

## BUILDING CREDIBLE INPUT MODELS

Lawrence M. Leemis

Department of Mathematics  
The College of William & Mary  
Williamsburg, VA 23187-8795, U.S.A.

### ABSTRACT

Most discrete-event simulation models have stochastic elements that mimic the probabilistic nature of the system under consideration. A close match between the input model and the true underlying probabilistic mechanism associated with the system is required for successful input modeling. The general question considered here is how to model an element (e.g., arrival process, service times) in a discrete-event simulation given a data set collected on the element of interest. For brevity, it is assumed that data is available on the aspect of the simulation of interest. It is also assumed that raw data is available, as opposed to censored data, grouped data, or summary statistics. This example-driven tutorial examines introductory techniques for input modeling. Most simulation texts (e.g., Law and Kelton 2000, Fishman 2001) have a broader treatment of input modeling than presented here. Nelson and Yamnitsky (1998) survey advanced techniques.

### 1 DATA COLLECTION

There are two approaches that arise with respect to the collection of data. The first is the classical approach, where a designed experiment is conducted to collect the data. The second is the exploratory approach, where questions are addressed by means of existing data that the modeler had no hand in collecting. The first approach is generally better in terms of control and the second approach is generally better in terms of cost.

Collecting data on the appropriate elements of the system of interest is one of the initial and pivotal steps in successful input modeling. An inexperienced modeler, for example, collects wait times on a single-server queue when waiting time is the measure of performance of interest. Although these wait times are valuable for model validation, they do not contribute to the input model. The appropriate data elements to collect for an input model for a single-server queue are typically arrival and service times. An

analysis of sample data collected on such a queue is given in Sections 3.1 and 3.2.

Even if the decision to sample the appropriate element is made correctly, Bratley, Fox, and Schrage (1987) warn that there are several things that can be “wrong” with a data set. Vending machine sales will be used to illustrate the difficulties.

- Wrong amount of aggregation. We desire to model daily sales, but have only monthly sales.
- Wrong distribution in time. We have sales for this month and want to model next month’s sales.
- Wrong distribution in space. We want to model sales at a vending machine in location A, but only have sales figures on a vending machine at location B.
- Censored data. We want to model *demand*, but we only have *sales* data. If the vending machine ever sold out, this constitutes a right-censored observation. The reliability and biostatistics literature contains techniques for accommodating censored data sets (Lawless 2003).
- Insufficient distribution resolution. We want the distribution of number the of soda cans sold at a particular vending machine, but our data is given in cases, effectively rounding the data up to the next multiple of 24.

### 2 INPUT MODELING TAXONOMY

Figure 1 contains a taxonomy illustrating the scope of potential input models available to simulation analysts. Modelers too often restrict their choice of input models to the top two branches. There is certainly no uniqueness in the branching structure chosen for the taxonomy. The branches under *stochastic processes*, for example, could have been *state* followed by *time*, rather than *time* followed by *state*, as presented.

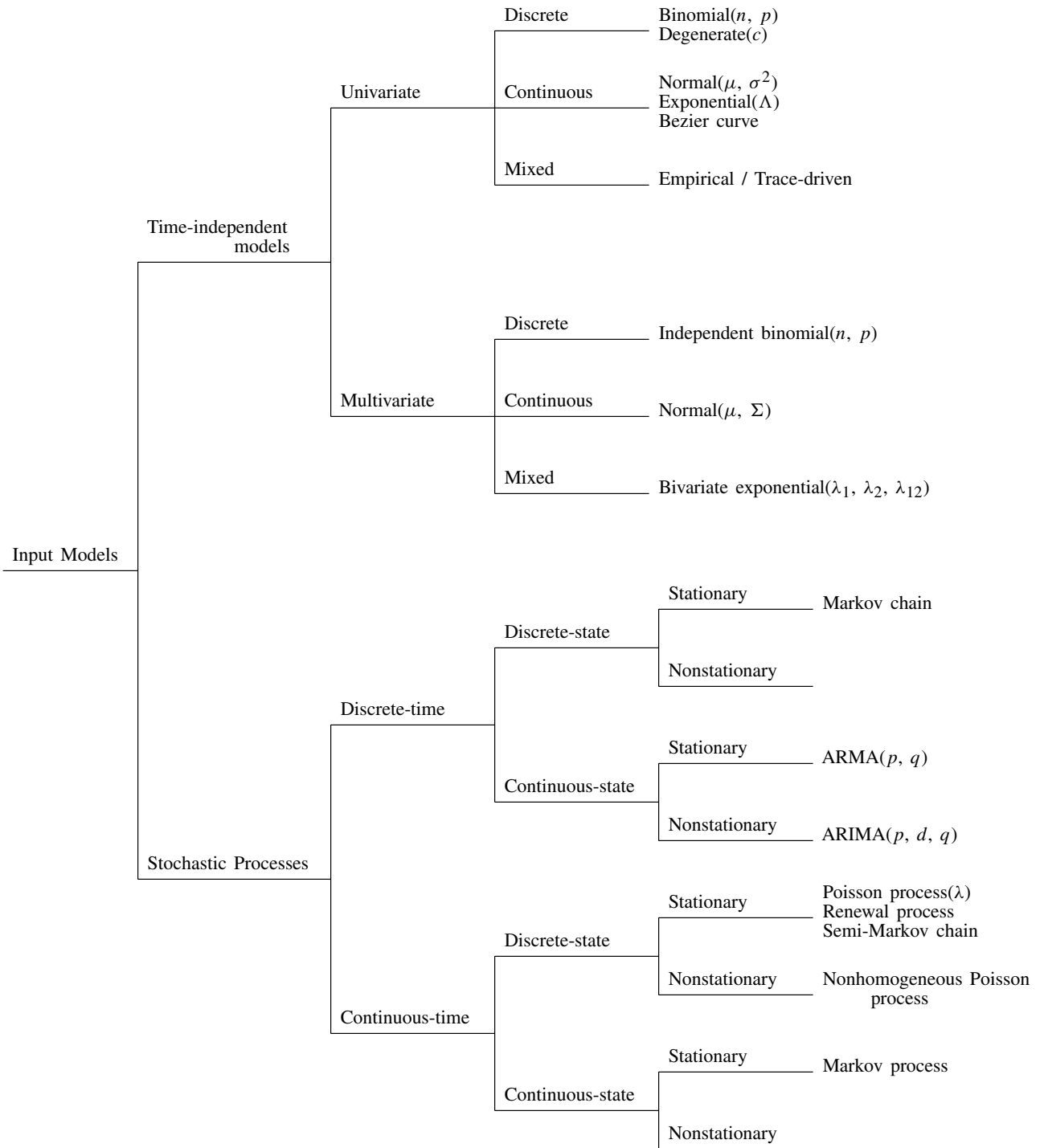


Figure 1: A Taxonomy for Input Models

Examples of specific models that could be placed on the branches of the taxonomy appear at the far right of the diagram. Mixed, univariate, time-independent input models have “empirical/trace-driven” given as a possible model. All of the branches include this particular model. A *trace-driven* input model simply generates a process that is identical to the collected data values so as not to rely on a parametric model. A simple example is a sequence of arrival times collected over a 24-hour time period. The trace-driven input model for the arrival process is generated by having arrivals occur at the same times as the observed values.

