Proceedings of the 1984
Winter Simulation Conference
S. Sheppard, U. Pooch, D. Pegden (eds.)

INPUT DATA COLLECTION AND ANALYSIS

W. David Kelton
Department of Industrial and Operations Engineering
The University of Michigan

ABSTRACT

This paper reviews perspectives and methods for specifying distribution and process forms and parameters from which observations are drawn to drive a stochastic simulation. All phases of input data analysis are covered, including data collection, choosing a modeling distribution, estimating parameters, and goodness-of-fit testing. A discussion is also presented concerning the debate over the desirability of fitting "standard" distributions to data as opposed to using a direct empirical distribution. Available software packages with a simulation orientation are also described.

INTRODUCTION

The modeling activity which stands at the beginning of a simulation study may be thought of as consisting of two sequential activities: **Structural** and **quantitative. Structural** modeling concerns the development of physical and logical relationships and interconnections composing the mechanism of the model; for example, queue disciplines, single-stage or tandem queues, and individual vs. batch arrivals. **Quantitative** modeling, on the other hand, concerns specification of the numerical attributes of the model, and usually occurs after structural modeling; examples include the maximum length allowed for a queue, the number of servers, and the times and sizes of arrivals of batches. In this paper it is assumed that the structural modeling has been done, and it remains to carry out the quantitative modeling.

One attribute of a simulation that has a great impact on the nature of quantitative modeling is whether there are any random inputs assumed to drive the model. In a **deterministic** model, there is no randomness in the inputs to the structural model; for example, we assume there are three servers, a group of 4 customers arrives every 3.8 minutes, and each customer requires 6.8 minutes of service. In this case the quantitative modeling task is simply to determine the values of all needed constants. **Stochastic** models allow for randomness in at least some of the inputs driving the structural model. Examples of input quantities that are typically modeled stochastically are the times of machine breakdowns, service times, and the number of items in a batch of random size. Quantitative modeling of a stochastic system is more involved than that for deterministic systems, since we must specify distribution (or process) forms, as well as the numerical values of the parameters needed to specify the chosen distributions and processes. This paper will address issues involved in quantitative modeling for stochastic simulations. Since this modeling activity determines the distributions and processes from which sampling will take place to drive the simulation, this topic is often referred to as **input data analysis.**

DATA COLLECTION

The data collection phase of input analysis may seem rather obvious, and in many cases will indeed be straightforward. There are, however, several issues that should be kept in mind.

First, care should be taken that the correct quantity or process be observed, and that all necessary information is collected. If only a surrogate for the desired quantity can be observed, inference about the desired quantity is risky, and its validity may be untestable. A prime example here would be simulation of a proposed facility that does not exist, and we attempt to collect data from "similar" systems elsewhere. Worse, it may be possible to collect the required data, but because of poor planning or lack of foresight, we do not do so. This could occur if data collection proceeds before there is a clear understanding of just what will be needed in the simulation; thus, it may be advisable to construct at least a preliminary version of the model before data are collected, to avoid overlooking any needed information.

Second, the mechanics of data collection should be thought about, and appropriate recording instruments, forms, etc. be supplied. For example, if interarrival times are to be collected, it is usually easier to note successive times of arrivals, then difference them later to obtain the desired interarrivals.

Third, since most of the statistical techniques used in input analysis assume that the data are independent and identically distributed (i.i.d.), care should be taken, if possible, to meet this assumption. In general, lack of independence can have a severe effect on statistical procedures assuming independence (as we have learned from simulation **output** analysis methodological research). As a pre-test, one might apply a test for independence (i.e., "randomness") before proceeding with a formal input analysis procedure; see, for example, Conover (1980).

Finally, there is a tacit assumption that the observations in a data set are **homogeneous**, i.e., are drawn from the same distribution or process. This becomes an issue if, for example, observations on the same quantity are gathered on several different days, in an attempt to increase the sample size. Before merging these separate days' data, a test for homogeneity might be applied to justify such action; one test for this purpose is the Kruskal-Wallis test (see Conover [1980]).

NO DATA?

While most of the literature and techniques concerning specification of sampling distributions and processes assume the availability of or opportunity to collect a reasonable amount of data, this is sometimes (perhaps often) an unaffordable or impossible luxury. In this case, we are left with no really good alternatives, but there may be some "rough" options to invoke, involving specific distributions.

