*Proceedings of the 1986 Winter Simulation Conference*
J. Wilson, J. Henriksen, S. Roberts (eds.)

# A TUTORIAL ON UNIFIT: AN INTERACTIVE COMPUTER PACKAGE FOR FITTING PROBABILITY DISTRIBUTIONS TO OBSERVED DATA

Averill M. Law
Stephen G. Vincent
Simulation Modeling And Analysis Company
P.O. Box 40996, Tucson, Arizona 85717, U.S.A.

## ABSTRACT

In this paper we present a tutorial on using the UNIFIT software package to fit probability distributions to observed data. In the first section we provide an example which demonstrates why selecting appropriate probability distributions is of particular interest to simulation analysts. A general overview of the UNIFIT software is presented in the second section. We then present a three-activity approach to fitting distributions to data and highlight the capabilities of UNIFIT which allow the analyst to perform these activities in a thorough and timely manner. The fourth section provides a discussion of additional features of the software. A number of graphical displays which are available in UNIFIT are included in the last section of the paper.

## 1. THE NEED FOR PROPER SELECTION OF INPUT PROBABILITY DISTRIBUTIONS

An important problem which occurs in many different disciplines is that of determining a probability distribution which is a good representation of an observed data set. For example, in building a simulation model of a manufacturing process or of a computer system, one needs to determine appropriate probability distributions for the input random variables. A common solution to this problem is to fit standard distributions (e.g., normal or gamma) to observed system data. However, since this fitting process is rather complicated and time consuming when done by hand, it is often performed in a superficial and incorrect manner. The net effect is, of course, that the selected distributions may not be good representations of the observed data.

We performed a small experiment to demonstrate the effect of distribution choice on simulation results. The system of interest was the single-server queueing system, where interarrival times were exponentially distributed. The objective of the experiment was to demonstrate the impact of the selection of service time distribution on the performance of the system. 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 simulation 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 | Proportion Of Delays At Least 20 |
|---|---|---|---|
| exponential | 6.71 | 6.78 | 0.064 |
| gamma | 4.54 | 4.60 | 0.019 |
| Weibull | 4.36 | 4.41 | 0.013 |
| lognormal | 7.19 | 7.30 | 0.078 |
| normal | 6.04 | 6.13 | 0.045 |

After a thorough analysis of the service time data using UNIFIT, 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. 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 chosen as an input probability distribution due to its familiarity, leads in this case to an average delay value which 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 proportion of delays which 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. From this evidence we believe it is clear that the choice of input

probability distribution can have a major impact on the results of a simulation and hence the choice of input probability distribution has a direct impact on the validity of a simulation study.

## 2. OVERVIEW OF THE UNIFIT SOFTWARE PACKAGE

UNIFIT is a state-of-the-art computer package for fitting probability distributions to observed data. By combining the latest statistical techniques with graphical displays, UNIFIT allows you to perform a comprehensive analysis in signficantly less time than would otherwise be possible. It also allows you to perform this comprehensive analysis on a wide variety of potential probability distributions. This significantly reduces the likelihood of your making a serious modeling error, of the type described in Section 1.

UNIFIT has been available for mainframe and minicomputers since 1983. A version for the IBM PC which features color graphics was introduced in 1985. There are now more than 60 organizations world-wide using the software to fit distributions to observed data.

## 3. A THREE-ACTIVITY APPROACH TO FITTING DISTRIBUTIONS TO OBSERVED DATA

The user can employ a three-activity approach for determining an appropriate distribution when using UNIFIT. The first activity involves using heuristic techniques such as histograms or sample moments to hypothesize one or more families of distributions which might be representative of the observed data. However, each of these families of distributions has several parameters which must be specified in order to have a completely determined distribution. Therefore, the second activity typically involves estimating the parameters of each hypothesized family from the data, thereby specifying a number of particular distributions. In the third activity we determine which of the fitted distributions, if any, is the best representation for the data using both heuristic techniques and goodness-of-fit tests.

### 3.1. Activity 1: Hypothesizing Families of Distributions Using UNIFIT

UNIFIT provides three procedures for summarizing the basic properties of a data sample which are of use in hypothesizing appropriate families of distributions.

