Proceedings of the 1980 Winter Simulation Conference
T.I. Ören, C.M. Shub, P.F. Roth (eds.)

A TUTORIAL ON STATISTICAL ANALYSIS OF SIMULATION OUTPUT DATA*

Averill M. Law†
Department of Industrial Engineering
and Mathematics Research Center
University of Wisconsin-Madison
Madison, WI 53706

In many simulation studies a large amount of time and money is spent on model development, but little effort is made to analyze the simulation output data in a proper manner. Since most simulation models use random variables as input, the output data are themselves random and care must therefore be taken in drawing conclusions about the system under study.

In this paper we present an up-to-date treatment of procedures which can be used for constructing confidence intervals for measures of performance of a simulated system. The emphasis will be on simple, easy-to-use procedures which have been shown to perform well in practice.

1. INTRODUCTION

It has been our observation that in many simulation studies a large amount of time and money is spent on model development and programming, but little effort is made to analyze the simulation output data in an appropriate manner. As a matter of fact, the most common mode of operation is to make a single simulation run of somewhat arbitrary length and then treat the resulting simulation estimates as being the "true" answers for the model. Since these estimates are random variables (r.v.'s) which may have large variances, these estimates may, in a particular simulation run, differ greatly from the corresponding true answers. The net effect is, of course, that there may be a significant probability of making erroneous inferences about the system under study.

One reason for the historical lack of definitive output data analyses is that simulation output data are rarely, if ever, independent. Thus, classical statistical analyses based on independent identically distributed (i.i.d.) observations are not directly applicable. At the present time, there are still several output analysis problems for which there is no completely accepted solution, and the solutions that do exist are often complicated to apply. Another impediment to getting accurate estimates of a model's true parameters or characteristics is the computer cost associated with collecting the necessary amount of simulation output data. Indeed, there are situations where an appropriate statistical procedure is available, but the cost of collecting the amount of data dictated by the procedure is prohibitive. We expect this latter problem to become less important as the cost of computer time continues to drop.

Our goal in this tutorial is to give an up-to-date treatment of statistical analyses for simulation output data, and to present the material with a practical focus which should be accessible by a reader having a basic understanding of statistics. The emphasis will be on statistical procedures which are relatively easy to understand, have been shown to perform well in practice, and have applicability to real-world problems. The remainder of the paper is organized as follows. In Section 2 we define the two types of simulations with regard to analysis of the output, namely, terminating and steady-state simulations. Section 3 contrasts measures of performance for these two

*This paper is adapted with permission from Averill M. Law and W. David Kelton, *Simulation Modeling and Analysis*, McGraw-Hill, Inc., New York (1981). All rights to this material are reserved by the authors and the publisher.

†Sponsored by the United States Army under Contract No. DAAG29-80-C-0041 and the Office of Naval Research under Contract N00014-76-C-0403 (NR 047-145).

types of simulations and Section 4 discusses the need to assess the accuracy of simulation output. In Sections 5 and 6 we discuss how to construct a confidence interval (c.i.) for a measure of performance in the terminating and steady-state cases, respectively.

A more comprehensive treatment of statistical analyses for simulation output data may be found in Law and Kelton (1981c).

2. TYPES OF SIMULATIONS WITH REGARD TO ANALYSIS OF THE OUTPUT

We begin by giving a precise definition of the two types of simulations with regard to analysis of the output data. A *terminating simulation* is one for which the desired measures of system performance are defined relative to the interval of simulated time $[0,T_E]$, where T_E is the instant in the simulation when a specified event E occurs. (Note that T_E may be a r.v.) The event E is specified before the simulation begins. Since measures of performance for terminating simulations explicitly depend on the state of the simulated system at time 0, care must be taken in choosing initial conditions. This point will be discussed further in the following examples of terminating simulations:

a) A retail establishment (e.g., a bank) closes each evening (physically terminating). If the establishment is open from 9 to 5, then the objective of a simulation might be to estimate some measure of the quality of customer service over the period beginning at 9 and ending when the last customer who entered before the doors closed at 5 has been served. In this case E = {at least 8 hours of simulated time have elapsed and the system is empty}, and reasonable initial conditions for the simulation might be that no customers are present at time 0.

b) Consider a telephone exchange which is always open (physically nonterminating). The objective of a simulation might be to determine the number of (permanent) telephone lines needed to service adequately incoming calls. Since the arrival rate of calls changes with the time of day, day of the week, etc., it is unlikely that a steady-state measure of performance (see Section 3), which is defined as a limit as time goes to infinity, will exist. A common objective in this case is to study the system during the period of peak loading, say, of length t hours, since the number of lines sufficient for this period will also do for the rest of the day. In this case, E = {t hours of simulated time have elapsed}. However, care must be taken in choosing the number of waiting calls at time 0, since the actual system will probably be quite congested at the beginning of the period of peak loading.

c) Consider a military confrontation between a defensive (fixed position) blue force and an offensive (attacking) red force. Relative to some initial force strengths, the objective of a simulation might be to estimate some function of the (final) force strengths at the time that the red force moves to within a certain specified distance from the blue force. In this case, E = {red force has moved to within a certain specified distance from the blue force}. The choice of initial conditions (e.g., the number of troops and tanks for each force) for the simulation is generally not a problem here since they are specified by the military scenario under consideration.