First, if we can somehow elicit values below which or above which it is felt that observations would never occur, one could (in the absence of other information) posit a uniform distribution between these two extreme values. This would apply in either the continuous or discrete case.

Second, if we can, in addition to a minimum and maximum, establish a "most likely" value (i.e., a mode) between them, then a triangular distribution would be determined; again, discrete and continuous cases are covered.

Finally, if we have a continuous distribution and one for which positive skewness is a reasonable assumption, then specification of a minimum, maximum, mode, and mean determines a beta distribution from which samples could be drawn (see Law and Kelton [1982]). This is the most sophisticated of these techniques, and requires the most information; it may not be easy to elicit different values for the mode and the mean, which would be required in this approach. This method is often used to specify durations of arcs in PERT networks. It should be mentioned that, while the beta option results in a smoother distribution than the uniform or triangular options, it may result in a distribution from which sampling during the simulation is considerably more difficult.

A PRIORI CHOICE

In some situations there may be sufficient information (or assumptions) to imply a distribution form, if not the numerical value(s) of its parameter(s). For example, if the quantity of interest is the sum (resp. product) of a large number of other i.i.d. quantities, then central limit theory considerations suggest a normal (resp. lognormal) distribution. If the quantity is the inter-event time of a process where events occur "at random" (i.e., at the same expected rate and independently of each other), then the exponential distribution is appropriate. Other such physical models are discussed in Bratley, Fox, and Schrage (1983). Even if we are fortunate enough to have such a priori information, estimation of the parameters of the implied distributions will probably still be necessary.

TO FIT OR NOT TO FIT?

Given a data set, there are two quite different routes possible in using it to specify a distribution from which to sample the corresponding quantity during the simulation:

(a) "Fit" one (or several) "standard" distribution(s) to the data, and use the resulting fitted distribution.

(b) Use the data values directly to define an empirical distribution which is then used in the simulation.

Route (a) has been the traditional approach, but has recently been challenged in favor of (b) by Fox (1981) and Bratley, Fox, and Schrage (1983). Most would agree that (b) is preferred in the case that no "standard" distribution can be found that adequately fits the observations. The debate occurs when (a) does result in at least one candidate distribution that adequately represents the data. Each side would appear to have its own set of merits, which might be summarized as follows:

In favor of (a):

1. There is less sensitivity to the particular data set obtained. For example, there could have been holes in the data set, i.e., intervals where little or no data appeared. While this may be characteristic, it may also have resulted from simple sampling fluctuations. Fitting a distribution will "smooth" out these features, which may be anomalies with respect to the underlying "true" process, while an empirical distribution will result in generating such holes in the data every time it is used in the simulation. This consideration may be especially relevant for small data sets.

2. Depending on how the empirical distribution is specified, it may confine the generated values in the simulation to the range covered by the observed data set. A fitted distribution need not have this restriction, if specified from a family with infinite range.

3. There may be a priori grounds for using one of the "standard" distributions (see above).

In favor of (b):

1. There is rarely reason to choose some particular distribution form (e.g., gamma) to fit to the data, other than the "spurious" similarity between the data and the distribution. Furthermore, the power of goodness-of-fit tests is generally low (see Bratley, Fox, and Schrage [1983]), so that failure to reject the null hypothesis that the data appear to have come from the distribution form in question is more a reflection of this lack of power than of a good fit.

2. A distribution that is mostly empirical can easily be specified, if desired, to have an infinite right tail, so that simulated observations can be generated beyond the range of the data (see Bratley, Fox, and Schrage [1983]).

3. Empirical distributions are very easy to generate from, needing only piecewise linear interpolation, possibly with the inversion of an exponential tail (see Bratley, Fox, and Schrage [1983, pp. 139-140]); route (a) can result in distributions from which generation is more difficult.

If route (b) is chosen, generation proceeds in a straightforward manner; the only computational consideration is the need to sort the data, but this is only done once. Thus, in the remainder of the paper the steps for implementing route (a) will be discussed.

STEPS IN FITTING A DISTRIBUTION

We assume that route (a) has been chosen, and that a univariate distribution is to be specified from the available data. The steps outlined below are covered in detail in Law and Kelton (1982). See also Johnson and Kotz (1969, 1970) for a wealth of information on many distributions and their properties.

