Proceedings of the 1990 Winter Simulation Conference
Osman Balci, Randall P. Sadowski, Richard E. Nance (eds.)

A TUTORIAL ON UniFit II: A SYSTEM FOR TOTAL SUPPORT OF SIMULATION INPUT MODELING

Stephen G. Vincent

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

Averill M. Law

Simulation Modeling and Analysis Company
440 North Alvernon Way, Suite 103
Tucson, Arizona 85711

ABSTRACT

This paper provides an overview of the UniFit II software package, which is used to select simulation input probability distributions. UniFit II is the successor to the original UniFit software package, which has offered state-of-the-art support for selection of distributions for nearly eight years. UniFit II provides more comprehensive and easier-to-use support for simulationists. UniFit II allows a simulationist to determine easily which of the standard probability distributions (e.g., exponential, gamma, normal, etc.) best represents a set of observations. The package then helps a simulationist determine the quality of the correspondence between the selected probability distribution and the observations. In most cases the correspondence will be good enough for simulation purposes, and UniFit II specifies how to generate values from the selected distribution in any of the major simulation packages. In the situation where the correspondence is not good, UniFit reports how to generate values directly from the observed data. UniFit II supports two different methods for choosing distributions when data are not available. The first method has been commonly recommended for this no-data case and may be used in almost any context. The second author developed the other method specifically for specifying the distributions of operating and repair time random variables needed for simulations of manufacturing systems.

1. INTRODUCTION

In the fall of 1978 the authors began joint research on the topic of fitting distributions to data as it related to simulation. After four years of research and development, the UniFit software for fitting distributions to observed data became available for mainframe computers (in 1982). Further development of the software resulted in the 1985 release of a version for personal computers. The UniFit package was designed to allow someone who was familiar with the process of fitting distributions to data to do so quickly and in a more comprehensive manner than was possible either with hand calculations or through the use of general statistical software. It is being used by simulationists, engineers, statisticians, and scientists in more than 150 organizations in a number of countries.

The recently released UniFit II software package is the result of a two year redesign process. One of the major goals during the redesign was to make the comprehensive array of tools available within the original UniFit package accessible and indeed simple to use by simulationists who have little formal training in the relevant statistical methods. The design of the package reflects this goal in a number of ways, but perhaps the most dramatic example of the redesign is the inclusion of probability distribution ranking facilities within the package. We have created a procedure for automatically evaluating and ranking alternative probability distributions. During the twelve years of developing and using software to fit distributions to data for simulation studies, we have invented and evaluated a number of heuristics for assessing the quality of a probability distribution. The automatic ranking process uses heuristic procedures that we have found to be the most effective for choosing distributions. This ranking procedure is discussed in more detail in Sections 2.3 and 2.4.

Another major goal was to make the software as "user-friendly" as possible. One problem with current software is the variety of interfaces employed; some packages are menu-based, others are not, and few menu-based packages operate in exactly the same way. We recognized that because a package like UniFit II may be used intensively by a simulationist during the early stages of a long simulation study but then infrequently until the next project commences, it is essential that little time be spent learning to use the interface. We have therefore used a straight forward menuing scheme which is as easy to use with a mouse as without one, and have avoided the common practice of requiring a user to employ special non-obvious keystrokes when performing basic tasks. Context-dependent on-line help is available in four categories to allow the user to find quickly the help information relevant to the task the simulationist is performing with UniFit II.

The third major goal for the redesign process was to increase the support for tasks performed by simulationists in selecting input distributions. The original version of UniFit was designed to be the state-of-the-art tool for fitting distributions to data. We have increased the applicability of UniFit by including options which support selecting distributions when no data are available. We have improved the usefulness of UniFit for simulationists by providing options that specify the exact usage of a probability distribution (e.g., the random value generator to use and its parameters) in the major simulation packages (languages or simulators).

2. SELECTING SIMULATION INPUT DISTRIBUTIONS

In this section we present an overview of the input distribution selection process. We begin with an examination of the need for proper input distribution selection. We then discuss the objective of the process and a provide a philosophy for selecting input distributions in the simulation context. The section concludes with an overview of the selection process, including an example.