A *steady-state simulation* is one for which the measures of performance are defined as limits as the length of the simulation goes to infinity. Since there is no natural event E to terminate the simulation, the length of *one* simulation is made large enough to get "good" estimates of the quantities of interest. Alternatively, the length of the simulation could be determined by cost considerations; however, this may not produce acceptable results (see Subsection 6.1). The following is an example of a steady-state simulation:

a) A computer manufacturer is constructing a simulation model of a proposed computer system. Rather than use data from the arrival process of an existing computer system as input to the model, he typically assumes that jobs arrive in accordance with a Poisson process (i.e., i.i.d. exponential interarrival times) with rate equal to the predicted arrival rate of jobs during the period of peak loading. (This is done because it is not clear that the arrival process of an existing system will be representative of that of the proposed system, or (alternatively) out of simplicity.) *He is interested in estimating the response time of a job after the system has been running long enough so that initial conditions (e.g., the number of jobs in the system at time 0) no longer have any effect.*

Because the arrival rate of jobs will vary with the time of day, etc., steady-state measures for real-world computer systems will probably not exist. However, by assuming that the arrival rate is constant over time in the model, this allows steady-state measures to exist. In performing a steady-state analysis of the proposed computer system, the model's developers are essentially trying to determine how the system will respond to a peak load of infinite duration.

3. MEASURES OF SYSTEM PERFORMANCE

3.1. Contrast of Measures of Performance

In this subsection we contrast measures of performance for terminating and steady-state simulations by means of an example for which the true measures of performance can be analytically computed. (This would not be possible, of course, for most complex real-world simulations.)

Consider the output process $\{D_i, i \geq 1\}$ for the $M/M/1$ queue, where D_i is the delay in queue (exclusive of service time) of the ith arriving customer. This is a single-server queueing system with i.i.d. exponential interarrival times with parameter $1/\lambda$, i.i.d. exponential service times with parameter $1/\mu$, and customers are served in a first-in, first-out manner. Assume that the traffic intensity $\rho = \lambda/\mu < 1$. The objective of a terminating simulation of the $M/M/1$ queue might be to estimate the *expected average delay of the first m customers* (i.e., the terminating event is $E = \{m$ customers have completed their delays$\}$) given some initial condition, say, that the number of customers in the system at time 0, $N(0)$, is zero. The desired quantity, which we denote by $d(m|N(0) = 0)$ (the vertical line is read "given that"), is then given by

$$d(m|N(0) = 0) = E\left[\sum_{i=1}^{m} D_i/m|N(0) = 0\right] . \qquad (1)$$

Although the expression on the right-hand side (r.h.s.) of (1) might seem imposing at a first glance, its interpretation is really quite simple: The r.v. $X = \sum_{i=1}^{m} D_i/m$ is just the average delay of the first m customers and we are interested in estimating $E(X)$ given that $N(0) = 0$. Since we may think of $E(X)$ as the average of the X's resulting from making a very large (infinite) number of independent simulation runs each of length m customers, one legitimate question is to ask how many independent runs of length m customers each are required to get a good estimate of $E(X)$. This issue is taken up in Section 5.

Note that measures of performance for terminating simulations explicitly depend on the state of the system at time 0. *In particular,* $d(m|N(0) = \ell_1) \neq d(m|N(0) = \ell_2)$ for $\ell_1 \neq \ell_2$.

The objective of a steady-state simulation of $\{D_i, i \geq 1\}$ for the $M/M/1$ queue would be to estimate the *steady-state expected average delay* d, which is given by

$$d = \lim_{m\to\infty} d(m|N(0) = \ell) \quad \text{for any} \quad \ell = 0,1,\ldots . \qquad (2)$$

If $\rho < 1$, as we assume, then d exists (i.e., the limit exists and is finite). (If, however, $\rho > 1$, then customers are arriving faster, on the average, then they are served. As time gets large, the length of the queue will get longer and longer, and the r.h.s. of (2) will diverge to plus infinity.)

Observe in (2) that d is independent of the state of the system at time 0, $N(0)$. In Figure 1 we plot $d(m|N(0) = 0)$ (see Heathcote and Winer (1969)) as a function of m. (The arrival rate $\lambda = 1$ and the service rate $\mu = 10/9$, so $\rho = 0.9$.) The horizontal line that $d(m|N(0) = 0)$ asymptotically approaches is at height $d = 8.1$ (see Gross and Harris 1974, p. 58). Note that $d(m|N(0) = 0)$ is small for small values of m because $N(0)$ was artifically chosen to be zero.