The upper half of the taxonomy contains models that are independent of time. These models could have been referred to as *Monte Carlo* models. Models are classified by whether there is one or several variables of interest, and whether the distribution of these random variables is discrete, continuous, or contains both continuous and discrete elements. Examples of univariate discrete models include the binomial distribution and a degenerate distribution with all of its mass at one value. Examples of continuous distributions include the normal distribution and an exponential distribution with a random parameter  $\Lambda$  (see, for example, Martz and Waller 1982). Bézier curves (Wagner and Wilson 1996) offer a unique combination of the parametric and nonparametric approaches. An initial distribution is fitted to the data set, then the modeler decides whether differences between the empirical and fitted models represent sampling variability or an aspect of the distribution that should be included in the input model.

Examples of  $k$ -variable multivariate input models (Johnson 1987, Wilson 1997) include a sequence of  $k$  independent binomial random variables, a multivariate normal distribution with mean  $\mu$  and variance-covariance matrix  $\Sigma$  and a bivariate exponential distribution (Barlow and Proschan 1981).

The lower half of the taxonomy contains stochastic process models. These models are often used to solve problems at the system level, in addition to serving as input models for simulations with stochastic elements. Models are classified by how time is measured (discrete/continuous), the state space (discrete/continuous) and whether the model is stationary in time. For Markov models, the discrete-state/continuous-state branch typically determines whether the model will be called a “chain” or a “process”, and the stationary/nonstationary branch typically determines whether the model will be preceded with the term “homogeneous” or “nonhomogeneous”. Examples of discrete-time stochastic processes include homogeneous, discrete-time Markov chains (Ross 2003) and ARIMA time series models (Box, Jenkins, and Reinsel 1994). Since point processes are counting processes, they have been placed on the continuous-time, discrete-space branch.

In conclusion, modelers are too often limited to univariate, stationary models since software is typically written for fitting distributions to these models. Successful input modeling requires knowledge of the full range of possible probabilistic input models.

### 3 EXAMPLES

Two simple examples illustrate the types of decisions that often arise in input modeling. The first example determines an input model for service times and the second example determines an input model for an arrival process.

#### 3.1 Service Time Model

Consider a data set of  $n = 23$  service times collected to determine an input model in a discrete-event simulation of a queuing system. The service times in seconds are

105.84	28.92	98.64	55.56	128.04	45.60
67.80	105.12	48.48	51.84	173.40	51.96
54.12	68.64	93.12	68.88	84.12	68.64
41.52	127.92	42.12	17.88	33.00	

[Although these service times come from the life testing literature (Caroni 2002; Lawless 2003), the same principles apply to both input modeling and survival analysis.]

The first step is to assess whether the observations are independent and identically distributed (iid). The data must be given in the order collected for independence to be assessed. Situations where the iid assumption would *not* be valid include:

- A new teller has been hired at a bank and the 23 service times represent a task that has a steep learning curve. The expected service time is likely to decrease as the new teller learns how to perform the task more efficiently.
- The service times represent 23 times to completion of a physically demanding task during an 8-hour shift. If fatigue is a significant factor, the expected time to complete the task is likely to increase with time.

If a simple linear regression of the observation numbers versus the service times shows a significant nonzero slope, then the iid assumption is probably not appropriate.

Assume that there is a suspicion that a learning curve is present, which makes a modeler suspect that the service times are decreasing. One appropriate hypothesis test is

$$H_0 : \beta_1 = 0$$

versus

$$H_1 : \beta_1 < 0$$

associated with the linear model (Kutner, Nachtsheim, Neter, Wasserman 2003)

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where  $X$  is the observation number,  $Y$  is the service time,  $\beta_0$  is the intercept,  $\beta_1$  is the slope, and  $\epsilon$  is an error term. Figure 2 shows a plot of the  $(x_i, y_i)$  pairs for  $i = 1, 2, \dots, 23$ , along with the estimated regression line. The  $p$ -value associated with the hypothesis test is 0.14, which is not enough evidence to conclude that there is a statistically significant learning curve present. The negative slope is likely due to sampling variability. The  $p$ -value may, however, be small enough to warrant further data collection.

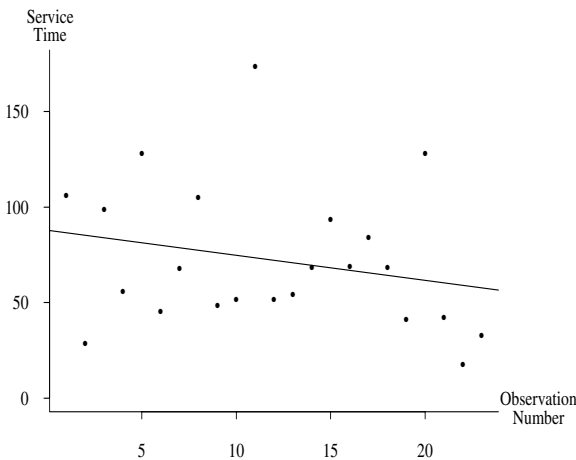


Figure 2: Service Time Vs. Observation Number

There are a number of other graphical and statistical methods for assessing independence. These include analysis of the sample autocorrelation function associated with the observations and a scatterplot of adjacent observations (Law and Kelton 2000). The sample autocorrelation function (ACF) for the service times is plotted in Figure 3 for the first ten lags. The sample ACF value at lag 1, for example, is the sample correlation for adjacent service times. The sample ACF value at lag 4, for example, is the sample correlation for service times four customers apart. The horizontal dotted lines at  $\pm 2/\sqrt{n}$  are 95% bounds used to determine whether the spikes in the ACF are statistically significant. None were statistically significant for the service time data. For this particular example, assume that we are satisfied that the observations are truly iid in order to perform a classical statistical analysis.

