*Proceedings of the 2016 Winter Simulation Conference*
*T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds*

# A TUTORIAL ON HOW TO SELECT SIMULATION INPUT PROBABILITY DISTRIBUTIONS

Averill M. Law

Averill M. Law & Associates, Inc.
4729 East Sunrise Drive, #462

Tucson, AZ 85718, USA

## ABSTRACT

An important, but often neglected, part of any sound simulation study is that of modeling each source of system randomness by an appropriate probability distribution. We first give some examples of data sets from real-world simulation studies, which is followed by a discussion of two critical pitfalls in simulation input modeling. The two major methods for modeling a source of randomness when corresponding data are available are delineated, namely, fitting a theoretical probability distribution to the data and the use of an empirical distribution. We then give a three-activity approach for choosing the theoretical distribution that best represents a set of observed data. This is followed by a discussion of how to model a source of system randomness when no data exist.

## 1 INTRODUCTION

To carry out a simulation using random inputs, we have to specify their probability distributions. For example, in the simulation of a single-server queueing system, we must give probability distributions for the interarrival times of customers and for the service times of customers at the server. Then, given that the input random variables to a simulation model follow particular distributions, the simulation proceeds through time by generating random values from these distributions. Our concern in this tutorial is how the analyst might go about specifying these input probability distributions.

Almost all real-world systems contain one or more sources of randomness. In Figures 1 through 3 we show histograms of three data sets taken from actual simulation projects. Figure 1 corresponds to 910 machine processing times (in minutes) for an automotive manufacturer. It can be seen than the histogram has a longer right tail (positive skewness) and that the minimum time is approximately 15 minutes. In Figure 2 we show a histogram for 122 repair times (in hours) for a component of a U.S. Navy weapons system, which is once again skewed to the right. Finally, in Figure 3 we display a histogram of 219 interarrival times (in minutes) to a drive-up bank. We will use this data set in our examples of Section 4. Looking at the three histograms, we see that none of them look like the density function of a normal distribution, which is *symmetric* about its mean. As a matter of fact, it might be said with some truth that, "The greatest application of the normal distribution is writing statistics books."

The remainder of this tutorial is organized as follows. Section 2 discusses two critical pitfalls in simulation input modeling. In Section 3 the two major methods are delineated for modeling a source of randomness when corresponding data are available, namely, fitting a theoretical probability distribution to the data and the use of an empirical distribution. Then in Section 4 we give a three-activity approach for choosing the standard theoretical distribution that best represents a set of observed data. This is followed

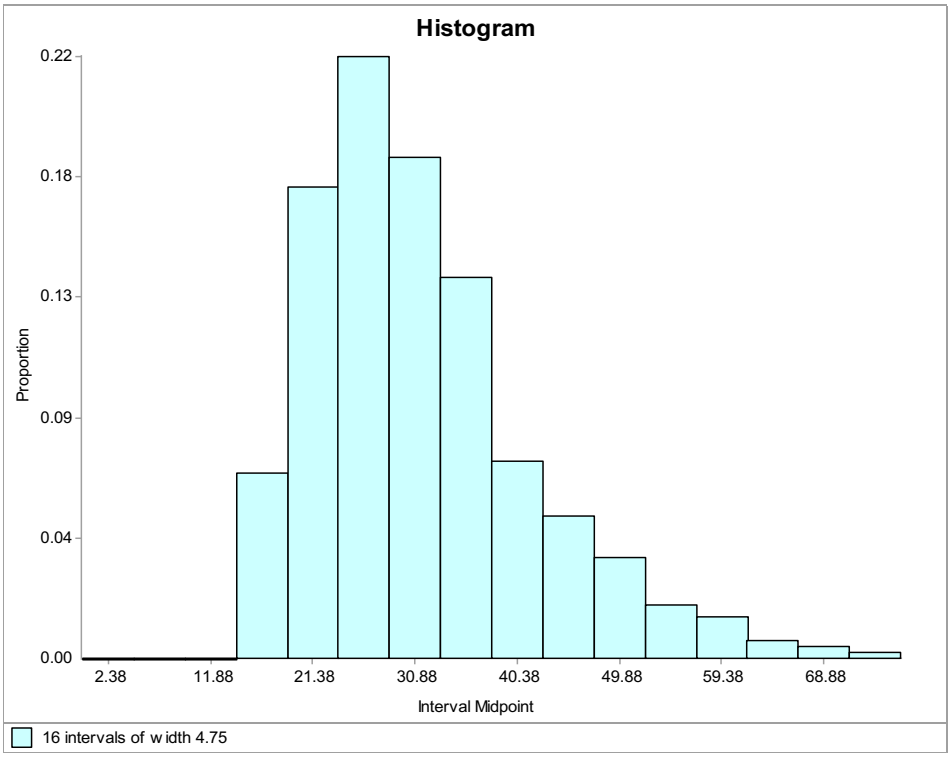


Figure 1: Histogram of 910 processing times for an automotive manufacturer.