3.2. The Meaning of Steady State

In the above queueing example, d was defined as the limit (as the number of customers m goes to infinity) of the expected average delay of the first m customers. (That definition was convenient there because it allowed us to relate d to $d(m|N(0) = \ell)$.) The difficulty with this definition is that it implies that d is, in effect, the average delay over an infinite number of simulation runs, each of infinite duration. We therefore give a more pragmatic definition of d. If d, as defined by (2), exists and is finite, then d is also given by the following expression:

$$d = \lim_{m\to\infty} \sum_{i=1}^{m} D_i/m \quad \text{(with probability 1) for any } N(0) = \ell . \qquad (3)$$

We now drop the adjective "expected" and call d the *steady-state average delay in queue*. What (3) says is that if one performs an infinite number of simulation runs, each resulting in a

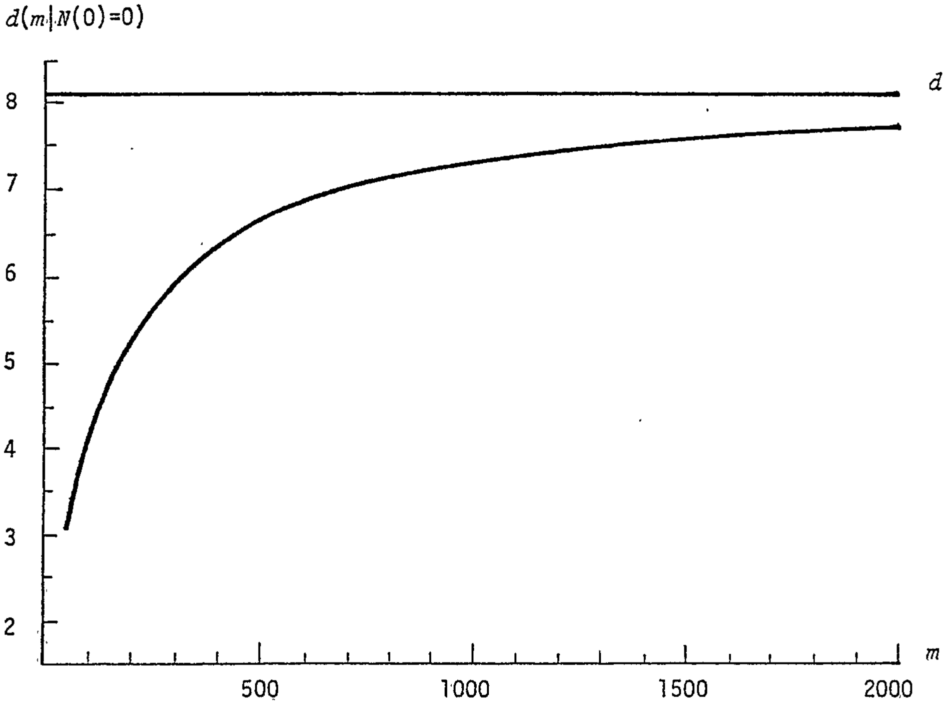

Figure 1. $d(m|N(0)=0)$ as a function of m for the $M/M/1$ queue with $\rho = 0.9$.

$\bar{D}(m) = \sum_{i=1}^{m} D_i/m$, and m is sufficiently large, then $\bar{D}(m)$ will be arbitrarily close to d for virtually (i.e., with probability one (w.p. 1)) all simulation runs. We can therefore think of d as the average delay resulting from making *one* sufficiently long simulation run. In the remainder of this paper, we will define steady-state measures of performance similarly to the manner in which d is defined in (3) (see Section 6).

For the queueing system considered above, we plot in Figure 2 $\bar{D}(m)$ as a function of m (computed from a single simulation run) and also d. Note that the convergence of $\bar{D}(m)$ to d as m goes to infinity is certainly not monotone as in Figure 1. Observe, in addition, that $\bar{D}(m)$ (as a function of m) still exhibits some random fluctuation even for m as large as 5000.

In many books and papers on simulation, a statement is made such as "It is desired to estimate some measure of performance for a system that is operating in steady state." Since we believe that this statement is not well understood, we now attempt to shed some light on the meaning of steady state. For the $M/M/1$ queue let

$$F_{i,\ell}(x) = P\{D_i \le x | N(0) = \ell\} .$$

We call $F_{i,\ell}(x)$ the *transient distribution of delay at time* i given $N(0) = \ell$. (The word "transient" means that there is a different distribution for each time i.) Now it can be shown that for any $x \ge 0$,

$$F(x) = \lim_{i\to\infty} F_{i,\ell}(x) \quad \text{for any } N(0) = \ell \tag{4}$$

exists, and we call $F(x)$ the *steady-state distribution of delay*. It follows from (4) that there exists a time index i' such that for all $i \ge i'$, $F_{i,\ell}(x) \approx F(x)$ for all $x \ge 0$. At the point in time when $F_{i,\ell}(x)$ is essentially no longer changing with i, we will intuitively say that the