1. One very useful heuristic is called **Summary Statistics** for the sample, which is a display showing the number of observations in the sample, the minimum observation, the maximum observation, the mean, the median, the variance, the coefficient of variation (a measure of variability), the skewness (a measure of symmetry), and the kurtosis (a measure of distribution "tail weight").
2. The **Histogram** is one of the most valuable and widely used tools for determining the shape of the underlying probability density function for a continuous data set or the shape of the underlying probability mass function for a discrete data set.
3. The **Quantile Summary And Box Plot** is a synopsis of the sample which is useful for determining whether the underlying density function or mass function is symmetric or skewed to the right or to the left.

### 3.2. Activity 2: Estimating The Parameters Of A Hypothesized Family Using UNIFIT

UNIFIT allows a user to fit thirteen continuous distributions to a continuous data set or five discrete distributions to a discrete data set. The thirteen continuous distributions are divided into three categories related to the values which a random variable can take on. Non-negative continuous models can take on values larger than a location parameter (typically zero). Unbounded continuous models can take on any finite value. Bounded continuous models can take on values between two fixed endpoints. The eighteen distributions supported by UNIFIT are shown in Table 2.

Table 2: Distributions Supported By UNIFIT

| Non-Negative Continuous | |
|---|---|
| exponential | Weibull |
| gamma | Pearson type 5 (inverted gamma) |
| inverse Gaussian | |
| lognormal | Pearson type 6 |
| **Unbounded Continuous** | |
| normal | extreme value (minimum) |
| logistic | extreme value (maximum) |
| **Bounded Continuous** | |
| uniform | beta |
| **Discrete** | |
| binomial | Poisson |
| geometric | uniform (discrete) |
| negative binomial | |

It should be noted that up to nine different distributions can be fit and compared simultaneously using UNIFIT. This is particulary useful, for example, when a user wishes to compare non-negative continuous models with a default location parameter and with an estimated location parameter.

Each of the eighteen families of distributions discussed above has one or more parameters which must be specified in order to have a completely determined distribution. Each parameter can be specified in up to three different ways. Some of the parameters have UNIFIT defaults which can be accepted by the user. For example, the location

parameters for all non-negative distributions have a default value of zero. Alternatively, if the user knows the value of some parameter, this can be so stated and the known value entered. Finally, if the value of a parameter is neither known nor the default value acceptable, then the parameter can be estimated from the observed data using, in most cases, the method of maximum likelihood.

In general, one or more of the parameters of a distribution will be estimated from the observed data. UNIFIT allows the user to make confidence intervals for the estimated parameters and to estimate the asymptotic variance-covariance matrix for the parameters. One of three types of confidence intervals is provided depending upon the distribution and the manner in which parameter values are specified. Exact confidence intervals are provided when available. If exact confidence intervals are not available, then either approximate confidence intervals or asymptotic confidence intervals based on the properties of maximum likelihood estimators are derived.

### 3.3. Activity 3: Determining The Representativeness Of Each Fitted Distribution Using UNIFIT

After the user has hypothesized one or more families of distributions (Activity 1) and specified their parameters (Activity 2), he must then compare and evaluate the fitted distributions in Activity 3 to determine which distributions, if any, are the best representations of the underlying distribution for the sample. UNIFIT provides both heuristic techniques and formal goodness-of-fit tests for this purpose.

UNIFIT provides seven different heuristic techniques for comparing and evaluating fitted distributions; five of these heuristics are graphical in nature.

1. The **Frequency Comparison** is a graphical display showing the observed proportion of observations from the sample and the expected proportion of observations from a particular fitted model for each histogram interval.

2. The **Density/Histogram Overplot** is a graphical display available for continuous data samples. It shows an estimate of the underlying density function derived from the sample histogram and the density function of a particular fitted model.

3. The **Cumulative Frequency Comparison** is a graphical comparison between a sample distribution function which is computed from the observed data and the distribution function of a fitted distribution.

4. The **Quantile-Quantile (Q-Q) Plot**, which is a graphical comparison between the quantiles of a particular distribution and the quantiles of the sample, is designed to amplify differences which exist between the tails of a fitted continuous distribution and the tails of the sample distribution function. If the fitted distribution is a good model for the sample, then the Q-Q plot would be approximately linear.

5. The **Probability-Probability (P-P) Plot**, which is a graphical comparison between the fitted distribution function and the sample distribution function, is designed to amplify differences between the "middles" of the two distribution functions. It will also be approximately linear if the fitted distribution is a good model for the sample.