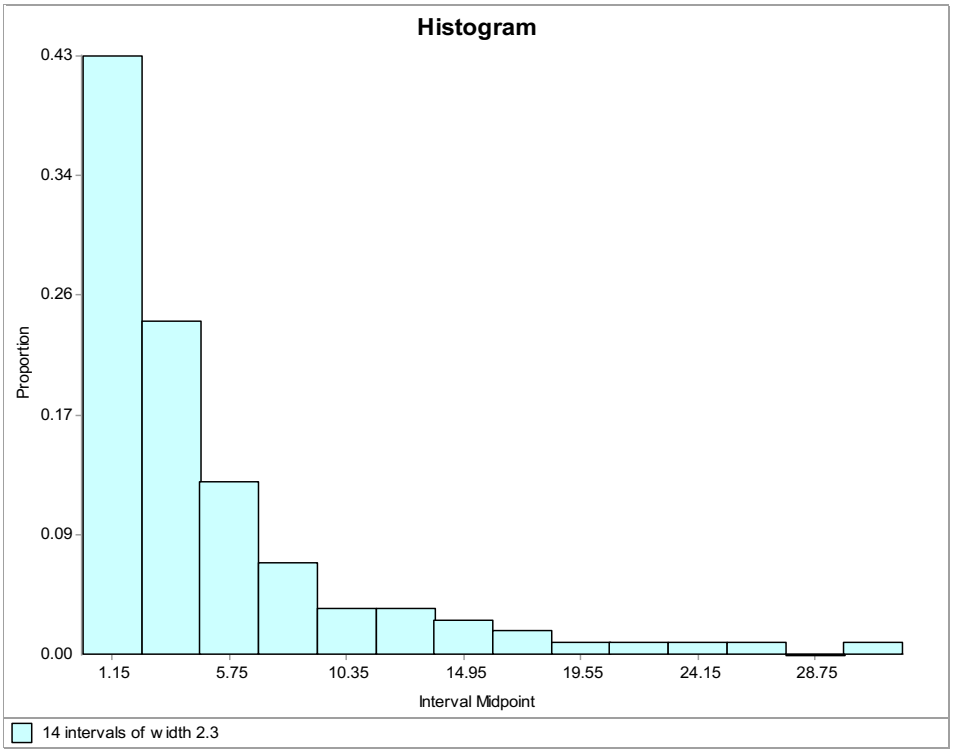


Figure 2: Histogram of 122 repair times for a U.S. Navy weapons system.

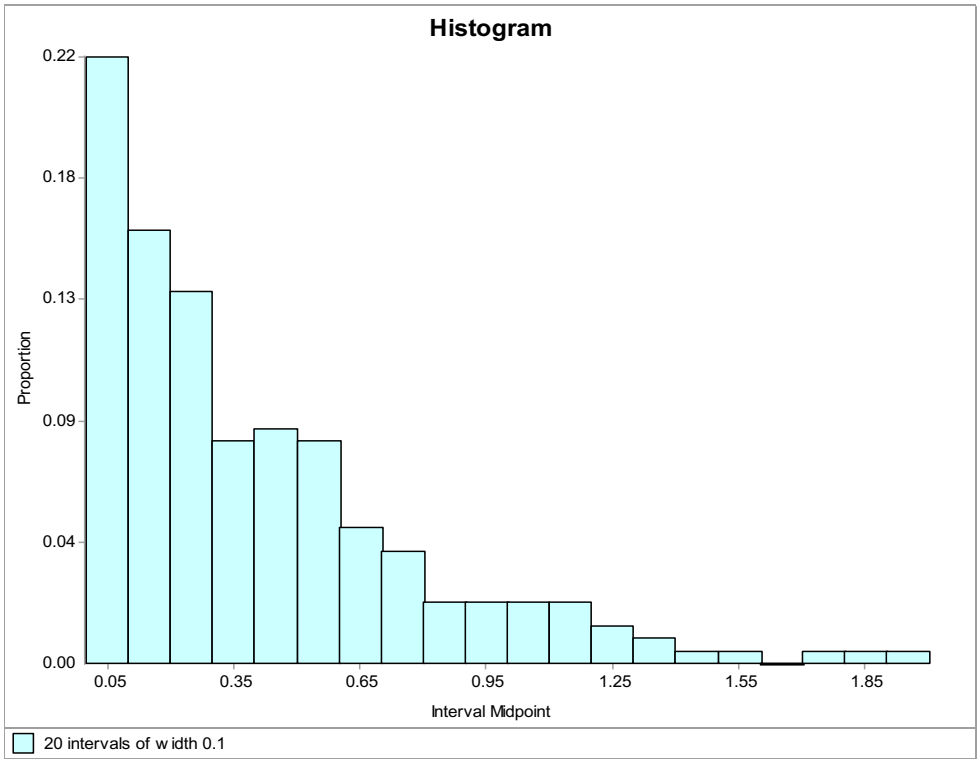


Figure 3: Histogram of 219 interarrival times to a drive-up bank.

in Section 5 by a discussion of how to model a source of system randomness when no data exist. Section 6 is a summary of this paper.

Portions of this paper are based on chapter 6 of Law (2015). Other references on simulation input modeling are Banks et al. (2010), Biller and Gunes (2010), and Kuhl et al. (2009). The graphical plots and goodness-of-fit tests presented in this paper were developed using the ExpertFit distribution-fitting software (see Averill M. Law & Associates (2016)).

## 2 TWO FUNDAMENTAL PITFALLS IN SIMULATION INPUT MODELING

We have identified a number of pitfalls that can undermine the success of a simulation study (see section 1.8 in Law (2015)). Two of these pitfalls that directly relate to simulation input modeling are discussed in the following sections.

### 2.1 Pitfall Number 1: Replacing a Distribution by its Mean

Simulation analysts have sometimes replaced an input probability distribution by the perceived value of its 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 delay 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, service, and transportation) systems.

### 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 (see Figures 1 through 3).

Suppose for the queueing system in Section 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. 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 (2015), the *Weibull distribution* actually provides the best overall model for the data.

We then made 100 independent simulation runs of length 1000 delays of the system using *each* of the fitted distributions. The overall average delay in the queue (i.e., based on 100,000 delays) for the Weibull distribution was 4.36 minutes, which should be close to the average delay in queue for the actual system. On the other hand, the average delay in queue for the normal distribution was 6.04 minutes, 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.

## 3 METHODS OF REPRESENTING RANDOMNESS GIVEN THAT SYSTEM DATA ARE AVAILABLE

Suppose that independent, identically distributed (IID) data $X_1, X_2, \ldots, X_n$ are available from a continuous distribution (e.g., service times) with distribution function $F(x)$. (Discrete distributions are discussed in Law (2015).) Our goal is to find a distribution that provides a sufficiently accurate *approximation* to $F(x)$ so that "valid" results are obtained from our simulation study. (We will probably never know $F(x)$ exactly.) There are two major approaches for trying to find a good approximation to $F(x)$, which are discussed in the following sections.