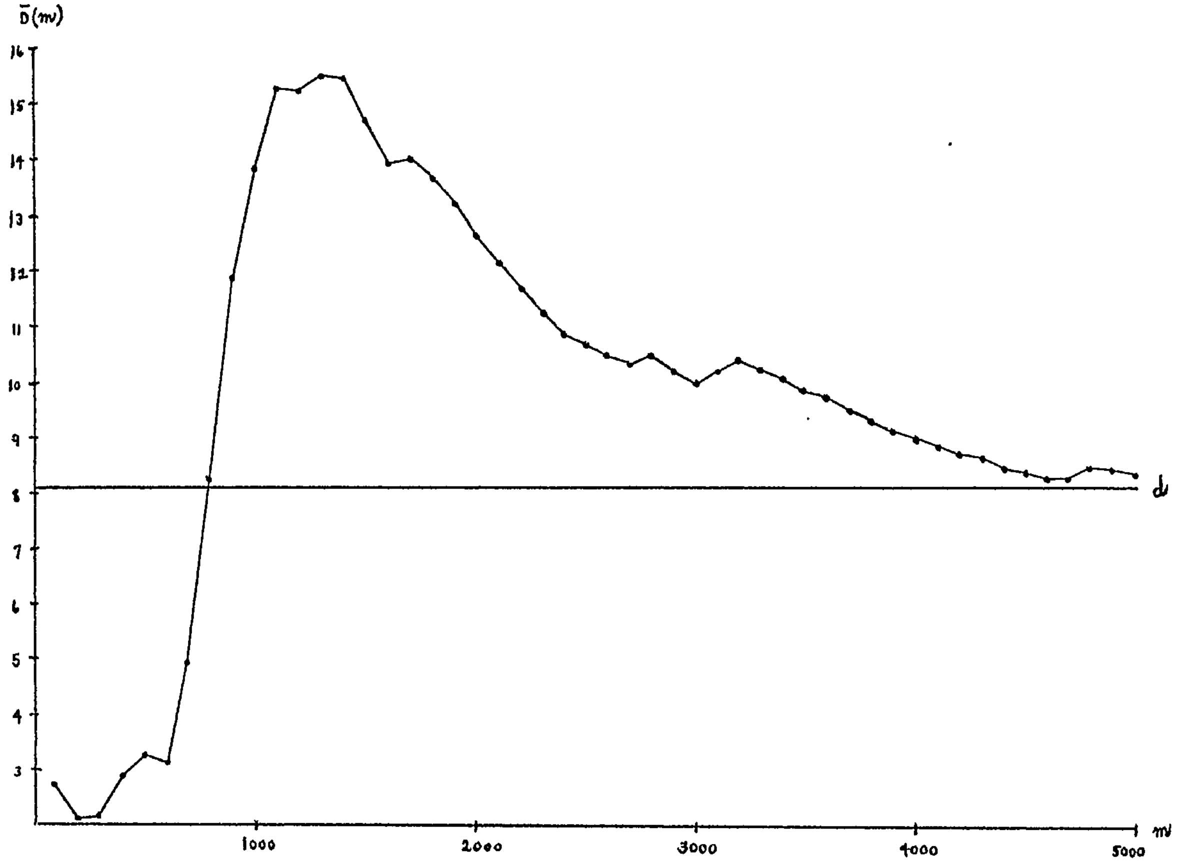

Figure 2. A realization of $\bar{D}(m)$ as a function of m for the $M/M/1$ queue with $\rho = 0.9$.

process $\{D_i, i \geq 1\}$ is in "steady state." Thus, steady state does *not* mean that the *actual* delays in a single realization (or run) of the simulation become constant after some point in time, but rather that the *distribution* of the delays becomes invariant.

Let D be the delay of a customer who arrives after the process $\{D_i, i \geq 1\}$ is in "steady state." Then it can also be shown that $E(D) = d$. Although we have given three definitions of d, we will (as stated above), henceforth use the one given by (3).

4. THE NEED FOR CONFIDENCE INTERVALS

Suppose we would like to estimate $d(25|N(0) = 0) = 2.124$ for the $M/M/1$ queue with $\rho = 0.9$. The following are ten independent realizations (i.e., different random numbers were used for each realization) of the r.v. $\sum_{i=1}^{25} D_i/25$ given $N(0) = 0$:

1.051, 6.438, 2.646, 0.805, 1.505, 0.546, 2.281, 2.822, 0.414, 1.307.

Note that the estimators range from a minimum of 0.414 to a maximum of 6.438 and that most of the estimators are not very close to the true answer, 2.124. We conclude that one realization, or replication, is generally not sufficient to obtain an acceptable estimate of a measure of performance and that a method is needed for ascertaining how close an estimator is to the true measure. The usual approach to assessing the accuracy of an estimator is to construct a c.i. for the true measure.

Although the above discussion was oriented toward terminating simulations, the same conclusion is valid for steady-state simulations; namely, that one needs a way of assessing the accuracy of an estimator and that a c.i. is the usual approach.

