and Anderson-Darling Tests

The critical values for the Kolmogorov-Smirnov and Anderson-Darling tests were calculated using Monte Carlo simulation techniques as explained by Stephens (1974), Stephens (1977) and Chandra et al. (1981).

Unlike the chi-square test, where the critical value is the same for all distributions, the Kolmogorov-Smirnov and Anderson-Darling tests include special cases for the Normal, Exponential, Weibull and Extreme Value distributions. For all other distributions, the critical values are estimated using an "all parameters known" test that is more conservative than the tests for specific distributions. BestFit marks estimated confidence levels with an asterisk (*) in the statistical reports.

For the Kolmogorov-Smirnov and Anderson-Darling tests, BestFit does not compare the critical value to the actual test statistics. Instead, BestFit modifies the critical value using formulas defined by Stephens (1974), Stephens (1977) and Chandra et al. (1981). These modifications take the number of data points into consideration when evaluating confidence levels.

2.4.3 Interpreting the Critical Value

The tests described in this article are very sensitive to the magnitude *n* (the number of data points). If *n* is small, the goodness of fit will only measure large difference between the input data and the distribution function. As *n* increases, the modified test statistics increase and the null hypothesis will be rejected more often. The results produced by these tests should be considered "guidelines" in selecting a fit. Always evaluate the results by comparing statistics and graphs before accepting or rejecting a fit.

3 ADVANCED FEATURES OF BESTFIT

BestFit includes many advanced features that make distribution fitting easier and more powerful.

3.1 Linking BestFit to Spreadsheet Data

BestFit can directly link to a range of data in a spreadsheet. Instead of creating a copy of your data in BestFit, the project is linked directly to the source of the data and automatically updates every time the data changes in the spreadsheet.

BestFit acts like an add-in by displaying a new toolbar icon in Microsoft Excel or Lotus 1-2-3 and creating a new BestFit command in the Tools menu. To use BestFit as an add-in, simply highlight a range of data in your spreadsheet and click on the BestFit command. BestFit opens with a link to the selected data, allowing you to immediately run an analysis.

3.2 Filtering

BestFit's filtering option defines outliers in your data which will be ignored during a BestFit analysis. Analyze a subset of your data without creating a new project file.

For example, you may wish to only analyze X values greater than zero. Or, you may wish to filter out tail values by only looking at data within a few standard deviations of the mean.

3.3 Transferring Results to Your Spreadsheet

All BestFit results, including graphs and statistical reports, can be transferred to a spreadsheet or other software at the click of a button. Statistical reports can be copied to the Windows clipboard and pasted into other software packages for further customization.

BestFit can transfer the data points of a graph to Lotus 1-2-3 or Excel and create a graph in the spreadsheet's native format. Or, an image of the graph can be copied to the Windows clipboard and pasted into other software packages.

3.4 Subjective Probability Assessment

To generate a probability distribution in the absence of historical data, BestFit interacts with RISKview (distribution viewing software from Palisade Corporation). RISKview can be purchased by itself or bundled with BestFit.

RISKview generates a generalized BETA distribution from subjective input using two special functions: BETA-SUBJECTIVE and BETA-PERT. The BETA-SUBJECTIVE function accepts estimated minimum, most likely, mean and maximum values as inputs. The function gives you the flexibility of a generalized function like the TRIANGULAR while producing a more realistic distribution. The BETA-PERT function uses the classic Project Management PERT function to create a BETA distribution using estimated minimum, most likely and maximum values as inputs.

RISKview offers other distribution assessment tools including distribution drawing and smoothing, function previews, statistics, target values and other subjective probability techniques.

REFERENCES

Anderson, T. W., and D. A. Darling. 1954. A test of goodness of fit. *American Statistical Association Journal.* 1954:765-769.

Chandra, M., N.D. Singpurwalla and M.A. Stephens. 1981. Kolmogorov statistics for tests of fit for the extreme value and weibull distribution. *Journal of the American Statistical Association.* 76:729-31.

Evans, Merran, Nicholas Hastings and Brian Peacock. 1993. *Statistical Distributions.* 2d ed. New York: John Wiley & Sons, Inc.

Law, A. M., and W. D. Kelton. 1982. *Simulation modeling and analysis.* 2d ed. New York: McGraw-Hill Book Company.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1988. *Numerical recipes in C: The art of scientific computing.* New York: Cambridge University Press.

Stephens, M.A. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association.* 69:730-737.

Stephens, M.A. 1977. Goodness of fit for the extreme value distribution. *Biometrika.* 64:583-588.

AUTHOR BIOGRAPHIES

LINDA JANKAUSKAS is a Software Engineer at Palisade Corporation. She is involved in developing software for decision analysis, including BestFit, RISKview and DecisionTree.

SAM MCLAFFERTY is the President of Palisade Corporation. He has been developing risk analysis software for over ten years and is currently involved in developing software for decision analysis including @RISK and TopRank.

This paper was presented at the 1995 Winter Simulation Conference in Crystal City, VA.