6. The **Relative Discrepancies Comparison** provides a measure of the linearity of the Q-Q and P-P plots for each fitted distribution. Relative discrepancies must lie between zero and one, with small relative discrepancies indicating that the corresponding distribution is a good representation of the underlying distribution for the sample.

7. The **Model Moment Comparison** is a comparison of the sample mean, variance, skewness, and kurtosis with the corresponding population moments for each fitted distribution.

Goodness-of-fit tests are used to examine formally whether there is any gross disagreement between the observed data and a fitted distribution. Specifically, these tests can be used to test the null hypothesis that the observed data are a random sample from the fitted distribution.

UNIFIT makes available to an analyst the chi-square test, the Kolmogorov-Smirnov (K-S) test, and the Anderson-Darling (A-D) tests, and also a heuristic for comparing fitted distributions based on these tests.

1. The **Chi-Square Test** is the most well-known and widely applicable goodness-of-fit test, being appropriate for any continuous or discrete distribution. It is, however, somewhat more complicated than what is stated in most books.

2. The **K-S Test** is not as widely applicable as the chi-square test, but is more powerful against many alternative distributions. In particular, it can only be validly performed for the exponential, lognormal, Weibull, normal, logistic, and extreme value distributions, and also for continuous distributions with all parameters known.

3. The **A-D Test** is applicable to the same distributions as the K-S test, but it is more powerful than either the chi-square or K-S tests against many alternatives.

4. Since the chi-square, K-S, and A-D test statistics are each measures of how well a hypothesized distribution fits the observed data, it is often informative to look at these statistics for all fitted distributions simultaneously. The three statistics should be "small" for a distribution which provides a good fit for the sample. The **Model Test Comparison** provides the values of all applicable test statistics for all of the fitted distributions.

It should be mentioned that these tests are often misstated in textbooks or incorrectly implemented in software packages.

## 4. ADDITIONAL FEATURES OF UNIFIT

Perhaps the most directly useful of the additional features of the UNIFIT package is the group of characteristics of a fitted distribution which can be calculated or graphed. This group contains the following capabilities.

1. Calculation of the distribution function for any value that can be taken on by the random variable.

2. Calculation of the percentage points of a continuous distribution for any percentage between zero and one hundred.

3. Calculation of an extensive table of percentage points of a continuous distribution which features central ranges of the random variable for common percentage values.

4. Calculation of the population moments.

5. Graphing of the density function for a continuous distribution.

UNIFIT also offers an extensive range of transformations which can be applied to a data sample. We have found this capability to be of great use when none of the standard distributions included in UNIFIT provides an adequate representation of the data. This also allows the analyst to employ more esoteric distributions which might be of particular interest in his discipline (e.g., the Rayleigh or log-logistic distributions).

With UNIFIT the analyst can select subsets of a data sample or merge a number of data samples together for analysis. This latter capability is of interest when observations of the same random variable from a number of different observation periods are available. To guarantee that such a merger is appropriate (e.g., that the values observed do indeed come from the same parent population), the user can first perform the Kruskal-Wallis test of homogeneity on the different data samples to test this assumption.

## 5. EXAMPLES OF THE GRAPHICAL DISPLAYS CREATED BY UNIFIT SOFTWARE

In our conference tutorial we shall show slides which demonstrate an analysis of a data set using UNIFIT. Included as Figures 1 through 4 are a frequency comparison, a density/histogram overplot, a cumulative frequency comparison, and a Q-Q plot. A Weibull distribution was fit to the continuous data sample and is referred to as model number 3 in the graphs. The displays are screen dumps produced on an IBM PC which had four color medium resolution graphics capabilities.

