Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

SIMULATION OUTPUT ANALYSIS

Marvin K. Nakayama

Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102, U.S.A.

ABSTRACT

We discuss methods for statistically analyzing the output from stochastic simulations. Both terminating and steady-state simulations are considered.

1 INTRODUCTION

Many simulations include some sort of randomness, which can arise in a variety of ways. For example, in a simulation of a manufacturing system, the processing times required at a station may have random variations or the arrival times of new jobs may not be known in advance. In a bank, customers arrive at random times and the amount of time spent at a teller is not known beforehand. In financial simulations, future returns are unknown.

Because of the randomness in the components driving a simulation, the output from the simulation is also random, so statistical techniques must be used to analyze the results. The data-analysis methods taught in an introductory statistics course typically assume that the data are independent and identically distributed (i.i.d.) with a normal distribution, but the output data from a simulation are often not i.i.d. normal. For example, consider customer waiting times before seeing a teller in a bank. If one customer has an unusually long waiting time, then the next customer probably also will, so the waiting times of the two customers are dependent. Moreover, customers arriving during the lunch hour will usually have longer waiting times than customers coming in at other times, so waiting times are not identically distributed throughout the day. Finally, waiting times are always positive and often skewed to the right, with a possible mode at zero, so waiting times are not normally distributed. For these reasons one often cannot analyze simulation output using the classical statistical techniques developed for i.i.d. normal data.

In this tutorial, we will examine some statistical methods for designing and analyzing simulation experiments. In the next section we begin by distinguishing between two types of performance measures: terminating (or transient) and steady-state (or infinite-horizon or long-run). These two types of measures require different statistical techniques to analyze the results, and Section 3 reviews methods for analyzing output from terminating simulations, while Section 4 covers techniques for steady-state simulations. In Section 5 we discuss the estimation of multiple performance measures, and Section 6 briefly covers some other useful methods for analyzing simulation output. Some concluding remarks are given in Section 7.

2 PERFORMANCE MEASURES

One of the first steps in any simulation study is choosing the *performance measure(s)* to calculate. In other words, what measures will be used to evaluate how "good" the system is? For example, we may measure how well a queueing system performs by its expected number of customers served in a day, or we may use the long-run average daily cost as a measure of the performance of a supply chain.

There are primarily two types of performance measures for stochastic systems, which we now briefly describe:

1. *Transient performance measures*, also known as *terminating* or *finite-horizon* measures, evaluate how the system evolves over a finite time horizon.
2. *Steady-state performance measures* describe how the system evolves over an infinite time horizon. These are also known as *long-run* or *infinite-horizon* measures.

A simulation in which a transient (resp., steady-state) measure is being estimated is called a *transient simulation* (resp., *steady-state simulation*). We now describe these concepts in more depth.

2.1 Transient Performance Measures

Definition: A *terminating simulation* is one for which there is a "natural" event B that specifies the length of time in which one is interested for the system. The event B often occurs either at a time point beyond which no useful information is obtained, or when the system is "cleaned out." For example, if we are interested in the performance of a system during the first 10 time units of operation of a day, then B would denote the event that 10 time units of system time have elapsed. If we want to determine the first time at which a queue has at least 8 customers, then B is the event of the first time the queue length reaching 8.

Since we are interested in the behavior of the system over only a finite time horizon, the "initial conditions" $\mathcal{I}$ (i.e., conditions under which the system starts) can have a large impact on the performance measure. For example, queueing simulations often start with no customers present, which would be the $\mathcal{I}$ in this setting.

In a transient simulation, we have the following
Goal: To calculate

$$\mu = E(X), \tag{1}$$

where X is a random variable representing the performance of the system over some finite horizon.

We now examine some examples of transient performance measures.
Example: Consider a bank vestibule containing an automatic teller machine (ATM). The vestibule is only open during normal banking business hours, which is 9:00am to 5:00pm, so customers can access the ATM only during those times. Any customers in the vestibule at 5:00pm will be allowed to complete their transactions, but no new customers will be allowed in. Let Z be the number of customers using the ATM in a day, and we may be interested in determining the following terminating performance measures:

- $E[Z]$, the expected value of Z. Here, to put things in the framework of (1), we set $X = Z$.
- $P\{Z \geq 500\} = E[I(Z \geq 500)]$, which is the probability that at least 500 customers use the ATM in a day, where $I(A)$ is the indicator function of the event A, which takes on the value 1 if A occurs, and 0 otherwise. In this case, in the notation of (1), $X = I(Z \geq 500)$.

The initial conditions $\mathcal{I}$ might be that the system starts out empty each day, and the terminating event B is that it is past 5:00pm and there are no more customers in the vestibule.

Alternatively we might define Z to be the average waiting time (in seconds) of the first 50 customers in a day. We can define the following performance measures:

- $E[Z]$, the expected value of Z. In this case, $X = Z$ in the notation of (1).
- $P\{Z \leq 30\} = E[I(Z \leq 30)]$, which is the probability that the average waiting time of the first 50 customers is no more than 30 seconds. Here, $X = I(Z \leq 30)$ in (1).

In this case we might specify the initial conditions $\mathcal{I}$ to be that the system starts out empty each day, and the terminating event B is that 50 customers have finished their transactions.

2.2 Steady-State Performance Measures

Now we consider steady-state performance measures. Let $\mathbf{Y} = (Y_1, Y_2, Y_3, \ldots)$ be a (discrete-time) stochastic process representing the output of a simulation. For example, if the vestibule containing the ATM in our previous example is now open 24 hours a day, then Y_i might represent the waiting time of the ith customer since the ATM was installed. Let $F_i(y|\mathcal{I}) = P(Y_i \leq y|\mathcal{I})$ for $i = 1, 2, \ldots$, where as before, $\mathcal{I}$ represents the initial conditions of the system at time 0. Observe that $F_i(\cdot\,|\mathcal{I})$ is the distribution function of Y_i given the initial conditions $\mathcal{I}$.
Definition: If

$$F_i(y|\mathcal{I}) \to F(y) \text{ as } i \to \infty \tag{2}$$

for all y and for any initial conditions $\mathcal{I}$, then $F(y)$ is called the *steady-state distribution* of the process $\mathbf{Y}$. If Y is a random variable with distribution F, we say that Y has the steady-state distribution, and we sometimes write this as $Y_i \xrightarrow{\mathcal{D}} Y$ as $i \to \infty$, which is read as "Y_i converges in distribution to Y."

