Proceedings of the 1991 Winter Simulation Conference
Barry L. Nelson, W. David Kelton, Gordon M. Clark (eds.)

UniFit II: TOTAL SUPPORT FOR SIMULATION INPUT MODELING

Stephen G. Vincent

School of Business Administration
University of Wisconsin-Milwaukee
P.O. Box 742
Milwaukee, Wisconsin 53201

Averill M. Law

Averill M. Law & Associates
P.O. Box 40996
Tucson, Arizona 85717

ABSTRACT

In this paper we explain the important 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 UniFit II. We use a set of real-world system data to demonstrate how the package automatically specifies, evaluates, and ranks candidate probability distributions, and then assists an analyst in deciding whether the "best" candidate probability distribution provides an adequate representation of the data. If no candidate probability distribution provides an adequate fit, then UniFit II can define an empirical distribution function. In either case, the probability model can be automatically expressed in the analyst's simulation software. We then consider the general case of selecting a probability distribution in the absence of data. As an example, we show how UniFit II can be used to create busy-time and downtime models for machines that are subject to random breakdowns.

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

In this section we shall describe simulation input modeling and show consequences that might result if this important, but sometimes neglected, activity is performed improperly. We then suggest that with the use of UniFit II any simulation analyst can perform simulation input modeling more quickly and with greater accuracy than would otherwise be possible.

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, operating times before failure, and repair times for a machine should usually be modeled by probability distributions.

In this paper we use the phrase "simulation input modeling" to mean the process of choosing a probability distribution for each random component of the system under study and expressing this representation in a form that can be used with the analyst's choice of simulation software. In Sections 2 and 3 we will demonstrate how an analyst can easily and accurately choose an appropriate probabilistic representation using the UniFit II software.

1.2 Two Pitfalls In Simulation Input Modeling

The second author has identified a number of pitfalls that can undermine the success of simulation studies (Law and Kelton 1991, Law and McComas 1989). Two of these pitfalls relate directly to simulation input modeling and are summarized in this section.

1.2.1 Pitfall Number 1: Replacing a Distribution by its Mean

Simulation analysts have sometimes replaced an input probability distribution by its mean in their simulation models. This practice may be caused by a lack of understanding 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 results, as is shown by the following example.

Consider 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 one minute and the mean processing time is 0.99 minute. Suppose further that the interarrival times and processing times actually 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 a source of randomness with a constant value. If we assume that each interarrival time is *exactly* one minute and each processing 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 impact on the congestion level in most queueing-type (e.g., manufacturing) systems. In Section 2 we shall show how use of UniFit II makes choosing an appropriate probability distribution a simple and easy process.

1.2.2 Pitfall Number 2: Incorrect Modeling of Random Machine Downtimes

The largest source of randomness for many manufacturing systems is that associated with random machine downtimes. An analyst is often faced with representing in a simulation model the random machine downtimes of a machine that has not yet been purchased. Data concerning the actual downtime behavior of machine tools is, thus, unavailable and the analyst must rely on estimates of reliability provided by vendors and engineers. Suppose, for example, that a vendor claims that a machine tool will be down 10 percent of the time, but is unwilling or unable to provide more information on its operating time before breakdown and its repair time. Given the limited available information, some simulation analysts account for downtimes by simply reducing the machine processing rate by 10 percent. Law and McComas (1989) compare through the use of simulation the described practice to a more accurate model that we shall demonstrate in Section 3. Although the two modeling approaches led to similar results for an average throughput measure of performance, the use of the reduced-production-rate model led to large errors with regard to measures such as average time in system and maximum number of jobs in queue. Accurate estimation of the latter performance measures is essential in many simulation studies. Thus, serious errors can result if an incorrect, simplified approach is taken. We will show in Section 3 how easy it is to obtain a more accurate model of random machine downtimes using UniFit II.

