Proceedings of the 2003 Winter Simulation Conference
S. Chick, P. J. Sanchez, D. Ferrin, and D. J. Morris, eds.

HOW THE EXPERTFIT DISTRIBUTION-FITTING SOFTWARE CAN MAKE YOUR SIMULATION MODELS MORE VALID

Averill M. Law
Michael G. McComas

Averill M. Law and Associates, Inc.
P.O. Box 40996
Tucson, AZ 85717, U.S.A.

ABSTRACT

In this paper, we discuss the critical role of simulation input modeling in a successful simulation study. Two pitfalls in simulation input modeling are then presented and we explain how any analyst, regardless of their knowledge of statistics, can easily avoid these pitfalls through the use of the ExpertFit distribution-fitting software. We use a set of real-world data to demonstrate how the software automatically specifies and ranks probability distributions, and then tells the analyst whether the "best" candidate distribution is actually a good representation of the data. If no distribution provides a good fit, then ExpertFit can define an empirical distribution. In either case, the selected distribution is put into the proper format for direct input to the analyst's simulation software.

1 THE ROLE OF SIMULATION INPUT MODELING IN A SUCCESSFUL SIMULATION STUDY

In this section we describe simulation input modeling and show the consequences of performing this critical activity improperly.

1.1 The Nature of Simulation Input Modeling

One of the most important activities in a successful simulation study is that of representing each source of system randomness by a probability distribution. For example in a manufacturing system, processing times, machine times to failure, and machine repair times should generally be modeled by probability distributions. If this critical activity is neglected, then one's simulation results are quite likely to be erroneous and any conclusions drawn from the simulation study suspect – in other words, "garbage in, garbage out."

In this paper, we use the phrase "simulation input modeling" to mean the process of choosing a probability distribution for each source randomness for the system under study and of expressing this distribution in a form that can be used in the analyst's choice of simulation software. In Sections 2 and 3 we discuss how an analyst can easily and accurately choose an appropriate probability distribution using the ExpertFit software. Section 4 discusses important features that have recently been added to ExpertFit.

1.2 Two Pitfalls in Simulation Input Modeling

We have identified a number of pitfalls that can undermine the success of a simulation study (see Law and Kelton 2000). Two of these pitfalls that directly relate to simulation input modeling are discussed in the following two sections [see our Web site <www.averill-law.com> ("ExpertFit Distribution-Fitting Software") for further discussion of pitfalls, and for a more comprehensive discussion of ExpertFit, in general].

1.2.1 Pitfall Number 1: Replacing a Distribution by its Mean

Simulation analysts have sometimes replaced an input probability distribution by its perceived mean in their simulation models. This practice may be caused by a lack of understanding of this issue on the part of the analyst or by lack of information on the actual form of the distribution (e.g., only an estimate of the mean of the distribution is available). Such a practice may produce completely erroneous simulation results, as is shown by the following example.

Consider a single-server queueing system (e.g., a manufacturing system consisting of a single machine tool) at which jobs arrive to be processed. Suppose that the mean interarrival time of jobs is 1 minute and that the mean service time is 0.99 minute. Suppose further that the interarrival times and service times each have an exponential distribution. Then it can be shown that the long-run mean number of jobs waiting in the queue is *approximately 98*. On the other hand, suppose we were to follow the dangerous practice of replacing each source of randomness with a constant value. If we assume that each interarrival time is *exactly* 1 minute and each service time is *exactly* 0.99 minute, *then each job is finished before the next arrives and no job ever waits in the queue*! The variability of the probability distributions, rather than just their means, has a significant effect on the congestion level in most queueing-type (e.g., manufacturing) systems.

1.2.2 Pitfall Number 2: Using the Wrong Distribution

We have seen the importance of using a distribution to represent a source of randomness. However, as we will now see, the actual distribution used is also critical. It should be noted that many simulation practitioners and simulation books widely use normal input distributions, even though in our experience this distribution will *rarely* be appropriate to model a source of randomness such as service times.

Suppose for the queueing system in Section 1.2.1 that jobs have exponential interarrival times with a mean of 1 minute. We have 200 service times that have been collected from the system, but their underlying probability distribution is unknown. Using ExpertFit, we fit the best Weibull distribution and the best normal distribution (and others) to the observed service-time data. However, as shown by the analysis in Section 6.7 of Law and Kelton (2000), the *Weibull distribution* actually provides the best overall model for the data.