In (2), we are considering a limit as $i \to \infty$. In practice, this often starts to approximately hold for finite values of i. Note that (2) states that for large i, the *distribution* of Y_i is close to F, not that the *values* of the Y_i are all the same. When Y is a random variable with distribution F, $E(Y)$ is a *steady-state performance measure*. It can be shown under appropriate moment conditions that $E(Y_i|\mathcal{I}) \to E(Y)$ as $i \to \infty$ for all initial conditions $\mathcal{I}$ when (2) holds. Figure 1 gives an example of density functions f_i approaching some limiting density f as i gets larger.

Many systems do not have a steady state. For example, consider our previous example of an ATM that is accessible only during business hours. Let Y_i be the waiting time of the ith customer to arrive since the ATM was installed. Then, the process $\mathbf{Y}$ does not have a steady state because the first customer of each day always has no wait, whereas other customers may have to wait. For example, suppose 500 customers are served on the first day, so day 2 begins with customer 501, who has no wait since there is no one ahead of him on that day. On the other hand, if the ATM were accessible 24 a day, then a steady state may exist.

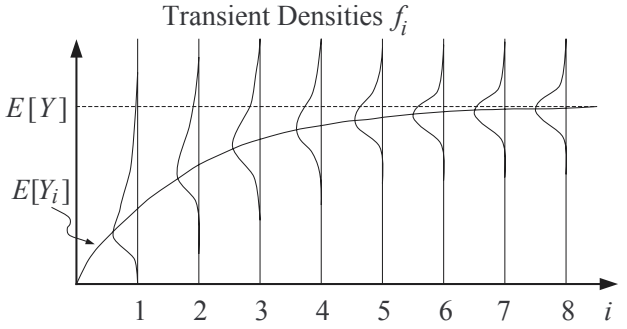

Figure 1: Densities f_i of an Output Process $(Y_1, Y_2, \ldots)$

In the above example where the ATM is only available from 9:00am to 5:00pm, we may be able to obtain a process $\mathbf{Y}$ that does have a steady state if we define the Y_i differently. In particular, suppose Y_i is the average waiting time of all the customers on the ith day since the ATM first became operational. Then, $\mathbf{Y}$ may have a steady state. (It still may not if the distribution of the number of customers in a day depends on the particular day of the week, or if there are seasonal variations.)

Example: Consider the ATM from before, but now suppose that it accessible all the time. Let Y_i be the number of customers served on the ith day of operation, and suppose that over time, the system "settles down" into steady state; i.e., $Y_i \xrightarrow{\mathcal{D}} Y$ as $i \to \infty$. We now may be interested in determining the following steady-state performance measures:

- $E[Y]$, which is the expected steady-state number of customers served in a day;
- $P\{Y \geq 400\} = E[I(Y \geq 400)]$, which is the steady-state probability that at least 400 customers are served in a day.

Again, we may let the initial conditions $\mathcal{I}$ denote that the system begins operations on the first day with no customers present, and over time, the effects of the initial conditions "wash away."

3 OUTPUT ANALYSIS FOR TRANSIENT SIMULATIONS

We now discuss how to analyze the output from a transient simulation. Recall our goal is to calculate $\mu = E(X)$, where X is a random variable representing the performance of the system over some finite horizon with initial conditions $\mathcal{I}$. The basic approach to estimate μ using simulation is as follows:

Method: Generate n i.i.d. replicates of X, say $X_1, X_2, \ldots, X_n$. Form the (point) estimator

$$\bar{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{3}$$

We generate i.i.d. replicates of X by running independent simulations of the system under study. In each simulation, start the system with initial conditions $\mathcal{I}$. We make the replicates independent by using non-overlapping streams of random numbers from the random-number generator.

If $E(|X|) < \infty$ (which is almost always the case in practice), the (strong) law of large numbers guarantees $\bar{X}(n) \to \mu$ as $n \to \infty$ with probability 1. Thus, if the sample size n is chosen large enough, $\bar{X}(n)$ will be close to μ. But how close is $\bar{X}(n)$ to μ? The central limit theorem (CLT) asserts that for all x,

$$P\left\{n^{1/2}\left(\frac{\bar{X}(n)-\mu}{\sigma}\right) \leq x\right\} \to P\{N(0,1) \leq x\}, \tag{4}$$

as $n \to \infty$, provided that $0 < \sigma^2 = \mathrm{Var}(X) < \infty$, where $N(a, b)$ denotes a normal random variable having mean a and variance b, and $\mathrm{Var}(Z)$ denotes the variance of random variable Z.

To interpret the CLT for $\bar{X}(n)$ in (4), it is convenient to write it in the form

$$n^{1/2}\left(\frac{\bar{X}(n)-\mu}{\sigma}\right) \overset{\mathcal{D}}{\approx} N(0,1), \tag{5}$$

where $\overset{\mathcal{D}}{\approx}$ means "has approximately the same distribution as," which implies (after a little algebra) that

$$\bar{X}(n) \overset{\mathcal{D}}{\approx} \mu + \frac{\sigma}{\sqrt{n}}N(0,1).$$

Hence, the *error* in $\bar{X}(n)$ is given (approximately) by

$$\frac{\sigma}{\sqrt{n}}N(0,1). \tag{6}$$

Note that the error depends on the sample size n only through the factor $1/\sqrt{n}$. In particular, this means that to obtain one additional significant figure of accuracy (i.e., increase accuracy by a factor of 10), we increase the sample size n by a factor of 100. So, the estimator $\bar{X}(n)$ converges to μ rather slowly. Moreover, the error depends on the model being simulated only through the *standard deviation* $\sigma = \sqrt{\mathrm{Var}(X)}$. The larger σ is, the harder μ is to estimate. Finally, note that the error given in (6) is normally distributed.

We now use the above facts to derive a *confidence interval* for μ. First define the *level* δ; typically, one chooses $\delta = 0.1$, 0.05 or 0.01. Then, we look up in a table the constant $z_{1-\delta/2}$ for which $P\{N(0,1) \leq z_{1-\delta/2}\} = 1 - \delta/2$, e.g., $z_{0.975} = 1.96$. The values of $z_{1-\delta/2}$ for various values of δ can be found in virtually any introductory statistics

book. Recall the CLT implies that for large n,

$$\begin{aligned}
&1-\delta \\
&\quad = P\{-z_{1-\delta/2} \le N(0,1) \le z_{1-\delta/2}\} \\
&\quad \approx P\left\{-z_{1-\delta/2} \le \frac{\sqrt{n}}{\sigma}(\bar{X}(n)-\mu) \le z_{1-\delta/2}\right\} \qquad (7)\\
&\quad = P\left\{-z_{1-\delta/2}\frac{\sigma}{\sqrt{n}} \le \bar{X}(n)-\mu \le z_{1-\delta/2}\frac{\sigma}{\sqrt{n}}\right\} \qquad (8)\\
&\quad = P\left\{\bar{X}(n)-\frac{z_{1-\delta/2}\sigma}{\sqrt{n}} \le \mu \le \bar{X}(n)+\frac{z_{1-\delta/2}\sigma}{\sqrt{n}}\right\}, \qquad (9)
\end{aligned}$$

giving us an approximate $100(1-\delta)\%$ confidence interval for μ, where the approximation in (7) follows from (5). Equation (9) states that the interval

$$\left[\bar{X}(n)-\frac{z_{1-\delta/2}\sigma}{\sqrt{n}},\ \bar{X}(n)+\frac{z_{1-\delta/2}\sigma}{\sqrt{n}}\right] \qquad (10)$$

has roughly probability $1-\delta$ of containing the true mean μ.