1.3 Advantages of Using UniFit II

With the assistance of UniFit II *any* analyst, regardless of 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 assists the analyst in deciding whether the fit is good. If none of the candidate distributions provides an adequate fit, then an empirical distribution function can be created by UniFit II. In either case, the representation of system randomness can be automatically expressed 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, UniFit II will determine appropriate busy-time and downtime probability distributions that match the system's behavior, *even if the machine is subject to blocking or starving.*

2 USING UniFit II WHEN SYSTEM DATA ARE AVAILABLE

We now consider the general case where an analyst has data corresponding to the source of randomness to be represented in the simulation model. Our intention is to highlight the capabilities of UniFit II rather than to show operation of the package on a keystroke-by-keystroke basis. Detailed examples of program operation are available in the documentation that accompanies the free demonstration version of the software, which is available from the second author. (The demonstration version differs from the full version primarily in that only specially formatted data files can be read.)

We will precede our example analysis with a general discussion of the structure and operation of UniFit II. The package features status information on the top and bottom lines of the screen, and scroll-bar menus are shown in the middle. Menus can be operated using a mouse, and complete on-line help is available for each menu entry, as well as for each result screen. Among the preliminary options are package configuration (typically used only when installing the package), logfile control (all text output can be selectively captured in a disk file), and choice of operating mode. UniFit II also provides on-line tutorial help on subjects such as creating good histograms. Three modes of operation are available for selecting probability models. The guided model selection mode and manual model selection mode options are used when data are available; otherwise, the no-data model selection mode option is used (see Section 3). The manual model selection mode option makes available to the experienced analyst a full set of tools for analyzing data sets. We have designed the guided model selection mode option to make accurate and thorough analysis of data sets easy for *any* analyst, regardless of prior knowledge of statistics. It is this mode that we will demonstrate.

Upon entering the guided mode, an analyst uses a menu selection to read a data sample from a disk file. The data set we have chosen for this example consists

of machine cycle times provided to us by a major automobile manufacturer. Immediately after reading the data file, UniFit II provided us with the summary shown in Figure 1. The data set has 890 observations with a range from 25.4 to 72.8 time units and a mean of about 38.7. The positive skewness and the fact that the mean is larger than the median both suggest that the underlying distribution has a longer right tail than left tail (positive skewness); this is typical of service-time data.

We recommend that every data analysis begin with the construction of a good histogram of the data. The parameters of such a histogram are "remembered" by the package and form the basis of a number of heuristics for assessing how well fitted models represent the data. UniFit II provides default values for most situations including histograms. After viewing the default histogram, we used available options to adjust the starting point, the interval width, and the number of intervals. The resulting histogram is not shown by itself in this paper; you can see the histogram in Figure 4, which is a histogram-based model evaluation heuristic that we introduce below. Note that the histogram is reasonably smooth, skewed to the right, and is definitely shifted away from the origin.

The basic information required by UniFit II to begin the fitting and evaluation process is a specification of the range of the underlying random variable. UniFit II features two menus at which the analyst chooses the minimum and maximum possible values of the underlying random variable. Although the smallest observed value is much larger than zero, we did not have any firm knowledge of a definite smallest possible value greater than zero. Similarly, we could not state the existence of a definite upper bound. At the two menus we selected appropriate options to characterize the stated range of the underlying random variable. UniFit II responded to our choices by fitting distributions with ranges starting at zero, distributions whose lower endpoint was estimated from the data itself, and distributions with no definite lower bound. These candidate models were then automatically evaluated. After a few seconds the result screen shown in Figure 2 was displayed.

UniFit II fit and ranked 16 candidate models, with the five best-fitting models listed on the screen along with their scores. The displayed scores are calculated by a proprietary evaluation scheme that is based on our 14 years of research in this area. Results from the heuristics that we have found to be the best indicators of good model fit are combined and the resulting numerical evaluation is normalized so that 100 indicates the best model and 0 indicates the worst model. These scores are *comparative* in nature and do not reflect an absolute assessment of the quality of fit. UniFit provides a separate evaluation of the adequacy of fit provided by the best-ranked model. In Figure 2 we see that the Pearson type 5 distribution (range starts at zero) and the (shifted) Weibull distribution (range starts at 25.3981) are tied for the best model. It should be noted that although the Pearson type 5 distribution may be unfamiliar to you, it is supported by most simulation packages since it can be generated as the inverse of a gamma random variable. It should also be noted that UniFit II completed the entire analysis without further input from the analyst; only the range had to be specified.