The next step in the analysis of this data set includes plotting a histogram and calculating the values of some sample statistics. A histogram of the observations is shown

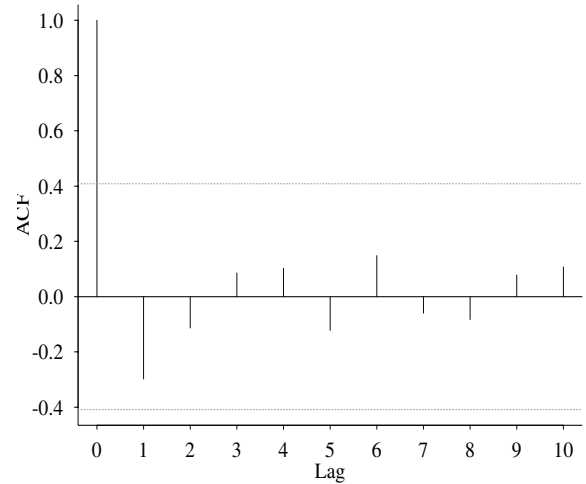


Figure 3: Sample Autocorrelation Function

in Figure 4. Although the data set is small, a skewed bell-shaped pattern is apparent. The largest observation lies in the far right-hand tail of the distribution, so care must be taken to assure that it is representative of the population. The sample mean, standard deviation, coefficient of variation, and skewness are

$$\bar{x} = 72.22 \quad s = 37.49 \quad \frac{s}{\bar{x}} = 0.52$$

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3 = 0.88.$$

Examples of the interpretations of these sample statistics are:

- A coefficient of variation  $s/\bar{x}$  close to 1, along with the appropriate histogram shape, indicates that the exponential distribution is a potential input model.
- A sample skewness close to 0 indicates that a symmetric distribution (e.g., a normal or uniform distribution) is a potential input model.

The next decision that needs to be made is whether a parametric or nonparametric input model should be used. One simple nonparametric model would repeatedly select one of the service times with probability  $1/23$ . A slightly more sophisticated nonparametric model is to use a piecewise linear cumulative distribution function estimate between the data values given by Law and Kelton (2000, p. 326):

$$\hat{F}(x) = \frac{i-1}{n-1} + \frac{x-x_{(i)}}{(n-1)(x_{(i+1)}-x_{(i)})}$$

for  $x_{(i)} \leq x < x_{(i+1)}$ ,  $i = 1, 2, \dots, n-1$ , where the sorted data values  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  are referred to as

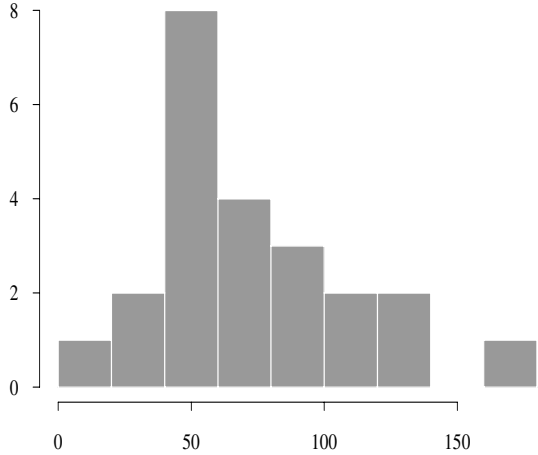


Figure 4: Histogram of Service Times

“order statistics”. In addition  $\hat{F}(x) = 0$  for  $x < x_{(1)}$ , and  $\hat{F}(x) = 1$  for  $x \geq x_{(n)}$ . This estimator is undefined for tied data values. Fortunately, this does not create problems with the associated variate generation algorithm. This estimator is still considered “nonparametric” because no parameters need to be estimated in order to arrive at the estimator.

The small size of the data set, the tied value, 68.64 seconds, and the observation in the far right-hand tail of the distribution, 173.40 seconds, tend to indicate that a parametric analysis is more appropriate. For this particular data set, a parametric approach is chosen.

There are dozens of choices for a univariate parametric model for the service times. These include general families of scalar distributions, modified scalar distributions and commonly-used parametric distributions (see, for example, Schmeiser 1990). Since the data is drawn from a continuous population and the support of the distribution is positive, a time-independent, univariate, continuous input model is chosen. The shape of the histogram indicates that the gamma, inverse Gaussian, log normal, and Weibull distributions (Lawless 2003) are good candidates. Derivation of the point and interval estimates for the Weibull distribution are given in detail here. Similar approaches apply to the other distributions.

Parameter estimates for the Weibull distribution can be found by least squares, the method of moments, and maximum likelihood. Due to desirable statistical properties, maximum likelihood is emphasized here. The Weibull distribution has probability density function

$$f(x) = \lambda^\kappa \kappa x^{\kappa-1} e^{-(\lambda x)^\kappa} \quad x \geq 0,$$

where  $\lambda$  is a positive scale parameter and  $\kappa$  is a positive shape parameter. Let  $x_1, x_2, \dots, x_n$  denote the data values.

The likelihood function is

$$L(\lambda, \kappa) = \prod_{i=1}^n f(x_i) = \lambda^{n\kappa} \kappa^n \left[ \prod_{i=1}^n x_i \right]^{\kappa-1} e^{-\sum_{i=1}^n (\lambda x_i)^\kappa}.$$

Since the natural logarithm (log) is a monotone function, the likelihood function and its logarithm achieve their maximum at the same values of  $\lambda$  and  $\kappa$ . The mathematics are typically more tractable for maximizing a log likelihood function, which, for the Weibull distribution, is

$$\log L(\lambda, \kappa) = n \log \kappa + \kappa n \log \lambda + (\kappa - 1) \sum_{i=1}^n \log x_i - \lambda^\kappa \sum_{i=1}^n x_i^\kappa.$$

The  $2 \times 1$  score vector has elements

$$\frac{\partial \log L(\lambda, \kappa)}{\partial \lambda} = \frac{\kappa n}{\lambda} - \kappa \lambda^{\kappa-1} \sum_{i=1}^n x_i^\kappa$$

and

$$\frac{\partial \log L(\lambda, \kappa)}{\partial \kappa} = \frac{n}{\kappa} + n \log \lambda + \sum_{i=1}^n \log x_i - \sum_{i=1}^n (\lambda x_i)^\kappa \log \lambda x_i.$$

When these equations are equated to zero, the simultaneous equations have no closed-form solution for the maximum likelihood estimators  $\hat{\lambda}$  and  $\hat{\kappa}$ :

$$\frac{\kappa n}{\lambda} - \kappa \lambda^{\kappa-1} \sum_{i=1}^n x_i^\kappa = 0$$

$$\frac{n}{\kappa} + n \log \lambda + \sum_{i=1}^n \log x_i - \sum_{i=1}^n (\lambda x_i)^\kappa \log \lambda x_i = 0.$$

To reduce the problem to a single unknown, the first equation can be solved for  $\lambda$  in terms of  $\kappa$  yielding

$$\lambda = \left( \frac{n}{\sum_{i=1}^n x_i^\kappa} \right)^{1/\kappa}.$$

Law and Kelton (2000, p. 305) give an initial estimate for  $\kappa$  and Qiao and Tsokos (1994) present a fixed-point algorithm for calculating the maximum likelihood estimators  $\hat{\lambda}$  and  $\hat{\kappa}$ . Their algorithm is guaranteed to converge for any positive initial estimate for  $\kappa$  for a complete data set.

