Proceedings of the 1994 Winter Simulation Conference
ed. J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila

INPUT MODELING

Lawrence M. Leemis

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

ABSTRACT

General guidelines for selecting probabilistic input models as part of a discrete-event simulation study are presented. Two short examples illustrating input modeling decisions are also presented, as opposed to a complete treatment of the subject.

1 INTRODUCTION

Discrete-event simulation models typically have stochastic components that mimic the probabilistic nature of the system under consideration. Successful input modeling requires a close match between the input model and the true underlying probabilistic mechanism associated with the system. 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.

Since time and space for this tutorial is limited, the following simplifying assumptions have been made.

- A reliable source of random numbers exists. Most introductory simulation textbooks (e.g., Law and Kelton 1991) consider random number generation algorithms.

- An algorithm is available for converting these random numbers to random variates associated with the input model to drive the simulation (Devroye 1986).

- Data is available on the aspect of the simulation of interest. For examples of input modeling in the absence of data, see Schmeiser and Deutsch (1977) or Law, McComas, and Vincent (1994).

With these assumptions limiting the scope of this tutorial, the focus turns to selecting the appropriate probabilistic models for the random components in a simulation model. Many simulation textbooks have a much broader treatment of input modeling than presented here (e.g., Law and Kelton 1991). These texts include more specific information on statistical tests for independence, graphical methods for model selection, parameter estimation techniques, and goodness-of-fit tests.

An input model can be specified in a variety of ways, such as a cumulative distribution function, hazard function, intensity function or a variate-generation algorithm. An input model characterizes each of the stochastic elements of a discrete-event simulation.

Figure 1 contains a taxonomy whose purpose is to illustrate the scope of potential input models that are available to simulation analysts. There is certainly no uniqueness in the branching structure of the taxonomy. The branches under *stochastic processes*, for example, could have been *state* followed by *time*, rather than *time* followed by *state*, as presented.

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 an 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 without relying 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 called *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 distribu-