5. CONFIDENCE INTERVALS FOR TERMINATING SIMULATIONS

Suppose we make n independent replications of a terminating simulation, where the length of each replication is determined by the specified event E and each replication is begun with the same initial conditions. The independence of replications is accomplished by using different random numbers for each replication. Assume for simplicity that there is a single performance measure of interest; the more general case is discussed in Law and Kelton (1981c). If X_j is the estimator of the measure of performance from the jth replication, then the X_j's are i.i.d. r.v.'s and classical statistical analysis may be used to construct a c.i. for $\mu = E(X)$. For the $M/M/1$ queue discussed above, X_j might be the average delay $\sum_{i=1}^{m} D_i/m$ from the jth replication. For an inventory system with a planning horizon of m months, X_j might be the average cost $\sum_{i=1}^{m} C_i/m$ from the jth replication (C_i is the total cost in the ith month).

5.1. Fixed Sample Size Procedure

The usual approach to constructing a c.i. for μ is to make a fixed number of replications $n(n \geq 2)$. If the estimators $X_1, X_2, \ldots, X_n$ are assumed to be normal r.v.'s, in addition to being i.i.d., then a $100(1 - \alpha)\%$ $(0 < \alpha < 1)$ c.i. for μ is given by

$$\bar{X}(n) \pm t_{n-1,1-\alpha/2} \sqrt{s^2(n)/n} \ , \tag{5}$$

where $\bar{X}(n) = \sum_{j=1}^{n} X_j/n$ is the sample mean, $s^2(n) = \sum_{j=1}^{n} [X_j - \bar{X}(n)]^2/(n - 1)$ is the sample variance, and $t_{n-1,1-\alpha/2}$ is the upper $1 - \alpha/2$ critical point for a t distribution with $n - 1$ degrees of freedom. Note that (5) is the same expression that is used in classical statistics to construct a c.i. for the mean of a population.

Suppose that we would like to construct a 90% c.i. for $d(25|N(0) = 0)$ in the case of the $M/M/1$ queue with $\rho = 0.9$. From the ten replications presented in Section 4, we obtained

$$\bar{X}(10) = 1.982 \ \text{ and } \ s^2(10) = 3.172 \ .$$

Then an approximate 90% c.i. for $d(25|N(0) = 0)$ is given by

$$\bar{X}(10) \pm t_{9,.95}\sqrt{s^2(10)/10} = 1.982 \pm 1.032 \ .$$

Thus, subject to the correct interpretation to be given to c.i.'s, we can claim with approximately 90% confidence that $d(25|N(0) = 0)$ is contained in the interval $[0.950, 3.014]$.

Suppose that we construct a very large (an infinite) number of $100(1 - \alpha)\%$ c.i.'s for μ using (5), with each c.i. being based on n replications. We call the proportion of c.i.'s which actually contain (cover) μ the *coverage* for the c.i. If the X_j's are normally distributed, then the coverage will be exactly $1 - \alpha$. Alternatively, if the number of replications n for each c.i. is "sufficiently large," then we know by the central limit theorem (c.l.t.) that the coverage will be very close to $1 - \alpha$. In practice, the X_j's which result from a simulation will rarely be exactly normally distributed nor will we know how to choose n sufficiently large. As a result, the actual coverage of the c.i. given by (5) may be somewhat *less* than the desired $1 - \alpha$; this is why we called the c.i. in the above example an *approximate* 90% c.i. If, however, an X_j is the average of a large number of individual data points (as in the example above), then our experience indicates that the degradation in coverage will not be very severe. Fortunately, many real-world simulations produce X_j's of this type. See Law (1980d) and (Law and Kelton 1981c) for further discussion.

5.2. Obtaining Confidence Intervals with a Specified Precision

One disadvantage of the fixed sample size approach to constructing a c.i. is that the simulator has no control over the c.i. half-length (i.e., $t_{n-1,1-\alpha/2}\sqrt{s^2(n)/n}$); for fixed n, the half-length will depend on the population variance of the X_j's, $\sigma^2(X)$. In the example of Subsection 5.1, the half-length of 1.032 (based on $n = 10$ replications) was probably too large to get an accurate idea of the true value of $d(25|N(0) = 0)$. In this subsection we briefly discuss procedures for determining the number of replications required to obtain a c.i. with a specified precision.

There are two principal ways of measuring the precision of a c.i. We will call the *actual* c.i. half-length the *absolute precision* of the c.i., and we will call the *ratio* of the c.i. half-length to the magnitude of the point estimator (i.e., $\bar{X}(n)$) the *relative precision* of the c.i. (Although not strictly correct, one can think of the relative precision as the "proportion" of μ by which $\bar{X}(n)$ differs from μ.) In Law (1980d) and Law and Kelton (1981c), procedures are given for obtaining a c.i. with a specified absolute precision or relative precision. These procedures are *sequential* in that they add new replications one at a time until a c.i. with the specified precision has been obtained.

