CHAPTER 9

Output Data Analysis for a Single System

Recommended sections for a first reading: 9.1 through 9.3, 9.4.1, 9.4.3, 9.5.1, 9.5.2, 9.8

9.1 INTRODUCTION

In many simulation studies a great deal of time and money is spent on model development and "programming," but little effort is made to analyze the simulation output data appropriately. As a matter of fact, a common mode of operation is to make a single simulation run of somewhat arbitrary length and then to treat the resulting simulation estimates as the "true" model characteristics. Since random samples from probability distributions are typically used to drive a simulation model through time, these estimates are just particular realizations of random variables that may have large variances. As a result, these estimates could, in a particular simulation run, differ greatly from the corresponding true characteristics for the model. The net effect is, of course, that there could be a significant probability of making erroneous inferences about the system under study.

Historically, there are several reasons why output data analyses have not been conducted in an appropriate manner. First, some users have the unfortunate impression that simulation is largely an exercise in computer programming, albeit a complicated one. Consequently some simulation "studies" begin with construction of an assumptions document (see Sec. 5.4.3) and subsequent "programming," and end with a single run of the simulation to produce the "answers." In fact, however, a simulation is a computer-based statistical sampling experiment. Thus if the results of a simulation study are to have any meaning, appropriate statistical techniques must be used to design and analyze the simulation experiments. A second reason for inadequate statistical analyses is that the output processes of virtually all simulations are nonstationary and autocorrelated (see Sec. 5.6). Thus, classical statistical techniques based on IID observations are not *directly* applicable. At present, there are still several output-analysis problems for which there is no completely accepted solution, and the methods that are available are often complicated to apply (see, for example, Sec. 9.5.4). Another impediment to obtaining precise estimates of a model's true parameters or characteristics is the computer time needed to collect the necessary amount of simulation output data. This difficulty often occurs in the simulation of large-scale military problems or high-speed communications networks.

We now describe more precisely the random nature of simulation output. Let Y_1, $Y_2, \ldots$ be an output stochastic process (see Sec. 4.3) from a *single* simulation run. For example, Y_i might be the throughput (production) in the ith hour for a manufacturing system. The Y_i's are random variables that will, in general, be neither independent nor identically distributed. Thus, most of the formulas of Chap. 4, which assume independence [e.g., the confidence interval given by (4.12)], do not apply *directly*.

Let $y_{11}, y_{12}, \ldots, y_{1m}$ be a realization of the random variables $Y_1, Y_2, \ldots, Y_m$ resulting from making a simulation run of length m observations using the random numbers $u_{11}, u_{12}, \ldots$. (The ith random number used in the jth run is denoted u_{ji}.) If we run the simulation with a different set of random numbers $u_{21}, u_{22}, \ldots$, then we will obtain a different realization $y_{21}, y_{22}, \ldots, y_{2m}$ of the random variables $Y_1, Y_2, \ldots, Y_m$. (The two realizations are not the same since the different random numbers used in the two runs produce different samples from the input probability distributions.) In general, suppose that we make n independent replications (runs) of the simulation (i.e., different random numbers are used for each replication, the statistical counters are reset at the beginning of each replication, and each replication uses the same initial conditions; see Sec. 9.4.3) of length m, resulting in the observations:

$$\begin{array}{ccccc} y_{11}, & \ldots, & y_{1i}, & \ldots, & y_{1m} \\ y_{21}, & \ldots, & y_{2i}, & \ldots, & y_{2m} \\ \vdots & & \vdots & & \vdots \\ y_{n1}, & \ldots, & y_{ni}, & \ldots, & y_{nm} \end{array}$$

The observations from a particular replication (row) are clearly not IID. However, note that $y_{1i}, y_{2i}, \ldots, y_{ni}$ (from the ith column) are IID observations of the random variable Y_i, for $i = 1, 2, \ldots, m$. This *independence across runs* (see Prob. 9.1) is the key to the relatively simple output-data-analysis methods described in later sections of this chapter. Then, roughly speaking, the goal of output analysis is to use the observations y_{ji} ($i = 1, 2, \ldots, m$; $j = 1, 2, \ldots, n$) to draw inferences about the (distributions of the) random variables $Y_1, Y_2, \ldots, Y_m$. For example, $\bar{y}_i(n) = \sum_{j=1}^{n} y_{ji}/n$ is an unbiased estimate of $E(Y_i)$.

EXAMPLE 9.1. Consider a bank with five tellers and one queue, which opens its doors at 9 A.M., closes its doors at 5 P.M., but stays open until all customers in the bank at 5 P.M. have been served. Assume that customers arrive in accordance with a Poisson process at rate 1 per minute (i.e., IID exponential interarrival times with mean 1 minute), that service times are IID exponential random variables with mean 4 minutes, and that customers are served in a FIFO manner. Table 9.1 shows several typical output statistics from 10 independent replications of a simulation of the bank, assuming that no

TABLE 9.1
Results for 10 independent replications of the bank model

Replication	Number served	Finish time (hours)	Average delay in queue (minutes)	Average queue length	Proportion of customers delayed < 5 minutes
1	484	8.12	1.53	1.52	0.917
2	475	8.14	1.66	1.62	0.916
3	484	8.19	1.24	1.23	0.952
4	483	8.03	2.34	2.34	0.822
5	455	8.03	2.00	1.89	0.840
6	461	8.32	1.69	1.56	0.866
7	451	8.09	2.69	2.50	0.783
8	486	8.19	2.86	2.83	0.782
9	502	8.15	1.70	1.74	0.873
10	475	8.24	2.60	2.50	0.779

customers are present initially. Note that results from various replications can be quite different. Thus, one run clearly does not produce "the answers."

Our goal in this chapter is to discuss methods for statistical analysis of simulation output data and to present the material with a practical focus that should be accessible to a reader having a basic understanding of probability and statistics. (Reviewing Chap. 4 might be advisable before reading this chapter.) We will discuss what we believe are all the important methods for output analysis; however, the emphasis will be on statistical procedures that are relatively easy to understand and implement, have been shown to perform well in practice, and have applicability to real-world problems.

In Secs. 9.2 and 9.3 we discuss types of simulations with regard to output analysis, and also measures of performance or parameters θ for each type. Sections 9.4 through 9.6 show how to get a point estimator $\hat{\theta}$ and confidence interval for each type of parameter θ, with the confidence interval typically requiring an estimate of the variance of $\hat{\theta}$, namely, $\widehat{\text{Var}}(\hat{\theta})$. Each of the analysis methods discussed may suffer from one or both of the following problems:

1. $\hat{\theta}$ is not an unbiased estimator of θ, that is, $E(\hat{\theta}) \neq \theta$; see, for example, Sec. 9.5.1.
2. $\widehat{\text{Var}}(\hat{\theta})$ is not an unbiased estimator of $\text{Var}(\hat{\theta})$; see, for example, Sec. 9.5.3.

Section 9.7 extends the above analysis to confidence-interval construction for several different parameters simultaneously. Finally, in Sec. 9.8 we show how time plots of important variables may provide insight into a system's dynamic behavior.

We will not attempt to give every reference on the subject of output-data analysis, since literally hundreds of papers on the subject have been written. A very comprehensive set of references up to 1983 is given in the survey paper by Law (1983); also see the survey paper by Pawlikowski (1990) and the book chapters by Alexopoulos and Seila (1998) and by Welch (1983). Most of the recent papers have been published in the journals *Transactions on Modeling and Computer Simulation, Operations Research, and INFORMS Journal on Computing*, or in the *Proceedings of the Winter Simulation Conference* (held every December).