One problem with the interval in (10) is that we typically do not know the value of the standard deviation σ. After all, we are using simulation to estimate the mean μ, so it is most likely the case that σ is also unknown. Thus, we will estimate σ^2 by the *sample variance*

$$S^2(n) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}(n)\right)^2, \qquad (11)$$

which one can show is an unbiased estimator of σ^2; i.e., $E[S^2(n)] = \sigma^2$ for any $n \ge 2$. Thus, our approach for constructing a confidence interval for μ is given below:

1. Specify a sample size n that is large. Typically, one should generally choose $n \ge 50$, but an appropriate value for n depends on the context.
2. Generate n i.i.d. replicates $X_1, X_2, \ldots, X_n$.
3. Using the n data points $X_1, X_2, \ldots, X_n$, calculate the sample mean $\bar{X}(n)$ using (3) and the sample variance $S^2(n)$ using (11).
4. The interval
$$\left[\bar{X}(n)-z_{1-\delta/2}\frac{S(n)}{\sqrt{n}},\ \bar{X}(n)+z_{1-\delta/2}\frac{S(n)}{\sqrt{n}}\right] \qquad (12)$$
is an approximate $100(1-\delta)\%$ confidence interval for μ.

If we construct the confidence interval (12) using these steps, the probability is approximately $1-\delta$ that the interval will contain μ. In other words, if we repeat these steps m independent times, this will give us m different confidence intervals. Some of them will contain (cover) μ, and others will not. The theory says that approximately $(1-\delta)m$ of the m intervals should cover μ. In practice, though, this does not always happen. The CLT only holds asymptotically as the sample size $n \to \infty$, so the coverage is only approximately $1-\delta$ for large n, i.e.,

$$\begin{aligned}
&P\left\{\mu \in \left[\ \bar{X}(n)-z_{1-\delta/2}\frac{S(n)}{\sqrt{n}},\ \bar{X}(n)+z_{1-\delta/2}\frac{S(n)}{\sqrt{n}}\ \right]\right\} \\
&\quad \approx \quad 1-\delta. \qquad (13)
\end{aligned}$$

The true probability that μ lies in the interval in (12) is known as the *coverage*.

It would be nice to know when the approximation in (13) is good, and when it is not. Fortunately, there is a theory available to help us here. Suppose that X satisfies $E(X^4) < \infty$ and X is a continuous random variable with some probability density function. Define

$$\hat{\sigma}^2(n) = n^{-1}\sum_{i=1}^{n}(X_i-\bar{X}(n))^2.$$

Note that $\hat{\sigma}^2(n)$ is not the unbiased sample variance $S^2(n)$ but is asymptotically unbiased (i.e., $E[\hat{\sigma}^2(n)] \to \sigma^2$ as $n \to \infty$). Also, for large n, $S^2(n)$ and $\hat{\sigma}^2(n)$ are essentially the same. Then, the following refinement of the CLT, the so-called Edgeworth expansion, holds:

$$\begin{aligned}
&P\left\{n^{1/2}\left(\frac{\bar{X}(n)-\mu}{\hat{\sigma}(n)}\right) \le x\right\} \\
&\quad = \quad P\{N(0,1) \le x\} \\
&\quad + \quad \frac{E[(X-\mu)^3]}{6\sigma^3\sqrt{n}}(2x^2+1)\frac{e^{-x^2/2}}{\sqrt{2\pi}} \ + \ O(1/n), \qquad (14)
\end{aligned}$$

where the $O(1/n)$ is a term that basically looks like a constant C divided by n, i.e., C/n; for details, see Hall (1987) or pp. 71–73 of Hall (1992).

Since for large sample sizes n the $O(1/n)$ term is small compared to the term before it in (14), the error in the CLT approximation is basically described by

$$\frac{E[(X-\mu)^3]}{6\sigma^3\sqrt{n}}(2x^2+1)\frac{e^{-x^2/2}}{\sqrt{2\pi}}, \qquad (15)$$

which decreases at rate $1/\sqrt{n}$ in the sample size n, which is quite slowly. Also, the error term given in (15) depends on the problem being simulated only through the quantity $\gamma \equiv E[(X-\mu)^3]/\sigma^3$, which is known as the *skewness* of X.

If X is symmetric about its mean, then the skewness γ is zero, so we can think of skewness as a measure of the *asymmetry* of X. The more symmetric X is, the better the CLT approximation is, and the more accurate the

approximation in (13) is. If X is highly asymmetric (as is typical of queueing simulations), the CLT approximation is not so good, and the coverage of the confidence interval in (12) may be significantly less than $1-\delta$. In fact, it is not unusual for confidence intervals that are supposed to have 90% coverage to cover μ, say, only 75% of the time.

3.1 Pre-Specifying Confidence Interval Widths

In the previous section we discussed so-called *fixed-sample-size methods* for estimating a transient performance measure $\mu = E[X]$, where X represents the random performance of the system over some finite time horizon. These methods are so named because the sample size is fixed before any simulations are run. However, before executing a simulation, we usually do not know how large the resulting error will be; i.e., we do not know how large the half width of the resulting confidence intervals will be.

The size of the error depends on

1. the variance of X, which is typically unknown and over which we often have no control;
2. the number n of replications to run, which we can control.

Thus, we would like a methodology that will allow us to determine the number of replications to simulate so that the resulting error (or half width of the confidence interval) is no larger than some pre-specified amount.

Goal: To refine the previously defined fixed-sample-size procedure so as to obtain an estimator for a transient performance measure $\mu = E(X)$ that satisfies one of the following:

1. **Absolute error criterion**: Here we define in advance a maximum acceptable absolute error value $\epsilon > 0$. We want our estimator $\bar{X}(n)$ to satisfy $|\bar{X}(n) - \mu| < \epsilon$ with a probability $1-\delta$. In other words, we want our estimator $\bar{X}(n)$ to be close to (i.e., within ϵ of) the true mean μ with high probability (i.e., $1-\delta$).
2. **Relative error criterion**: In this case, we define in advance a maximum acceptable relative error value $\epsilon > 0$. Now, we want our estimator $\bar{X}(n)$ to satisfy $|\bar{X}(n) - \mu| < \epsilon|\mu|$ with probability $1-\delta$; i.e., the estimator should be within $100\epsilon\%$ of the correct value with high probability.