5.3. Recommended Use of the Procedures

We now make our recommendations on the use of the fixed sample size and sequential procedures for terminating simulations. If one is performing an exploratory experiment where the precision of the c.i. may not be overwhelmingly important, then we recommend using the fixed sample size procedure. However, if the X_j's are highly nonnormal and the number of replications n is too small, then the actual coverage of the constructed c.i. may be somewhat lower than that desired.

From an exploratory experiment consisting of n replications, one can estimate the cost per replication and the population variance of the X_j's, and then estimate from formulas in Law and Kelton (1981c) the number of replications required for a desired absolute precision or relative precision. Sometimes the precision desired might have to be tempered by the cost associated with the required number of replications. If it is finally decided to construct a c.i. with a small absolute or relative precision, then the sequential procedures mentioned in Subsection 5.2 are recommended. It should be noted that almost all of the statistical analyses for terminating simulations thus far discussed in Section 5 can be automatically performed in SIMSCRIPT II.5 using a library routine called STAT.R (see Law 1979c).

Depending on the complexity of the system of interest, the cost of making one replication of a simulation model may range from less than one dollar per replication to an extreme of $500 or even more. Thus precise c.i.'s may simply not be affordable. Regardless of the cost per replication, we recommend always making at least three replications of the simulation to assess the variability of the X_j's. (With two replications it is possible to get X_1 and X_2 very close together even though the X_j's are highly variable.) *If X_1, X_2, and X_3 are not very close together, then additional replications must be made or any conclusions derived from the simulation study will probably be of doubtful validity.*

6. CONFIDENCE INTERVALS FOR STEADY-STATE SIMULATIONS

Let $Y_1, Y_2, \ldots$ be an output process resulting from a *single* simulation run. (For example, Y_i might be the delay of the ith customer, D_i, for a queueing system or the total cost in the ith month, C_i, for an inventory system.) Then define the *steady-state average response* ν of $\{Y_i, i \geq 1\}$ (when it exists) by

$$\nu = \lim_{m\to\infty} \sum_{i=1}^{m} Y_i/m \quad (\text{w.p. } 1) .$$

(This definition is consistent with the definition of d given by (3).) We also assume that the limit ν is independent of the state of the simulation at time zero.

There have been two general approaches suggested in the simulation literature for constructing a c.i. for ν:

(i) Fixed sample size procedures - A simulation run of an *arbitrary* fixed length is performed and then one of a number of available procedures is used to construct a c.i. from the available data.

(ii) Sequential procedures - The length of a simulation is sequentially increased until an "acceptable" c.i. can be constructed. There are several techniques for deciding when to stop collecting data.

These two general approaches are discussed in further detail in the next two subsections.

6.1. Fixed Sample Size Procedures

There have been five fixed sample procedures suggested in the literature (see Law and Kelton 1979b for a survey). In this subsection we discuss two of these five procedures, namely, batch means and replication. Both procedures break the output data $Y_1, Y_2, \ldots$ into (approximately) i.i.d. "observations" to which classical statistical analyses can be applied to construct a c.i. for ν.

Batch Means

Assume temporarily that $\{Y_i, i \geq 1\}$ is a covariance stationary process with $E(Y_i) = \nu$ for all i. (For a covariance stationary process, all observations have the same mean and the same variance, and the covariance between any two observations depends only on the separation between the observations.) Suppose we make a simulation run of length m and then divide the resulting observations $Y_1, Y_2, \ldots, Y_m$ into n batches of length ℓ. (Assume that $m = n \cdot \ell$.) Thus, batch 1 consists of observations $Y_1, \ldots, Y_\ell$, batch 2 consists of observations $Y_{\ell+1}, \ldots, Y_{2\ell}$, etc. Let $\bar{Y}_j(\ell)$ $(j = 1,2,\ldots,n)$ be the sample (or batch) mean of the ℓ observations in the jth batch and let $\bar{\bar{Y}}(n,\ell) = \sum_{j=1}^{n} \bar{Y}_j(\ell)/n = \sum_{i=1}^{m} Y_i/m$ be the grand sample mean. We will use $\bar{\bar{Y}}(n,\ell)$ as our point estimator for ν. (The $\bar{Y}_j(\ell)$'s will eventually play the same role for batch means as did the X_j's for the fixed sample size c.i. in Subsection 5.1.)

If we choose the batch size ℓ large enough, then it can be shown that the $\bar{Y}_j(\ell)$'s will be approximately uncorrelated (see Law and Carson 1979). Suppose we can choose ℓ large enough so that, in addition, the $\bar{Y}_j(\ell)$'s are approximately normally distributed. This is not implausible since there are c.l.t.'s for certain types of correlated stochastic processes (see Anderson 1971, p. 427). Also it can be shown that the sample mean of the first ℓ delays, $\bar{D}(\ell)$, for the $M/M/1$ queue will be approximately normally distributed if ℓ is large (see Law 1974a). However, if the $\bar{Y}_j(\ell)$'s are both uncorrelated and normally distributed, then it can be shown that the $\bar{Y}_j(\ell)$'s are independent and normally distributed. Denote these two properties by (P1).