9.2 TRANSIENT AND STEADY-STATE BEHAVIOR OF A STOCHASTIC PROCESS

Consider the output stochastic process $Y_1, Y_2, \ldots$. Let $F_i(y \mid I) = P(Y_i \le y \mid I)$ for $i = 1, 2, \ldots$, where y is a real number and I represents the initial conditions used to start the simulation at time 0. [The conditional probability $P(Y_i \le y \mid I)$ is the probability that the event $\{Y_i \le y\}$ occurs *given* the initial conditions I.] For a manufacturing system, I might specify the number of jobs present, and whether each machine is busy or idle, at time 0. We call $F_i(y \mid I)$ the *transient distribution* of the output process at (discrete) time i for initial conditions I. Note that $F_i(y \mid I)$ will, in general, be different for each value of i and each set of initial conditions I. The density functions for the transient distributions corresponding to the random variables Y_{i_1}, Y_{i_2}, Y_{i_3}, and Y_{i_4} are shown in Fig. 9.1 for a particular set of initial conditions I and increasing time indices i_1, i_2, i_3, and i_4, where it is assumed that the random variable Y_{i_j} has density function $f_{Y_{i_j}}$. The density $f_{Y_{i_j}}$ specifies how the random variable Y_{i_j} can vary from one replication to another. In particular, suppose that we make a very large number of replications, n, of the simulation and observe the stochastic process $Y_1, Y_2, \ldots$ on each one. If we make a histogram of the n observed values of the random variable Y_{i_j}, then this histogram (when appropriately scaled) will look very much like the density $f_{Y_{i_j}}$.

For fixed y and I, the probabilities $F_1(y \mid I), F_2(y \mid I), \ldots$ are just a sequence of numbers. *If* $F_i(y \mid I) \to F(y)$ as $i \to \infty$ for all y and for any initial conditions I, then $F(y)$ is called the *steady-state distribution* of the output process $Y_1, Y_2, \ldots$. Strictly speaking, the steady-state distribution $F(y)$ is only obtained in the limit as $i \to \infty$. In practice,

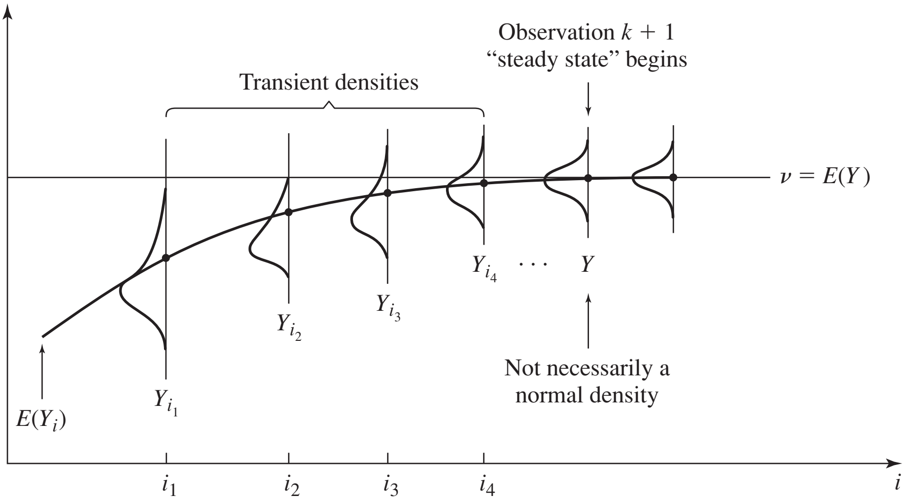


FIGURE 9.1
Transient and steady-state density functions for a particular stochastic process $Y_1, Y_2, \ldots$ and initial conditions I.

however, there will often be a finite time index, say, $k + 1$, such that the distributions from this point on will be approximately the same as each other; "steady state" is figuratively said to start at time $k + 1$ as shown in Fig. 9.1. Note that steady state does *not* mean that the random variables $Y_{k+1}, Y_{k+2}, \ldots$ will all take on the same value in a particular simulation run; rather, it means that they will all have approximately the same *distribution*. Furthermore, these random variables will not be independent, but will approximately constitute a covariance-stationary stochastic process (see Sec. 4.3). See Welch (1983) for an excellent discussion of transient and steady-state distributions.

The steady-state distribution $F(y)$ does not depend on the initial conditions I; however, the rate of convergence of the transient distributions $F_i(y|I)$ to $F(y)$ does, as the following example shows.

EXAMPLE 9.2. Consider the stochastic process $D_1, D_2, \ldots$ for the $M/M/1$ queue with $\rho = 0.9$ ($\lambda = 1$, $\omega = 10/9$), where D_i is the delay in queue of the ith customer. In Fig. 9.2 we plot the convergence of the transient mean $E(D_i)$ to the steady-state mean

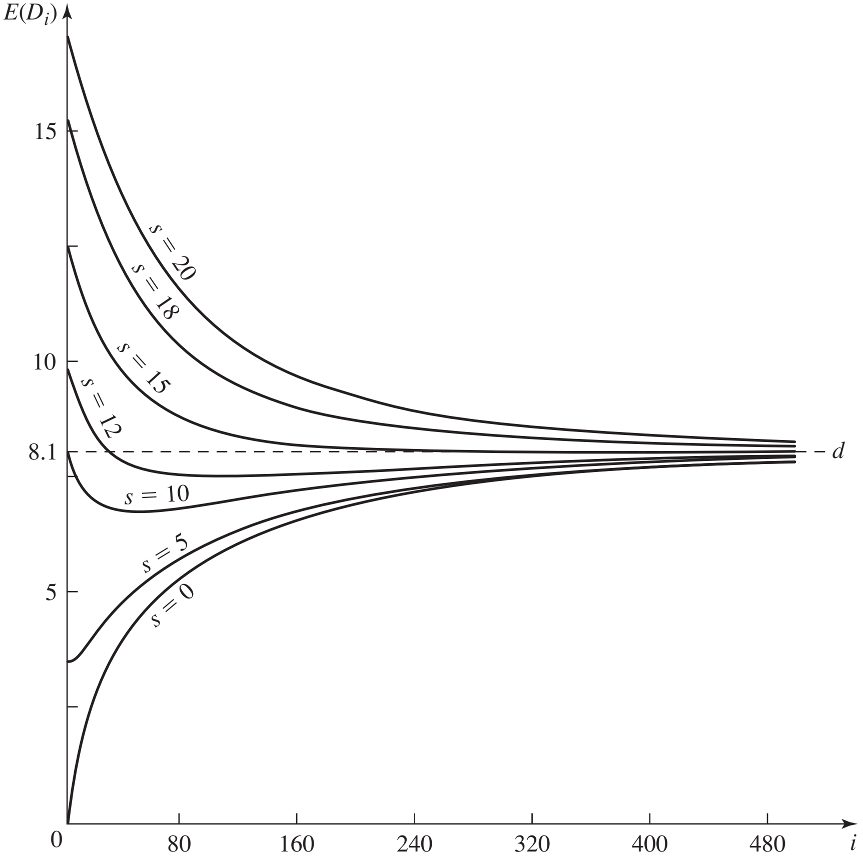


FIGURE 9.2
$E(D_i)$ as a function of i and the number in system at time 0, s, for the $M/M/1$ queue with $\rho = 0.9$.