We then made a *very long* simulation run of the system using *each* of the fitted distributions. The average number of jobs in the queue for the Weibull distribution was 4.41, which should be close to the average number in queue for the actual system. On the other hand, the average number in queue for the normal distribution was 6.13, corresponding to a *model output error of 39 percent*. It is interesting to see how poorly the normal distribution works, given that it is the most well-known distribution.

We will see in Section 2 how the use of ExpertFit makes choosing an appropriate probability distribution a quick and easy process.

1.3 Advantages of Using ExpertFit

With the assistance of ExpertFit, an analyst, regardless of their prior knowledge of statistics, can avoid the two pitfalls introduced above. When system data are available, a complete analysis with the package takes just minutes. The package identifies the "best" of the candidate probability distributions, and also tells the analyst whether the fitted distribution is good enough to actually use in the simulation model. If none of the candidate distributions provides an adequate fit, then ExpertFit can construct an empirical distribution. In either case, the selected distribution can be represented automatically in the analyst's choice of simulation software. Appropriate probability distributions can also be selected when no system data are available. For the important case of machine breakdowns, ExpertFit will specify time-to-failure and time-to-repair distributions that match the system's behavior, even if the machine is subject to blocking or starving.

2 USING EXPERTFIT WHEN SYSTEM DATA ARE AVAILABLE

We consider first the case where data are available for the source of randomness to be represented in the simulation model. Our goal is to give an overview of the capabilities of ExpertFit – a demo disk with a thorough discussion of program operation is available from the authors.

We have designed ExpertFit based on our 25 years of research and experience in selecting simulation input distributions to be easy to use but without sacrificing technical correctness. The user interface employs four tabs that are typically used sequentially to perform an analysis. Furthermore, the options in each tab have default settings to promote ease of use. There are many illuminating graphs available and multiple distributions can be plotted on each.

There are two modes of operation that allow the analyst to configure ExpertFit to their particular needs. *Standard Mode* contains features sufficient for 95 percent of all analyses and focuses the user on those features that are really important. *Advanced Mode* contains numerous additional features for the sophisticated user.

ExpertFit has the most extensive documentation in the simulation industry, which includes 450 pages of context-sensitive help for *all* menus and *all* results tables/graphs, an online feature index and tutorials, and a user's guide with 8 complete examples.

The first data-analysis tab has options for obtaining the data set and for displaying its characteristics. An analyst can read a data file, manually enter a data set, paste in a data set from the Clipboard, or import a data set from Excel. Once a data set is available, a number of graphical and tabular data summaries can be created, including histograms, sample statistics (e.g., mean, variance, skewness, etc.), and plots designed to assess the independence of the observations.

The data set we have chosen for this example consists of 856 ship-loading times, which were provided to us by a major oil company.

At the second tab distributions are fit to the data set. For the recommended automated-fitting option, the only information required by ExpertFit to begin the fitting and evaluation process is a specification of the range of the underlying random variable. Since all we know about the data is that the values are non-negative, we accepted the default limits of "zero" and "infinity." ExpertFit responds by fitting distributions with a range starting at zero and also distributions whose lower endpoint was estimated from the data itself. These candidate models were then automatically evaluated and the results screen shown in Figure 1 was displayed.

ExpertFit fit and ranked 26 candidate models, with the three best-fitting models and their estimated parameters being displayed on the screen, along with their relative scores. The displayed scores are calculated using a proprietary evaluation scheme that is based on our 25 years of experience and research in this area, including the analysis of 35,000 computer-generated data sets. Results from the heuristics that we have found to be the best indicators of a good model fit are combined and the resulting numerical evaluation is normalized so that 100 indicates the best possible model and 0 indicates the worst possible model. These scores are *comparative* in nature and do not give an overall assessment of the quality of fit. ExpertFit provides a separate *absolute* evaluation of the quality of the representation provided by the best-ranked model. This absolute evaluation is critical because, perhaps, one third of all data sets are not well represented by a standard theoretical distribution. *Furthermore, ExpertFit is the only software package that provides such a definitive absolute evaluation.*

In Figure 1 we see that the log-logistic distribution (with a range starting at zero) is the best model for the ship-loading time data. Furthermore, the Absolute Evaluation is "Good," which indicates that this distribution is good enough to use in a simulation model. Although the log-logistic distribution may be unfamiliar to you, it occurs widely in practice and is easy to use in most simulation packages.