Example: Suppose we want to estimate the expected daily withdrawals from an ATM.

1. If we want the estimator to be within \$500 of the correct value with probability $1-\delta$, then we use an absolute-error criterion with $\epsilon = 500$.
2. If we want an estimator which is within 10% of the correct value with probability $1-\delta$, then we use a relative-error criterion with $\epsilon = 0.10$.

We first consider the absolute-error criterion. Note that (8) implies that the estimator $\bar{X}(n)$ satisfies $|\bar{X}(n) - \mu| \leq z_{1-\delta/2}\sigma/\sqrt{n}$ with probability approximately $1-\delta$. Thus, if we choose the sample size n sufficiently large so that

$$z_{1-\delta/2}\frac{\sigma}{\sqrt{n}} \approx \epsilon,$$

then we find that

$$P\left\{ -\epsilon \ \leq \bar{X}(n) - \mu \ \leq \epsilon \right\} \approx 1-\delta,$$

and we have found an estimator that satisfies our desired absolute-error criterion. More precisely, let

$$N_a(\epsilon) = \left\lceil \frac{z_{1-\delta/2}^2 \sigma^2}{\epsilon^2} \right\rceil,$$

where $\lceil x \rceil$ is the least integer greater than or equal to x, which is called the *ceiling* of x. Then,

$$P\left\{ -\epsilon \ \leq \bar{X}(N_a(\epsilon)) - \mu \ \leq \epsilon \right\} \to 1-\delta$$

as $\epsilon \to 0$.

One problem is that we typically do not know the value of σ^2. Our solution is to use a *two-stage procedure* in which we first estimate σ^2 from *trial* or *pilot runs* (*first stage*), and then use the estimated variance to calculate the sample size for the *production runs* (*second stage*). The following is a variation of a procedure developed by Stein (1945).

Two-Stage Procedure for Absolute-Precision Confidence Intervals

1. Select n_0, a sample size for the set of trial runs. (In practice, one should specify $n_0 \geq 50$). Also, select an absolute precision ϵ. (In practice, one should specify the error ϵ to be "small," the meaning of which depends on the context.)
2. Generate n_0 (independent) trial runs, yielding samples $X_1, X_2, \ldots, X_{n_0}$.
3. Calculate

$$S_1^2(n_0) = \frac{1}{n_0 - 1}\sum_{i=1}^{n_0}(X_i - \bar{X}(n_0))^2,$$

which is our estimator for $\sigma^2 = \mathrm{Var}(X)$ formed from the trial runs.

4. Calculate

$$N_a(\epsilon) = \left\lceil \frac{z^2_{1-\delta/2} S_1^2(n_0)}{\epsilon^2} \right\rceil.$$

5. Generate $N_a(\epsilon)$ (independent) production runs that are independent of $X_1, X_2, \ldots, X_{n_0}$. The samples from the production runs are denoted $X_{n_0+1}, X_{n_0+2}, \ldots, X_{n_0+N_a(\epsilon)}$.
6. Set

$$\tilde{X}(\epsilon) = \frac{1}{N_a(\epsilon)} \sum_{j=n_0+1}^{n_0+N_a(\epsilon)} X_j$$

and

$$\tilde{S}^2(\epsilon) = \frac{1}{N_a(\epsilon)-1} \sum_{j=n_0+1}^{n_0+N_a(\epsilon)} (X_j - \tilde{X}(\epsilon))^2,$$

which are the sample mean and sample variance of only the values from the production runs.
7. Then

$$\left[\tilde{X}(\epsilon) - z_{1-\delta/2} \frac{\tilde{S}(\epsilon)}{\sqrt{N_a(\epsilon)}},\ \tilde{X}(\epsilon) + z_{1-\delta/2} \frac{\tilde{S}(\epsilon)}{\sqrt{N_a(\epsilon)}} \right]$$

is an approximate $100(1-\delta)\%$ confidence interval for μ, the half-width of which should be *approximately* ϵ.

If ϵ is small (as is usual in applications), then $N_a(\epsilon) \gg n_0$ (i.e., $N_a(\epsilon)$ will be much larger than n_0) so that throwing away the first n_0 observations in forming the estimators $\tilde{X}(\epsilon)$ and $\tilde{S}^2(\epsilon)$ is not going to affect the procedure much. One advantage of the above approach is that the estimators formed in Step 6 are based on the second-stage random variables $X_{n_0+1}, \ldots, X_{n_0+N_a(\epsilon)}$, which are both identically distributed and (conditionally) independent. Hence, our discussion on fixed sample size procedures basically carries over to this setting.

If we based our estimators on all the data gathered (i.e., $X_1, X_2, \ldots, X_{n_0+N_a(\epsilon)}$), fixed-sample-size methods do not really apply, since the sample size $n_0 + N_a(\epsilon)$ is a random variable correlated with some of the data gathered. In particular, the sample size and the sample mean are correlated. This can cause certain hard-to-handle problems. For example,

$$E\left[\frac{1}{n_0 + N_a(\epsilon)} \sum_{i=1}^{n_0+N_a(\epsilon)} X_i \right] \neq E(X),$$

i.e., the sample mean based on all the observations $X_1, X_2, \ldots, X_{n_0+N_a(\epsilon)}$ is *biased* for μ. On the other hand,

$$E\left[\frac{1}{N_a(\epsilon)} \sum_{i=n_0+1}^{n_0+N_a(\epsilon)} X_i \right] = E(X),$$

i.e., the sample mean based on the observations collected over the production runs is *unbiased* for μ.

We can also define a two-stage procedure to construct relative-precision confidence intervals.

