Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

INPUT MODELING

Lawrence Leemis

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

ABSTRACT

Discrete-event simulation models typically have stochastic elements 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. 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. Most simulation texts (e.g., Law and Kelton 2000) have a broader treatment of input modeling than presented here. Nelson et al. (1995) and Nelson and Yamnitsky (1998) survey advanced techniques.

1 COLLECTING DATA

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 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 performance measure 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” about the 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 biostatistical literature contains techniques for accommodating censored data sets (Lawless 1982).
- Insufficient distribution resolution. We want the distribution of number 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 that illustrates the scope of potential input models that are 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.

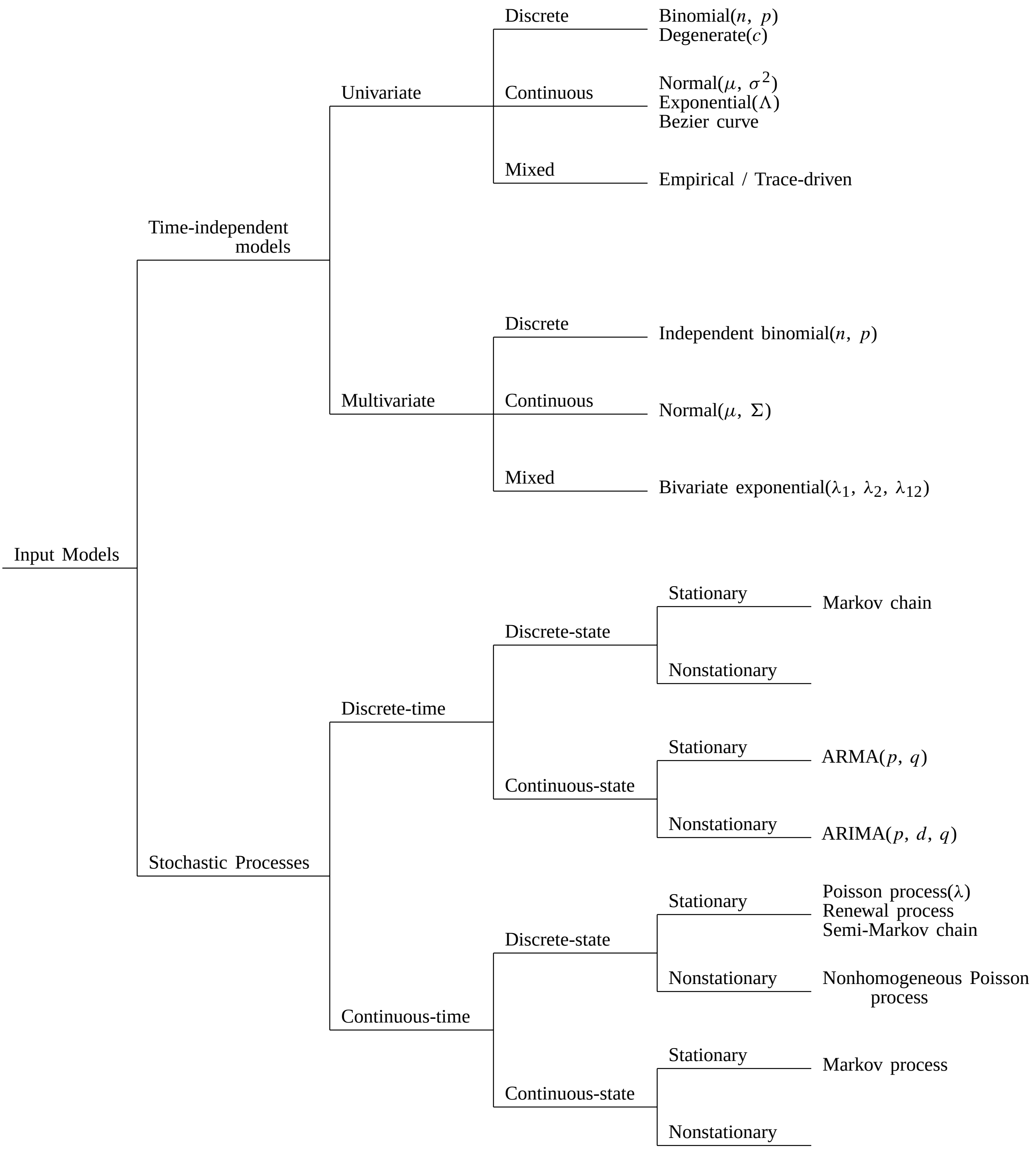

Figure 1: A Taxonomy for Input Models

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 Λ (see, for example, Martz and Waller 1982). Bézier curves (Flanigan–Wagner and Wilson 1993) 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 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 1997) 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.