### 3.1 Fitting Standard Theoretical Distributions to the Data

With this approach we "fit" various standard theoretical distributions (e.g., exponential, lognormal, or Weibull) to our data with the goal of finding one that provides a good approximation to $F(x)$. What it means to fit a distribution to data and how we determine the quality of the representation are discussed in Section 4. The major drawback of this approach is that for some data sets we simply cannot find a theoretical distribution that provides a good representation for our data. Two possible reasons for this are that our data are actually from two or more heterogeneous populations or that the data have been significantly rounded (e.g., service times that have been rounded to the nearest hour), effectively discretizing the data in the latter case.

### 3.2 Using an Empirical Distribution Constructed from the Data

With this approach we construct an empirical distribution $F(x)$ from our data, which is used as an ap-

proximation to $F(x)$. Let $X_{(i)}$ denote the $i$th smallest of the $X_j$'s, so that $X_{(1)} \le X_{(2)} \le \quad \le X_{(n)}$. Then we define $F(x)$ as follows:

$$F(x) = \begin{cases} 0 & \text{if } x < X_{(1)} \\ \dfrac{i-1}{n-1} + \dfrac{x - X_{(i)}}{(n-1)(X_{(i+1)} - X_{(i)})} & \text{if } X_{(i)} \le x < X_{(i+1)} \text{ for } i = 1, 2, \quad , n-1 \\ 1 & \text{if } X_{(n)} \le x \end{cases}$$

An illustration for $n = 5$ is given in Figure 4.

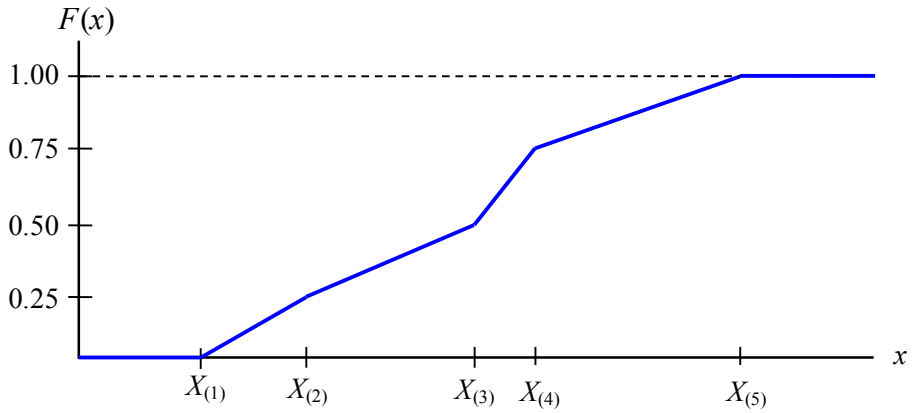


Figure 4: Continuous, piecewise-linear empirical distribution function

The major disadvantage of using the empirical distribution function $F(x)$ is that values outside of the range of the observed data, namely, $[X_{(1)}, X_{(n)}]$ cannot be generated in the simulation, which is a problem if $n$ is "small." Another problem with using an empirical distribution is that $2n$ values (i.e., the $n$ $X_{(i)}$'s and their corresponding cumulative probabilities) have to be entered into the simulation model, which may be problematic for "large" $n$.

### 3.3 Deciding which Approach to Use

If a standard theoretical distribution can be found that provides a good representation of our data (see Section 4.3), then we believe that this approach is preferable over the use of an empirical distribution, because of its shortcomings of the latter approach noted above. Also, a theoretical distribution provides a compact representation of our data that smoothes out any "irregularities." If a good theoretical distribution cannot be found, then an empirical distribution should be used. As the sample size $n$ get gets larger, $F(x)$ will converge to $F(x)$, but there is still the problem of entering the $2n$ values into the simulation model.

## 4 FINDING THE THEORETICAL PROBABILITY DISTRIBUTION THAT BEST REPRESENTS A DATA SET

In this section we discuss the three basic activities in specifying a theoretical distribution on the basis of the observed data $X_1, X_2, \ldots, X_n$.

### 4.1 Activity I: Hypothesizing Families of Distributions

The first step in selecting a particular input distribution is to decide what general families (e.g., exponential, gamma, Weibull, normal, or lognormal) appear to be appropriate on the basis of their shapes, without worrying (yet) about the specific parameter values for these families.

Some distributions are characterized at least partially by functions of their *true* parameters. In Table 1 we give a number of these functions, formulas to estimate these functions from IID data (these estimates are called *summary or descriptive statistics*), and comments about their interpretation or use. These functions might be used in some cases to suggest an appropriate distribution family. For a symmetric continuous distribution (e.g., normal), the mean $\mu$ is equal to the median $x_{0.5}$. Thus, if the estimates $\bar{X}(n)$ and $\hat{x}_{0.5}$ are almost "equal," then this is some indication that the underlying distribution may be symmetric. If $\bar{X}(n) > \hat{x}_{0.5}$, then it is often (but not always) true that the underling density function has a longer right tail than left tail, and vice versa.