Two-Stage Procedure for Relative-Precision Confidence Intervals

1. Select n_0, a sample size for the set of trial runs. (In practice, one should specify $n_0 \geq 50$). Also, select a relative precision ϵ. (In practice, one should specify $\epsilon \leq 0.10$.)
2. Generate n_0 trial runs, yielding samples $X_1, X_2, \ldots, X_{n_0}$.
3. Calculate

$$\hat{\mu} = \frac{1}{n_0} \sum_{i=1}^{n_0} X_i$$

and

$$S_1^2(n_0) = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (X_i - \bar{X}(n_0))^2,$$

which are the sample mean and sample variance from the trial runs.
4. Calculate

$$N_r(\epsilon) = \left\lceil \frac{z^2_{1-\delta/2} S_1^2(n_0)}{\hat{\mu}^2 \epsilon^2} \right\rceil.$$

5. Generate $N_r(\epsilon)$ production runs, yielding samples $X_{n_0+1}, X_{n_0+2}, \ldots, X_{n_0+N_r(\epsilon)}$.
6. Set

$$\tilde{X}(\epsilon) = \frac{1}{N_r(\epsilon)} \sum_{j=n_0+1}^{n_0+N_r(\epsilon)} X_j$$

and

$$\tilde{S}^2(\epsilon) = \frac{1}{N_r(\epsilon)-1} \sum_{j=n_0+1}^{n_0+N_r(\epsilon)} (X_j - \tilde{X}(\epsilon))^2,$$

which are the sample mean and sample variance of only the values from the production runs.

7. Then

$$\left[\tilde{X}(\epsilon) - z_{1-\delta/2}\frac{\tilde{S}(\epsilon)}{\sqrt{N_r(\epsilon)}},\ \tilde{X}(\epsilon) + z_{1-\delta/2}\frac{\tilde{S}(\epsilon)}{\sqrt{N_r(\epsilon)}}\right]$$

is an approximate $100(1-\delta)\%$ confidence interval for μ, the half-width of which should be *approximately* $\epsilon|\mu|$.

4 OUTPUT ANALYSIS FOR STEADY-STATE SIMULATIONS

We now discuss the estimation of steady-state performance measures. There are two cases to consider:

1. Discrete-time process: $\mathbf{Y} = (Y_i : i = 1, 2, \ldots)$ is an output process with a integer-valued time index, and our goal is to estimate (and produce confidence intervals for) ν, where ν is defined such that

$$\frac{1}{m}\sum_{i=1}^{m} Y_i \to \nu \tag{16}$$

as $m \to \infty$.

2. Continuous-time process: $\mathbf{Y} = (Y(t) : t \geq 0)$ is an output process with a continuous-valued time index, and we want to estimate (and produce confidence intervals for) ν, where ν is defined such that

$$\frac{1}{t}\int_0^t Y(s)ds \to \nu \tag{17}$$

as $t \to \infty$.

We previously saw in Section 2.2 some examples of steady-state measures for a discrete-time process. For example, Y_i could be the waiting time of the ith customer to a queueing system, so ν represents the steady-state expected waiting time. We now give an example of a continuous-time process.

Example: Suppose that the ATM from before is accessible 24 hours a day, and let $Y(t)$ denote the number of customers waiting in line at time t. We define the continuous-time stochastic process $\mathbf{Y} = (Y(t) : t \geq 0)$, and assuming that $\mathbf{Y}$ has a steady state (which may not be the case since the distribution of the number of customers waiting may depend on the time of day), then we may be interested in calculating ν defined in (17), which in this case is the long-run time-average number of customers waiting. Another possible measure is

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t I(Y(s) \geq a)\, ds,$$

which is the long-run fraction of time that at least a customers are waiting.

4.1 The Difficulties of Output Analysis of Steady-State Simulations

We will concentrate on discrete-time processes (continuous-time processes can be handled in a similar manner). Our goal is to estimate and produce confidence intervals for the steady-state parameter ν. First, we examine how to produce an estimator for ν. As we can see in (16), the parameter ν can be viewed as the long-run average level of Y_i. Thus, if we set

$$\bar{Y}(m) = \frac{1}{m}\sum_{i=1}^{m} Y_i,$$

then $\bar{Y}(m)$ is a *consistent* estimator for ν, in the sense that

$$\bar{Y}(m) \to \nu$$

as $m \to \infty$. In other words, running a "long" simulation will result in an estimator that is "close" to ν. Hence, the problem of constructing an estimator for ν is easily solved.

However, the task of constructing a confidence interval for ν is more delicate. For virtually all reasonably behaved systems possessing a unique steady state, one can show that a central limit theorem for $\bar{Y}(m)$ is valid; i.e., there exists a constant $\bar{\sigma}$ such that

$$\sqrt{m}(\bar{Y}(m) - \nu) \xrightarrow{\mathcal{D}} \bar{\sigma}N(0, 1) \tag{18}$$

as $m \to \infty$. Hence, just as in the setting of terminating simulations, we have that

$$\bar{Y}(m) \overset{\mathcal{D}}{\approx} \nu + \frac{\bar{\sigma}}{\sqrt{m}}N(0, 1).$$

Thus, the error decreases as the inverse of the square root of the sample size m. In particular, this means that to add an additional significant figure of accuracy to our current estimate of ν, we need to increase the run length m by a factor of 100. Furthermore, the error is normally distributed. Finally, the constant $\bar{\sigma}$ reflects the "difficulty" of the steady-state simulation: the larger $\bar{\sigma}$ is, the longer we need to simulate to get an accurate answer.

The basic idea behind constructing confidence intervals for ν based on the CLT in (18) is straightforward. Suppose $z_{1-\delta/2}$ is selected so that $P\{N(0, 1) \leq z_{1-\delta/2}\} = 1 - \delta/2$. Then the CLT asserts that

$$P\left\{\nu \in \left[\ \bar{Y}(m) - \frac{z_{1-\delta/2}\bar{\sigma}}{\sqrt{m}},\quad \bar{Y}(m) + \frac{z_{1-\delta/2}\bar{\sigma}}{\sqrt{m}}\ \right]\right\} \to\ 1 - \delta$$

as $m \to \infty$. Hence, for large m, the interval

$$\left[\bar{Y}(m) - \frac{z_{1-\delta/2}\bar{\sigma}}{\sqrt{m}},\ \bar{Y}(m) + \frac{z_{1-\delta/2}\bar{\sigma}}{\sqrt{m}} \right]$$

is an approximate $100(1-\delta)\%$ confidence interval for ν.

Unfortunately, this interval is useless from a practical standpoint since we typically do not know the value of $\bar{\sigma}$. We cannot use the standard variance estimator as in (11) applied to the $Y_1, Y_2, \ldots$ to estimate $\bar{\sigma}$ since the standard variance estimator is only for i.i.d. data. The $Y_1, Y_2, \ldots$ are *dependent*.

To get around this, we would like to do something similar to what we did in the terminating simulation setting. Specifically, if we could find a variance estimator $V(m)$ with the property that $V(m) \to \bar{\sigma}^2$ as $m \to \infty$, then we would basically be done. To construct a confidence interval for ν, we would compute $\tilde{S}(m) = \sqrt{V(m)}$, and use

$$\left[\bar{Y}(m) - \frac{z_{1-\delta/2}\tilde{S}(m)}{\sqrt{m}},\quad \bar{Y}(m) + \frac{z_{1-\delta/2}\tilde{S}(m)}{\sqrt{m}} \right]$$

as our approximate $100(1-\delta)\%$ confidence interval for ν. So now the problem is to find a *consistent estimator* for $\bar{\sigma}^2$.