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 (Lawless 1982, p. 228), 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 regressed against 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 (Neter, Wasserman, and Kutner 1989)

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

where X is the observation number, Y is the service time, β_0 is the intercept, β_1 is the slope, and ϵ is an error term. Figure 2 shows a plot of the (x_i, y_i) pairs for $i = 1, 2, \ldots, 23$, 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.

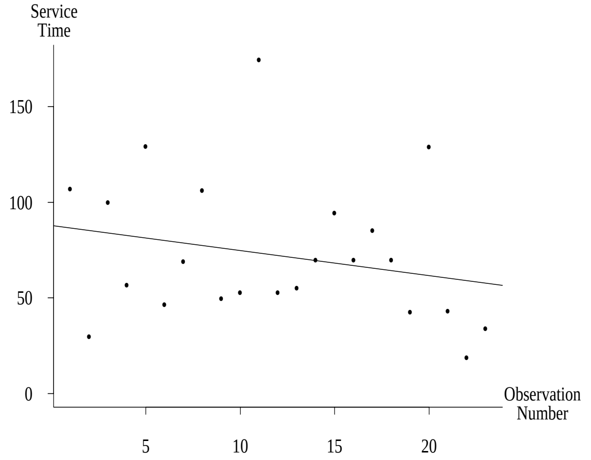

Figure 2: Service Time Vs. Observation Number

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. 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 in Figure 3. 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

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

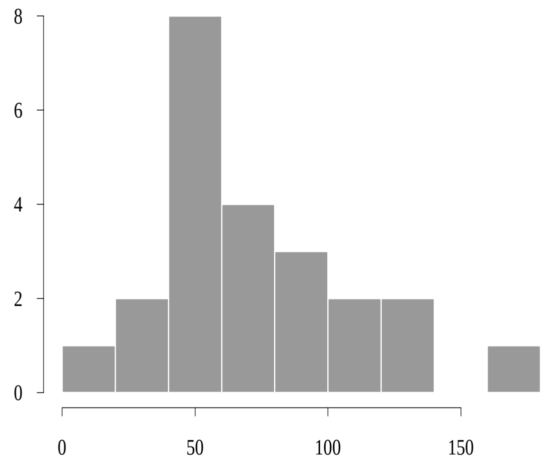

Figure 3: Histogram of Service Times

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

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 (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 $1/23$. 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 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 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 $x_1, x_2, \ldots, x_n$ 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 λ and κ. The mathematics are typically more tractable for maximizing log likelihood functions, 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 2×1 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 MLEs $\hat{\lambda}$ 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 (2000, p. 305) give an initial estimate for κ 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 κ for a complete data set.

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 4 shows the empirical cumulative distribution function (a step function with a step of height $1/n$ at each data point) along with the Weibull fit to the data.

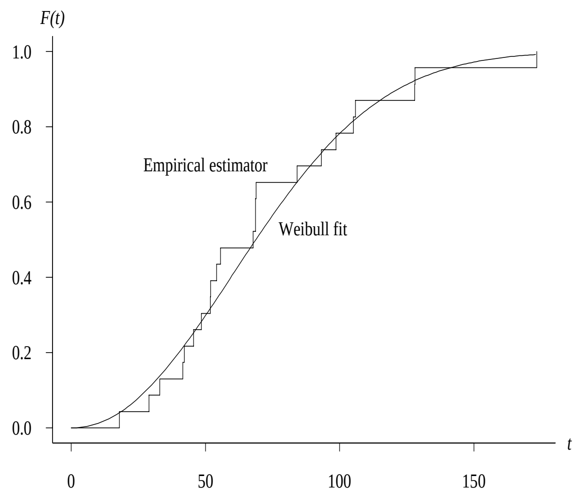

Figure 4: Empirical and Fitted Cumulative Distribution Functions for the Service Times

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 λ and κ. Using the fact that the likelihood ratio statistic, $2[\log L(\hat{\lambda}, \hat{\kappa}) - \log L(\lambda, \kappa)]$, is asymptotically χ^2 distributed in n 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 5. 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.