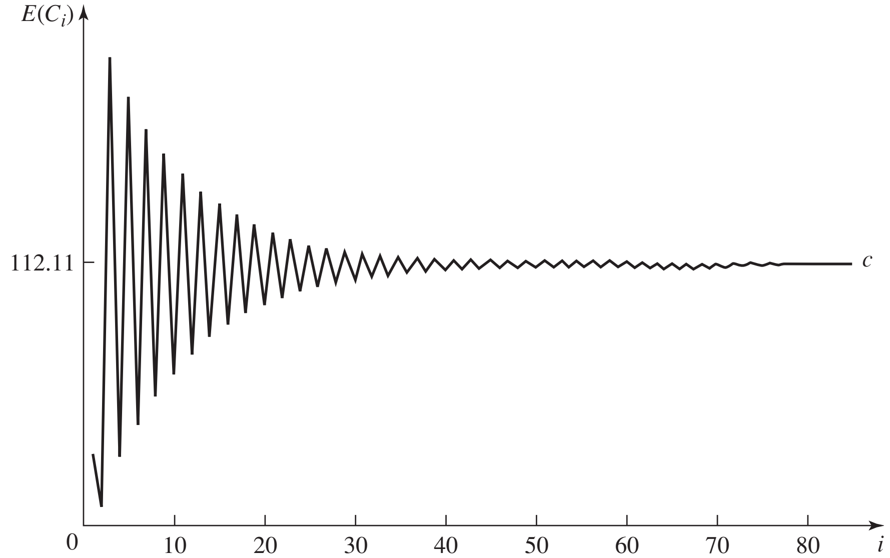


FIGURE 9.3
$E(C_i)$ as a function of i for the (s, S) inventory system.

$d = E(D) = 8.1$ as i gets large for various values of the number in system at time 0, s. (The random variable D has the steady-state delay in queue distribution.) Note that the convergence of $E(D_i)$ to d is, surprisingly, much faster for $s = 15$ than for $s = 0$ (see Prob. 9.11). The values for $E(D_i)$ were derived in Kelton and Law (1985); see also Kelton (1985) and Murray and Kelton (1988). The distribution function of D is given by (4.15) in App. 4A.

EXAMPLE 9.3. Consider the stochastic process $C_1, C_2, \ldots$ for the inventory problem of Example 4.23, where C_i is the total cost in the ith month. In Fig. 9.3 we plot the convergence of $E(C_i)$ to the steady-state mean $c = E(C) = 112.11$ [see Wagner (1969, p. A19)] as i gets large for an initial inventory level of 57. Note that the convergence is clearly not monotone.

In Examples 9.2 and 9.3 we plotted the convergence of the *expected value* $E(Y_i)$ to the steady-state mean $E(Y)$. It should be remembered, however, that the entire *distribution* of Y_i is also converging to the distribution of Y as i gets large.

9.3 TYPES OF SIMULATIONS WITH REGARD TO OUTPUT ANALYSIS

The options available in designing and analyzing simulation experiments depend on the type of simulation at hand, as depicted in Fig. 9.4. Simulations may be either terminating or nonterminating, depending on whether there is an obvious way for

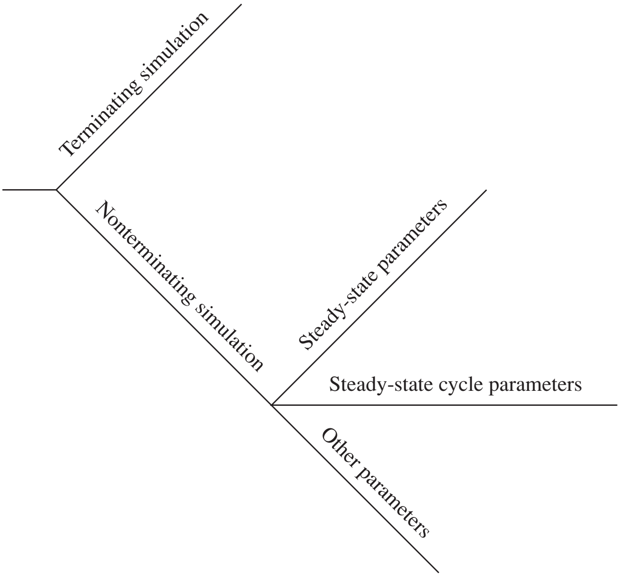


FIGURE 9.4
Types of simulations with regard to output analysis.

determining the run length. Furthermore, measures of performance or parameters for nonterminating simulations may be of several types, as shown in the figure. These concepts are defined more precisely below.

A *terminating simulation* is one for which there is a "natural" event E that specifies the length of each run (replication). Since different runs use independent random numbers and the same initialization rule, this implies that comparable random variables from the different runs are IID (see Sec. 9.4). The event E often occurs at a time point when the system is "cleaned out" (see Example 9.4), at a time point beyond which no useful information is obtained (see Example 9.5), or at a time point specified by management mandate (see Example 9.8). It is specified before any runs are made, and the time of occurrence of E for a particular run may be a random variable. Since the *initial conditions for a terminating simulation generally affect the desired measures of performance*, these conditions should be representative of those for the actual system (see Sec. 9.4.3).

EXAMPLE 9.4. A retail/commercial establishment, e.g., a bank, closes each evening. If the establishment is open from 9 A.M. to 5 P.M., the objective of a simulation might be to estimate some measure of the quality of customer service over the period beginning at 9 A.M. and ending when the last customer who entered before the doors closed at 5 P.M. has been served. In this case $E = \{$at least 8 hours of simulated time have elapsed and the system is empty$\}$, and the initial conditions for the simulation are the number of customers present at time 0 (see Sec. 9.4.3).

EXAMPLE 9.5. Consider a military ground confrontation between a blue force and a red force. Relative to some initial force strengths, the goal of a simulation might be to determine the (final) force strengths when the battle ends. In this case $E = \{$either the blue force or the red force has "won" the battle$\}$. An example of a condition that would end the battle is one side losing 30 percent of its force, since this side would no longer be considered viable. 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.

EXAMPLE 9.6. An aerospace manufacturer receives a contract to produce 100 airplanes, which must be delivered within 18 months. The company would like to simulate various manufacturing configurations to see which one can meet the delivery deadline at least cost. In this case $E = \{100$ airplanes have been completed$\}$.

EXAMPLE 9.7. Consider a manufacturing company that operates 16 hours a day (two shifts) with work in process carrying over from one day to the next. Would this qualify as a terminating simulation with $E = \{16$ hours of simulated time have elapsed$\}$? No, since this manufacturing operation is essentially a continuous process, with the ending conditions for one day being the initial conditions for the next day.

EXAMPLE 9.8. A company that sells a single product would like to decide how many items to have in inventory during a planning horizon of 120 months (see Sec. 1.5). Given some initial inventory level, the objective might be to determine how much to order each month so as to minimize the expected average cost per month of operating the inventory system. In this case $E = \{120$ months have been simulated$\}$, and the simulation is initialized with the current inventory level.

A *nonterminating simulation* is one for which there is no natural event E to specify the length of a run. This often occurs when we are designing a new system or changing an existing system, and we are interested in the behavior of the system in the long run when it is operating "normally." Unfortunately, "in the long run" doesn't naturally translate into a terminating event E. A measure of performance for such a simulation is said to be a *steady-state parameter* if it is a characteristic of the steady-state distribution of some output stochastic process $Y_1, Y_2, \ldots$. In Fig. 9.1, if the random variable Y has the steady-state distribution, then we might be interested in estimating the steady-state mean $\nu = E(Y)$ or a probability $P(Y \le y)$ for some real number y.

EXAMPLE 9.9. Consider a company that is going to build a new manufacturing system and would like to determine the long-run (steady-state) mean hourly throughput of their system after it has been running long enough for the workers to know their jobs and for mechanical difficulties to have been worked out. Assume that:

(*a*) The system will operate 16 hours a day for 5 days a week.
(*b*) There is negligible loss of production at the end of one shift or at the beginning of the next shift (see Prob. 9.3).
(*c*) There are no breaks (e.g., lunch) that shut down production at specified times each day.

This system could be simulated by "pasting together" 16-hour days, thus ignoring the system idle time at the end of each day and on the weekend. Let N_i be the number of parts manufactured in the ith hour. If the stochastic process $N_1, N_2, \ldots$ has a steady-state distribution with corresponding random variable N, then we are interested in estimating the mean $\nu = E(N)$ (see Prob. 9.4).