The score vector has a mean of  $\mathbf{0}$  and a variance-covariance matrix  $I(\lambda, \kappa)$  given by the  $2 \times 2$  Fisher information matrix

$$I(\lambda, \kappa) = \begin{bmatrix} E \left[ \frac{-\partial^2 \log L(\lambda, \kappa)}{\partial \lambda^2} \right] & E \left[ \frac{-\partial^2 \log L(\lambda, \kappa)}{\partial \lambda \partial \kappa} \right] \\ E \left[ \frac{-\partial^2 \log L(\lambda, \kappa)}{\partial \kappa \partial \lambda} \right] & E \left[ \frac{-\partial^2 \log L(\lambda, \kappa)}{\partial \kappa^2} \right] \end{bmatrix}.$$

The observed information matrix

$$O(\hat{\lambda}, \hat{\kappa}) = \begin{bmatrix} \frac{-\partial^2 \log L(\hat{\lambda}, \hat{\kappa})}{\partial \lambda^2} & \frac{-\partial^2 \log L(\hat{\lambda}, \hat{\kappa})}{\partial \lambda \partial \kappa} \\ \frac{-\partial^2 \log L(\hat{\lambda}, \hat{\kappa})}{\partial \kappa \partial \lambda} & \frac{-\partial^2 \log L(\hat{\lambda}, \hat{\kappa})}{\partial \kappa^2} \end{bmatrix},$$

can be used to estimate  $I(\lambda, \kappa)$ .

For the 23 service times, the fitted Weibull distribution has maximum likelihood estimators  $\hat{\lambda} = 0.0122$  and  $\hat{\kappa} = 2.10$ . The log likelihood function evaluated at the maximum likelihood estimators is  $\log L(\hat{\lambda}, \hat{\kappa}) = -113.691$ . Figure 5 shows the empirical cumulative distribution function (a step function with a step of height  $1/23$  at each data point) along with the Weibull fit to the data.

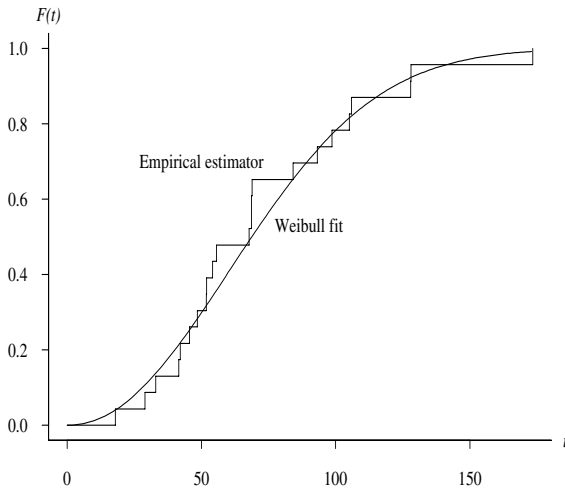


Figure 5: Empirical and Fitted Cumulative Distribution Functions for the Service Times

The observed information matrix is

$$O(\hat{\lambda}, \hat{\kappa}) = \begin{bmatrix} 681,000 & 875 \\ 875 & 10.4 \end{bmatrix},$$

revealing a positive correlation between the elements of the score vector. We now consider interval estimators for  $\lambda$  and  $\kappa$ . Using the fact that the likelihood ratio statistic,  $2[\log L(\hat{\lambda}, \hat{\kappa}) - \log L(\lambda, \kappa)]$ , is asymptotically  $\chi^2$  distributed in  $n$  with 2 degrees of freedom and that  $\chi_{2,0.05}^2 = 5.99$ , a 95% confidence region for the parameters is all  $\lambda$  and  $\kappa$  satisfying

$$2[-113.691 - \log L(\lambda, \kappa)] < 5.99.$$

The maximum likelihood estimators and 95% confidence region are shown in Figure 6. The line  $\kappa = 1$  is not interior to the region, indicating that the exponential distribution is not an appropriate model for this particular data set.

As further proof that  $\kappa$  differs significantly from 1, the standard errors of the distribution of the parameter

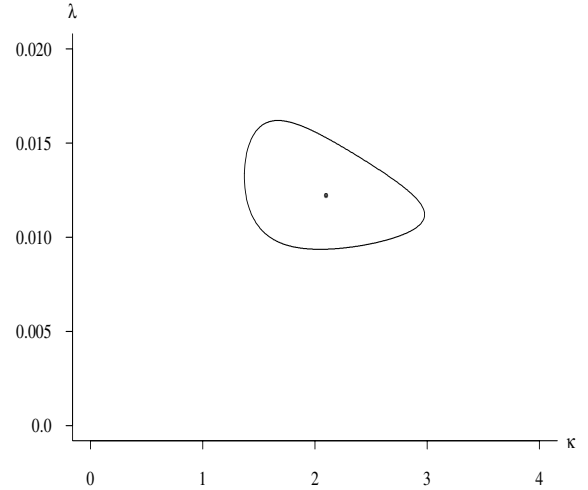


Figure 6: 95% Confidence Region Based on the Likelihood Ratio Statistic

estimators can be computed by using the inverse of the observed information matrix

$$O^{-1}(\hat{\lambda}, \hat{\kappa}) = \begin{bmatrix} 0.00000165 & -0.000139 \\ -0.000139 & 0.108 \end{bmatrix}.$$

This is the asymptotic variance-covariance matrix for the parameter estimators  $\hat{\lambda}$  and  $\hat{\kappa}$ . The standard errors of the parameter estimators are the square roots of the diagonal elements

$$\hat{\sigma}_{\hat{\lambda}} = 0.00128 \quad \hat{\sigma}_{\hat{\kappa}} = 0.329.$$

Thus an asymptotic 95% confidence interval for  $\kappa$  is

$$2.10 - (1.96)(0.329) < \kappa < 2.10 + (1.96)(0.329)$$

or

$$1.46 < \kappa < 2.74,$$

since  $z_{0.025} = 1.96$ . Since this confidence interval does not contain 1, the inclusion of the Weibull shape parameter  $\kappa$  is justified.