Specifying a Distribution Form

The first step in fitting a distribution is to decide what form, or "family," is to be considered. For example, interarrival times may be assumed to have an exponential distribution, or times between failures to have a Weibull distribution. Short of a priori determination (see above), there is no complete way to arrive at a choice of distribution form. The usual method is to rely on various heuristics to aid in an informed choice. Such heuristics are:

1. Range Considerations. This is not really a heuristic, and serves principally to rule out some distributions on the basis of their range. For example, if X is the proportion of time a customer's service is to be attended to by a particular server, then clearly $0 \leq X \leq 1$, so using an exponential distribution for X is inappropriate in view of the lack of an upper limit on this distribution. Such considerations, incidentally, make the normal distribution inappropriate (strictly speaking) for modeling activity durations, since any normal distribution allows the possibility of negative values.

2. Histograms. In a certain sense, a histogram is an unbiased estimator of the shape of a density function (in the continuous case) or the probability mass function (in the discrete case) of the underlying distribution. Thus, a histogram is drawn and compared with the density or mass functions of candidate distributions.

3. Point Statistics. Sample means, variances, and ratios of these may be compared with what would be expected from various distributions, particularly whether the mean or variance (or standard deviation) is larger. This may serve as a rough guide, but it should be remembered that such statistics may be quite variable.

4. Probability Plots. With a few exceptions, these apply only in the continuous case, and measure the similarity between the empirical cumulative distribution function and the distribution function of the candidate distributions. These plots do not require an interval choice (as do histograms), but must be re-plotted for each candidate distribution. Further, if the candidate distribution has shape parameters, they may have to be pre-estimated. Without appropriate software, probability plots may be difficult to construct for some distributions. See in addition Hahn and Shapiro (1967) for more information on alternative forms of probability plots.

If none of the standard distributions appears reasonable at this point, consideration should be given to using an empirical distribution. If it appears that some distribution(s) may fit, the next step is parameter estimation.

Parameter Estimation

With a particular distributional form in mind, the parameter(s) of it need to be estimated; this is the actual "fit." Many methods of estimation are possible (least squares, unbiased, method of moments, etc.), but the most widely used and accepted method is that of **maximum likelihood**; this method also has the advantage of at least partially justifying one of the later goodness-of-fit tests. Basically, the principle states that the best numerical values to assign to the parameters are those which maximize the probability, under the assumed distributional form, of obtaining the particular data values which were observed (or near those observed in the continuous case). Maximum likelihood estimators (m.l.e.'s) enjoy several nice statistical properties, such as asymptotic (as the sample size grows) unbiasedness and normality.

Using the m.l.e. technique, one can also construct a confidence interval (c.i.) for the parameter being estimated (see Law and Kelton [1982, p. 191]); this is quite useful in simulation, in the following sense. As a model is being developed, there is always uncertainty about the values of the parameters used. If, on the basis of available data, we have a c.i. for some true parameter, the simulation could be run with this parameter set once at the lower endpoint and again at the upper endpoint. If the simulation output measures do not appear to be sensitive to this change, then we would accept the m.l.e. as the parameter; if significant sensitivity is displayed, however, we might want to collect more data to get a better estimate of this critical parameter.

The particular method of finding m.l.e.'s depends wholly on the distributional form; sometimes it is trivial and sometimes quite difficult. In particular, numerical methods may be required to maximize the likelihood function for a given data set. Sometimes there are tables available to aid in this task, and some of the software (see below) internally finds m.l.e's for the available distributions.

Assessing the Fit

The final step is to perform goodness-of-fit tests to check on the adequacy of the chosen distributional form. The best-known and most widely applicable test is the classical chi-squared test; it applies in either the discrete or continuous case, but is only asymptotically (as the sample size, not the number of test intervals, grows) valid. Further, care must be taken to choose the degrees of freedom correctly, in view of the number of parameters being estimated.

One troublesome aspect of the chi-squared goodness-of-fit test is the need to choose intervals for grouping the data. The Kolmogorov-Smirnov (K-S) test avoids this grouping decision. Whereas the chi-squared test may be thought of as a comparison of the empirical and fitted densities, the K-S test measures the (maximum vertical) distance between the empirical and fitted cumulative distribution functions. In its original form, the K-S test applies only to continuous distributions whose fit did not use any of the data; this is rarely the case in practice, so the test has been extended to allow testing fit of several distributions with estimated parameters. Care should be taken to note that the K-S test with estimated parameters requires a separate table of critical values for each different distribution; this point seems to have been frequently missed.