It should be mentioned that stochastic processes for most *real* systems do not have steady-state distributions, since the characteristics of the system change over time. For example, in a manufacturing system the production-scheduling rules and the facility layout (e.g., number and location of machines) may change from time to time. On the other hand, a simulation model (which is an abstraction of reality) may have steady-state distributions, since characteristics of the *model* are often assumed not to change over time. When we have new information on the characteristics of the system, we can redo our steady-state analysis.

If, in Example 9.9, the manufacturing company wanted to know the time required for the system to go from startup to operating in a "normal" manner, this would be a terminating simulation with terminating event $E = \{$simulated system is running "normally"$\}$ (if such can be defined). *Thus, a simulation for a particular system might be either terminating or nonterminating, depending on the objectives of the simulation study.*

EXAMPLE 9.10. Consider a simulation model for a communications network that does not currently exist. Since there are typically no representative data available on the arrival mechanism for messages, it is common to assume that messages arrive in accordance with a Poisson process with *constant* rate equal to the *predicted* arrival rate of messages during the period of peak loading. (When the system is actually built, the arrival rate will vary as a function of time, and the period of peak loading may be relatively short.) Since the state of the system during "normal operation" is unknown, initial conditions must be chosen somewhat arbitrarily (e.g., no messages present at time 0). Then the goal is to run the simulation long enough so that the arbitrary choice of initial conditions is no longer having a significant effect on the estimated measures of performance (e.g., mean end-to-end delay of a message).

In performing the above steady-state analysis of the proposed communications network, we are essentially trying to determine how the network will respond to a peak load of infinite duration. If, however, the peak period in the actual network is short or if the arrival rate before the peak period is considerably lower than the peak rate, our analysis may overestimate the congestion level during the peak period in the network. This might result in purchasing a network configuration that is more powerful than actually needed.

Consider a stochastic process $Y_1, Y_2, \ldots$ for a nonterminating simulation that does not have a steady-state distribution. Suppose that we divide the time axis into equal-length, contiguous time intervals called *cycles*. (For example, in a manufacturing system a cycle might be an 8-hour shift.) Let Y_i^C be a random variable defined on the ith cycle, and assume that $Y_1^C, Y_2^C, \ldots$ are comparable. Suppose that the process $Y_1^C, Y_2^C, \ldots$ has a steady-state distribution F^C and that $Y^C \sim F^C$. Then a measure of performance is said to be a *steady-state cycle parameter* if it is a characteristic of Y^C such as the mean $\nu^C = E(Y^C)$. Thus, a steady-state cycle parameter is just a steady-state parameter of the appropriate cycle process $Y_1^C, Y_2^C, \ldots$.

EXAMPLE 9.11. Suppose for the manufacturing system in Example 9.9 that there is a half-hour lunch break at the beginning of the fifth hour in each 8-hour shift. Then the process of hourly throughputs $N_1, N_2, \ldots$ has no steady-state distribution (see Prob. 9.6). Let N_i^C be the average hourly throughput in the ith 8-hour shift (cycle). Then we might be interested in estimating the steady-state expected average hourly throughput over a cycle, $\nu^C = E(N^C)$, which is a steady-state cycle parameter.

EXAMPLE 9.12. Consider a call center for an airline. Suppose that the arrival rate of calls to the system varies with the time of day and day of the week, but assume that the pattern of arrival rates is identical from week to week. Let D_i be the delay experienced by the ith arriving call. The stochastic process $D_1, D_2, \ldots$ does not have a steady-state distribution. Let D_i^C be the average delay over the ith week. Then we might be interested in estimating the steady-state expected average delay over a week, $\nu^C = E(D^C)$.

For a nonterminating simulation, suppose that the stochastic process $Y_1, Y_2, \ldots$ does not have a steady-state distribution, and that there is no appropriate cycle definition such that the corresponding process $Y_1^C, Y_2^C, \ldots$ has a steady-state distribution. This can occur, for example, if the parameters for the model continue to change over time. In Example 9.12, if the arrival rate of calls changes from week to week and from year to year, then steady-state (cycle) parameters will probably not be well defined. In these cases, however, there will typically be a fixed amount of data describing how input parameters change over time. This provides, in effect, a terminating event E for the simulation and, thus, the analysis techniques for terminating simulations in Sec. 9.4 are appropriate. This is why we do not treat this situation as a separate case later in the chapter. Measures of performance or parameters for such simulations usually change over time and are included in the category "Other parameters" in Fig. 9.4.

EXAMPLE 9.13. Consider the manufacturing system of Example 5.26. There was a 3-month build schedule available from marketing, which described the types and numbers of computers to be produced each week. The schedule changed from week to week because of changing sales and the introduction of new computers. In this case, weekly or monthly throughputs did not have steady-state distributions. We therefore performed a terminating simulation of length 3 months and estimated the mean throughput for each week.

9.4 STATISTICAL ANALYSIS FOR TERMINATING SIMULATIONS

Suppose that we make n independent replications of a terminating simulation, where each replication is terminated by the event E and is begun with the "same" initial conditions (see Sec. 9.4.3). The independence of replications is accomplished by using different random numbers for each replication. (For a discussion of how this can easily be accomplished if the n replications are made in more than one execution, see Sec. 7.2.) Assume for simplicity that there is a single measure of performance of interest. (This assumption is dropped in Sec. 9.7.) Let X_j be a random variable defined on the jth replication for $j = 1, 2, \ldots, n$; it is assumed that the X_j's are comparable for different replications. Then the X_j's are IID random variables. For the bank of Examples 9.1 and 9.4, X_j might be the average delay $\sum_{i=1}^{N} D_i/N$ over a day (see column 4 in Table 9.1) from the jth replication, where N (a random variable) is the number of customers served in a day. For the combat model of Example 9.5, X_j might be the number of red tanks destroyed on the jth replication. Finally, for the inventory system of Example 9.8, X_j could be the average cost $\sum_{i=1}^{120} C_i/120$ from the jth replication.

9.4.1 Estimating Means

Suppose that we would like to obtain a point estimate and confidence interval for the mean $\mu = E(X)$, where X is a random variable defined on a replication as described above. Make n independent replications of the simulation and let X_1, $X_2, \ldots, X_n$ be the resulting IID random variables. Then, by substituting the X_j's into (4.3) and (4.12), we get that $\overline{X}(n)$ is an unbiased point estimator for μ, and an approximate $100(1 - \alpha)$ percent ($0 < \alpha < 1$) confidence interval for μ is given by

$$\overline{X}(n) \pm t_{n-1,1-\alpha/2}\sqrt{\frac{S^2(n)}{n}} \tag{9.1}$$

where the sample variance $S^2(n)$ is given by Eq. (4.4). We will call the confidence interval based on (9.1) the *fixed-sample-size procedure*. [See also the Willink confidence interval given by (4.13).]

EXAMPLE 9.14. For the bank of Example 9.1, suppose that we want to obtain a point estimate and an approximate 90 percent confidence interval for the expected average delay of a customer over a day, which is given by

$$E(X) = E\left(\frac{\sum_{i=1}^{N} D_i}{N}\right)$$

(Note that we estimate the expected *average* delay, since each delay has, in general, a different mean.) From the 10 replications given in Table 9.1 we obtained

$$\overline{X}(10) = 2.03, \qquad S^2(10) = 0.31$$

and

$$\overline{X}(10) \pm t_{9,0.95}\sqrt{\frac{S^2(10)}{10}} = 2.03 \pm 0.32$$

Thus, subject to the correct interpretation to be given to confidence intervals (see Sec. 4.5), we can claim with approximately 90 percent confidence that $E(X)$ is contained in the interval [1.71, 2.35] minutes.