The model adequacy should now be assessed. Since the chi-square goodness-of-fit test has arbitrary interval limits, it should not be applied to small data sets (e.g.,  $n = 23$ ), such as the service times being considered here. The Kolmogorov–Smirnov, Cramer–von Mises, or Anderson–Darling goodness-of-fit tests (Lawless 2003) are appropriate here. The Kolmogorov–Smirnov test statistic, which is the maximum vertical difference between the empirical and fitted cumulative distribution functions, is 0.151 for this data set with a Weibull fit. This test statistic corresponds to a  $p$ -value of approximately 0.15 (Law and Kelton 2000, p. 366), so the Weibull distribution provides a reasonable

model for these service times. The Kolmogorov–Smirnov test statistic values for several models are shown in Table 1, including four that are superior to the Weibull with respect to fit.

Table 1: Kolmogorov–Smirnov Test Statistics for Models Fitted to Service Time Data

Model	Test statistic
Exponential	0.307
Weibull	0.151
Gamma	0.123
Arctangent	0.094
Log normal	0.090
Inverse Gaussian	0.088

Many of the discrete-event simulation packages exhibited at the *Winter Simulation Conference* have the capability of determining maximum likelihood estimators for several popular parametric distributions. Most packages also perform goodness-of-fit tests, such as the Kolmogorov–Smirnov or chi-square test, so that the distribution that best fits the data set can quickly be determined.

P–P (probability–probability) and Q–Q (quantile–quantile) plots can also be used to assess model adequacy. A P–P plot, for example, is a plot of the fitted cumulative distribution function at the  $i$ th order statistic  $x_{(i)}$ ,  $\hat{F}(x_{(i)})$ , versus the adjusted empirical cumulative distribution function,  $\tilde{F}(x_{(i)}) = \frac{i-0.5}{n}$ , for  $i = 1, 2, \dots, n$ . A plot where the points fall close to the line passing through the origin and (1, 1) indicates a good fit. For the 23 service times, a P–P plot for the Weibull fit is shown in Figure 7, along with a line connecting (0, 0) and (1, 1). P–P plots should be constructed for all competing models.

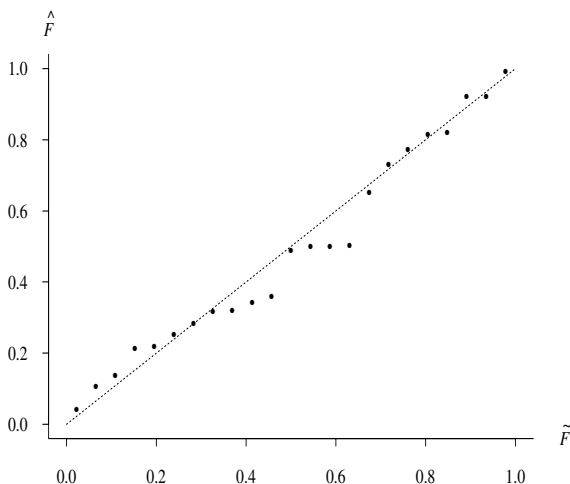


Figure 7: A P–P Plot for the Service Times Using the Weibull Model

### 3.2 Arrival Time Model

Accurate input modeling requires a careful evaluation of whether a stationary (no time dependence) or nonstationary model is appropriate. Modeling arrivals to a lunch wagon is used to illustrate the decision-making process.

Customer arrival times to a lunch wagon between 10:00 AM and 2:30 PM are collected on three days. The realizations were generated from a hypothetical arrival process given by Klein and Roberts (1984). A total of  $n = 150$  arrival times were observed, including  $n_1 = 56$ ,  $n_2 = 42$  and  $n_3 = 52$  on the  $k = 3$  days. Defining  $(0, 4.5]$  to be the time interval of interest (in hours) the three realizations are

0.2152 0.3494 0.3943 ... 4.175 4.248,

0.3927 0.6211 0.7504 ... 4.044 4.374,

and

0.4499 0.5495 0.6921 ... 3.643 4.357.

One preliminary statistical issue concerning this data is whether the three days represent processes drawn from the same population. External factors such as the weather, day of the week, advertisement, and workload should be fixed. For this particular example, we assume that these factors have been fixed and the three processes are representative of the population of arrival processes to the lunch wagon.

The input model for the process comes from the lower branch (stochastic processes) of the taxonomy in Figure 1. Furthermore, the arrival times constitute realizations of a continuous-time, discrete-state stochastic process, so the remaining question concerns whether or not the process is stationary.

If the process proves to be stationary, the techniques from the previous example, such as drawing a histogram, and choosing a parametric or nonparametric model for the *interarrival* times, are appropriate. This results in a Poisson or renewal process model. On the other hand, if the process is nonstationary, a nonhomogeneous Poisson process might be an appropriate input model. A nonhomogeneous Poisson process is governed by an intensity function  $\lambda(t)$  which gives an arrival rate [e.g.,  $\lambda(2) = 10$  means that the arrival rate is 10 customers per hour at time 2] that can vary with time. The next paragraph describes a nonparametric procedure for estimating the cumulative intensity function  $\Lambda(t) = \int_0^t \lambda(\tau) d\tau$  from  $k$  realizations.

The cumulative intensity function is to be estimated on  $(0, S]$ , where  $S$  is a known constant which equals 4.5 in this case. The interval  $(0, S]$  may represent the time a system allows arrivals (e.g., 9 AM to 5 PM at a bank) or one period of a cycle (e.g., one day at an emergency room). Let  $n_i, i = 1, 2, \dots, k$  be the number of observations in the  $i$ th realization,  $n = \sum_{i=1}^k n_i$ , and let  $t_{(1)}, t_{(2)}, \dots, t_{(n)}$  be

the order statistics of the superposition of the  $k$  realizations,  $t_{(0)} = 0$  and  $t_{(n+1)} = S$ . The piecewise-linear estimator of the cumulative intensity function between the time values in the superposition is

$$\hat{\Lambda}(t) = \frac{in}{(n+1)k} + \left[ \frac{n(t-t_{(i)})}{(n+1)k(t_{(i+1)}-t_{(i)})} \right]$$

for  $t_{(i)} < t \leq t_{(i+1)}$ ;  $i = 0, 1, 2, \dots, n$ , which is given in Leemis (1991) and extended to nonoverlapping intervals in Arkin and Leemis (2000). Asymptotic confidence intervals and variate generation via inversion are also contained in these references. This estimator (solid line), along with 95% confidence bounds (dashed lines), for the customer arrival times to the lunch wagon are given in Figure 8. The cumulative intensity function estimator at time  $4.5 = 150/3 = 50$ ,