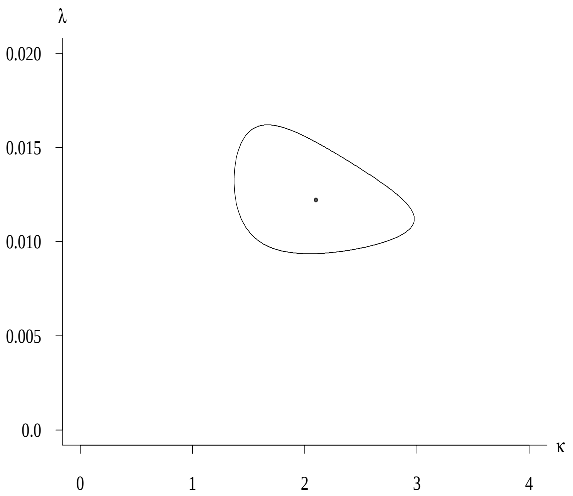

Figure 5: 95% Confidence Region Based on the Likelihood Ratio Statistic

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

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

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

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

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.

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 1982) are appropriate here. The Kolmogorov–Smirnov test statistic for this data set with a Weibull fit is 0.151, which measures the maximum vertical difference between the empirical and fitted cumulative distribution functions. This test statistic corresponds to a p-value of approximately 0.15 (Law and Kelton 2000, page 366), so the Weibull distribution provides a reasonable model for these service times. The Kolmogorov–Smirnov test statistic values for several models are shown below, including four that are superior to the Weibull with respect to fit.

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. 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 (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 ith order statistic $x_{(i)}$, i.e., $\hat{F}(x_{(i)})$, versus the adjusted empirical cumulative distribution function, i.e., $\tilde{F}(x_{(i)}) = \frac{i-0.5}{n}$, for $i = 1, 2, \ldots, n$. A plot where the points fall close to the line passing thru 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 6, along with a line connecting (0, 0) and (1, 1). P–P plots should be constructed for all competing models.

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.

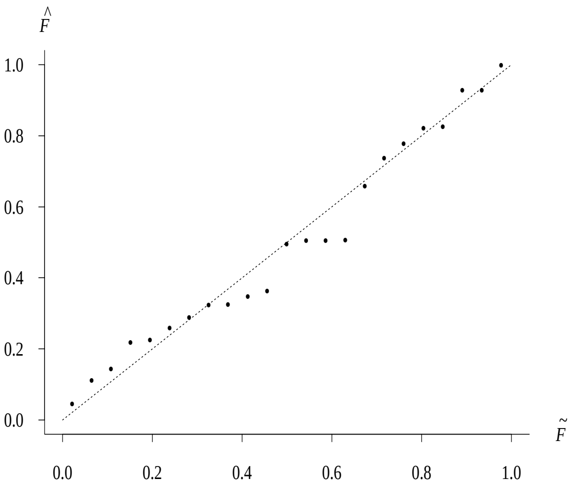

Figure 6: A P–P Plot for the Service Times Using the Weibull Model

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 $(0, 4.5]$ 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, $i = 1, 2, \ldots, k$ be the number of observations in the ith realization, $n = \sum_{i=1}^{k} n_i$, and let $t_{(1)}, t_{(2)}, \ldots, t_{(n)}$ 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 $t_{(i)} < t \le t_{(i+1)}$; $i = 0, 1, 2, \ldots, n$, 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), are given in Figure 7. 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