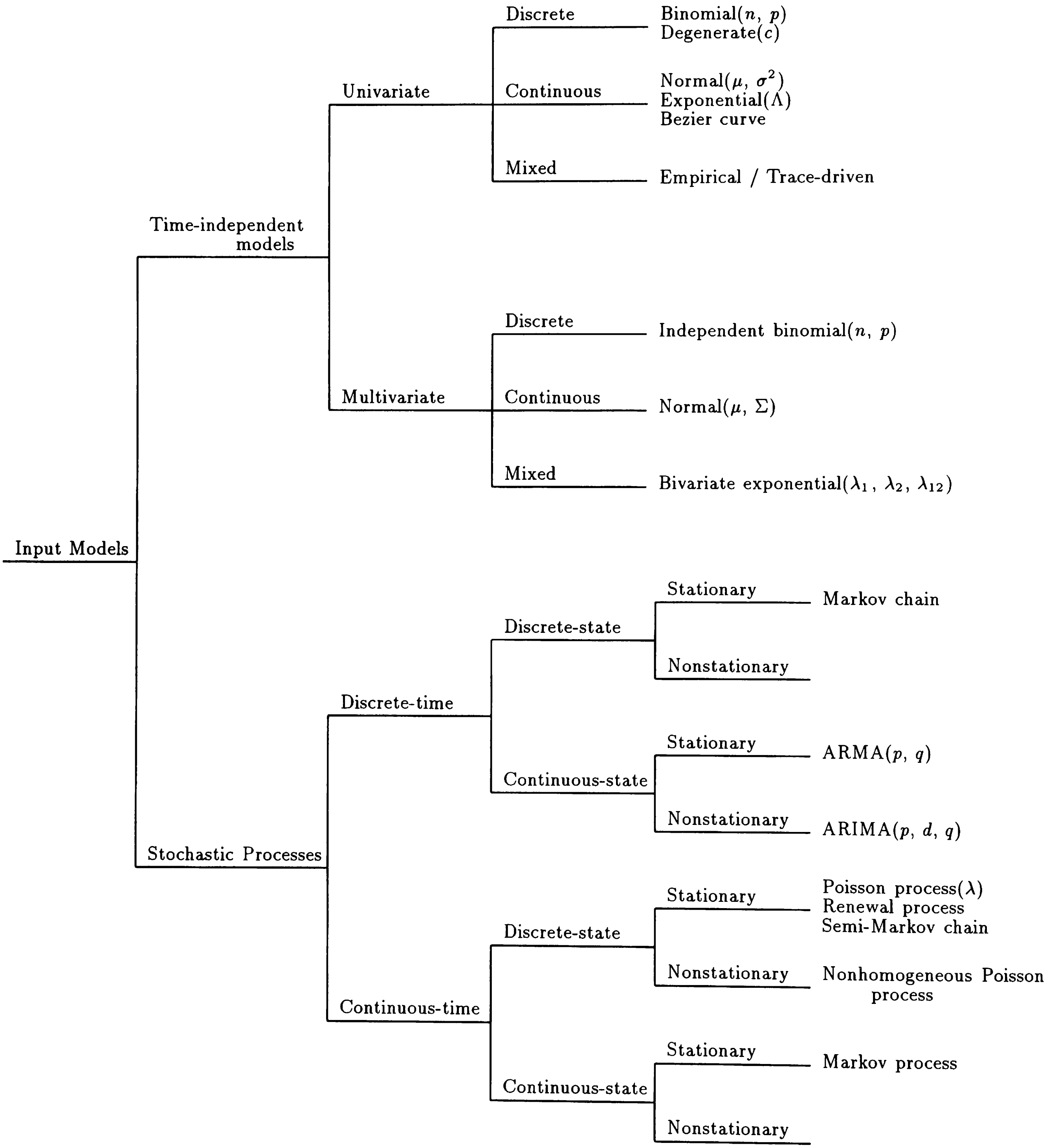

Figure 1: A Taxonomy for Input Models

tion and a degenerate distribution with all of its mass at one value. Examples of continuous distributions include the normal distribution, an exponential distribution with a random parameter Γ (see, for example, Martz and Waller 1982) and Bézier curves (Flanigan-Wagner and Wilson 1993). Bézier curves 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 (chance variation) or an aspect of the distribution that should be included in the input model.

Examples of k-variable multivariate input models (see Johnson 1987) include a sequence of k independent binomial random variables, a multivariate normal distribution with mean μ and variance-covariance matrix Σ 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 1993) and ARIMA time series models (Box and Jenkins 1976). Since point processes are counting processes, they have been placed on the continuous-time, discrete-space branch. Although the Poisson, renewal and nonhomogeneous Poisson processes are all pure birth processes, more general point processes, such as one to model the number of customers in a queue, can be placed on one of the continuous time, discrete-space branches.

2 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.

2.1 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 ordered service times in seconds are

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

[Although these service times come from the life testing literature (Lieblein and Zelen 1956), 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 completion times 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 number regressed on the service times shows a significant nonzero slope, the the iid assumption is probably not appropriate. 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. 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 to 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 in Figure 2. 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

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

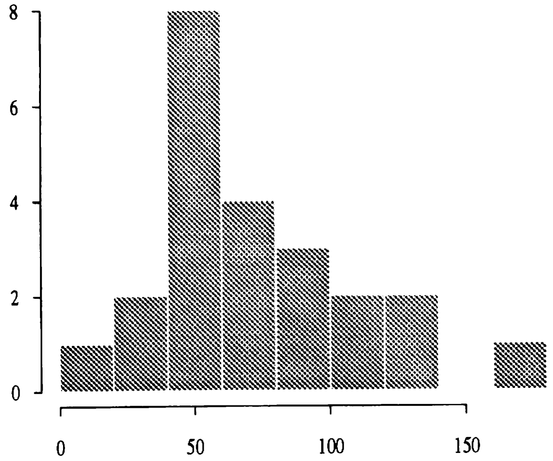

Figure 2: Histogram of Service Times

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

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 is an appropriate 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. The small size of the data set, the tied value, 68.64, and the observation in the far right-hand tail of the distribution, 173.40, tend to indicate that a parametric analysis is more appropriate. Since the input model is for service times, the accurate modeling of the right-hand tail of the distribution is critical. These long service times significantly impact queuing statistics. 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 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 Weibull, gamma, log normal, and log logistic distributions (Lawless 1982) are good candidates. The Weibull distribution is analyzed 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

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

where λ is a positive scale parameter and κ is a positive shape parameter. Let $x_1, x_2, \ldots, x_n$ be the failure times. The likelihood function is

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

The 2×1 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 $\hat{\lambda}$ 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 λ in terms of κ yielding

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

Law and Kelton (1991, p. 334) give an initial estimate for κ that can be used in Newton's method to numerically solve for the maximum likelihood estimators. The score vector has a mean of 0 and a variance-covariance matrix $I(\lambda,\kappa)$ given by the 2×2 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 3 shows the empirical cumulative distribution function along with the Weibull fit to the data.