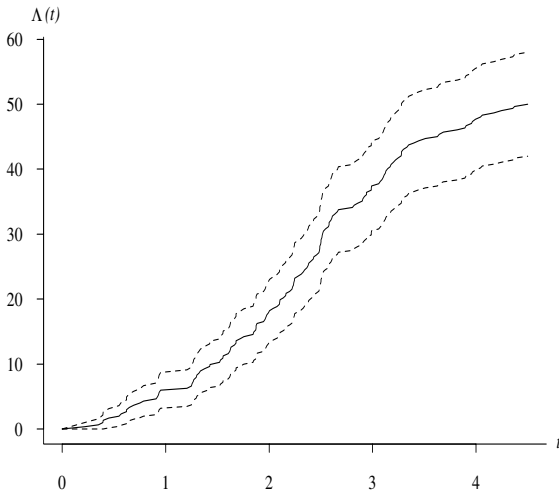


Figure 8: Point and 95% Confidence Interval Estimators for the Cumulative Intensity Function

the point estimator for the expected number of arriving customers per day. If  $\hat{\Lambda}(t)$  is linear, a stationary model is appropriate. Since customers are more likely to arrive to the lunch wagon between 12:00 ( $t = 2$ ) and 1:00 ( $t = 3$ ) than at other times and the cumulative intensity function estimator has an S-shape, a nonstationary model is indicated. More specifically, a nonhomogeneous Poisson process is a reasonable model for the arrival process.

The next question to be answered is whether a parametric or nonparametric model should be chosen for the process. Figure 8 indicates that the intensity function increases initially, remains fairly constant during the noon hour, then decreases. This may be difficult to model parametrically, so a nonparametric approach, possibly using  $\hat{\Lambda}(t)$  in Figure 8 might be appropriate. Process generation for simulation is straightforward (Leemis 1991).

There are many potential parametric models for non-stationary arrival processes. The next paragraph describes the procedure for fitting a *power law process*, where the intensity function has the same parametric form as the hazard function for the Weibull distribution. Other models can be fit in a similar fashion.

The likelihood function for estimating the vector of unknown parameters  $\theta = (\theta_1, \theta_2, \dots, \theta_p)'$  from a single realization on  $(0, S]$  is

$$L(\theta) = \left[ \prod_{i=1}^n \lambda(t_i) \right] \exp \left[ - \int_0^S \lambda(t) dt \right].$$

Maximum likelihood estimators can be determined by maximizing  $L(\theta)$  or its logarithm with respect to all unknown parameters. Confidence intervals for the unknown parameters can be found in a similar manner to the service time example. Owing to the additive property of the intensity function for multiple realizations, the likelihood function for the case of  $k$  realizations is

$$L(\theta) = \left[ \prod_{i=1}^n k \lambda(t_i) \right] \exp \left[ - \int_0^S k \lambda(t) dt \right],$$

where  $t_1, t_2, \dots, t_n$  are the arrival times that occurred during the  $k$  realizations. The power law process has intensity function

$$\lambda(t) = \lambda^\kappa \kappa t^{\kappa-1} \quad t > 0,$$

for  $\lambda > 0$  and  $\kappa > 0$ . Thus the likelihood function for  $k$  realizations is

$$L(\lambda, \kappa) = k^n \lambda^{n\kappa} \kappa^n e^{-k(\lambda S)^\kappa} \prod_{i=1}^n t_i^{\kappa-1}.$$

The log likelihood function is

$$\log L(\lambda, \kappa) = n \log(k\kappa) + n\kappa \log \lambda - k(\lambda S)^\kappa + (\kappa - 1) \sum_{i=1}^n \log t_i.$$

The  $2 \times 1$  score vector has elements

$$\frac{\partial \log L(\lambda, \kappa)}{\partial \lambda} = \frac{\kappa n}{\lambda} - k S^\kappa \kappa \lambda^{\kappa-1}$$

and

$$\frac{\partial \log L(\lambda, \kappa)}{\partial \kappa} = n \log \lambda + \frac{n}{\kappa} + \sum_{i=1}^n \log t_i - k(\lambda S)^\kappa \log \lambda S.$$

When the score is equated to zero, the analytic expressions for  $\lambda$  and  $\kappa$  are

$$\hat{\kappa} = \frac{n}{n \log S - \sum_{i=1}^n \log t_i} \quad \hat{\lambda} = \frac{1}{S} \left( \frac{n}{\hat{\kappa}} \right)^{1/\hat{\kappa}}.$$

Substituting the arrival times into these formulas yields maximum likelihood estimators  $\hat{\lambda} = 4.86$  and  $\hat{\kappa} = 1.27$ .



The cumulative intensity function for the power law process

$$\Lambda(t) = (\lambda t)^\kappa \quad t > 0,$$

is plotted along with the nonparametric estimator in Figure 9. Note that due to the peak in customer arrivals around the noon hour, the power law process is not an appropriate model since it is not able to adequately approximate the intensity function.

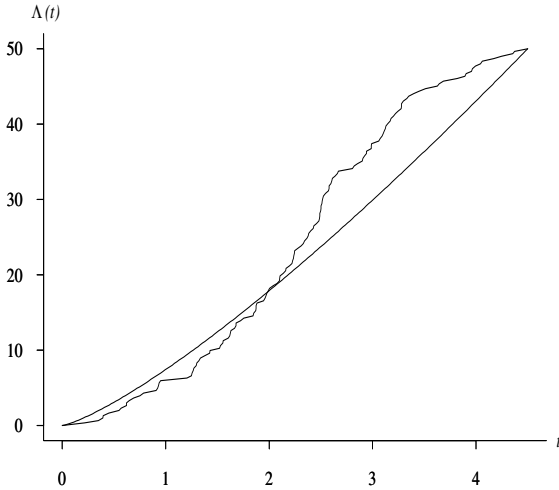


Figure 9: Empirical and Fitted Power Law Estimators for the Cumulative Intensity Function

Since the intensity function is analogous to the hazard function for time-independent models, an appropriate 2-parameter distribution to consider would be one with a hazard function that increases initially, then decreases. A log-logistic process, for example, with intensity function (Lawless 2003)

$$\lambda(t) = \frac{\lambda\kappa(\lambda t)^{\kappa-1}}{1 + (\lambda t)^\kappa} \quad t > 0,$$

for  $\lambda > 0$  and  $\kappa > 0$ , would certainly be more appropriate. More generally, the EPTMP (exponential-polynomial-trigonometric function with multiple periodicities) model, originally given by Lee, Wilson and Crawford (1991) and generalized by Kuhl, Damerджи and Wilson (1998) with intensity function

$$\lambda(t) = \exp \left[ \sum_{i=0}^m \alpha_i t^i + \sum_{k=1}^p \gamma_k \sin(\omega_k t + \phi_k) \right] \quad t > 0$$

can model a nonmonotonic intensity function. Goodness-of-fit tests are given in Rigdon and Basu (2000).