2.1 The Need For Proper Input Distribution Selection

When performing a simulation study, the sources of randomness for the system under consideration must be represented properly. (A source of randomness in the real-world system is typically called a *random variable.*) In many manufacturing systems, for example, correctly modeling machine operating times and repair times is critical to obtaining meaningful simulation results.

In order to demonstrate the dramatic effect that the choice of input distribution may have on the performance of a simulated system, we performed a small experiment involving a simple system. The system of interest was the single-server queueing system, where interarrival times were exponentially distributed. The experiment had five cases corresponding to the choice of the service-time distribution. Five distributions (exponential, gamma, Weibull, lognormal, and normal) were fit to a set of observed service times. We then made 100 replications of the system for each choice of service-time distribution, where each replication was run until the 1000th delay in queue was observed. The results of the experiment are summarized in Table 1. Each value in the table is the average of the measure of performance over the 100 replications for the appropriate distribution.

Table 1. Empirical Results From 100 Replications For Each Distribution

Distribution	Average Delay In Queue	Average Number In Queue	Percentage of Delays At Least 20
exponential	6.71	6.78	6.4%
gamma	4.54	4.60	1.9%
Weibull	**4.36**	**4.41**	**1.3%**
lognormal	7.19	7.30	7.8%
normal	6.04	6.13	4.5%

After a thorough analysis of the service-time data using UniFit II, which included distributions not shown in Table 1, we concluded that the Weibull distribution provided the best representation of the data, and the results produced by this distribution will be used as reference points in the discussion which follows. (An overview of this analysis is presented in Section 2.4.) The values for the average delay in queue for different service-time distributions highlight the impact of the choice of distribution on simulation results. In particular, note that the normal distribution, which has often been used as an input probability distribution due to its familiarity, leads in this case to an average delay value that differs by almost 39 percent from that produced by the reference Weibull distribution. What is more surprising is that the result produced by the lognormal distribution (which can have a shape very similar to that of the Weibull) differs from the reference by 65 percent. Similar results occur with respect to the average number in queue measure of system performance. We would expect that differences in simulation results should be the greatest when we consider the likelihood of extreme values occurring, because the service-time distributions considered in Table 1 differ most in their "tails." This expectation is borne out by the output measure reporting the percentage of delays that are at least 20. Here the result produced by the normal distribution differs from that of the reference Weibull by 246 percent. An even more striking discrepancy from the reference of 500 percent occurs with the result produced by the lognormal distribution.

In many simulation studies little attention has been paid to the process of selecting input distributions. Some simulationists have used distributions such as the normal to represent sources of randomness in their simulations simply because they were most familiar with the normal. It should be evident from the example that such a practice calls into question the validity of the overall simulation results and, thus, any conclusions based upon those results.

2.2 The Desired Outcome of the Input Distribution Selection Process

In the simulation context the desired outcome of the input distribution selection process is the choice of a mechanism for generating random values from a probability distribution. (In the simulation literature such mechanisms are known as *random variate generators*.) Most simulation packages offer two different types of random value generators which differ in their overall approach to modeling a particular random variable. An empirical function generator generates values on the basis of a parameter list containing example values and their associated probabilities. This type of generator is given its name because of the common practice of using the sorted observations as the example values and values computed from the empirical (sample) distribution function as the set of associated probabilities. Generators of the second type produce random values from a specific standard probability distribution. These generators require specification of the parameter values corresponding to the specific probability distribution (e.g., the mean of the exponential). Most simulation packages include an empirical function generator as well as a number of generators for standard probability distributions.

The simulationist must select an appropriate generator to represent a real-world source of randomness in a simulation model and then must specify the list of parameters required by the random value generator. We believe that a generator for a standard probability distribution should be selected whenever possible, for practical as well as more theoretical reasons. (The second author is participating in a panel discussion at this conference that addresses different approaches for specifying input probability distributions. His position paper, included elsewhere in these proceedings, discusses the reasons for our preference.)