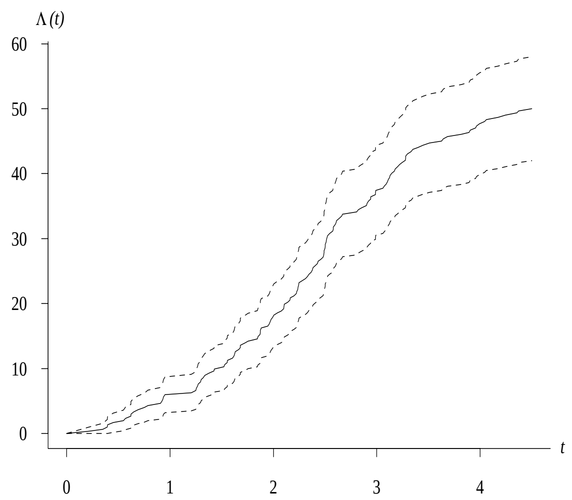

Figure 7: Point and 95% Confidence Interval Estimators for the Cumulative Intensity Function

estimator has an S-shape, a nonstationary model is indicated. More specifically, a nonhomogeneous Poisson process is an appropriate model for the arrival process.

The next question to be determined is whether a parametric or nonparametric model should be chosen for the process. Figure 7 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 7 might be appropriate.

There are many potential parametric models for nonstationary 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.

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

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

MLEs 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

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

The power law process has intensity function

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

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 2×1 score vector has elements

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

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 λ and κ are

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

Substituting the arrival times into these formulas yields MLEs $\hat{\lambda} = 4.86$ and $\hat{\kappa} = 1.27$. The cumulative intensity function for the power law process is

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

which is plotted along with the nonparametric estimator in Figure 8. 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.

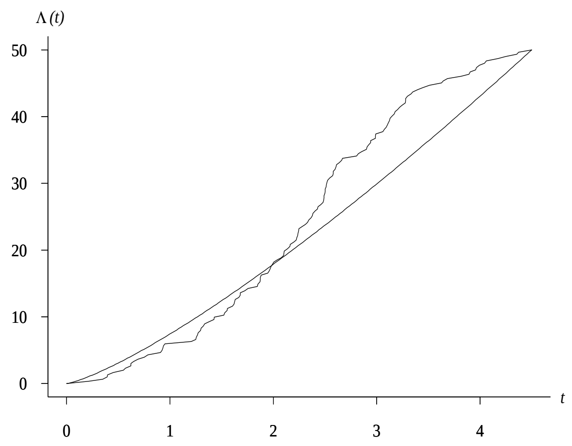

Figure 8: Empirical and Fitted Power Law Estimators for the Cumulative Intensity Function

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 1982)

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

for $\lambda > 0$ and $\kappa > 0$, would certainly be more appropriate. More generally, the EPTF (exponential-polynomial-trigonometric function) model given by Lee, Wilson and Crawford (1991) with intensity function

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

can model a nonmonotonic intensity function.

ACKNOWLEDGMENTS

The author thanks Steve Tretheway for his help in developing Figure 1, and Diane Evans, Steve Park & 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., and G. Jenkins. 1976. *Time series analysis: forecasting and control*. Oakland, California: Holden–Day.

Bratley, P., B. L. Fox, and L. E. Schrage. 1987. *A Guide to Simulation*. 2d ed. New York: Springer–Verlag.

Flanigan–Wagner, M., and J. R. Wilson. 1993. Using univariate Bézier distributions to model simulation input processes. In *Proceedings of 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, Piscataway, New Jersey.

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. 2000. *Simulation modeling and analysis*. 3d ed. New York: McGraw–Hill.

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 having cyclic behavior. *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.

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

Nelson, B. L., P. Ware, M. C. Cario, C. A. Harris, S. A. Jamison, J. O. Miller, J. Steinbugl, J. Yang. 1995. Input modeling when simple models fail. In *Proceedings of the 1995 Winter Simulation Conference*, ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman, 93–100. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

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. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Neter, J., W. Wasserman, and M. H. Kutner. 1989. *Applied Linear Regression Models*. 2d ed. Boston: Irwin.

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

Ross, S. M. 1997. *Introduction to probability models*. 6th 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.

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. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

AUTHOR BIOGRAPHY

LAWRENCE M. LEEMIS is a professor and chair of the Mathematics Department at the College of William & 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, 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>`.