Definition: The parameter $\bar{\sigma}^2$ is called the *time-average variance constant* of the steady-state simulation.

If we return to our point estimator for ν, we find that

$$\bar{Y}(m) = \frac{1}{m}\sum_{i=1}^{m} Y_i,$$

where the Y_i are dependent random variables. Since $\bar{Y}(m)$ is an average of dependent observations, there is no obvious way to construct a variance estimator $V(m)$ here. The standard sample variance estimator only works for calculating the variance of an estimator that is the sample mean of *independent* and *identically distributed* random variables.

4.2 Method of Multiple Replications

The *method of multiple replications* offers one escape from this difficulty of estimating $\bar{\sigma}^2$. Suppose that rather than simulating one long replicate of length m, we simulate n' *independent* replications, each of length $m' = m/n'$. We achieve independence of the replications by using non-overlapping streams of random numbers for the different replications. Because we now have n' independent observations, we can form a sample variance across the replications. This is the basic idea underlying the method of multiple replications.

For $j = 1, \ldots, n'$, let $Y_{j,1}, Y_{j,2}, \ldots, Y_{j,m'}$ be the m' observations from the jth (independent) replication, and let

$$X'_j = \frac{1}{m'}\sum_{i=1}^{m'} Y_{j,i}$$

be the sample mean formed from the jth replication. Then the X'_j, $j = 1, 2, \ldots, n'$, are i.i.d. observations with $E(X'_j) \approx \nu$ if m' is sufficiently large. So we can use classical statistics to form a point estimator and confidence interval using the observations $X'_1, X'_2, \ldots, X'_{n'}$. Specifically, let

$$\bar{X}'(n') = \frac{1}{n'}\sum_{j=1}^{n'} X'_j$$

and

$$S^2(n') = \frac{1}{n'-1}\sum_{j=1}^{n'}(X'_j - \bar{X}'(n'))^2$$

be the sample mean and sample variance of the X'_j. Then, an approximate $100(1-\delta)\%$ confidence interval for ν is given by

$$\left[\bar{X}'(n') - t_{n'-1,1-\delta/2}\frac{S(n')}{\sqrt{n'}},\ \bar{X}'(n') + t_{n'-1,1-\delta/2}\frac{S(n')}{\sqrt{n'}} \right],$$

where $t_{n'-1,1-\delta/2}$ is the upper $1-\delta/2$ point of a t-distribution with $n'-1$ degrees of freedom. (Here we use the critical point from a t-distribution rather than a standard normal distribution because the number n' of replications is often small.)

A major problem with the method of multiple replications is that, while the technique permits simple estimation of the variance, the multiple replicate estimator $\bar{X}'(n')$ can be significantly contaminated by the presence of *initialization bias*. Specifically, the law of large numbers guarantees that

$$X'_j = \frac{1}{m'}\sum_{i=1}^{m'} Y_{j,i} \to \nu$$

as $m' \to \infty$. However, since each replicate is typically started with an initial condition $\mathcal{I}$ that is atypical of the steady state (e.g., queueing simulations are often started with no customers present), it follows that for any finite m',

$$E\left[\frac{1}{m'}\sum_{i=1}^{m'} Y_{j,i}\right] \neq \nu.$$

Thus, we conclude that if the number of replicates n' is large relative to the run length m' of each replication, then the estimator $\bar{X}'(n')$ may be significantly biased.

A partial solution to this problem is to use *initial-data deletion*, which we now describe. Suppose that we somehow can determine the first c observations of the simulation are significantly contaminated, i.e., not very representative of steady state. Also, suppose all observations Y_i with $i > c$ are not significantly contaminated. Then in each replication, we can delete the first c observations when calculating the sample mean of the replication. Specifically, let

$$X_j = \frac{1}{m' - c} \sum_{i=c+1}^{m'} Y_{j,i}$$

be the sample mean of the observations $Y_{j,c+1}, Y_{j,c+2}, \ldots, Y_{j,m'}$. After simulating the n' replications, compute

$$\bar{X}(n') = \frac{1}{n'} \sum_{j=1}^{n'} X_j$$

and

$$S^2(n') = \frac{1}{n' - 1} \sum_{j=1}^{n'} (X_j - \bar{X}(n'))^2,$$

which are the sample mean and sample variance of the X_j. Then, an approximate $100(1 - \delta)\%$ confidence interval for ν is given by

$$\left[\bar{X}(n') - t_{n'-1,1-\delta/2} \frac{S(n')}{\sqrt{n'}},\ \bar{X}(n') + t_{n'-1,1-\delta/2} \frac{S(n')}{\sqrt{n'}} \right].$$

For more details on initial-data deletion, including some heuristics on determining c, see Section 9.1 of Law and Kelton (2000).

One problem with initial-data deletion is that in each of the n' replications, we have to delete c observations. Thus, we are throwing away a total of $n'c$ observations over all of the replications. If we used a *single replicate algorithm* (i.e., one with $n' = 1$), then we would only delete a total of c observations.

4.3 Single-Replicate Methods

Typically in practice, when k is large, Y_i and Y_{i+k} are almost independent for each i. For example, suppose Y_i is the waiting time of the ith customer in a queueing system. Then we would expect that the waiting time of the 100th customer to be almost independent of the 10th customer's waiting time.

Now suppose that we run a simulation of length m, giving us observations $Y_1, Y_2, \ldots, Y_m$. Suppose we group the m observations into large, non-overlapping *batches*, where the first batch consists of the first k observations, the second batch consists of the next k observations, and so on. If k is chosen to be large, then most of the observations in one batch should be almost independent of most of the observations in any other batch. The only dependence that essentially exists is between observations in two adjacent batches. Observations in batches that are not adjacent are almost independent. Moreover, if we compute the sample mean of each of the batches, then the sample means should be almost independent when the batch size k is large. Also, each sample mean will be close to normally distributed for large k, since it is a sample mean and so a corresponding central limit theorem holds. Using the above observations, we now present the *method of batch means*.