EXAMPLE 9.15. For the inventory system of Sec. 1.5 and Example 9.8, suppose that we want to obtain a point estimate and an approximate 95 percent confidence interval for the expected average cost over the 120-month planning horizon, which is given by

$$E(X) = E\left(\frac{\sum_{i=1}^{120} C_i}{120}\right)$$

We made 10 independent replications and obtained the following X_j's:

129.35	127.11	124.03	122.13	120.44
118.39	130.17	129.77	125.52	133.75

which resulted in

$$\bar{X}(10) = 126.07, \qquad S^2(10) = 23.55$$

and the 95 percent confidence interval

$$126.07 \pm 3.47 \qquad \text{or, alternatively,} \qquad [122.60, 129.54]$$

Note that the estimated coefficient of variation (see Table 6.5), a measure of variability, is 0.04 for the inventory system and 0.27 for the bank model. Thus the X_j's for the bank model are inherently more variable than those for the inventory system.

EXAMPLE 9.16. For the bank of Example 9.1, suppose that we would like to obtain a point estimate and an approximate 90 percent confidence interval for the expected proportion of customers with a delay less than 5 minutes over a day, which is given by

$$E(X) = E\left(\frac{\sum_{i=1}^{N} I_i(0, 5)}{N}\right)$$

where the *indicator function* $I_i(0, 5)$ is defined as

$$I_i(0, 5) = \begin{cases} 1 & \text{if } D_i < 5 \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, 2, \ldots, N$. From the last column of Table 9.1, we obtained

$$\bar{X}(10) = 0.853, \qquad S^2(10) = 0.004$$

and the 90 percent confidence interval

$$0.853 \pm 0.036 \qquad \text{or} \qquad [0.817, 0.889]$$

The correctness of the confidence interval given by (9.1) (in terms of having coverage close to $1 - \alpha$) depends on the assumption that the X_j's are normal random variables (or on n being "sufficiently large"); this is why we called the confidence intervals in Examples 9.14, 9.15, and 9.16 *approximate*. Since this assumption will rarely be satisfied in practice, we now use several simple stochastic models with *known* means to investigate empirically the robustness of the confidence interval to departures from normality. Our goal is to provide the simulation practitioner with some guidance as to how well the confidence interval will perform, in terms of coverage, in practice.

We first performed 500 independent simulation experiments for the $M/M/1$ queue with $\rho = 0.9$. For each experiment we considered $n = 5, 10, 20, 40$, and for each n we used (9.1) to construct an approximate 90 percent confidence interval for

$$d(25|s = 0) = E\left(\left.\frac{\sum_{i=1}^{25} D_i}{25}\right| s = 0\right) = 2.12$$

TABLE 9.2
Fixed-sample-size results for $d(25|s = 0) = 2.12$ based on 500 experiments, $M/M/1$ queue with $\rho = 0.9$

n	Estimated coverage	Average of (confidence-interval half-length)/$\overline{X}(n)$
5	0.880 ± 0.024	0.67
10	0.864 ± 0.025	0.44
20	0.886 ± 0.023	0.30
40	0.914 ± 0.021	0.21

where s is the number of customers present at time 0 [see Kelton and Law (1985) and Example 9.2]. Table 9.2 gives the proportion, $\hat{p}$, of the 500 confidence intervals that covered the true $d(25|s = 0)$, a 90 percent confidence interval for the true coverage p [the proportion of a very large number of confidence intervals that would cover $d(25|s = 0)$], and the average value of the confidence-interval half-length [that is, $t_{n-1,1-\alpha/2}\sqrt{S^2(n)/n}$] divided by the point estimate $\overline{X}(n)$ over the 500 experiments, which is a measure of the precision of the confidence interval; see below for further discussion. The 90 percent confidence interval for the true coverage is computed from

$$\hat{p} \pm z_{0.95}\sqrt{\frac{\hat{p}(1-\hat{p})}{500}}$$

and is based on the fact that $(\hat{p} - p)/\sqrt{\hat{p}(1-\hat{p})/500}$ is approximately distributed as a standard normal random variable [see, e.g., Hogg and Craig (1995, pp. 254–255)]. It is recommended that this confidence interval for p only be used if $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$. If this is not the case in a particular situation, then the *score* confidence interval [Devore (2008, p. 266)] might be used instead.

From Table 9.2 it can be seen that 86.4 percent of the 500 confidence intervals based on $n = 10$ replications covered $d(25|s = 0)$, and we know with approximately 90 percent confidence that the true coverage for $n = 10$ is between 0.839 and 0.889. Considering that a simulation model is always just an approximation to the corresponding real-world system, we believe that the estimated coverages presented in Table 9.2 are close enough to the desired 0.9 to be useful. Note also from the last column of the table that four times as many replications are required to increase the precision of the confidence interval by a factor of approximately 2. This is not surprising since there is a $\sqrt{n}$ in the denominator of the expression for the confidence-interval half-length in (9.1).

To show that the confidence interval given by (9.1) does not always produce coverages close to $1 - \alpha$, we considered a second example. A reliability model consisting of three components will function as long as component 1 works and either component 2 or 3 works. If G is the time to failure of the whole system and G_i is the time to failure of component i (where $i = 1, 2, 3$), then $G = \min\{G_1, \max\{G_2, G_3\}\}$. We further assume that the G_i's are independent random variables and that each G_i has a Weibull distribution with shape parameter 0.5 and scale parameter 1 (see Sec. 6.2.2). This particular Weibull distribution is extremely skewed and

TABLE 9.3
Fixed-sample-size results for $E(G \mid \text{all components new}) = 0.78$ based on 500 experiments, reliability model

n	Estimated coverage	Average of (confidence-interval half-length)/$\overline{X}(n)$
5	0.708 ± 0.033	1.16
10	0.750 ± 0.032	0.82
20	0.800 ± 0.029	0.60
40	0.840 ± 0.027	0.44

non-normal. Once again we performed 500 independent simulation experiments; for each experiment we considered $n = 5, 10, 20, 40$, and for each n we used (9.1) to construct a 90 percent confidence interval for $E(G \mid \text{all components new}) = 0.78$ (which was calculated by analytic reasoning). The results from these experiments are given in Table 9.3. Note that for small values of n there is significant coverage degradation. Also, as n gets large, the coverage appears to be approaching 0.9, as guaranteed by the central limit theorem. The Willink confidence interval could possibly be used for this model unless n is "large."

We can see from Tables 9.2 and 9.3 that the coverage actually obtained from the confidence interval given by (9.1) depends on the simulation model under consideration (actually, on the distribution of the resulting X_j's) and also on the sample size n. It is therefore natural to ask why the confidence interval worked better for the $M/M/1$ queue than it did for the reliability model. To answer this question, we first performed 500 independent simulation experiments for the $M/M/1$ queue with $\rho = 0.9$ and $s = 0$ ($n = 1$), and we observed $\sum_{i=1}^{25} D_i/25$ on each replication. A histogram of the 500 average delays is given in Fig. 9.5, and the sample skewness was 1.64 (see

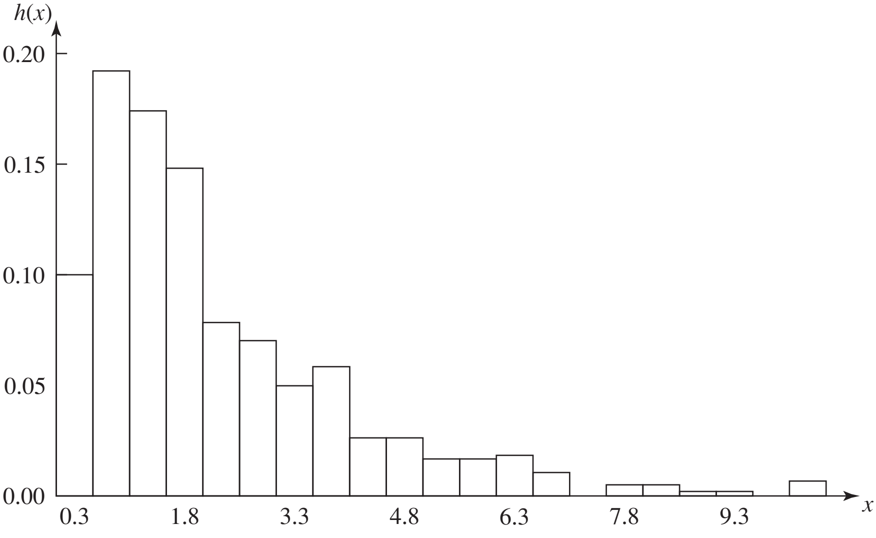


FIGURE 9.5
Histogram of 500 average delays (each based on 25 individual delays) for the $M/M/1$ queue with $\rho = 0.9$.