2.3 A Philosophy to Guide The Input Distribution Selection Process

In some disciplines analysts have used UniFit to meet the objective of determining whether a specified distribution (e.g., normal) is the *true* underlying distribution for their observed data. However, it is not clear that this is the correct objective for simulation applications, since none of the standard distributions is probably *exactly* correct for most simulation input random variables. Instead, we feel that an appropriate goal is to find a standard probability distribution that provides a representation that can be assessed as being good enough for the purposes of the simulation study.

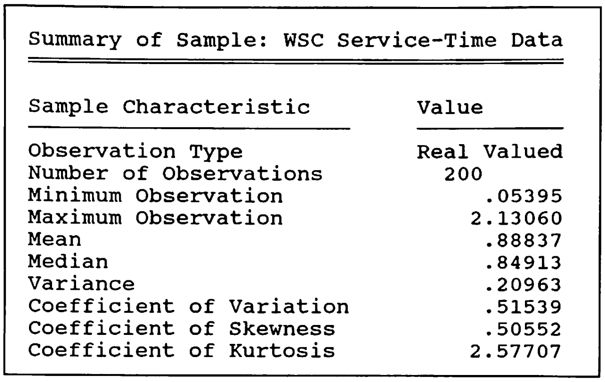
We recommend a two-phase approach for choosing a probability distribution that is "representative" of a simulation input random variable. In the *selection phase* the alternative probability distributions available within UniFit II are "fit" to the observed data and then compared in order to determine which one best represents the observed data. UniFit II provides many graphical and tabular options for comparing fitted distributions to the observed data. A simulationist may use any or all of these comparisons to select one of the distributions as being the most representative, or may elect to have UniFit II make the selection with a ranking methodology. The automatic ranking methodology utilizes the results of comparisons based upon goodness-of-fit test statistics as well as the results of comparisons based upon metrics computed from graphical procedures. We incorporated into the ranking methodology those comparisons that we have found to be most effective in discriminating between distributions during the past twelve years.

In the *confirmation phase* the probability distribution identified in the selection phase is evaluated using formal goodness-of-fit tests and graphical comparisons to determine if it represents the observed data well enough to be used in a simulation model. In our experience goodness-of-fit tests do not necessarily provide by themselves a definitive assessment of the quality of fit of a distribution to be used for simulation. We prefer to use such tests as an indication of the amount of further evidence required in order to confirm the quality of a probability distribution: better representations as indicated by formal goodness-of-fit tests require less confirming visual evidence from comparison plots. Plots employed in the confirmation phase are demonstrated in the example found in the next section.

2.4 An Example of the Process of Selecting Input Distributions

In many situations it is possible to collect a sample from the system random variable of interest. We shall first consider this case and later discuss the situation where data are unavailable. The following discussion is based upon the process we recommend a simulationist follow when using the UniFit II software and will be illustrated with example output from the package. We shall use the service-time data referenced in the example of Section 2.1 as our set of observations.

Immediately after using an option in the UniFit II package to read observations recorded in a data file, the sample summary shown as Figure 1 is displayed. The sample has two hundred observations, ranging from .05395 to 2.13060, and a mean of .88837.



```
Summary of Sample: WSC Service-Time Data
========================================

Sample Characteristic          Value
_____________________________  ___________

Observation Type               Real Valued
Number of Observations          200
Minimum Observation                .05395
Maximum Observation               2.13060
Mean                               .88837
Median                             .84913
Variance                           .20963
Coefficient of Variation           .51539
Coefficient of Skewness            .50552
Coefficient of Kurtosis           2.57707
```

Figure 1. Sample Summary of Example Data

In addition to a set of observations, the simulationist should have a basic understanding of the random variable that produced the sample. The most important single piece of information is the range of the possible values of the random variable, for it identifies the class of candidate probability distributions. Most data sets collected for simulation purposes will be strictly non-negative, that is, the values must be greater than zero. This is true, for example, for operating times, repair times, and service times. The example data are service times and are therefore strictly non-negative.