However, it is generally desirable to confirm the quality of the representation using the third tab. It should also be noted that ExpertFit completed the entire analysis without any further input from the analyst. After automated fitting, the analyst is automatically transferred to the third tab, where the specified models can be compared to the sample to confirm the quality of fit (if additional confirmation is desired). Two of our favorite comparisons are the Density/Histogram Overplot and the Distribution-Function-Differences Plot, which are shown in Figures 2 and 3, respectively. In the former case, the density function of the log-logistic distribution has been plotted over a histogram of the data (a graphical estimate of the true density function). This plot indicates that the log-logistic distribution is a good model for the observed data. The Distribution-Function-Differences Plot graphs the differences between a sample distribution function (a graphical estimate of the true distribution function) and the distribution function of the log-logistic

Relative Evaluation of Candidate Models

Model	Relative Score	Parameters	
1 – Log-Logistic	100.00	Location Scale Shape	0.00000 0.82199 8.84027
2 – Pearson Type VI	91.00	Location Scale Shape #1 Shape #2	0.00000 0.25314 99.97455 31.06366
3 - Pearson Type V	88.00	Location Scale Shape	0.00000 19.18409 23.78474

26 models are defined with scores between 0.00 and 100.00

Absolute Evaluation of Model 1 – Log-Logistic

Evaluation: Good
Suggestion: Additional evaluations using Comparisons Tab might be informative.

Additional Information About Model 1 – Log-logistic

"Error" in the model mean relative to the sample mean: 2.8975e-3 = 0.34%

Figure 1: Evaluation of the Candidate Models for the Ship-Loading Time Data

distribution. Since these vertical differences are small (i.e., within the horizontal error bounds), this also suggests that the log-logistic distribution is a good representation for the data. Note that the third tab also allows the analyst to *correctly* perform goodness-of-fit tests such as the chi-square, Kolmogorov-Smirnov, and Anderson-Darling tests. ExpertFit includes an option in the fourth tab for displaying the representation of the log-logistic distribution using different simulation packages. We show in Figure 4 the representations for four of the simulation packages supported by ExpertFit.

For some data sets, no candidate model provides an adequate representation. In this case we recommend the use of an empirical distribution. Note that ExpertFit allows an empirical distribution to be based on all data values or on a histogram to reduce the information that is needed for specification. We show a histogram-based representation (with 20 intervals) for two simulation packages in Figure 5.

3 USING EXPERTFIT WHEN NO DATA ARE AVAILABLE

Sometimes a simulation analyst must model a source of randomness for which no system data are available. ExpertFit provides two types of analyses for this situation. A general task time (e.g., a service time) can be modeled in ExpertFit by using a triangular or beta distribution. In the case of a triangular distribution, the analyst specifies the distribution by giving subjective estimates of the minimum, maximum, and most-likely task times.

ExpertFit will also help the analyst specify time-to-failure and time-to-repair distributions for a machine that randomly breaks down. In this case, the analyst gives, for example, subjective estimates for the percentage of time that the machine is operational (e.g., 90 percent) and for the mean repair time.

4 NEW FEATURES IN EXPERTFIT

The following are new ExpertFit features:

- There are now two levels of precision available: Normal and High. *Normal Precision* (the default) provides, for many data sets, good estimates of the parameters of a distribution and has a small execution time. *High Precision* provides much-better parameter estimates for most data sets, but has a larger execution time. High Precision provided better-fitting distributions for 84 percent of the 69 real-world data sets tested.
- ExpertFit now has an online feature index, which allows the user to find the location of any software feature quickly.
- It is now possible to use the Weibull and lognormal distributions to model a task time in the absence of data – the user just specifies the minimum task time, the most-likely task time, and, say, the 90th percentile of task time. Up to now, one was forced to use the triangular distribution, which has a number of fundamental shortcomings (e.g., the inability to model task times whose density function has a long right tail).