Immediately after presenting the summary screen shown in Figure 2, UniFit provided a heuristic plot of our own devising that can be used to assess how well a candidate model represents the data. The heuristic is called a DF (distribution function) difference plot and is constructed as follows. For each observation X in the sample, UniFit II calculates a Y value as the difference between the probability that a candidate distribution takes on a value no larger than X and the proportion of observations that are no larger than X.

Summary of Sample: Machine Cycle Times

Sample Characteristic	Value
Observation Type	Real Valued
Number of Observations	890
Minimum Observation	25.4000
Maximum Observation	72.8000
Mean	38.7142
Median	37.7000
Variance	62.3921
Skewness	.80496

Figure 1: Sample Summary of the Cycle-Time Data

```
Guided Selection Model Rankings                    Sample: Machine Cycle Times

During the fitting process UniFit considers distributions having any
reasonable range (not just the specified range), provided they produce
values in the specified range at least 99% of the time.

Specified random variable range     At least 0.

                          Score     Random Variable Range
Models                    (0-100)   (if different from that specified)
-------------------------------------------------------------------------
1-Pearson Type 5          93.3
2-Weibull (E)             93.3      At least 25.3981
3-Extreme Value Type B    86.7      Unrestricted
4-Inverse Gaussian        81.7
5-Lognormal               78.3

In addition, 11 other models were considered having scores from   .0 to 65.0.

UniFit Evaluation         Based on a heuristic evaluation, there is no
                          current evidence for not using the primary model.
```

Figure 2: Evaluation of Candidate Models for the Cycle-Time Data

The plot is made by connecting the (X, Y) pairs produced in this manner. When a distribution provides a good representation of a data set, all of the Y values will be small; large Y values indicate major discrepancies between the proposed model and the observed sample. UniFit II automatically provided such a plot for the best model. Instead of including the plot displayed at this point, we have included as Figure 3 a plot we created a short time later. In Figure 3 both the Pearson type 5 and Weibull distributions are shown. This plot demonstrates how heuristic plots in UniFit can be used to assess the quality of fit of a single distribution as well as to compare the fits provided by competing models. The small vertical differences (errors) suggest that both models provide a good fit; there is little reason to choose one model over the other on the basis of this heuristic.

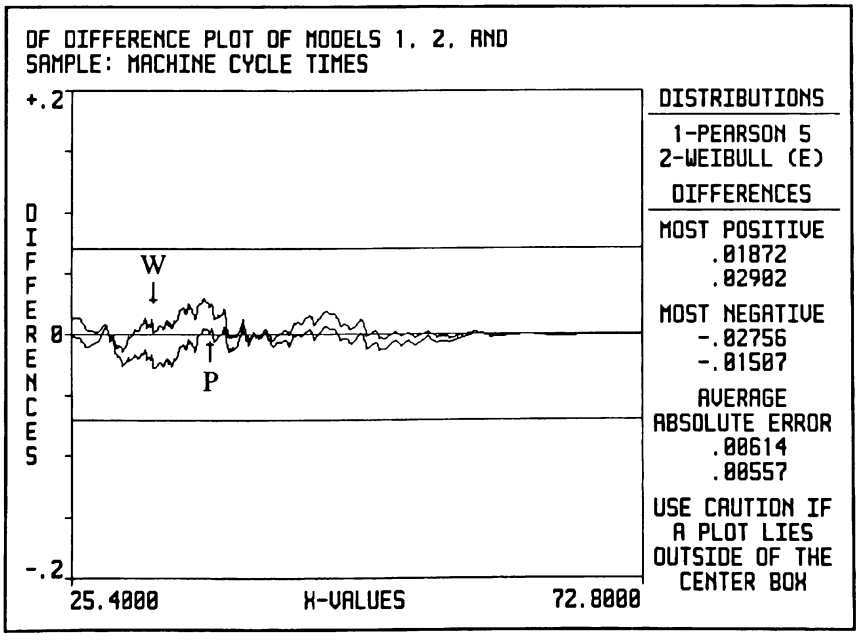

Figure 3: DF Differences Plot for the Cycle-Time Data

After viewing the plot automatically provided by UniFit II, a menu of confirmation options was presented. Three of the options in this menu deserve some attention although they will not be demonstrated. An option for making distribution function comparisons allows us to make the DF difference plot shown in Figure 3, as well as a plot of sample and model cumulative frequencies. Goodness-of-fit tests, including the chi-square, Kolmogorov-Smirnov, and Anderson-Darling, are available through another option. It is also possible to change to manual selection mode in order to use a number of other options for assessing how well candidate models represent the data. We continued our analysis by selecting an option that allows for histogram-based comparisons. With this option we can modify our basic histogram as well as request a number of different plots and result screens. We show a density/histogram overplot with both candidate models in Figure 4. Here the densities for the two candidates have been plotted over the histogram (which is an estimate of the true density function). This plot seemed to indicate that the Pearson type 5 distribution provides a slightly better fit, particularly in the center of the distribution.