In addition to the chi-squared and K-S tests, there is the Anderson-Darling test, as well as various specialized tests for specific distributions, such as normal and uniform. Finally, with so many possible distributions and the possibility of applying several tests, one could end up with a matrix full of test statistics or p-values, with rows corresponding to distributions and columns corresponding to types of tests. This should serve only as a rough comparative guideline on assessing goodness of fit, and no overall significance level should be ascribed to these values in a formal hypothesis-testing sense, due to the (possibly severe) effect of the multiple comparisons problem.

CORRELATION AND MULTIVARIATE MODELING

The above discussion has centered on choosing a univariate distribution for a scalar random variable in the simulation. If all the random variables in a simulation are independent (or assumed to be so), this will suffice since the univariate distribution choice is just repeated for each random variable. If we want to model dependencies among variables or specify a complete multivariate distribution, however, other considerations come into play. The importance of explicitly including correlation in simulation input modeling has been demonstrated by Mitchell, Paulson, and Beswick (1977) who showed that ignoring correlation between service times in a tandem queueing system can result in serious errors in the output.

In the simplest case, we may feel that some random variable X is correlated with some other random variable Y, and that both has its own (marginal univariate) distribution. No other assumptions are modeled, i.e., we are not seeking to specify the complete joint distribution of (X,Y). A direct approach here is simply to fit distributions (or use empirical distributions) for X and Y, and from the same data estimate the correlation between X and Y. While this presents no special difficulties, attention must be paid in the simulation context to the need to generate these correlated pairs (X,Y), which limits the choice of marginal distributions. Of course, normal (and thus lognormal) is a possibility, but aside from this the choices are few due to limitations on the ability to generate correlated random variables. In particular, correlated gamma pairs are possible (Schmeiser and Lal [1982]), as are correlated exponentials as a result.

The most general goal would be to specify a complete joint distribution for a random vector $(X_1, \ldots, X_k)$ used as input to a simulation. This is an ambitious task; indeed, outside the normal (or lognormal) case, there is not general agreement on just what a joint distribution should be defined as, even in such simple cases as marginal exponentials. In addition, data requirements would probably be demanding. In the multivariate normal case, specification of all pairwise correlations determines the joint distribution anyway, in which case we are in the situation of the preceding paragraph. Thus, it would seem that in most cases, the difficulties associated with modeling an entire joint distribution would preclude its use in simulation. Schmeiser and Lal (1980) survey available continuous multivariate models amenable to simulation, and provide numerous references. A general reference on continuous multivariate distributions is Johnson and Kotz (1972).

NONSTATIONARY EVENT OCCURRENCES

A situation that seems to come up often enough in applications to warrant special mention here is that of a random process of events of some sort through time which do **not** occur at a constant expected rate. Important examples include customer arrivals to a queueing facility throughout a period characterized by "peak loads" and other times of relative inactivity, traffic engineering that must deal with rush hours, and reliability studies where machines are subjected to varying levels of stress resulting in fluctuating breakdown rates.

If such events are modeled as having i.i.d. interevent times (e.g., exponential for a Poisson process), the desired nonstationarity will not result. A particular process model that has been found useful is a **nonstationary** Poisson process (see Cinlar [1975]) which assumes, basically, that events occur independent of one another but at an expected rate that may be a function of time; this rate function may thus be chosen to reflect rush hours, etc., and its specification completely determines the structure of the process. Estimating the rate function from data, however, is problematic since we need to specify an entire function as opposed to just a few parameters; see Law and Kelton (1982) for a rough but practical approach, or Lewis and Shedler (1976) for a more sophisticated method assuming a particular functional form with parameters to be estimated.

Although not really part of input analysis, it should be mentioned that generating observations from a nonstationary Poisson process is not as trivial as it might seem. In particular, it is not valid to generate the next interevent time as being exponential with parameter being determined by the current value of the rate function. Instead, the thinning algorithm of Lewis and Shedler (1979) should be employed.

SOFTWARE

While conceptualizing the methodology of fitting distributions may seem straightforward, implementing many of the steps involved can lead quickly into some involved computational difficulties. For example, making a probability plot for a candidate gamma distribution may require inverting its cumulative distribution function, requiring special techniques and thus special software. Moreover, most well-known, general-purpose statistical analysis packages are not really oriented to dealing with the wide variety of distributions we would typically like to consider in quantitative modeling for simulation; attention in such packages usually focuses on normal distribution theory. In this section three packages are mentioned that were specifically designed for application to some of the problems of distribution fitting in the simulation context.