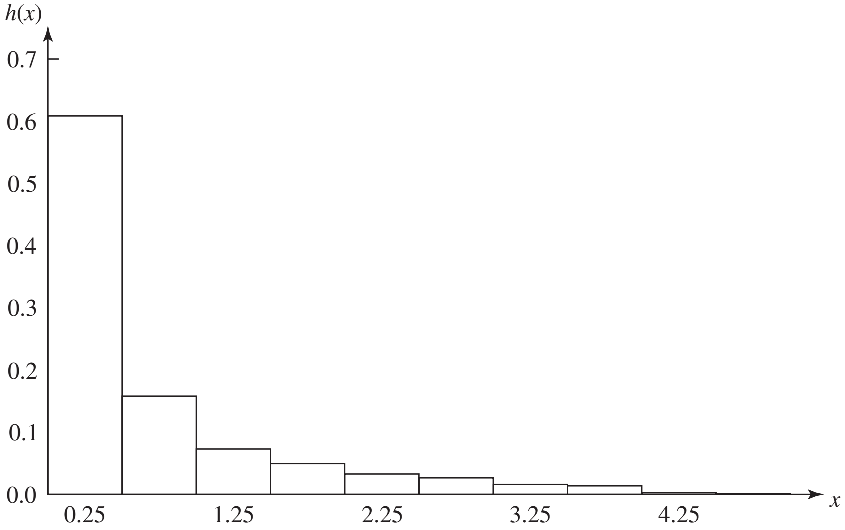


FIGURE 9.6
Histogram of 500 times to failure for the reliability model.

Tables 4.1 and 6.5). Although the histogram indicates that average delay is not normally distributed, it does show that the distribution of average delay is not extremely skewed. (For example, an exponential distribution has a skewness of 2.) We next performed 500 independent experiments for the reliability model and observed the time to failure G on each replication. A histogram of the 500 values of G is given in Fig. 9.6, and the estimated skewness was 3.64. Thus, the distribution of time to failure is considerably more non-normal than the distribution of average delay. These results shed some light on why the coverages for the $M/M/1$ queue are closer to 0.9 than for the reliability model.

The reader might wonder why average delay is more normally distributed than time to failure. Note that an X_j for the $M/M/1$ queue is actually an average of 25 individual delays, while an X_j for the reliability model is computed from the three individual times to failure by a formula involving a minimum and a maximum. There are central limit theorems for certain types of correlated data that state that averages of these data become approximately normally distributed as the number of points in the average gets large. To show this for the $M/M/1$ queue, we performed 500 independent experiments and observed $\sum_{i=1}^{6400} D_i/6400$ on each replication. A histogram of the 500 average delays (each based on 6400 individual delays) is given in Fig. 9.7, and the estimated skewness was 1.07. (The skewness of a normal distribution is 0.) Clearly, the histogram in Fig. 9.7 is closer to a normal distribution than the histogram in Fig. 9.5.

We therefore expect that if X_j is the average of a large number of individual observations (even though correlated), the degradation in coverage of the confidence interval may not be severe. Our experience indicates that many real-world simulations produce X_j's of this type.

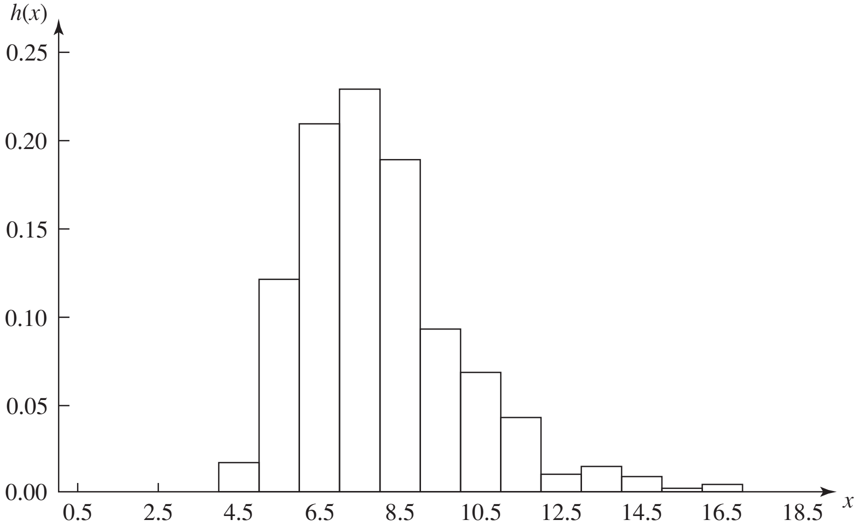


FIGURE 9.7
Histogram of 500 average delays (each based on 6400 individual delays) for the *M*/*M*/1 queue with $\rho = 0.9$.

Obtaining a Specified Precision

One disadvantage of the fixed-sample-size procedure based on *n* replications is that the analyst has no control over the confidence-interval half-length [or the precision of $\overline{X}(n)$]; for fixed *n*, the half-length will depend on Var(*X*), the population variance of the X_j's. In what follows we discuss procedures for determining the number of replications required to estimate the mean $\mu = E(X)$ with a specified error or precision.

We begin by defining two ways of measuring the error in the estimate $\overline{X}$. (The dependence on *n* is suppressed, since the number of replications may be a random variable.) If the estimate $\overline{X}$ is such that $|\overline{X} - \mu| = \beta$, then we say that $\overline{X}$ has an *absolute error* of β. If we make replications of a simulation until the half-length of the $100(1 - \alpha)$ percent confidence interval given by (9.1) is less than or equal to β (where $\beta > 0$), then

$$\begin{aligned} 1 - \alpha &\approx P(\overline{X} - \text{half-length} \le \mu \le \overline{X} + \text{half-length}) \\ &= P(|\overline{X} - \mu| \le \text{half-length}) \\ &\le P(|\overline{X} - \mu| \le \beta) \end{aligned}$$

[If *A* and *B* are events with *A* being a subset of *B*, then $P(A) \le P(B)$.] Thus, $\overline{X}$ has an absolute error of at most β with a probability of approximately $1 - \alpha$. In other words, if we construct 100 independent 90 percent confidence intervals using the above stopping rule, we would expect $\overline{X}$ to have an absolute error of at most β in about 90 out of the 100 cases; in about 10 cases the absolute error would be greater than β.