Table 1. Useful summary statistics.

| Function | Sample estimate (summary statistic) | Comments |
|---|---|---|
| Mean $\mu$ | $\bar{X}(n)$ | Measure of central tendency |
| Median $x_{0.5}$ | $\hat{x}_{0.5}(n) = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ [X_{(n/2)} + X_{((n/2)+1)}]/2 & \text{if } n \text{ is even} \end{cases}$ | Alternative measure of central tendency |
| Variance $\sigma^2$ | $S^2(n)$ | Measure of variability |
| Coefficient of variation, $\text{cv} = \dfrac{\sqrt{\sigma^2}}{\mu}$ | $\text{cv}(n) = \dfrac{\sqrt{S^2(n)}}{\bar{X}(n)}$ | Alternative measure of variability |
| Skewness, $\nu = \dfrac{E[(X-\mu)^3]}{\left(\sigma^2\right)^{3/2}}$ | $\hat{\nu}(n) = \dfrac{n^2}{(n-1)(n-2)} \dfrac{\sum_{i=1}^{n} [X_i - X(n)]^3 / n}{\left[S^2(n)\right]^{3/2}}$ | Measure of symmetry |

The *coefficient of variation* cv can sometimes provide useful information about the form of a continuous distribution. In particular, cv = 1 for the exponential distribution. The *skewness* $\nu$ is a measure of the symmetry of a distribution. For symmetric distributions like the normal, $\nu = 0$. If $\nu > 0$, the distribution is skewed to the right (i.e., the density has a longer right tail than left tail); if $\nu < 0$, the distribution is skewed to the left. Thus, the estimated skewness $\hat{\nu}(n)$ can be used to ascertain the shape of the underlying density function. See section 6.4.1 of Law (2015) for additional uses of summary statistics.

A histogram of the data is one of the most useful tools for determining the shape of the underlying density function, since it is essentially a graphical estimate of the density. However, a fundamental problem with making a histogram is in choosing the interval width *w*, and we recommend selecting the smallest interval width *w* that gives us a reasonably "smooth" histogram.

**Example 1.** Consider the 219 interarrival times of cars to a drive-up bank in Figure 3. The summary statistics for these data are given in Table 2. Since $\bar{X}(219) = 0.399 > 0.270 = \hat{x}_{0.5}(219)$ and $\hat{v}(219) = 1.478$, this suggests that the underlying distribution is skewed to the right, rather than symmetric. Furthermore, $\text{cv}(219) = 0.953$, which is close to the theoretical value of 1 for the expo-nential distribution. A smooth histogram of the data with $w = 0.1$ was given in Figure 3. In Figure 5

Table 2: Summary statistics for the interarrival time data.

| Summary statistic | Value |
|---|---|
| Mean | 0.399 |
| Median | 0.270 |
| Variance | 0.144 |
| Coefficient of variation | 0.953 |
| Skewness | 1.478 |

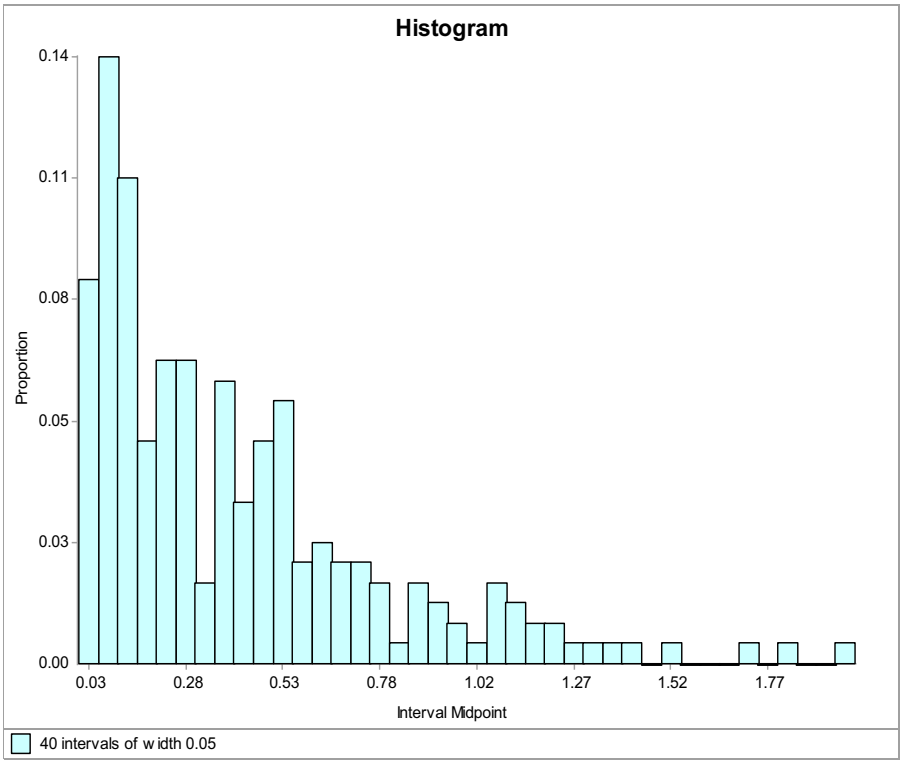


Figure 5: Histogram of 219 interarrival times to a drive-up bank with an interval width of 0.05.

we give a histogram of the data when the interval width is $w = 0.05$, and we see that this histogram is fairly "jagged." (A histogram with an interval width of 0.15 is also smooth.) Thus, the smooth histogram with the smallest interval width corresponds to $w = 0.1$ and its shape resembles that of an exponential density.

## 4.2 Activity II: Estimation of Parameters

After one or more candidate families of distributions have been hypothesized in Activity I, we must somehow specify the values of their parameters in order to have completely specified distributions for possible use in our simulation model. (For example, the exponential distribution has one parameter $\beta$ that is its mean.) Our IID data $X_1, X_2, \ldots, X_n$ were used to help us hypothesize distributions, and these same data can also be used to estimate their parameters. When data are used directly in this way to specify a numerical value for an unknown parameter, we say that we are *estimating* that parameter from the data.

An *estimator* is a numerical function of the data. There are many ways to specify the form of an estimator for a particular parameter of a given distribution, and many ways to evaluate the quality of an estimator. We shall consider only one type, *maximum-likelihood estimators* (MLEs), for three reasons: (1) MLEs have several desirable properties often not enjoyed by alternative methods of estimation, (2) the use of MLEs turns out to be important in justifying the chi-square and Kolmogorov-Smirnov goodness-of-fit tests, and (3) the central idea of maximum-likelihood estimation has a strong intuitive appeal.