#### 4 DISCRETE-EVENT SIMULATION MODELING FRAMEWORK

This section contains a description of a diagram that has been developed for describing the process of constructing a discrete-event simulation model. The purpose of providing the description of the diagram here is to: (i) show where input modeling fits into the simulation modeling process, and (ii) isolate various sources of error involved in simulation modeling. The diagram depicting a high-level, abstract framework of a discrete-event simulation modeling process for analyzing an existing or proposed system (labeled “System” in the diagram) given in Figure 10 is adapted from Schmeiser (2001) and Nelson (1987).

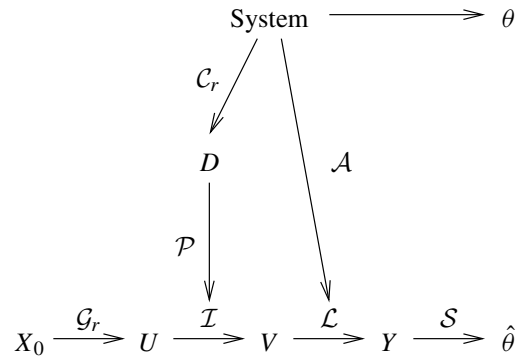


Figure 10: A Framework for Discrete-Event Simulation

The upper-case letters  $X_0$ ,  $U$ ,  $V$ ,  $Y$ ,  $\hat{\theta}$ ,  $\theta$ , and  $D$  denote ordered sets containing one or more numbers. To avoid writing “one or more numbers” in our descriptions of these sets, we assume that there are multiple numbers in the sets. The descriptions of these ordered sets follows.

- $X_0$  is a set of seeds for a random number generator, one for each stream used in the implementation of the discrete-event simulation model.
- $U$  is a set of random numbers created by using the random number generator  $\mathcal{G}_r$  to transform the seeds in the set  $X_0$  to random numbers. The random numbers in  $U$  are partitioned by the associated stream when multiple streams are employed.
- $V$  is a set of input data (“variates”) created by applying the input model  $\mathcal{I}$  to the set of random numbers  $U$ .
- $Y$  is a set of output data generated by applying the logic model  $\mathcal{L}$  to the set of input data  $V$ . The output data are typically dependent, although the probability model for each individual observation is often identical for a steady-state analysis once the simulation model warms up.
- $\hat{\theta}$  is a set of point estimators for the unknown system measures of performance  $\theta$ , calculated as

a function of the output data  $Y$ . In general, there is some error present, i.e.,  $\hat{\theta} \neq \theta$ .

- $\theta$  is the corresponding set of measures of performance associated with the system of interest.
- $D$  is a set of system data values collected on appropriate elements of the system of interest in order to build an input model  $\mathcal{I}$ .

Although Figure 10 conceptually lumps the thousands or millions of random numbers into a set  $U$ , the next-event approach to simulation allows us to generate them one at a time in order to save memory and CPU time.

The calligraphic letters  $\mathcal{G}_r$ ,  $\mathcal{I}$ ,  $\mathcal{L}$ ,  $\mathcal{S}$ ,  $\mathcal{C}_r$ ,  $\mathcal{P}$  and  $\mathcal{A}$  in Figure 10 are all associated with arrows. These are the seven sources of error associated with the discrete-event simulation modeling process. These letters denote transformations, probability models, data collection methods, assumptions, etc., as described below.

- $\mathcal{G}_r$  is a random number generator used to transform the seeds in the set  $X_0$  to random numbers in the set  $U$ .
- $\mathcal{I}$  is the input model used to transform the set of random numbers  $U$  to the set of input data  $V$ . The process of transforming  $U$  to  $V$  is known as *random variate generation*. The input model is often determined by analyzing a set of data  $D$ , although in rare cases an input model is determined in the absence of data using expert opinion, bypassing the set  $D$  entirely.
- $\mathcal{L}$  is the logic model that captures assumptions made about the system into transformations (often formulated as algorithms) that are used to transform the set of input data  $V$  to the set of output data  $Y$ .
- $\mathcal{S}$  is a statistical estimation procedure. The  $\mathcal{S}$  connecting the set of output data  $Y$  and the set of point estimates of the measures of performance  $\hat{\theta}$  involves computing statistics, which are functions of the set of output data  $Y$  (e.g., sample mean, sample median, or sample variance). Confidence intervals for measures of performance are often-times computed to give a sense of the accuracy of the point estimates.
- $\mathcal{C}_r$  denotes the data collection procedures from the system of interest. It is crucial to collect the appropriate data elements from the system. Also, the data should be collected in an appropriate and representative fashion using standard sampling techniques.
- $\mathcal{P}$  involves the process of formulating a probabilistic input model that adequately describes the set of data collected in  $D$ . The  $\mathcal{P}$  connecting the set of system data values  $D$  and the input model  $\mathcal{I}$  involves either resampling the data (i.e., the

trace-driven or nonparametric approach) or fitting a parametric model to the data set. The process of formulating  $\mathcal{I}$  is the focus of this tutorial.

- $\mathcal{A}$  denotes assumptions made on the system of interest. These assumptions are used to create the logic model  $\mathcal{L}$  describing the operation of the system. Incorrect or simplifying assumptions lead to modeling error.

What part of Figure 10 describes the discrete-event simulation model? The simulation model consists of the combination of the probabilistic input model  $\mathcal{I}$  and the logical model  $\mathcal{L}$ . Once the simulation model,  $\mathcal{I}$  and  $\mathcal{L}$ , has been determined, the sequence of four arrows leading from  $X_0$  to  $\hat{\theta}$  is a sequence of four deterministic functions for a particular random number generator  $\mathcal{G}_r$  and choice of sample statistics collected  $\mathcal{S}$ . All that is needed to arrive at  $\hat{\theta}$  are the random number seeds in the set  $X_0$ .

Error can occur in any of the arrows labeled by a calligraphic letter. There is no letter on the arrow attaching the system of interest to the measures of performance  $\theta$  because there is no error associated with this transition. The values of the measures of performance are unknown, which typically necessitates the use of a discrete-event simulation analysis for a complex system. If the model could be simulated for an infinite length of time and an infinitely large data set could be collected on the system of interest, then the error between  $\theta$  and  $\hat{\theta}$  would be a constant value induced only by “logic-modeling error”. “Sampling error”, on the other hand, stems from the random sampling variability inherent in  $\mathcal{G}_r$  and  $\mathcal{C}_r$ . Thus the mean square error:

$$\begin{aligned} E[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2] \\ &= E[\hat{\theta}^2] - E[2\hat{\theta}\theta] + E[\theta^2] \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\ &= V[\hat{\theta}] + (E[\hat{\theta}] - \theta)^2, \end{aligned}$$

captures the sampling error in the first term  $V[\hat{\theta}]$  and the modeling error in the second term  $(E[\hat{\theta}] - \theta)^2$ . The mean square error can only be computed on simple “toy” systems where the values in  $\theta$  are known.