Suppose that we have constructed a confidence interval for μ based on a fixed number of replications n. If we assume that our estimate $S^2(n)$ of the population variance will not change (appreciably) as the number of replications increases, an *approximate* expression for the total number of replications, $n_a^*(\beta)$, required to obtain an absolute error of β is given by

$$n_a^*(\beta) = \min\left\{ i \geq n\text{: } t_{i-1,1-\alpha/2}\sqrt{\frac{S^2(n)}{i}} \leq \beta \right\} \tag{9.2}$$

(The colon ":" is read "such that.") We can determine $n_a^*(\beta)$ by iteratively increasing i by 1 until a value of i is obtained for which $t_{i-1,1-\alpha/2}\sqrt{S^2(n)/i} \leq \beta$. [Alternatively, $n_a^*(\beta)$ can be approximated as the smallest integer i satisfying $i \geq S^2(n)(z_{1-\alpha/2}/\beta)^2$.] If $n_a^*(\beta) > n$ and if we make $n_a^*(\beta) - n$ additional replications of the simulation, then the estimate $\overline{X}$ based on all $n_a^*(\beta)$ replications should have an absolute error of approximately β. The accuracy of Eq. (9.2) depends on how close the variance estimate $S^2(n)$ is to Var(X).

EXAMPLE 9.17. For the bank of Example 9.14, suppose that we would like to estimate the expected average delay with an absolute error of 0.25 minute and a confidence level of 90 percent. From the 10 available replications, we get

$$n_a^*(0.25) = \min\left\{ i \geq 10\text{: } t_{i-1,0.95}\sqrt{\frac{0.31}{i}} \leq 0.25 \right\} = 16$$

We now discuss another way of measuring the error in $\overline{X}$. Assume now that $\mu \neq 0$. If the estimate $\overline{X}$ is such that $|\overline{X} - \mu|/|\mu| = \gamma$, then we say that $\overline{X}$ has a *relative error* of γ, or that the *percentage error* in $\overline{X}$ is 100γ percent. Suppose that we make replications of a simulation until the half-length of the confidence interval given by (9.1), divided by $|\overline{X}|$, is less than or equal to $\gamma(0 < \gamma < 1)$. This ratio is an estimate of the actual relative error. Then

$$\begin{aligned}
1 - \alpha &\approx P(|\overline{X} - \mu|/|\overline{X}| \leq \text{half-length}/|\overline{X}|) \\
&\leq P(|\overline{X} - \mu| \leq \gamma|\overline{X}|) && [(\text{half-length}/|\overline{X}|) \leq \gamma] \\
&= P(|\overline{X} - \mu| \leq \gamma|\overline{X} - \mu + \mu|) && (\text{add, subtract } \mu) \\
&\leq P(|\overline{X} - \mu| \leq \gamma(|\overline{X} - \mu| + |\mu|)) && (\text{triangle inequality}) \\
&= P((1 - \gamma)|\overline{X} - \mu| \leq \gamma|\mu|) && (\text{algebra}) \\
&= P(|\overline{X} - \mu|/|\mu| \leq \gamma/(1 - \gamma)) && (\text{algebra})
\end{aligned}$$

Thus, $\overline{X}$ has a relative error of at most $\gamma/(1 - \gamma)$ with a probability of approximately $1 - \alpha$. In other words, if we construct 100 independent 90 percent confidence intervals using the above stopping rule, we would expect $\overline{X}$ to have a relative error of at most $\gamma/(1 - \gamma)$ in about 90 of the 100 cases; in about 10 cases the relative error would be greater than $\gamma/(1 - \gamma)$. Note that we get a relative error of $\gamma/(1 - \gamma)$ rather than the desired γ, since we *estimate* $|\mu|$ by $|\overline{X}|$.

Suppose once again that we have constructed a confidence interval for μ based on a fixed number of replications n. If we assume that our estimates of both the population mean and population variance will not change (appreciably) as the number of replications increases, an *approximate* expression for the number of replications, $n_r^*(\gamma)$, required to obtain a relative error of γ is given by

$$n_r^*(\gamma) = \min\left\{ i \geq n\colon \frac{t_{i-1,1-\alpha/2}\sqrt{S^2(n)/i}}{|\overline{X}(n)|} \leq \gamma' \right\} \tag{9.3}$$

where $\gamma' = \gamma/(1 + \gamma)$ is the "adjusted" relative error needed to get an *actual* relative error of γ. {Again, $n_r^*(\gamma)$ is approximated as the smallest integer i satisfying $i \geq S^2(n)[z_{1-\alpha/2}/(\gamma'\overline{X}(n))]^2$.} If $n_r^*(\gamma) > n$ and if we make $n_r^*(\gamma) - n$ additional replications of the simulation, then the estimate $\overline{X}$ based on all $n_r^*(\gamma)$ replications should have a relative error of approximately γ.

EXAMPLE 9.18. For the bank of Example 9.14, suppose that we would like to estimate the expected average delay with a relative error of 0.10 and a confidence level of 90 percent. From the 10 available replications, we get

$$n_r^*(0.10) = \min\left\{ i \geq 10\colon \frac{t_{i-1,0.95}\sqrt{0.31/i}}{2.03} \leq 0.09 \right\} = 27$$

where $\gamma' = 0.1/(1 + 0.1) = 0.09$.

The difficulty with using Eq. (9.3) directly to obtain an estimate $\overline{X}$ with a relative error of γ is that $\overline{X}(n)$ and $S^2(n)$ may not be precise estimates of their corresponding population parameters. If $n_r^*(\gamma)$ is greater than the number of replications actually required, then a significant number of unnecessary replications may be made, resulting in a waste of computer resources. Conversely, if $n_r^*(\gamma)$ is too small, then an estimate $\overline{X}$ based on $n_r^*(\gamma)$ replications may not be as precise as we think. We now present a *sequential* procedure (new replications are added one at a time) for obtaining an estimate of μ with a specified relative error that takes only as many replications as are actually needed. The procedure assumes that $X_1, X_2, \ldots$ is a sequence of IID random variables that need not be normal.

The specific objective of the procedure is to obtain an estimate of μ with a relative error of $\gamma (0 < \gamma < 1)$ and a confidence level of $100(1 - \alpha)$ percent. Choose an initial number of replications $n_0 \geq 10$ and let

$$\delta(n, \alpha) = t_{n-1,1-\alpha/2}\sqrt{\frac{S^2(n)}{n}}$$

be the usual confidence-interval half-length. Then the sequential procedure is as follows:

0. Make n_0 replications of the simulation and set $n = n_0$.
1. Compute $\overline{X}(n)$ and $\delta(n, \alpha)$ from $X_1, X_2, \ldots, X_n$.
2. If $\delta(n, \alpha)/|\overline{X}(n)| \leq \gamma'$, use $\overline{X}(n)$ as the point estimate for μ and stop. Equivalently,

$$I(\alpha, \gamma) = [\overline{X}(n) - \delta(n, \alpha), \overline{X}(n) + \delta(n, \alpha)] \tag{9.4}$$

is an approximate $100(1 - \alpha)$ percent confidence interval for μ with the desired precision. Otherwise, replace n by $n + 1$, make an additional replication of the simulation, and go to step 1.

Note that the procedure computes a new estimate of $\text{Var}(X)$ after *each* replication is obtained, and that the total number of replications required by the procedure is a random variable.

EXAMPLE 9.19. For the bank of Example 9.14, suppose that we would like to obtain an estimate of the expected average delay with a relative error of $\gamma = 0.1$ and a confidence level of 90 percent. Using the previous $n_0 = 10$ replications as a starting point, we obtained

$$\text{Number of replications at termination} = 74$$

$$\bar{X}(74) = 1.76, \qquad S^2(74) = 0.67$$

$$\text{90 percent confidence interval: } [1.60, 1.92]$$

Note that the number of replications actually required, 74, is considerably larger than the 27 predicted in Example 9.18, due mostly to the imprecise variance estimate based on 10 replications.