We began the selection phase by choosing an option that requested UniFit II to estimate all of the parameters for the seven distributions appropriate for non-negative data sets. An eighth probability distribution, the normal, was also fit to the data using a separate option.

Once the distributions were selected and parameters estimated, the ranking methodology described in the previous section was invoked. The top two probability distributions found automatically by UniFit II were the Weibull and the gamma, in that order. After seeing the summary ranking, the default option provided by UniFit II is to reorder the fitted distributions to reflect their rankings; this option was selected. (All probability distributions are given numbers that serve as abbreviations for their complete names.)

We began the confirmation phase by performing several goodness-of-fit tests with the Weibull distribution. The results of the tests, which are not shown here, indicated that the Weibull distribution was a good representation for the random variable. Following the recommended procedure for choosing an input probability distribution outlined in Section 2.3, only minimal additional confirming evidence concerning the quality of the correspondence between the Weibull distribution and the sample was required before accepting the distribution.

We first chose to perform a frequency comparison because it provides a very easy way to get a visual impression of how well a probability distribution represents a sample. In a frequency comparison a subrange of the possible values is divided up into equal-width intervals and the proportion observed in the sample and that expected from the probability distribution are plotted for comparison. UniFit II provides default values wherever choices may be made, such as for the specification of intervals. We displayed the frequency comparison with the default intervals and then modified them slightly to obtain the comparison shown in Figure 2. (Please note that for clarity of reproduction, the multiple-color plots produced by UniFit II have been printed in black and white only.) In a frequency comparison like that shown in Figure 2 the outside (hollow) bars represent the observed frequencies (proportions) and the inside (full) bars represent the frequencies expected from the distribution being evaluated. This plot indicated that there was general close agreement between the sample and fitted Weibull distribution frequencies.

Another way to assess the quality of the representation is to compare the density function of a probability distribution and the sample density (this is just a rescaling of the sample histogram and thus has the same shape). We created such a comparison using the Weibull and exponential distributions which is shown in Figure 3. (The curve for the exponential density is the one with a large value at an x-value of zero). It should be noted that the exponential was rated by the ranking methodology as the *worst* of the candidate probability distributions and was chosen for this plot to show how easily bad probability distributions can be ascertained using UniFit II plots. This plot provided clear evidence of the bad representation offered by the exponential distribution. The close match between the sample density and that of the Weibull distribution offered more evidence to its being a good choice to represent the sample.

There are a number of additional plots available in UniFit II for assessing the quality of fit provided by distributions. We chose to create one more plot called the distribution function difference plot, which is shown in Figure 4. In this type of plot the differences between the empirical distribution function (based upon the sample) and a specified probability distribution function are graphed. A good fit is indicated by small differences being detected across the range of data values. The plot provided another demonstration of the quality of fit of the Weibull distribution and the lack of quality of fit by