In summary, both distributions appear to be a good model for the cycle-time data. However, we prefer the Pearson type 5 distribution because its density function matches the histogram more closely. From the confirmation options menu we selected an option that allowed us to display the representation of the model using different software packages. We show in Figure 5 the representations for two of the software packages supported by UniFit II.

With some data samples, no candidate model provides an adequate representation. In this case we recommend the use of an empirical distribution function. One useful feature of UniFit II is that in addition to using all of the sample values in the simulation software representation, it is possible to reduce the amount of information required through the use of a histogram-based empirical distribution function. We show a histogram-based representation, with 20 intervals, for two simulation software packages in Figure 6.

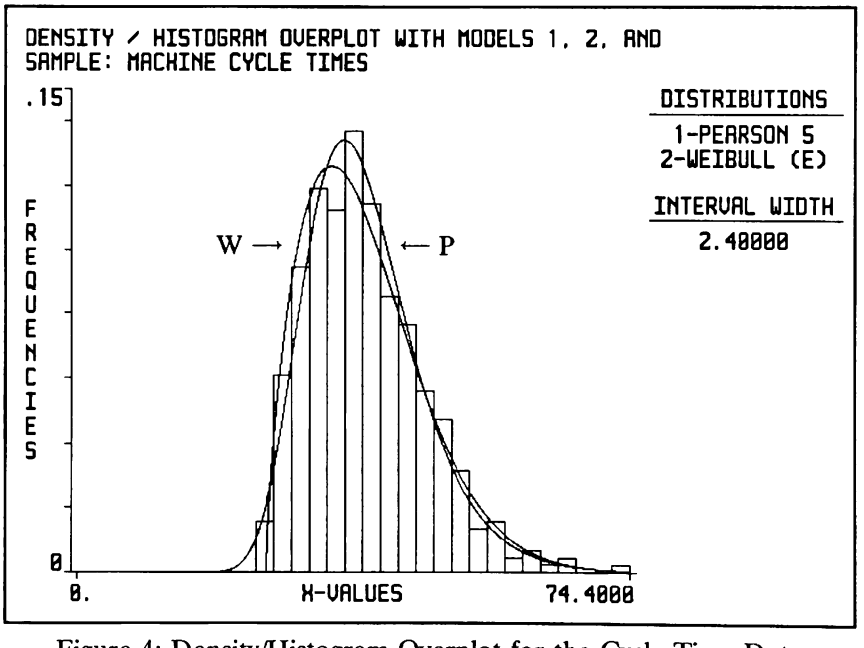

Figure 4: Density/Histogram Overplot for the Cycle-Time Data

Simulation Software	Representation	
AutoMod	1./(gamma 26.3976, .00102)	
SIMFACTORY II.5	Generator	PEARSON V
	shape	26.3976
	scale	983.031
	offset	.00000

Figure 5: Simulation Software Representation of the Pearson Type 5 Distribution

Simulation Software	Representation
GPSS/H	<name> FUNCTION RN<stream>,C21 .0000,26.51/.0900,29.7915/.1900,33.073/.3600,36.3545/ .6400,39.636/.8000,42.9175/.8500,46.199/.9100,49.4805/ .9100,52.762/.9100,56.0435/.9200,59.325/.9300,62.6065/ .9300,65.888/.9400,69.1695/.9400,72.451/.9400,75.7325/ .9600,79.014/.9600,82.2955/.9600,85.577/.9800,88.8585/ 1.0000,92.14
ProModel	D<number> .00,26.51 9.00,29.7915 19.00,33.073 36.00,36.3545 64.00,39.636 80.00,42.9175 85.00,46.199 91.00,49.4805 91.00,52.762 91.00,56.0435 92.00,59.325 93.00,62.6065 93.00,65.888 94.00,69.1695 94.00,72.451 94.00,75.7325 96.00,79.014 96.00,82.2955 96.00,85.577 98.00,88.8585 100.00,92.14,<stream>