Suppose that we have hypothesized a continuous distribution for our data that has one unknown parameter $\theta$. Let $f_\theta(x)$ denote the probability density function for this distribution, so that the parameter $\theta$ is part of the notation. *Given that we have already observed* the IID data $X_1, X_2, \ldots, X_n$, we define the likelihood function $L(\theta)$ as follows:

$$L(\theta) = f_\theta(X_1) f_\theta(X_2) \quad f_\theta(X_n)$$

$L(\theta)$, which is just the joint probability density function since the data are independent, can be *thought of* as giving the probability (likelihood) of obtaining our observed data if $\theta$ is the value of the unknown parameter (see problem 6.26 in Law (2015) for a justification). Then the MLE of the unknown value of $\theta$, which we denote by $\hat{\theta}$, is defined to be that value of $\theta$ that maximizes $L(\theta)$; that is, $L(\hat{\theta}) \geq L(\theta)$ for all possible values of $\theta$. Thus, $\hat{\theta}$ "best explains" the data that we have collected.

**Example 2.** For the exponential distribution that appeared to be good candidate distribution in Example 1, $\theta = \beta$ ($\beta > 0$) and

$$f_\beta(x) = \frac{1}{\beta} e^{-x/\beta} \text{ for } x \geq 0$$

The likelihood function is

$$L(\beta) = \left(\frac{1}{\beta} e^{-X_1/\beta}\right)\left(\frac{1}{\beta} e^{-X_2/\beta}\right) \quad \left(\frac{1}{\beta} e^{-X_n/\beta}\right) = \beta^{-n} \exp\left(-\frac{1}{\beta}\sum_{i=1}^{n} X_i\right)$$

and we seek the value of $\beta$ that maximizes $L(\beta)$ over all $\beta > 0$. The task is more easily accomplished if, instead of working directly with $L(\beta)$, we work with its logarithm. Thus, we define the *log-likelihood function* $l(\beta)$ as

$$l(\beta) = \ln L(\beta) = -n \ln \beta - \frac{1}{\beta}\sum_{i=1}^{n} X_i$$

Since the logarithm is strictly increasing, maximizing $L(\beta)$ is equivalent to maximizing $l(\beta)$, which is much easier. Standard differential calculus can be used to maximize $l(\beta)$ by setting its derivative to zero and solving for $\beta$. That is,

$$\frac{dl}{d\beta} = \frac{-n}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} X_i$$

which equals zero if and only if

$$\beta = \sum_{i=1}^{n} X_i / n = \bar{X}(n)$$

To make sure that that $\beta = \bar{X}(n)$ is a maximizer of $l(\beta)$ (as opposed to a minimizer or an inflection point), a sufficient (but not necessary) condition is that $\frac{d^2 l}{d\beta^2}$, evaluated at $\beta = \bar{X}(n)$, be negative, which is the case here. Notice that the MLE is quite natural here, since $\beta$ is the mean of the hypothesized distribution and the MLE is the *sample* mean, which is an unbiased estimator of $\beta$. For the data of Example 1, $\hat{\beta} = \bar{X}(219) = 0.399$.

### 4.3 Activity III: Determining How Representative the Fitted Distributions Are

After determining one or more probability distributions that might fit our observed data in Activities I and II, we must now closely examine these distributions to see how well they represent the true underlying distribution for our data. If several of these distributions are "representative," we must determine which distribution provides the best fit. Remember that in general, none of our fitted distributions will probably be *exactly* correct. What we are really trying to do is to determine a distribution that is accurate enough for the intended purposes of the model.

In this section we discuss both graphical procedures and goodness-of-fit hypothesis tests for determining the "quality" of our fitted distributions.

#### 4.3.1 Graphical Procedures

We discuss two heuristic graphical procedures for comparing fitted distributions with the true underling distribution.

**Density-Histogram Plots**

For continuous data, a *density-histogram plot* can be made by plotting $w\hat{f}(x)$ over the histogram and looking for similarities, where $\hat{f}(x)$ is the density function of a fitted distribution. (Note that the area under a histogram is *w*, while the area under a density is 1.)

**Example 3.** For the interarrival-time data of Example 1, we hypothesized an exponential distribution and obtained the MLE $\hat{\beta} = 0.399$ in Example 2. Thus, the density function of the fitted distribution is

$$\hat{f}(x) = \begin{cases} 2.506e^{-x/0.399} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

For the histogram in Figure 3, we give a density-histogram plot in Figure 6.