Since $Y_1, Y_2, \ldots$ is assumed to be covariance stationary with $E(Y_i) = \nu$, it easily follows that the $\bar{Y}_j(\ell)$'s have the same mean ν and the same variance; denote these properties by (P2).

It follows from (P1) and (P2) that the $\bar{Y}_j(\ell)$'s are normal r.v.'s with the same mean and variance. Since a normal r.v. is completely determined by its mean and variance, it in turn follows that the $\bar{Y}_j(\ell)$'s are identically distributed with mean ν, which we denote by (P3). Therefore, if the batch size ℓ is large enough, it follows from (P1) and (P3) that it is not unreasonable to treat the $\bar{Y}_j(\ell)$'s as if they were i.i.d. normal r.v.'s with mean ν and to construct an approximate $100(1 - \alpha)\%$ c.i. for ν from

$$\bar{\bar{Y}}(n,\ell) \pm t_{n-1,1-\alpha/2} \sqrt{s^2_{\bar{Y}_j(\ell)}(n)/n} \,, \tag{6}$$

where

$$s^2_{\bar{Y}_j(\ell)}(n) = \sum_{j=1}^{n} [\bar{Y}_j(\ell) - \bar{\bar{Y}}(n,\ell)]^2/(n - 1) \,.$$

Expression (6) is analogous to the c.i. given by (5) for terminating simulations.

There are three potential sources of error when one uses (6) to construct a c.i. for ν:

(i) The fact that $Y_1, Y_2, \ldots$ will rarely, if ever, be covariance stationary in practive. However, if ν exists, then in general $Y_{k+1}, Y_{k+2}, \ldots$ will be approximately covariance stationary if k is large enough. (Thus it may be prudent to delete some data from the beginning of the simulation run before applying batch means.)

(ii) If ℓ is not large enough, then the $\bar{Y}_j(\ell)$'s may not be approximately normally distributed.

(iii) If ℓ is not large enough, then the $\bar{Y}_j(\ell)$'s may be highly correlated and $s^2_{\bar{Y}_j(\ell)}(n)/n$ will be a severely biased estimator of $\sigma^2[\bar{\bar{Y}}(n,\ell)]$; see Law (1977b). In particular, if the Y_i's are positively correlated (as is often the case in practice), then the $\bar{Y}_j(\ell)$'s will be too, resulting in the variance estimate's being biased low and the c.i.'s being too small.

In order to see how well the method of batch means works in practice, Law and Kelton (1979b) simulated the $M/M/1$ queue with $\rho = 0.8$ and a model of a time-shared computer system. The measures of performance of interest were, respectively, the steady-state average delay and the steady-state average response time, both of which can be computed analytically. They performed 200 independent simulation experiments for each stochastic model and in each experiment their goal was to construct a 90% c.i. for the desired measure of performance. Not knowing how to choose definitively the total sample size m and the number of batches n, they arbitrarily chose m = 320, 640, 1280, 2560 and n = 5, 10, 20, 40. Thus, for each experiment with each model they constructed sixteen different c.i.'s using batch means. They then computed the proportion of the 200 c.i.'s which covered the known measure of performance in each of the sixteen cases for each model. They found that if the total sample size m is chosen too small, then the actual coverages for batch means may be considerably lower than the desired 0.90. For example, when $m = 1280$ and $n = 10$ for the $M/M/1$ queue, the estimated coverage was 0.74. Their results also suggested that if m was chosen large enough, then batch means would produce coverages close to the desired level; however, the "appropriate" choice of m appeared to be extremely model dependent.

At the same time that the above experiments were being performed, Law and Kelton also tested the other fixed sample size procedures (with the exception of replication). They found that these procedures also did not perform well in terms of coverage if the total sample size m is chosen too small.

Replication

The reader may have wondered why a simulator could not do a terminating simulation-type analysis of an output process $Y_1, Y_2, \ldots$ to estimate the steady-state average response ν. To illustrate the danger of making independent replications (each starting from the same initial conditions) and of doing a terminating analysis of a system for which we *really* want to estimate a *steady-state* measure of performance, consider the $M/M/1$ queue with $\rho = 0.9$. Suppose we want to estimate the steady-state average delay $d = 8.1$, and make n independent replications each of length $m = 320$ customers and each with $N(0) = 0$. Since $E(X_j) = d(320|N(0) = 0) = 6.01$ (see Figure 1), $E[\bar{X}(n)] = 6.01$ and $\bar{X}(n)$ is a biased estimator of d no matter how many replications are made. (Here X_j is the average delay of the 320 customers in the jth replication.) Furthermore, as n gets large the length of the c.i. constructed from (5) will become smaller and smaller, and the coverage of the c.i. eventually approaches zero (see Law 1977b). We are actually constructing a c.i. for $d(320|N(0) = 0)$, not d.