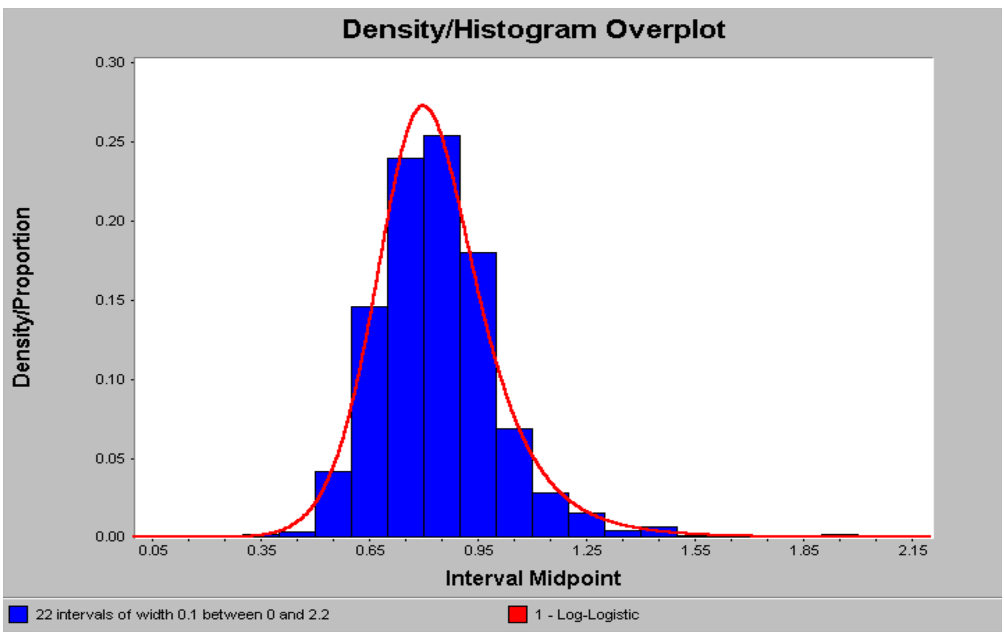

Figure 2: Density/Histogram Overplot for the Ship-Loading Time Data

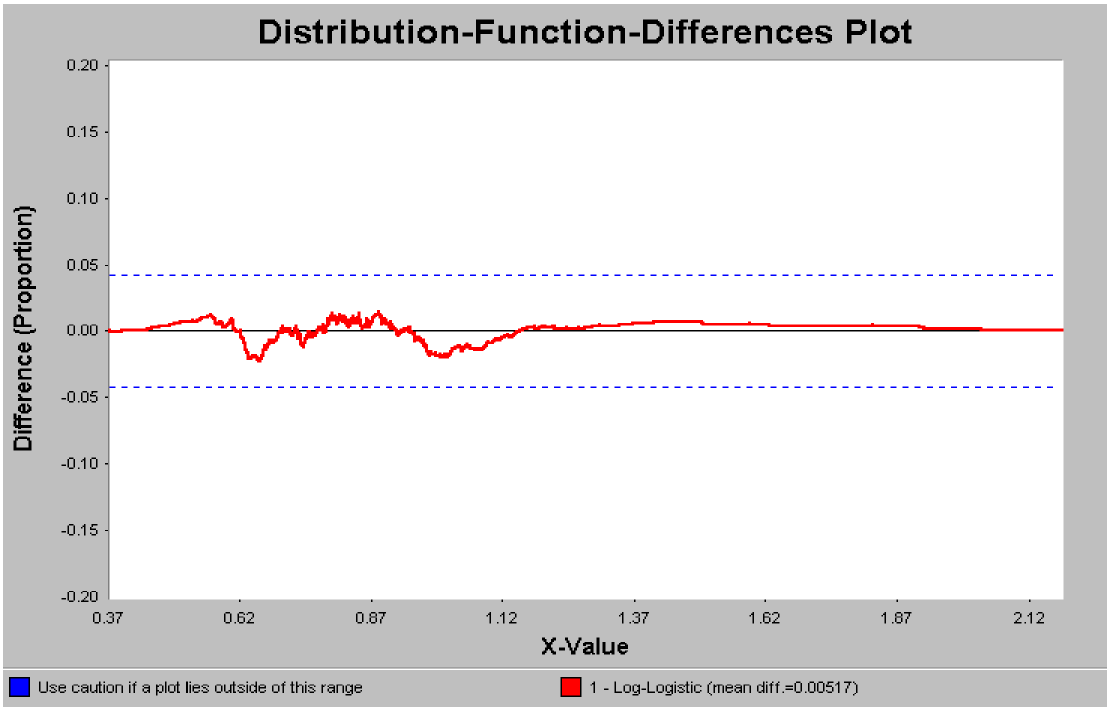

Figure 3: Distribution-Function-Differences Plot for the Ship-Loading Time Data

Simulation Software	Representation	
Extend	Block	Generator (DE) or Input Random Number (Generic)
	Distribution	LogLogistic
	Scale	0.821990
	Shape	8.840875
	Location	0.000000
ProModel	LogLogistic(8.840875, 0.821990, <stream>, 0.000000) (ExpertFit provides the LogLogistic generator as an add-in function.)	
Flexsim	Log-Logistic: Location 0, Scale 0.821990, Shape 8.840875	
WITNESS	RVLogLog(0.000000, 0.821990, 8.840875, <stream>) (ExpertFit provides the RVLogLog generator as an add-in function.)	