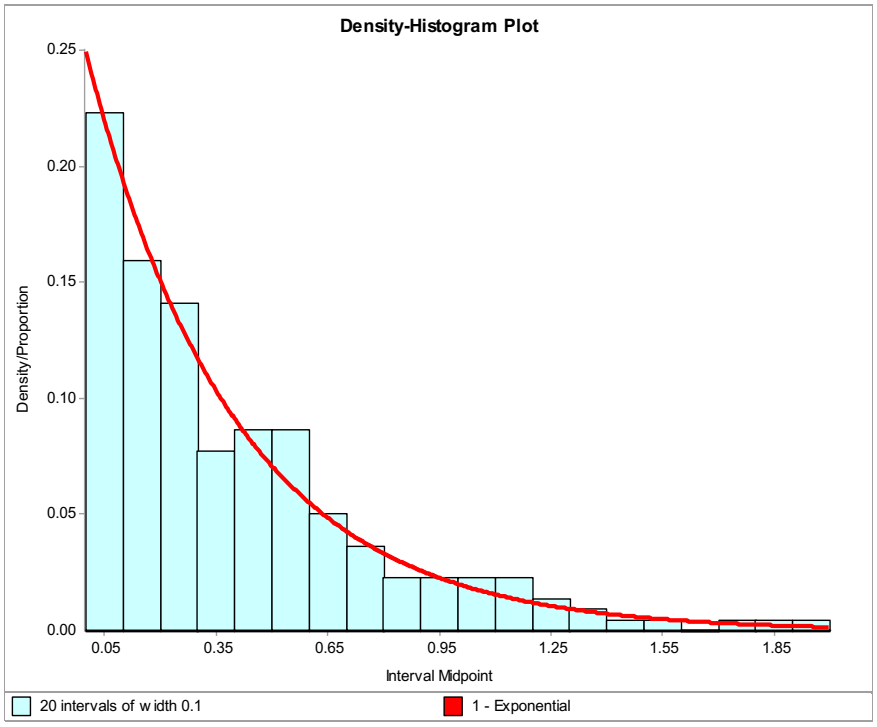


Figure 6: Density-histogram plot for the fitted exponential distribution and the interarrival-time data.

**Distribution-Function-Differences Plots**

The density-histogram plot can be thought of as a comparison of the individual probabilities of the fitted distribution and of the individual probabilities of the true underlying distribution. We can also make a graphical comparison of cumulative probabilities (distribution functions). Define a sample distribution function $F_n(x)$ as follows:

$$F_n(x) = \frac{\text{number of } X_i\text{'s} \le x}{n}$$

which is the proportion of observations that are less than or equal to $x$. Let $\hat{F}(x)$ be the distribution function of the fitted distribution. A *distribution-function-differences plot* is a plot of the differences between $\hat{F}(x)$ and $F_n(x)$, over the range of the data. If the fitted distribution is a perfect fit and the sample size is infinite, then this plot will be a horizontal line at height 0. Thus, the greater the vertical deviations from this line, the worse the quality of fit.

**Example 4.** A distribution-function-differences plot for the interarrival-time data of Example 1 and the fitted exponential distribution is given in Figure 7. This plot indicates a good fit except possibly at the lower end of the range of the observed data.

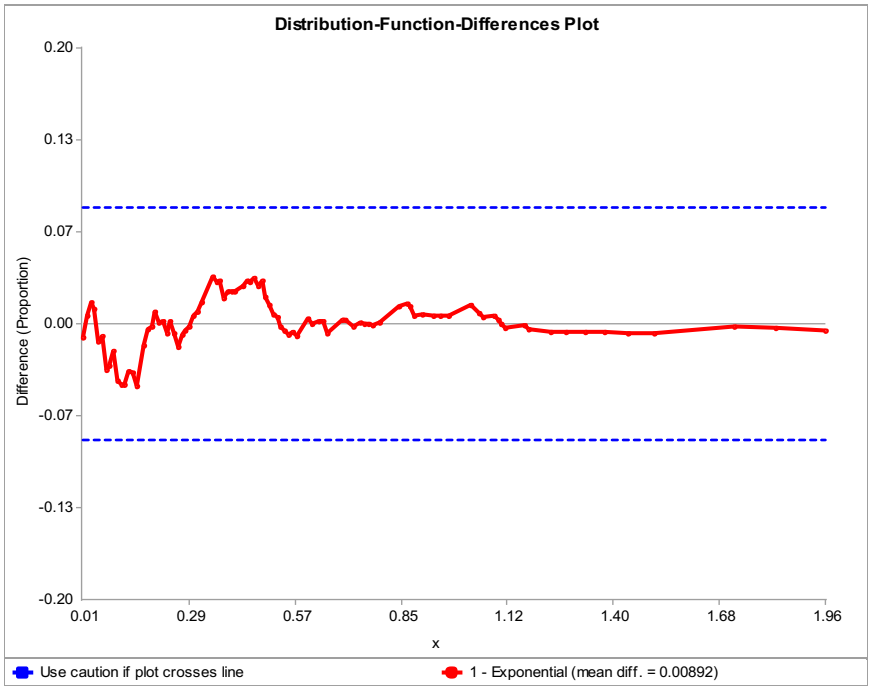


Figure 7: Distribution-function-differences plot for the fitted exponential distribution and the interarrival-time data.

### 4.3.2. Goodness-of-Fit Tests

A *goodness-of fit test* is a statistical hypothesis test (see, for example, Devore (2016)) that is used to assess formally whether the observations $X_1, X_2, \ldots, X_n$ are an independent sample from a particular distribution with distribution function $\hat{F}$. That is, a goodness-of fit test can be used to test the following null hypothesis:

$$\mathrm{H}_0\text{: The } X_i\text{'s are IID random variables with distribution function } \hat{F}$$

We begin our discussion with the *chi-square test*, which can be considered a more formal comparison of a histogram with the fitted density function. To compute the chi-square test statistic, we must first divide the entire range of the fitted distribution into *k* adjacent intervals $[a_0, a_1), [a_1, a_2), \quad , [a_{k-1}, a_k)$. (For Example 5 below, $a_0 = 0$ and $a_k = \infty$.) Then we tally

$$N_j = \text{number of } X_i\text{'s in the } j\text{th interval } [a_{j-1}, a_j)$$

for $j = 1, 2, \quad , k$. (Note that $\sum_{j=1}^{k} N_j = n$.) Next, we compute the expected *proportion* $p_j$ of the $X_i$'s that would fall in the *j*th interval if we were sampling from the fitted distribution, which is

$$p_j = \int_{a_{j-1}}^{a_j} \hat{f}(x)\,dx$$

Finally, we compute the test statistic

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

Since $np_j$ is the expected number of the $n$ $X_i$'s that would fall in the $j$th interval if $H_0$ were true, we would expect $\chi^2$ to be small if the fit were good. Therefore, we reject $H_0$ if $\chi^2$ is too large.