In looking at Figure 1, it becomes clear why the above terminating analysis does not perform well in the steady-state case. Because of a simulator's inability to start the simulation off at time 0 in a state which is representative of the steady-state behavior of the system (see Subsection 3.2), the output data at the beginning of the simulation are not "good" estimates of the steady-state average response ν. (This difficulty has been called the "startup problem" or the problem of the "initial transient" in the simulation literature.) This suggests "warming up" the simulation for some amount of time, say k observations, before beginning data collection. The difficulty is in knowing how to choose k; a survey paper by Gafarian, Ancker, and Morisaku (1978) indicates that no published procedure performs at all well in practice.

6.2. Sequential Procedures

We saw in the last subsection that fixed sample size procedures cannot, in general, be relied upon to produce c.i.'s with coverages close to the desired level. The results were encouraging, however, in that they indicated these procedures would perform well provided that enough data were available. In this section we discuss briefly sequential procedures for constructing a c.i. for a steady-state average response ν which determine the amount of data required during the course of a simulation run.

In (1978a), Law and Kelton surveyed the published sequential procedures and found that only two of these procedures performed well when tested on a variety of stochastic models with a known value of ν. One procedure, which was developed by Fishman (1977c), is based on the regenerative method (see Crane and Iglehart 1974a, 1974b, 1975c, 1975d, Fishman 1973a, 1974b, 1978d, and Crane and Lemoine 1977) and, thus, we feel has limited applicability to most real-world problems at the present time. The other procedure (see Law and Carson 1979), which is based on idea of batch means, appears to have greater applicability. It also performed well for each of the 13 stochastic models on which it was tested.

REFERENCES

Anderson, T. W. (1971), *The Statistical Analysis of Time Series*, Wiley, New York.

Crane, M. A. and D. L. Iglehart (1974a), "Simulating stable stochastic systems, I: General multiserver queues," *J. Assoc. Comput. Mach. 21*, pp. 103-113.

Crane, M. A. and D. L. Iglehart (1974b), "Simulating stable stochastic systems, II: Markov chains," *J. Assoc. Comput. Mach. 21*, pp. 114-123.

Crane, M. A. and D. L. Iglehart (1975c), "Simulating stable stochastic systems, III: Regenerative processes and discrete-event simulations," *Operations Res. 23*, pp. 33-45.

Crane, M. A. and D. L. Iglehart (1975d), "Simulating stable stochastic systems, IV: Approximation techniques," *Management Science 21*, pp. 1215-1224.

Crane, M. A. and A. J. Lemoine (1977), *An Introduction to the Regenerative Method for Simulation Analysis,* Lecture Notes in Control and Information Sciences, Volume 4, Springer-Verlag, New York.

Fishman, G. S. (1973a), "Statistical analysis for queueing simulations," *Management Science 20*, pp. 363-369.

Fishman, G. S. (1974b), "Estimation in multiserver queueing simulations," *Operations Res. 22*, pp. 72-78.

Fishman, G. S. (1977c), "Achieving specific accuracy in simulation output analysis," *Comm. Assoc. Comput. Mach. 20*, pp. 310-315.

Fishman, G. S. (1978d), *Principles of Discrete Event Simulation*, Wiley, New York.

Gafarian, A. V., C. J. Ancker, Jr., and T. Morisaku (1978), "Evaluation of commonly used rules for detecting "steady state" in computer simulation," *Naval Res. Logist. Quart. 25*, pp. 511-529.

Gross, D. and C. M. Harris (1974), *Fundamentals of Queueing Theory*, Wiley, New York.

Heathcote, C. R. and P. Winer (1969), "An approximation for the moments of waiting times," *Operations Res. 17*, pp. 175-186.

Law, A. M. (1974a), *Efficient estimators for simulated queueing systems*, ORC 74-7, Operations Research Center, University of California, Berkeley.

Law, A. M. (1977b), "Confidence intervals in discrete event simulation: A comparison of replication and batch means," *Naval Res. Logist. Quart. 24*, pp. 667-678.

Law, A. M. (1979c), "Statistical Analysis of Simulation Output Data with SIMSCRIPT II.5," CACI, Inc., Los Angeles.

Law, A. M. (1980d), "Statistical analysis of the output data from terminating simulations," *Naval Res. Logist. Quart. 27*, pp. 131-143.

Law, A. M. and J. S. Carson (1979), "A sequential procedure for determining the length of a steady-state simulation," *Operations Res. 27*, pp. 1011-1025.

Law, A. M. and W. D. Kelton (1978a), *Confidence intervals for steady-state simulations, II: A survey of sequential procedures*, Technical Report No. 78-6, Department of Industrial Engineering, University of Wisconsin, Madison.

Law, A. M. and W. D. Kelton (1979b), *Confidence intervals for steady-state simulations, I: A survey of fixed sample size procedures*, Technical Report No. 78-5, Department of Industrial Engineering, University of Wisconsin, Madison.

Law, A. M. and W. D. Kelton (1981c), *Simulation Modeling and Analysis*, to be published by McGraw-Hill, Inc., New York.