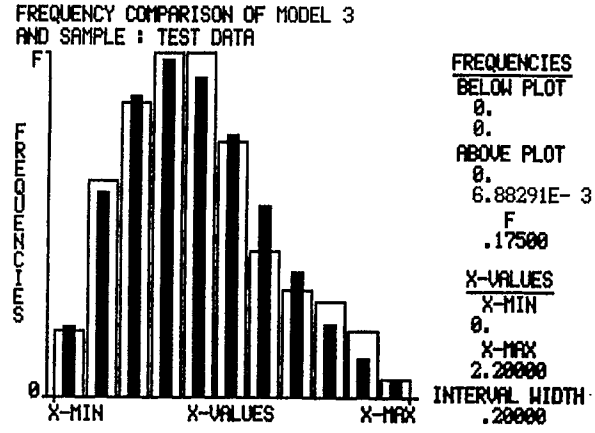


Figure 1: An Example Frequency Comparison Using A Weibull Distribution

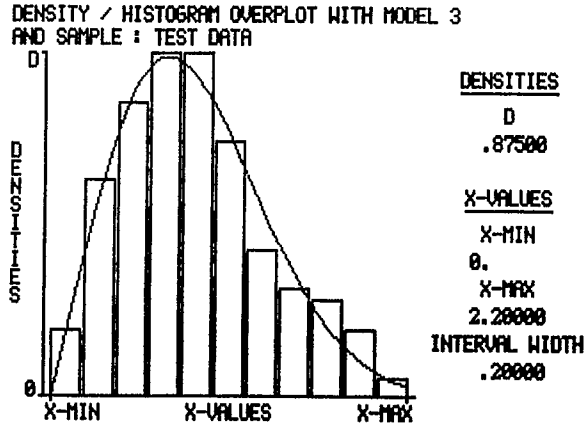


Figure 2: An Example Density/Histogram Overplot Using A Weibull Distribution

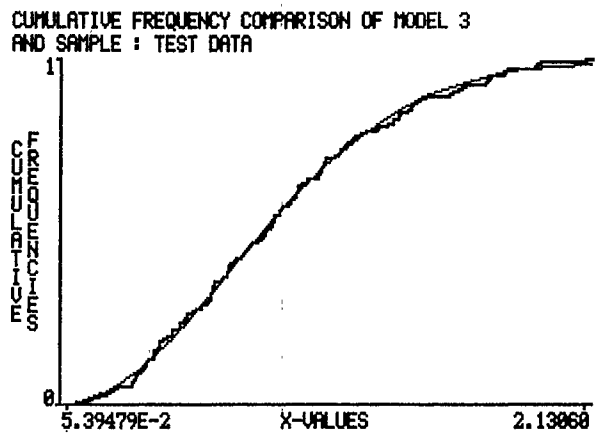


Figure 3: An Example Cumulative Frequency Comparison Using A Weibull Distribution

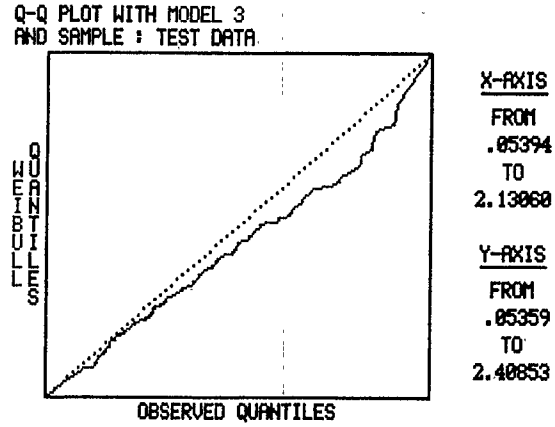


Figure 4: An Example Q-Q Plot Using A Weibull Distribution

AUTHORS' BIOGRAPHIES

AVERILL M. LAW is President of Simulation Modeling and Analysis Company and Professor at the University of Arizona. His book Simulation Modeling and Analysis (coauthored with David Kelton) is widely used by industry and universities. He is also the author of three other books and numerous papers on simulation, operations research, and statistics. Dr. Law was the editor (and an author) of Industrial Engineering Magazine's recent series on the simulation of manufacturing systems. His seminars "Simulation of Manufacturing Systems" and "Simulation Modeling and Analysis" have been attended by approximately 2000 people during the past nine years, and he has been a simulation consultant to more than 35 organizations. He received his Ph.D. in Industrial Engineering and Operations Research from the University of California at Berkeley.

Dr. Averill M. Law
Simulation Modeling and Analysis Company
P.O. Box 40996
Tucson, Arizona 85717
(602) 299-8441

STEPHEN G. VINCENT is a Vice President of Simulation Modeling and Analysis Company, in charge of software development. He has more than seven years of experience in developing software for simulation and statistics applications. He has B.S. and M.S. degrees in Industrial Engineering from the University of Wisconsin.

Stephen G. Vincent
Simulation Modeling and Analysis Company
P.O. Box 40996
Tucson, Arizona 85717
(602) 299-8441