Figure 6: Simulation Software Representation of the Empirical Distribution Function

3 USING UniFit II WHEN NO DATA ARE AVAILABLE

Quite often a simulation analyst must model a source of randomness for which no data are available. In this section we show how UniFit II can be used to assist in the case of modeling random machine downtimes. UniFit II supports accurate modeling of systems with or without significant blocking or starving. For the example in this section we will assume that the machine of interest is never blocked or starved.

Consider a machine that has an efficiency of 0.9; that is, it is actually producing parts 90 percent of the time. When the machine goes down, the average downtime is 60 minutes. However the minimum downtime is 10 minutes. This information is specified to UniFit II through a sequence of easy-to-use menus. After all of the required information has been specified, the average number of downs (actually the average number of busy-time/downtime cycles) per 8-hour shift is calculated by the package to be 0.8. This makes sense since the average length of a busy-time/downtime cycle is 10 hours. A menu then allows various characteristics of the busy-time and downtime distributions to be displayed. We show in Figure 7 the fully specified busy-time and downtime distributions and in Figure 8 corresponding simulation software representations for four packages.

Busy-Time Model: Gamma Distribution		Downtime Model: Gamma Distribution	
Location Parameter	0.	Location Parameter	10.0000
Scale Parameter	771.429	Scale Parameter	35.7143
Shape Parameter	.70000	Shape Parameter	1.40000

Figure 7: Specified Busy-Time and Downtime Models

Simulation Software	Busy-Time and Down-Time Representations
SIMAN IV	GAMMA(771.429, .70000, <stream>) 10.0000 + GAMMA(35.7143, 1.40000, <stream>)
SIMSCRIPT II.5	GAMMA.F(540.000, .70000, <stream>) 10.0000 + GAMMA.F(50.0000, 1.40000, <stream>)
SLAM	GAMA(771.429, .70000, <stream>) 10.0000 + GAMA(35.7143, 1.40000, <stream>)
Witness Version 7	GAMMA(.70000, 771.429, <stream>) 10.0000 + GAMMA(1.40000, 35.7143, <stream>)

Figure 8: Simulation Software Representations of Busy-Time and Downtime Models

REFERENCES

Law, A.M. and W.D. Kelton. 1991. *Simulation Modeling and Analysis,* Second Edition. New York: McGraw-Hill.

Law, A.M. and M.G. McComas. 1989. Secrets of successful simulation studies. *Industrial Engineering* **21**: 28-31 (May 1989).

AUTHOR BIOGRAPHIES

STEPHEN G. VINCENT is an Assistant Professor in the School of Business Administration at the University of Wisconsin-Milwaukee, where he teaches courses in the areas of software engineering and simulation. He was Vice President of Simulation Modeling and Analysis Company in charge of software development until 1987 during which time he developed the UniFit software package with Averill Law. He received his Ph.D. in Management Information Systems from the University of Arizona and has B.S. and M.S. degrees in Industrial Engineering from the University of Wisconsin-Madison.

AVERILL M. LAW is President of Averill M. Law & Associates (Tucson, Arizona). He has been a simulation consultant to such organizations as General Motors, IBM, AT&T, ALCOA, General Electric, 3M, Nabisco, Xerox, NASA, and the Army. He has presented more than 180 simulation seminars in 10 countries. He is the author (or coauthor) of three books and more than 30 papers on simulation, manufacturing, operations research, and statistics, including the widely used textbook *Simulation Modeling and Analysis*. His series of papers on the simulation of manufacturing systems won the 1988 Institute of Industrial Engineers' best publication award. He is the codeveloper of the UniFit II software package for fitting probability distributions to observed data, and he developed a four-hour videotape on simulation with the Society of Manufacturing Engineers. Dr. Law writes a regular column on simulation for *Industrial Engineering* magazine. He has taught simulation at the University of Arizona and the University of Wisconsin. Dr. Law has a Ph.D. in Industrial Engineering and Operations Research from the University of California at Berkeley.