1. Select a total run length m, which is large. Also, select a number of batches n. (Schmeiser 1982 suggest choosing $10 \leq n \leq 30$. If m is small, then n should probably be chosen in the lower end of the range since the key to the validity of the method of batch means is that the batches are "large" so that most of the observations in one batch are independent of most of the observations in any other batch.)
2. Run a simulation generating a total of m observations. This results in observations $Y_1, Y_2, \ldots, Y_m$.
3. Then group the m observations into n batches, each of size $k = m/n$. For $j = 1, 2, \ldots, n$, the jth *batch mean* is calculated as

$$\bar{Y}_j(k) = \frac{1}{k} \sum_{l=(j-1)k+1}^{jk} Y_l.$$

   Note that $\bar{Y}_j(k)$ is the sample mean of the k observations in the j-th batch.
4. We then treat $\bar{Y}_1(k), \bar{Y}_2(k), \ldots, \bar{Y}_n(k)$ as i.i.d. observations (note that they are not, but should be reasonably close for large values of k) and use classical statistics to construct a confidence interval. Specifically, our point estimate for ν is

$$\bar{\bar{Y}}(n, k) = \frac{1}{n} \sum_{j=1}^{n} \bar{Y}_j(k),$$

and an approximate $100(1-\delta)\%$ confidence interval for ν is

$$\left[\bar{\bar{Y}}(n,k) - \frac{t_{n-1,1-\delta/2} S(n,k)}{\sqrt{n}}, \bar{\bar{Y}}(n,k) + \frac{t_{n-1,1-\delta/2} S(n,k)}{\sqrt{n}} \right],$$

where

$$S^2(n,k) = \frac{1}{n-1} \sum_{j=1}^{n} \left(\bar{Y}_j(k) - \bar{\bar{Y}}(n,k) \right)^2$$

is the sample variance of the n batch means.

It can be shown that in virtually all situations arising in practice, the method of batch means will produce valid confidence intervals as $m \to \infty$ with n fixed. More specifically,

$$P\left\{ \nu \in \left[\bar{\bar{Y}}(n,k) - \frac{t_{n-1,1-\delta/2} S(n,k)}{\sqrt{n}}, \bar{\bar{Y}}(n,k) + \frac{t_{n-1,1-\delta/2} S(n,k)}{\sqrt{n}} \right] \right\} \to 1-\delta$$

as $m \to \infty$ with n fixed.

We can easily modify the above procedure to incorporate initial-data deletion by instead collecting a total of $m+c$ observations and removing the first c contaminated observations. Then apply the method of batch means with the remaining m data points. When using a single-replicate method such as batch means, we only need to delete a total of c observations, as opposed to $n'c$ when using the method of multiple replications with n' replications. Whitt (1991) provides a mathematical analysis that basically yields the following:

Rule of thumb: Single replicate procedures tend to be better (as measured by the mean square error of the steady-state estimator) than multiple replicate procedures.

There has been a lot of recent work on improvements to the batch-means method described above. See Schmeiser and Song (1996) for a survey.

4.4 Other Methods

There are numerous other methods for statistically analyzing simulation output in the steady-state context. These include spectral methods (e.g., Anderson 1994), regenerative methods (Crane and Iglehart 1975, Shedler 1993), and standardized time series (Schruben 1983), but these techniques require more sophisticated mathematics to understand and can be somewhat more difficult to implement. For an overview of these other techniques, see Bratley, Fox and Schrage (1987) or Law and Kelton (2000). Finally, Nakayama (1994) presents two-stage procedures for obtaining fixed-width confidence intervals in steady-state simulations.

5 ESTIMATING MULTIPLE PERFORMANCE MEASURES

Consider our previous example of an ATM that is accessible only between 9:00am and 5:00pm, and suppose that we want to calculate

- μ_1, the expected number of customers served in a day;
- μ_2, the probability that the number served in a day is at least 1000;
- μ_3, the expected amount of money withdrawn from the ATM in a day.

These are all transient performance measures, and we can use the same simulation to estimate all 3 measures by running n independent replications. Let $X_{1,i}$ denote the number of customers served in the ith replication. Let $X_{2,i}$ be 1 if at least 1000 customers are served in the ith replication, and 0 otherwise. Let $X_{3,i}$ be the amount of money withdrawn on the ith replication.

After running n replications, we construct a 95% confidence interval for each μ_s, $s = 1, 2, 3$. Let I_s denote the 95% confidence interval for μ_s, so if we ran enough replications so that the asymptotics of the CLT roughly hold, then $P\{\mu_s \in I_s\} \approx 0.95$ for each $s = 1, 2, 3$. But what can we say about the *joint* coverage of the 3 confidence intervals; i.e., what is $P\{\mu_s \in I_s, \text{ for all } s = 1, 2, 3\}$?

More generally, suppose that we are estimating k means μ_s, $s = 1, 2, \ldots, k$, and for each μ_s, we constructed a $100(1-\delta_s)\%$ confidence interval I_s. What can we say about $P\{\mu_s \in I_s, \text{ for all } s = 1, 2, \ldots, k\}$? Bonferroni's inequality can be used to provide a lower bound for this probability:

$$P\{\mu_s \in I_s, \text{ for all } s = 1, 2, \ldots, k\} \geq 1 - \sum_{s=1}^{k} \delta_s.$$

Thus, in our previous example in which we had three 95% confidence intervals, we can say that the joint probability that all three confidence intervals contain their respective true means is at least 85%. If we wanted the joint confidence to be at least 95%, then we might set $\delta_s = 0.01$ for each s. This would yield individual 99% confidence intervals, with the joint probability being at least 0.97.

In the case that the k confidence intervals are independent, then the joint probability that all k confidence intervals contain their respective true means is exactly $\prod_{s=1}^{k}(1-\delta_s)$. However, for the confidence intervals to be independent,

each confidence interval must be constructed using a simulation that is independent of all of the other simulations used to construct the other confidence intervals. In other words, the simulations used to construct the various confidence intervals must use non-overlapping streams of random numbers. If this is not the case, then without any assumed dependence structure, the best that we can say about the joint probability is the lower bound provided by Bonferroni's inequality.

Often, one wants to compare different systems to see which one is the "best." For example, we may have 5 possible designs for a manufacturing system, and we want to determine which has the highest expected production per day. There is substantial literature on this topic, mainly in the areas of so-called *selection procedures* and *multiple-comparison procedures*. For an overview of these methods, see Goldsman and Nelson (2001).