Suppose that we would like to perform a test at level $\alpha$, where $\alpha$ is typically 0.05 or 0.10. Let $\chi^2_{k-1,1-\alpha}$ be the upper $1-\alpha$ critical point for a chi-square distribution with $k-1$ degrees of freedom (see, for example, Table T2 on page 723 in Law (2015)). Then we reject the null hypothesis $H_0$ at level $\alpha$ if $\chi^2 > \chi^2_{k-1,1-\alpha}$, and we fail to reject $H_0$ otherwise.

The most troublesome aspect of carrying out the chi-square test is choosing the number and size of the intervals. This is a difficult problem, and no definitive prescription can be given that is guaranteed to produce good results in terms of validity (actual level of the test close to the desired level $\alpha$) and high power (ability to discriminate between $\hat{F}$ and the distribution that is really true) for all hypothesized distributions and all sample sizes. There are, however, a few guidelines that are often followed. First, some of the ambiguity in interval selection is eliminated if the intervals are chosen so that $p_1 = p_2 = \quad = p_k$, which is called the *equiprobable approach*. (Thus, under this approach, equal-sized histogram intervals would *not* be used.) For the equiprobable approach, it is also recommended that $k \geq 3$ and $np_j \geq 5$ for all $j$. However, these recommendations are not completely definitive. For example, in the case of the $n$ = 219 interarrival times of Example 1, these rules would say that $k$ should be between 3 and 44, which is a large range of values. The lack of a clear prescription for interval selection is the major drawback of the chi-square test. In some situations entirely different conclusions can be reached from the *same* data set depending on how the intervals are specified. The chi-square test nevertheless remains in wide use, since it can be applied to any hypothesized distribution.

**Example 5.** We now use the chi-square test with level $\alpha = 0.05$ to compare the $n$ = 219 interarrival times of Example 1 with the fitted exponential distribution having distribution function $\hat{F}(x) = 1 - e^{-x/0.399}$ for $x \geq 0$. If we form, say, $k$ = 20 intervals with $p_j = 1/k = 0.05$ for $j = 1, 2, \ , 20$, then $np_j = (219)(0.05) = 10.95$, so this satisfies the guidelines that the intervals be chosen with equal $p_j$'s and $np_j \geq 5$. The computations for the test are given in section 6.6.2 of Law (2015) and the value of the test statistic turns out to be $\chi^2 = 22.188$. Referring to Table T2 in Law (2015), we see that $\chi^2_{19,0.95} = 30.144$, which is not exceeded by $\chi^2$, so we do not reject $H_0$ at level $\alpha = 0.05$. Thus, this test gives us no reason to conclude that our data are poorly fitted by an exponential distribution with $\beta = 0.399$.

We now consider the Kolmogorov-Smirnov (K-S) test, which does not have the troublesome interval specification of the chi-square test. However, it does have its own drawbacks as we will see below. To define the K-S statistic, recall the sample distribution function $F_n(x)$ from Section 4.3.1. If $\hat{F}(x)$ is the fitted distribution function, a natural assessment of goodness of fit is some kind of measure of the closeness of the functions $F_n$ and $\hat{F}$. The K-S test statistic $D_n$ is simply the *largest* (vertical) distance between $F_n(x)$ and $\hat{F}(x)$ *for all values of* $x$, and it can be computed from the following formulas:

$$D_n^+ = \max_{1 \le i \le n} \left\{ \frac{1}{n} - \hat{F}(X_{(i)}) \right\}, \quad D_n^- = \max_{1 \le i \le n} \left\{ \hat{F}(X_{(i)}) - \frac{i-1}{n} \right\}$$

and

$$D_n = \max\left\{ D_n^+, D_n^- \right\}$$

Clearly, a large value of $D_n$ indicates a poor fit, so that the form of the test is to reject the null hypothesis $H_0$ if $D_n$ exceeds some constant $d_{n,1-\alpha}$, where $\alpha$ is the specified level of the test. The problem is that values of $d_{n,1-\alpha}$ are available for only certain *continuous* distributions and the values are different for each applicable distribution. In particular, values of $d_{n,1-\alpha}$ are available for five cases: (1) all parameters of $\hat{F}$ are known (i.e., none of the parameters of $\hat{F}$ are estimated in any way from the data, which includes the $U(0,1)$ distribution), (2) normal (lognormal) distribution, (3) exponential distribution, (4) Weibull distribution, and (5) logistic (log-logistic) distribution. Moreover, in the latter three cases parameters of the fitted distributions have to be estimated by the method of maximum likelihood. Unfortunately, these limitations of the K-S test are not at all well known, and people routinely apply the K-S test to all continuous and discrete distributions using the values of $d_{n,1-\alpha}$ that are only applicable to the all-parameters-known case. This results in a precipitous drop in the power (discriminating ability) of the K-S test. More details about the K-S test can be found in Law (2015).

**Example 6.** We now perform the K-S test at level $\alpha = 0.05$ to determine whether the $n = 219$ interarrival times are well fit by the exponential distribution having distribution function $\hat{F}(x) = 1 - e^{-x/0.399}$ for $x \ge 0$. Using the above formulas we got a test statistic of $D_{219} = 0.047$. From Table 6.15 in Law (2015) we computed that $d_{219,0.95} = 0.073$, which is not exceeded by the test-statistic value of 0.047. Therefore, the K-S test gives us no reason to reject the fitted exponential distribution at level $\alpha = 0.05$.

It should be mentioned that there is another goodness-of-fit test, called the Anderson-Darling test, which has higher power than the K-S test against many alternative distributions, as discussed in Stephens (1974) and Law (2015).

We conclude this section with some general comments about the efficacy of goodness-of-fit tests. In particular, the following are some drawbacks of these tests:

- The null hypothesis $H_0$ is often false.
- The power of these tests is low for small to moderate sample sizes.
- The power of these tests approaches 1 as the sample size gets large, causing the null hypothesis to be rejected unless the fitted distribution is exactly correct.