The discussion here assumes an ideal system that does not change with time. Most real-world systems are changing with time, however, so an infinite sample drawn from the system is about how the system performed in the past, not how it will perform in the future.

The  $r$  subscript denotes a step in the discrete-event modeling process where error from random sampling variability is present. Both the random number generator  $\mathcal{G}_r$  and the data collection procedures  $\mathcal{C}_r$  involve random sampling variability. An “unlucky” single random number seed on a good generator  $\mathcal{G}_r$  could, for example, produce a sequence

of unusually small random numbers  $U$  whose average is significantly less than  $1/2$ . Likewise, an “unlucky” random sample on a legitimate data collection procedure  $C_r$  could, for example, produce a sequence of unusually large data values in  $D$ . The error induced by random sampling variability can be minimized by making numerous long simulation replications (in the case of  $\mathcal{G}_r$ ) and by collecting large system data sets (in the case of  $C_r$ ). Almost universally, the former is cheaper than the latter.

The other sources of error are associated with the calligraphic letters in the diagram are:

- using a poor random number generator  $\mathcal{G}_r$ ,
- making poor modeling decisions in  $\mathcal{P}$  resulting in a poor probabilistic input model  $\mathcal{I}$ ,
- using incorrect system data sampling procedures  $C_r$ ,
- making incorrect or simplifying assumptions about the system in  $\mathcal{A}$  resulting in a poor logic model  $\mathcal{L}$ ,
- making poor choices in  $\mathcal{S}$  when analyzing the set of output data  $Y$ .

Why do we simulate? An “analytic” model is appropriate when mathematics can be used to find the exact values of the measures of performance in  $\theta$ . For many real-world systems, however, the transformation from  $U \rightarrow V \rightarrow Y \rightarrow \hat{\theta}$  is so mathematically complex that the axiomatic approach to probability results in mathematically intractable expressions for the elements in the set  $\theta$ . Equivalently, the numbers in the set  $Y$  are drawn from an unknown or mathematically intractable probability model.

## ACKNOWLEDGMENTS

The author thanks Steve Tretheway for his help in developing Figure 1, Bruce Schmeiser and Barry Nelson for their help in developing Figure 10, and Diane Evans & Sigrún Andradóttir for reading a draft of this tutorial.

## REFERENCES

Arkin, B. L., and L. M. Leemis. 2000. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process from overlapping realizations. *Management Science* 46:989–998.

Barlow, R. E., and F. Proschan. 1981. *Statistical theory of reliability and life testing: Probability models*. Silver Springs, Maryland: To begin with.

Box, G., G. M. Jenkins, and G. Reinsel. 1994. *Time series analysis: Forecasting and control*. 3rd ed. Englewood Cliffs, New Jersey: Prentice–Hall.

Bratley, P., B. L. Fox, and L. E. Schrage. 1987. *A guide to simulation*. 2nd ed. New York: Springer–Verlag.

Caroni, C. 2002. The correct “ball bearings” data. *Lifetime Data Analysis* 8:395–399.

Fishman, G. 2001. *Discrete-event simulation: Modeling, programming, and analysis*. New York: Springer–Verlag.

Johnson, M. E. 1987. *Multivariate statistical simulation*. New York: John Wiley & Sons.

Klein, R. W., and S. D. Roberts. 1984. A time-varying Poisson arrival process generator. *Simulation* 43:193–195.

Kuhl, M. E., H. Damerджи, and J. R. Wilson. 1998. Least squares estimation of nonhomogeneous Poisson processes. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 637–645. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Kutner, M. H., C. J. Nachtsheim, J. Neter, W. Wasserman. 2003. *Applied linear regression models*. 4th ed. New York: McGraw–Hill.

Law, A. M., and W. D. Kelton. 2000. *Simulation modeling and analysis*. 3rd ed. New York: McGraw–Hill.

Lawless, J. F. 2003. *Statistical models & methods for lifetime data*. 2nd ed. New York: John Wiley & Sons.

Lee, S., J. R. Wilson, and M. M. Crawford. 1991. Modeling and simulation of a nonhomogeneous Poisson process having cyclic behavior. *Communications in Statistics — Simulation and Computation* 20:777–809.

Leemis, L. M. 1991. Nonparametric estimation of the cumulative intensity function for a nonhomogeneous Poisson process. *Management Science* 37:886–900.

Martz, H. F., and R. A. Waller. 1982. *Bayesian reliability analysis*. New York: John Wiley & Sons.

Nelson, B. L. 1987. A perspective on variance reduction in dynamic simulation experiments,” *Communications in Statistics* B16: 385–426.

Nelson, B. L., and M. Yamnitsky. 1998. Input modeling tools for complex problems. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 105–112. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Qiao, H., and C. P. Tsokos. 1994. Parameter estimation of the Weibull probability distribution. *Mathematics and Computers in Simulation* 37:47–55.

Rigdon, S. E., and A. P. Basu. 2000. *Statistical methods for the reliability of repairable systems*. New York: John Wiley & Sons.

Ross, S. M. 2003. *Introduction to probability models*. 8th ed. Boston: Academic Press.

Schmeiser, B. 1990. Simulation experiments. In *Handbooks in OR & MS*, ed. D. P. Heyman and M. J. Sobel, 296–330. New York: Elsevier Science Publishers.

- Schmeiser, B.W. 2001. Some myths and common errors in simulation experiments. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B.A. Peters, J.S. Smith, D.J. Medeiros, M.W. Rohrer, 39–46. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.
- Wagner, M. A. F., and J. R. Wilson. 1996. Using univariate Bézier distributions to model simulation input processes. *IIE Transactions* 28:699–711.
- Wilson, J. R. 1997. Modeling dependencies in stochastic simulation inputs. In *Proceedings of the 1997 Winter Simulation Conference*, ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, 47–52. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

## AUTHOR BIOGRAPHY

**LAWRENCE M. LEEMIS** is a professor in the Mathematics Department at the College of William & Mary. He received his BS and MS degrees in Mathematics and his Ph.D. in Industrial Engineering from Purdue University. He has also taught at Baylor University, The University of Oklahoma, and Purdue University. His consulting, short course, and research contract work includes contracts with AT&T, NASA/Langley Research Center, Delco Electronics, Department of Defense (Army, Navy), Air Logistic Command, ICASE, Komag, Federal Aviation Administration, Tinker Air Force Base, Woodmizer, Magnetic Peripherals, and Argonne National Laboratory. His research and teaching interests are in reliability and simulation. He is a member of ASA, IIE, and INFORMS. His email and web addresses are <leemis@math.wm.edu> and <www.math.wm.edu/~leemis>.