Figure 4: Simulation-Software Representations of the Log-Logistic Distribution

Simulation Software	Representation
Arena	CONT(0.0000,0.367360, 0.0035,0.457985, 0.0152,0.548610, 0.1168,0.639235, 0.2617,0.729860, 0.4895,0.820485, 0.7103,0.911110, 0.8703,1.001735, 0.9346,1.092360, 0.9591,1.182985, 0.9766,1.273610, 0.9836,1.364235, 0.9860,1.454860, 0.9907,1.545485, 0.9930,1.636110, 0.9942,1.726735, 0.9942,1.817360, 0.9965,1.907985, 0.9977,1.998610, 0.9988,2.089235, 1.0000,2.179860))
AutoMod	continuous(0.0000:0.367360,0.0035:0.457985,0.0152:0.548610,0.1168:0.639235,0.2617:0.729860,0.4895:0.820485,0.7103:0.911110,0.8703:1.001735,0.9346:1.092360,0.9591:1.182985,0.9766:1.273610,0.9836:1.364235,0.9860:1.454860,0.9907:1.545485,0.9930:1.636110,0.9942:1.726735,0.9942:1.817360,0.9965:1.907985,0.9977:1.998610,0.9988:2.089235,1.0000:2.179860)

Figure 5: Simulation-Software Representations of the Empirical Distribution Function

5 SUMMARY

ExpertFit can help you develop more valid simulation models than if you use a standard statistical package, an input processor built into a simulation package, or hand calculations to determine input probability distributions. ExpertFit uses a sophisticated algorithm to determine the best-fitting distribution and, furthermore, has 40 built-in standard theoretical distributions and 30 different types of graphs. On the other hand, a typical simulation package contains roughly 10 distributions.

ExpertFit can represent most of its 40 distributions in 26 different simulation packages such as Arena, AutoMod, eM-Plant, Extend, Flexsim, GPSS/H, HyPerformix/*workbench*, Micro Saint, OPNET Modeler, ProModel, SIMPROCESS, SIMUL8, SLX, and WITNESS, *even though the distribution may not be explicitly available in the simulation package itself.*

Note that ExpertFit has pioneered virtually every major development in distribution-fitting software – first such product, first with automated fitting, first with an absolute evaluation for a distribution, first with batch mode, etc. Furthermore, certain advanced ExpertFit features were funded by contracts with Accenture, NIST, and Oak Ridge National Lab.

ExpertFit is bundled with Flexsim and SIMPROCESS simulation products.

REFERENCE

Law, A. M. and W. D. Kelton. 2000. *Simulation Modeling and Analysis.* 3d ed. New York: McGraw-Hill.

AUTHOR BIOGRAPHIES

AVERILL M. LAW is President of Averill M. Law & Associates, a company specializing in simulation training, consulting, and software. He has been a simulation consultant to numerous organizations including Accenture, ARCO, Boeing, Compaq, Defense Modeling and Simulation Office, Kimberly-Clark, M&M/Mars, 3M, U.S. Air Force, and U.S. Army. He has presented more than 360 simulation short courses in 17 countries. He has written or coauthored numerous papers and books on simulation, operations research, statistics, and manufacturing including the book *Simulation Modeling and Analysis* that is used by more than 85,000 people and widely considered to be the "bible" of the simulation industry. This book contains the most comprehensive and practical discussion of simulation input modeling that is available. He developed the ExpertFit distribution-fitting software and also several videotapes on simulation modeling. He has been the keynote speaker at simulation conferences worldwide. He wrote a regular column on simulation for *Industrial Engineering* magazine. He has been a tenured faculty member at the University of Wisconsin-Madison and the University of Arizona. He has a Ph.D. in industrial engineering and operations research from the University of California at Berkeley. His E-mail address is <averill@ix.netcom.com> and his Web site is <www.averill-law.com>.

MICHAEL G. MCCOMAS is Vice President of Averill M. Law & Associates for Consulting Services. He has considerable simulation modeling experience in application areas such as manufacturing, oil and gas distribution, transportation, defense, and communications networks. His educational background includes an M.S. in systems and industrial engineering from the University of Arizona. He is the coauthor of seven published papers on applications of simulation.