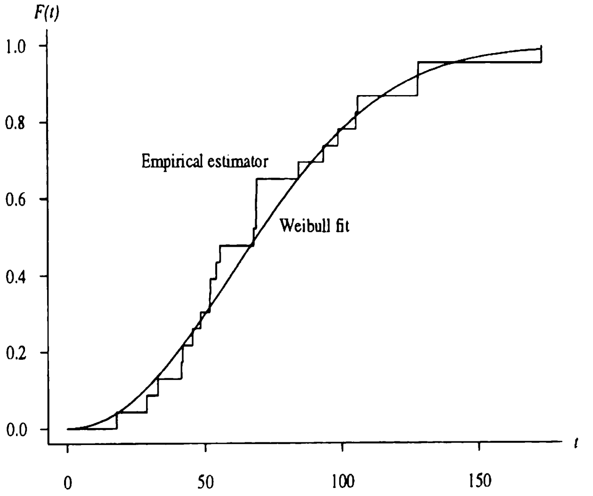

Figure 3: Empirical and Fitted Cumulative Distribution Functions for the Service Times

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. Using the fact that the likelihood ratio statistic, $2[\log L(\hat{\lambda},\hat{\kappa}) - \log L(\lambda,\kappa)]$, is asymptotically χ^2 with 2 degrees of freedom and that $\chi^2_{2,0.05} = 5.99$, a 95% confidence region for the parameters is all λ and κ satisfying

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

The 95% confidence region is shown in Figure 4. 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 κ is significantly different from 1, the standard errors of the distribution of the parameter estimators can be computed by using the inverse of the observed information matrix

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

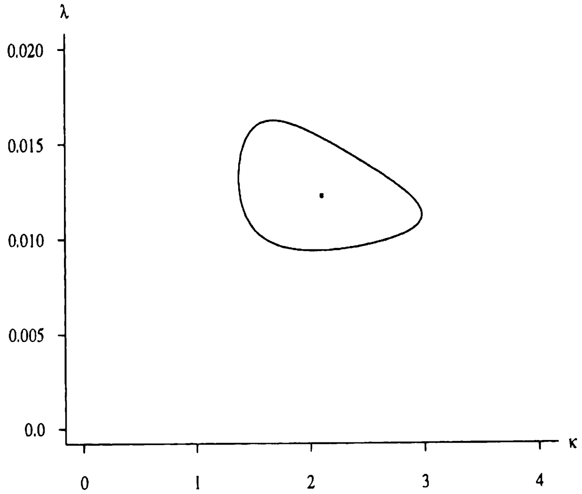

Figure 4: 95% Confidence Region Based on the Likelihood Ratio Statistic

This matrix 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

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

Thus an asymptotic 95% confidence interval for κ 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 κ is justified.

At this point, model adequacy should be assessed. Since the chi-square goodness-of-fit test suffers from arbitrary interval limits and can not be applied to small data sets, the Kolmogorov-Smirnov, Cramer-von Mises or Anderson-Darling goodness-of-fit tests are appropriate here (Lawless 1982). The Kolmogorov-Smirnov test statistic, for example, for this data set is 0.152, which measures the maximum difference between the empirical and fitted cumulative distribution functions. This test statistic corresponds to a P-value of approximately 0.15 (Law and Kelton 1991, page 391), so the Weibull distribution provides a reasonable model for these service times. Other models should also be assessed and compared to the Weibull model.

Many of the discrete-event simulation packages exhibited at the *Winter Simulation Conference* have the capability of determining maximum likelihood estimators for several parametric distributions. If the package also performs a goodness-of-fit test such as the Kolmogorov-Smirnov or chi-square test, the distribution that best fits the data set can quickly be determined. P-P and Q-Q plots can also be used to assess model adequacy.

2.2 Arrival Process Model

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 $(0, 4.5]$ be the time interval of interest (in hours) the three realizations are

$$0.2152 \quad 0.3494 \quad 0.3943 \quad \cdots \quad 4.175 \quad 4.248,$$

$$0.3927 \quad 0.6211 \quad 0.7504 \quad \cdots \quad 4.044 \quad 4.374,$$

and

$$0.4499 \quad 0.5495 \quad 0.6921 \quad \cdots \quad 3.643 \quad 4.357.$$

One preliminary statistical question 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 kept fixed. For this particular example, 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 is 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. On the other hand, if the process is nonstationary, a nonhomogeneous Poisson process might be an input appropriate model.