6 OTHER USEFUL METHODS

We now briefly discuss some other techniques that can be useful for simulations. *Variance-reduction techniques* (VRTs), which are also known as *efficiency-improvement techniques*, can lead to simulation estimators with smaller error (variance) by typically either collecting additional information from the simulation run(s) or changing or controlling the way in which simulation is run. Some of the more widely used VRTs include the following:

- *Common random numbers* (e.g., see Section 11.2 of Law and Kelton 2000) can improve simulations comparing two or more systems by running the simulations of the various systems using the same stream of random numbers. This generally induces positive correlation among the resulting estimators, which can be advantageous when estimating differences of performance measures between systems.
- *Antithetic variates* (e.g., see Section 11.3 of Law and Kelton 2000) can improve results from simulating a single system by inducing negative correlations between runs.
- The method of *control variates* (e.g., see Section 11.4 of Law and Kelton 2000) collects additional data during the simulation, where the mean of the extra collected data is known. For example, in a queueing simulation, one often knows the mean of the service-time distribution, and so one might additionally collect the random service times that are generated during the simulation. The data collected typically is correlated with the simulation output, and this correlation can be exploited to obtain an estimator with lower variance than the standard estimator.
- *Importance sampling* (Hammersley and Handscomb 1965, Glynn and Iglehart 1989) is often used in rare-event simulations, such as for analyzing buffer overflows in communication networks and system failures of fault-tolerant systems. In these settings, the event of interest, typically some kind of failure, occurs very rarely, and importance sampling changes the dynamics of the system to cause the event to occur more frequently. Unbiased estimators are recovered by multiplying by a correction factor known as the likelihood ratio. Heidelberger (1995) and Nicola, Shahabuddin and Nakayama (2001) review importance-sampling methods for rare-event simulations of queueing and reliability systems.

Other VRTs include stratified sampling, conditional Monte Carlo, and splitting. These and other methods are described in Chapter 11 of Law and Kelton (2000) and Chapter 2 of Bratley, Fox and Schrage (1987).

One is often interested in estimating derivatives of performance measures with respect to system parameters. For example, in a reliability system, one may want to know the derivative of the mean time to failure with respect to a component's failure rate. This information can be useful in designing systems by identifying components on which to focus to improve overall performance. Also, derivative information can be used with some simulation optimization methods (e.g., Andradóttir 1998). Techniques for estimating derivatives using simulation include perturbation analysis (Glasserman 1991, Ho and Cao 1991, Fu and Hu 1997) and the likelihood-ratio or score-function method (Reiman and Weiss 1989, Rubinstein 1989, Glynn 1990).

7 CONCLUSIONS

We have described some techniques for statistically analyzing the output from a simulation. It is important to keep in mind that the methods presented here are all *asymptotically* valid, so large run lengths are needed to ensure that valid inferences are drawn.

In addition to the references given throughout the paper, other resources covering simulation-output analysis include Banks (1998), Banks et al. (2001), Fishman (2001), Melamed and Rubinstein (1998), and Ross (2002).

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under grants DMI-9624469 and DMI-9900117.

REFERENCES

Anderson, T. W. 1994. *The Statistical Analysis of Time Series*, New York: Wiley.

Andradóttir, S. 1998. Simulation optimization. Chapter 9 in *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*. Ed. J. Banks. New York: John Wiley and Sons.

Banks, J. 1998. *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*. New York: John Wiley and Sons.

Banks, J., J. S. Carson, II, B. L. Nelson, and D. M. Nicol. 2001. *Discrete-Event System Simulation, 3rd edition*. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Bratley, P., B. L. Fox, and L. E. Schrage. 1987. *A Guide to Simulation, Second Edition*, New York: Springer-Verlag.

Crane, M. and D. L. Iglehart. 1975. Simulating stable stochastic systems, III: Regenerative processes and discrete-event simulations. *Operations Research* 23: 33–45.

Fishman, G. S. 2001. *Discrete-Event Simulation: Modeling, Programming, and Analysis*. New York: Springer-Verlag.

Fu, M., and J. Q. Hu. 1997. *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Boston: Kluwer Academic Publishers.

Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis*. Boston: Kluwer Academic Publishers.

Glynn, P. W. 1990. Likelihood ratio derivative estimators for stochastic systems. *Communications of the ACM* 33: 75–84.

Glynn, P. W., and D. L. Iglehart. 1989. Importance sampling for stochastic simulations. *Management Science* 35: 1367–1392.

Goldsman, D., and B. L. Nelson. 2001. Statistical selection of the best system. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, 139–146, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Hall, P. 1987. Edgeworth expansion for Student's t statistic under minimal moment conditions. *Annals of Probability* 15: 920–931.

Hall, P. 1992. *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag

Hammersley, J. M., and D. C. Handscomb. 1964. *Monte Carlo Methods*. London: Methuen.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation* 5: 43–85.

Ho, Y. C., and X. R. Cao. 1991. *Discrete Event Dynamic Systems and Perturbation Analysis*. Boston: Kluwer Academic Publishers.

Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd edition. New York: McGraw-Hill.

Melamed, B., and R. Y. Rubinstein. 1998. *Modern Simulation and Modeling*. New York: John Wiley & Sons.

Nakayama, M. K. 1994. Two-stage stopping procedures based on standardized time series. *Management Science* 40: 1189–1206.

Nicola, V. F., P. Shahabuddin and M. K. Nakayama. 2001. Techniques for fast simulation of models of highly dependable systems. *IEEE Transactions on Reliability* 50: 246–264.

Reiman, M. I., and A. Weiss. 1989. Sensitivity analysis for simulations via likelihood ratios. *Operations Research* 37: 830–844.

Ross, S. M. 2002. *Simulation, 3rd Edition*. Boston: Academic Press.

Rubinstein, R. Y. 1989. Sensitivity analysis and performance extrapolation for computer simulation models. *Operations Research* 37: 72–81.

Schmeiser, B. W. 1982. Batch size effects in the analysis of simulation output. *Operations Research* 30: 556–568.

Schmeiser, B. W., and W. T. Song. 1996. Batching methods in simulation output analysis: what we know and what we don't. *Proceedings of the 1996 Winter Simulation Conference*, ed. J. M. Charnes, D. M. Morrice, D. T. Brunner, and J. J. Swain, 122–127, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Schruben, L. W. 1983. Confidence interval estimation using standardized time series. *Operations Research* 31: 1090–1108.

Shedler, G. S. 1993. *Regenerative Stochastic Simulation*. San Diego: Academic Press.

Stein, C. 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* 16: 243–258.

Whitt, W. 1991. The efficiency of one long run versus independent replications in steady-state simulation. *Management Science* 37: 645–666.

AUTHOR BIOGRAPHY

MARVIN K. NAKAYAMA is an associate professor in the Department of Computer Science at the New Jersey Institute of Technology. He received a Ph.D. in operations research from Stanford University. He won second prize in the 1992 George E. Nicholson Student Paper Competition sponsored by INFORMS and is a recipient of a CAREER Award from the National Science Foundation. He is the area editor for the Stochastic Modeling Area of *ACM Transactions on Modeling and Computer Simulation* and an associate editor for *INFORMS Journal on Computing*. His research interests include applied probability, statistics, simulation and modeling.