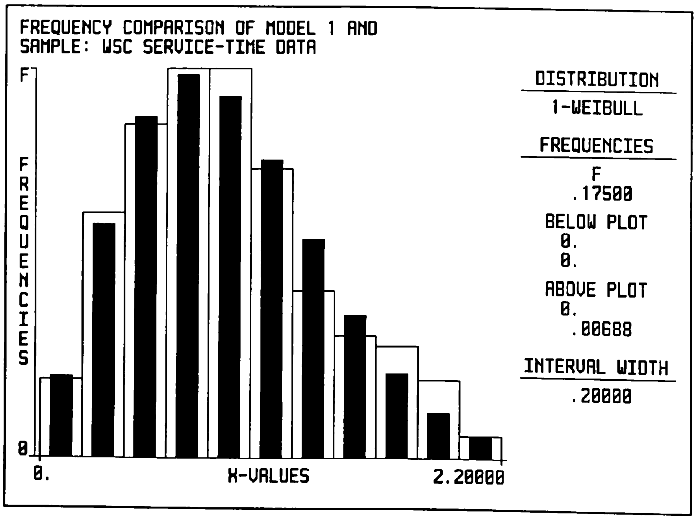


Figure 2. Example Frequency Comparison

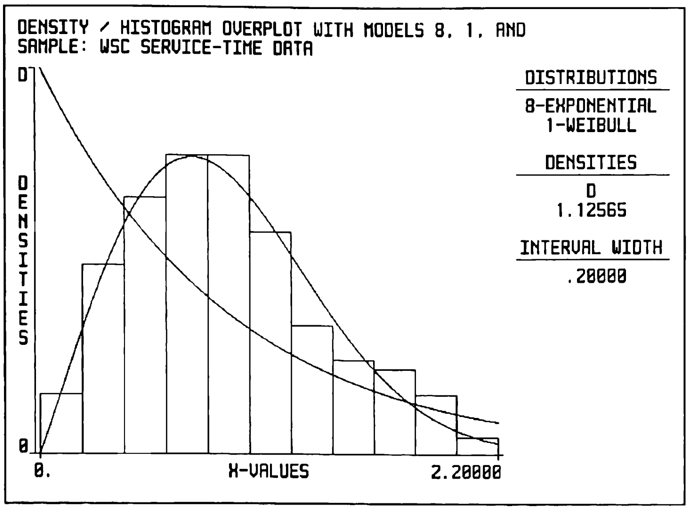


Figure 3. Example Density / Histogram Overplot

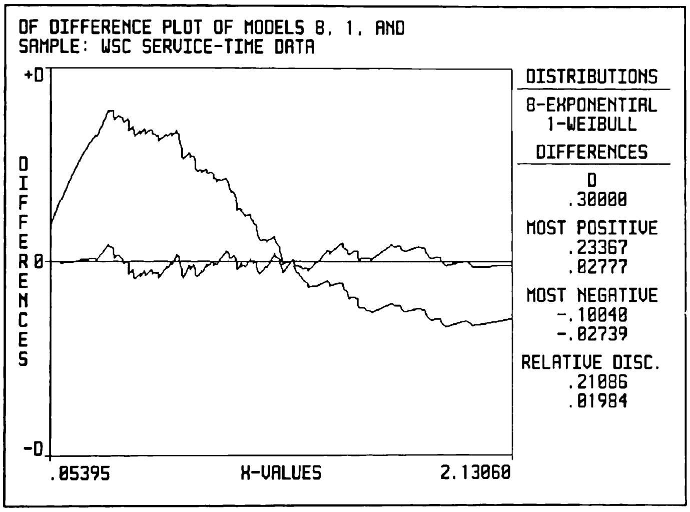


Figure 4. Example Distribution Function Difference Plot

the exponential.

All of the evidence suggested that the Weibull distribution provided an excellent representation for the service-time random variable. The last required step in the input distribution selection process was the specification of a corresponding random value generator for the simulation. An option in UniFit II can perform this step for the major simulation packages. The representations of the Weibull distribution for the SIMAN, SIMSCRIPT II.5, and SLAM languages were requested individually, and are shown together in Figure 5.

The example demonstrates the process of selecting a simulation input distribution when data are available, but simulationists must often specify input distributions without data. Data may be unavailable because the system does not exist or because there is not time for data collection and analysis. In these cases, knowledge possessed by people most familiar with the system under consideration or similar systems must be used in place of a data sample. UniFit II supports two approaches to this problem. The first approach is well known and involves specification of minimum, maximum, and most likely values, which are then used in the triangular distribution. The other approach was developed by the second author and is appropriate in the context of simulation of manufacturing systems. The method is used to specify the distributions of operating and repair time random variables.

```
SIMAN Representation:

   Parameter Value(s)   1.00701, 2.04472
   Model Usage          WE(IP, IS)
   where IP is the parameter set
   and IS is the stream


SIMSCRIPT II.5 Representation:

   WEIBULL.F(2.04472, 1.00342, IS)
   where IS is the stream


SLAM Representation:

   WEIBL(1.00701, 2.04472, IS)
   where IS is the stream
```

Figure 5. Simulation Language Representations