AID, available from Pritsker & Associates of West Lafayette, Indiana, is an interactive graphics package taking as input a data set to which a distribution is to be fitted. The user specifies the desired distribution form, a fit is made, and chi-squared and K-S tests are available. Two discrete distributions (discrete uniform and Poisson) are allowed, and there is provision for ten continuous distributions (uniform, triangular, normal, lognormal, exponential, Erlang, gamma, Weibull, beta, and beta-PERT [taking as input the minimum, maximum, and mode, for PERT network applications]). Much of the output is in graphical form, for example, for the chi-squared goodness-of-fit

test the density of the chi-squared distribution is plotted with the test statistic and critical values noted by vertical lines, and a plot is provided of the sample cumulative distribution function with bands around it showing how far a permissible deviation is allowed under the assumed distribution.

UNIFIT, available from Simulation Modeling and Analysis Company of Tucson, Arizona, is an interactive package providing aids for hypothesizing a distribution form (including histograms, probability plots, and point statistics), carries out a fit according to several alternative user-selectable criteria (including known parameters and maximum likelihood estimators), and performs several goodness of fit tests, including chi-squared, K-S, and Anderson-Darling, providing p-values as well as test statistics. Five discrete distributions are available (binomial, geometric, negative binomial, Poisson, and discrete uniform), as well as thirteen continuous distributions (exponential, gamma, inverse Gaussian, lognormal, Weibull, Pearson type 5 and 6, extreme value type A and B, logistic, normal, uniform, and beta). Any number of candidate distributions may be fitted to a given data set, for comparison of goodness of fit. Other capabilities, such as file management and the Kruskal-Wallis test for homogeneity of different data sets, are also present.

In the appendix of Solomon (1983) are listings of FORTRAN programs composing a package called SIMSTAT which provide support for distribution fitting, including summary statistics and goodness of fit testing for exponential, normal, uniform, and Poisson distributions. One of the routines of the package also deals with individual customer records from observation of a queueing system, for analysis of interarrival and service time patterns.

CONCLUSIONS

The problem of specification of distributions and processes from which samples will be generated during a simulation is an important part of simulation modeling, and can affect the output and conclusions from a simulation study in important ways. Although this quantitative modeling may not be as visible or as easily understood as the more physical structural modeling, it may be as important in determining model validity (i.e., the degree to which the model as a whole resembles the system being simulated) and thus the ultimate usefulness of the simulation project.

REFERENCES

Bratley, P., B.L. Fox, and L.E. Schrage. A Guide to Simulation, Springer-Verlag, New York, 1983.

Cinlar, E. Introduction to Stochastic Processes. Prentice-Hall, New York, 1975.

Conover, W.J. Practical Nonparametric Statistics, second edition, Wiley, New York, 1980.

Fox, B.L., "Fitting 'Standard' Distributions to Data is Necessarily Good: Dogma or Myth?," Proceedings of the 1981 Winter Simulation Conference, 305-307, 1981.

Hahn, G.J. and S.S. Shapiro. Statistical Models in Engineering, Wiley, New York, 1967.

Johnson, N.L. and S. Kotz. Discrete Distributions, Wiley, New York, 1969.

Johnson, N.L. and S. Kotz. Continuous Univariate Distributions, 2 volumes, Wiley, New York, 1970.

Johnson, N.L. and S. Kotz. Continuous Multivariate Distributions, Wiley, New York, 1972.

Law, A.M. and W.D. Kelton. Simulation Modeling and Analysis, McGraw-Hill, New York, 1982.

Lewis, P.A.W. and G.S. Shedler, "Statistical Analysis of Non-Stationary Series of Events in a Data Base System," IBM J. Research and Development, 20, 465-482, 1976.

Lewis, P.A.W. and G.S. Shedler, "Simulation of Nonhomogeneous Poisson Processes by Thinning," Naval Research Logist. Quart., 26, 403-413, 1979.

Mitchell, C.R., A.S. Paulson, and C.A. Beswick, "The Effect of Correlated Exponential Service Times on Single Server Tandem Queues," Naval Research Logist. Quart., 24, 95-112, 1977.

Schmeiser, B. and R. Lal, "Multivariate Modeling in Simulation: A Survey," ASQC Technical Conference Transactions, 1980.

Schmeiser, B. and R. Lal, "Bivariate Gamma Random Vectors," Operations Research, 30, 355-374, 1982.

Solomon, S.L. Simulation of Waiting-Line Systems, Prentice-Hall, New York, 1983.