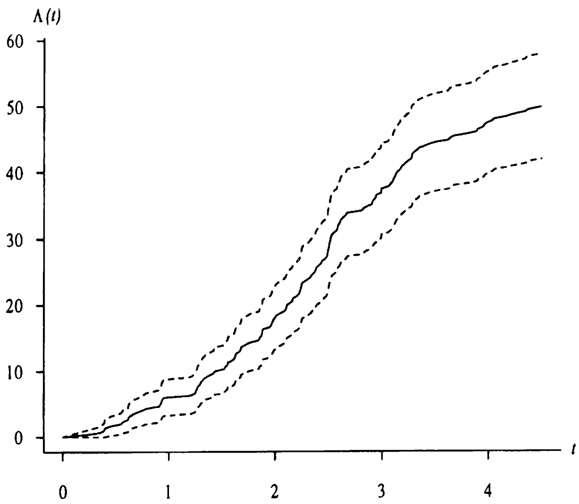

Figure 5: Point and 95% Confidence Interval Estimators for the Cumulative Intensity Function

Figure 5 contains a plot of the empirical cumulative intensity function estimator suggested by Leemis (1991) for the three realizations. The solid line denotes the point estimator for the cumulative intensity function $\Lambda(t)$ and the dashed lines denote 95% confidence intervals. The cumulative intensity function estimator at time 4.5 is $150/3 = 50$, the point estimator for the expected number of arriving customers per day. If $\hat{\Lambda}(t)$ is linear, a stationary model is appropriate. Since people 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 will be used to model the arrival process.

The next question to be determined is whether a parametric or nonparametric model should be chosen for the process. Figure 5 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 5 might be appropriate.

There are many potential parametric models for nonstationary arrival processes. The Weibull, or power law process has intensity function

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

where λ and κ are positive parameters. This popular model would *not* be appropriate for this data set since the intensity function can only increase, decrease or remain constant, and can not model an intensity function that increases, then decreases. 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

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

for $\lambda > 0$ and $\kappa > 0$, would certainly be an improved choice. A more general EPTF (exponential-polynomial-trigonometric function) model is given by Lee, Wilson and Crawford (1991) with intensity function

$$\lambda(t) = \exp\left[\sum_{i=0}^{m} \alpha_i t^i + \gamma \sin(\omega t + \phi)\right] \qquad t > 0.$$

The trigonometric function is capable of modeling the intensity function that increases, then decreases.

In all of the parametric models, the likelihood function for the vector of unknown parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ from a single realization on $(0, c]$ is

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

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.

ACKNOWLEDGMENTS

The author would like to thank Steve Tretheway for his help in developing Figure 1 while at The University of Oklahoma. Thanks also goes to Steve Park for reading a draft of this tutorial. The figures have been done using the Splus and xfig programs.

REFERENCES

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

Box, G., and G. Jenkins. 1976. *Time series analysis: forecasting and control.* Oakland, California: Holden-Day.

Devroye, L. 1986. *Non-uniform random variate generation.* New York: Springer-Verlag.

Flanigan-Wagner, M., and J. R. Wilson. 1993. Using univariate Bézier distributions to model simulation input processes. In *Proceedings ot the 1993 Winter Simulation Conference*, ed. G. W. Evans, M. Mollaghasemi, E. C. Russell, and W. E. Biles, 365–373. Institute of Electrical and Electronics Engineers, San Francisco, California.

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.

Law, A. M., and W. D. Kelton. 1991. *Simulation modeling and analysis.* 2d ed. New York: McGraw-Hill.

Law, A. M., M. G. McComas, and S. G. Vincent. 1994. The crucial role of input modeling in successful simulation studies. *Industrial Engineering* 26:55–59.

Lawless, J. F. 1982. *Statistical models & methods for lifetime data.* New York: John Wiley & Sons.

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

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

Lieblein, J., and M. Zelen. 1956. Statistical investigation of the fatigue life of deep-groove ball bearings. *Journal of Research of the National Bureau of Standards* 57:273–316.

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

Ross, S. M. 1993. *Introduction to probability models.* 5d 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. and S. J. Deutsch. 1977. A versatile four-parameter family of probability distributions, suitable for simulation. *AIIE Transactions* 2:176–182.

AUTHOR BIOGRAPHY

LAWRENCE M. LEEMIS is an associate professor in the Mathematics Department at the College of William and Mary. He received his BS and MS degrees in Mathematics and his PhD in Industrial Engineering from Purdue University. He has also taught courses 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, ICASE, Federal Aviation Administration, Tinker Air Force Base, and Argonne National Laboratory. His research and teaching interests are in reliability and simulation. He is a member of **ASA**, **IIE**, and **ORSA**.