## 5 SELECTING A DISTRIBUTION IN THE ABSENCE OF DATA

In some simulation studies it may not be possible to collect data on the random variables of interest, so the techniques of Section 4 are not applicable to the problem of selecting corresponding probability distributions. For example, if the system being studied does not currently exist in some form, then collecting data from the system is obviously not possible. This difficulty can also arise for existing systems, if the number of required probability distributions is large and the time available for the simulation study prohibits the necessary data collection and analysis.

Let us assume that the random quantity of interest is a continuous random variable *X*. It will also be useful to think of this random variable as being the time to perform some task, e.g., the time required to repair a piece of equipment when it fails. One approach in this case would be to use a triangular distribution,

which we describe next. The first step in using the triangular distribution approach is to identify an interval $[a,b]$ (where $a$ and $b$ are real numbers such that $a < b$) in which it is felt that $X$ will lie with probability close to 1; that is, $P(a \le X \le b) \approx 1$. To obtain *subjective* estimates of $a$ and $b$, subject-matter experts (SMEs) are asked for their most optimistic and pessimistic estimates, respectively, of the time to perform the task. We next ask the SMEs for their subjective estimate of the most-likely time to perform the task, $m$, which is the mode of the distribution of $X$. Given $a$, $b$, and, $m$, the random variable $X$ is then considered to have a triangular distribution on the interval $[a,b]$ with mode $m$, as shown in Figure 8. The height of the triangle above $m$ is chosen to make the area under the density function equal to 1.

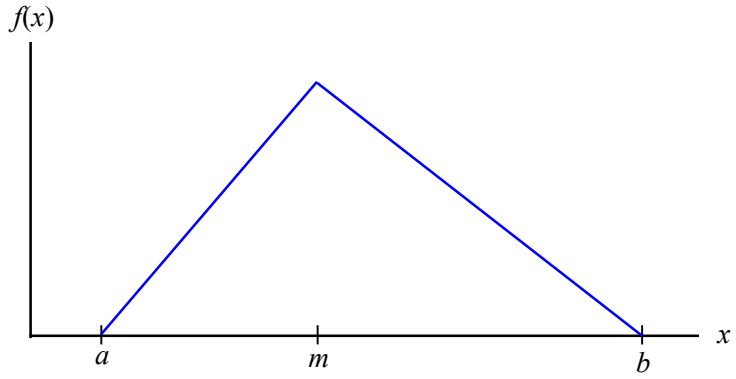


Figure 8: Triangular density function on the interval $[a,b]$ with mode $m$.

## 6 SUMMARY

We have seen in Section 2 the danger of replacing a probability distribution by its perceived mean value or of using an inappropriate distribution. For the case where data are available, we discussed the two main approaches for representing a source of system randomness, namely, fitting standard theoretical distributions and the use of empirical distributions, and we gave recommendations for when to use each approach. Finally, we showed how the triangular distribution can be used to model a source of randomness such as a task time in the absence of data.

There is an extensive amount of material available on selecting simulation input probability distributions, and further details on all of the topics covered in this tutorial can be found in Law (2015).

## REFERENCES

Averill M. Law & Associates. 2016. "ExpertFit Distribution-Fitting Software," Version 8. Tucson, Arizona.

Banks, J., J. S. Carson, B. L. Nelson, and D. M. Nicol. 2010. *Discrete-Event System Simulation*. 5th ed. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Biller, B., and C. Gunes. 2010. "Introduction to Simulation Input Modeling." In *Proceedings of the 2010 Winter Simulation Conference,* Edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, E. Yucësan, 49-58. Piscataway, New Jersey: Institute of Electrical and Electronic Engineers.

Devore, J. L. 2016. *Probability and Statistics for Engineering and the Sciences*. 9th ed. Boston, MA: Cengage Learning.

Kuhl, M. E., N. M. Steiger, E. K. Lada, M. A. Wagner, and J. R. Wilson. 2009. "Introduction to Modeling and Generating Probabilistic Input Processes for Simulation." In *Proceedings of the 2009 Winter Simulation Conference,* Edited by M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin, and R. G. Ingalls, 184-202. Piscataway, New Jersey: Institute of Electrical and Electronic Engineers.

Law, A. M. 2015. *Simulation Modeling & Analysis.* 5th ed. New York: McGraw-Hill, Inc.
Stephens, M. A. 1974. "EDF Statistics for Goodness of Fit and Some Comparisons." *J. American Statist. Assoc.* 69: 730-737.

## AUTHOR BIOGRAPHIES

**AVERILL M. LAW** is President of Averill M. Law & Associates, a company specializing in simulation seminars, simulation consulting, and software. He has presented more than 550 simulation and statistics short courses in 20 countries, including onsite seminars for AT&T, Boeing, Caterpillar, Coca-Cola, CSX, GM, IBM, Intel, Lockheed Martin, Los Alamos National Lab, NASA, NSA, NATO (Netherlands), Sasol Technology (South Africa), 3M, UPS, U.S. Air Force, U.S. Army, U.S. Navy, and Verizon. Dr. Law has been a simulation consultant to more than 50 organizations including Booz Allen & Hamilton, Conoco/Phillips, Defense Modeling and Simulation Office, Kimberly-Clark, M&M/Mars, Oak Ridge National Lab, U.S. Air Force, U.S. Army, U.S. Marine Corps, and U.S. Navy. He has written or coauthored numerous papers and books on simulation, operations research, statistics, manufacturing, and communications networks, including the book *Simulation Modeling and Analysis* that has more than 163,000 copies in print and 16,000 citations. He developed the ExpertFit® distribution-fitting software and also several videotapes on simulation modeling. He was awarded the INFORMS Simulation Society Lifetime Professional Achievement Award in 2009. Dr. Law 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@simulation.ws> and his website is <www.averill-law.com>.