Although the sequential procedure described above is intuitively appealing, the question naturally arises as to how well it performs in terms of producing a confidence interval with coverage close to the desired $1 - \alpha$. In Law, Kelton, and Koenig (1981), it is shown that if $\mu \neq 0$ [and $0 < \text{Var}(X) < \infty$], then the coverage of the confidence interval given by Eq. (9.4) will be arbitrarily close to $1 - \alpha$, provided the desired relative error is sufficiently close to 0. Based on sampling from a large number of stochastic models and probability distributions (including the *M/M/*1 queue and the above reliability model) for which the true values of μ are known, our recommendation is to use the sequential procedure with $n_0 \geq 10$ and $\gamma \leq 0.15$. It was found that if these recommendations are followed, the estimated coverage (based on 500 independent experiments for each model) for a desired 90 percent confidence interval was never less than 0.864.

Analogous to the sequential procedure described above is a sequential procedure due to Chow and Robbins (1965) for constructing a $100(1 - \alpha)$ percent confidence interval for μ with a small absolute error β. Furthermore, it can be shown that the coverage actually produced by the procedure will be arbitrarily close to $1 - \alpha$ provided the desired absolute error β is sufficiently close to 0. However, since the meaning of "*absolute error* sufficiently small" is extremely model-dependent, and since the coverage results in Law (1980) indicate that the procedure is very sensitive to the choice of β, we do not recommend the use of the Chow and Robbins procedure in general.

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 confidence interval may not be overwhelmingly important, we recommend using the fixed-sample-size procedure. However, if the X_j's are highly non-normal and the number of replications n is too small, the actual coverage of the constructed confidence interval may be somewhat lower than desired. In this case consider the use of the Willink confidence interval.

From an exploratory experiment consisting of n replications, one can estimate the execution time per replication and the population variance of the X_j's, and then obtain from Eq. (9.2) a *rough estimate* of the number of replications, $n_a^*(\beta)$, required to estimate μ with a desired absolute error β. Alternatively, one can obtain from Eq. (9.3) a *rough estimate* of the number of replications, $n_r^*(\gamma)$, required to estimate μ with a desired relative error γ. Sometimes the choice of β or γ may have to be tempered by the execution time associated with the required number of replications. If it is finally decided to construct a confidence interval with a small relative error γ, we recommend use of the sequential procedure with $\gamma \leq 0.15$ and $n_0 \geq 10$. If one wants a confidence interval with a relative error γ greater than 0.15, we recommend several successive applications of the fixed-sample-size approach. In particular, one might estimate $n_r^*(\gamma)$, collect, say $[n_r^*(\gamma) - n]/2$ more replications, and then use (9.1) to construct a confidence interval based on the existing $[n + n_r^*(\gamma)]/2$ replications. If the estimated relative error of the resulting confidence interval is still greater than γ', then $n_r^*(\gamma)$ can be reestimated based on a new variance estimate, and some portion of the necessary additional replications may be collected, etc. To construct a confidence interval with a small absolute error β, we once again recommend several successive applications of the fixed-sample-size approach.

Regardless of the time per replication, we recommend always making at least three to five replications of a stochastic simulation to assess the variability of the X_j's. If this is not possible due to time considerations, then the simulation study should probably not be done at all.

9.4.2 Estimating Other Measures of Performance

In this section we discuss estimating measures of performance other than means. As the following example shows, comparing two or more systems by some sort of mean system response may result in misleading conclusions.

EXAMPLE 9.20. Consider the bank of Example 9.14, where the utilization factor $\rho = \lambda/(5\omega) = 0.8$. We compare the policy of having one queue for each teller (and jockeying) with the policy of having one queue feed all tellers on the basis of *expected average delay in queue* (see Example 9.14) and *expected time-average number of customers in queue*, which is defined by

$$E\left[\frac{\int_0^T Q(t)\,dt}{T}\right]$$

where $Q(t)$ is the number of customers in queue at time t and T is the bank's operating time ($T \geq 8$ hours). Table 9.4 gives the results of making one simulation run of each policy. [These simulation runs were performed so that the time of arrival of the ith customer ($i = 1, 2, \ldots, N$) was identical for both policies and so that the service time of

TABLE 9.4
Simulation results for the two bank policies: averages

	Estimates	
Measure of performance	**Five queues**	**One queue**
Expected operating time, hours	8.14	8.14
Expected average delay, minutes	5.57	5.57
Expected average number in queue	5.52	5.52

the *i*th customer to begin service ($i = 1, 2, \ldots, N$) was the same for both policies.] Thus, on the basis of "average system response," it would appear that the two policies are equivalent. However, this is clearly not the case. Since customers need not be served in the order of their arrival with the multiqueue policy, we would expect this policy to result in greater variability of a customer's delay. Table 9.5 gives estimates, computed from the same two simulation runs used above, of the expected proportion of customers with a delay in the interval [0, 5) (in minutes), the expected proportion of customers with a delay in [5, 10), . . . , the expected proportion of customers with a delay in [40, 45) for both policies. (We did not estimate variances from these runs since, as pointed out in Sec. 4.4, variance estimates computed from correlated simulation output data are highly biased.) Observe from Table 9.5 that a customer is more likely to have a large delay with the multiqueue policy than with the single-queue policy. In particular, if 480 customers arrive in a day (the expected number), then 33 and 6 of them would be expected to have delays greater than or equal to 20 minutes for the five-queue and one-queue policies, respectively. (For larger values of ρ, the differences between the two policies would be even greater.) This observation together with the greater equitability of the single-queue policy has probably led many organizations, e.g., banks and airlines, to adopt this policy.

We conclude from the above example that comparing alternative systems or policies on the basis of average system behavior alone can sometimes result in misleading conclusions and, furthermore, that proportions can be a useful measure of system performance. In Example 9.16 we showed how to obtain a point estimate

TABLE 9.5
Simulation results for the two bank policies: proportions

	Estimates of expected proportions of delays in interval	
Interval (minutes)	**Five queues**	**One queue**
[0, 5)	0.626	0.597
[5, 10)	0.182	0.188
[10, 15)	0.076	0.107
[15, 20)	0.047	0.095
[20, 25)	0.031	0.013
[25, 30)	0.020	0
[30, 35)	0.015	0
[35, 40)	0.003	0
[40, 45)	0	0

and a confidence interval for an expected proportion. In this section we show how to perform similar analyses for probabilities and quantiles in the context of terminating simulations.

Let X be a random variable defined on a replication as described in Sec. 9.4.1. Suppose that we would like to estimate the probability $p = P(X \in B)$, where B is a set of real numbers. {For example, B could be the interval $[20, \infty)$ in Example 9.20.} Make n independent replications and let $X_1, X_2, \ldots, X_n$ be the resulting IID random variables. Let S be the number of X_j's that fall in the set B. Then S has a binomial distribution (see Sec. 6.2.3) with parameters n and p, and an unbiased point estimator for p is given by

$$\hat{p} = \frac{S}{n}$$

Furthermore, if n is "sufficiently large," then an approximate $100(1 - \alpha)$ percent confidence interval for p is given by

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

[see Devore (2008, p. 266) for an alternative confidence interval and also Prob. 9.9].

EXAMPLE 9.21. For the bank of Example 9.14, suppose that we would like to get a point estimate and approximate 90 percent confidence interval for

$$p = P(X \le 15) \qquad \text{where } X = \max_{0 \le t \le T} Q(t)$$

In this case, $B = [0, 15]$. We made 100 independent replications of the bank simulation and obtained $\hat{p} = 0.77$. Thus, for approximately 77 out of every 100 days, we expect the maximum queue length during a day to be less than or equal to 15 customers. We also obtained the following approximate 90 percent confidence interval for p:

$$0.77 \pm 0.07 \qquad \text{or, alternatively,} \qquad [0.70, 0.84]$$

Suppose now that we would like to estimate the q-quantile ($100q$th percentile) x_q of the distribution of the random variable X (see Sec. 6.4.3 for the definition). For example, the 0.5-quantile is the median. If $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the order statistics corresponding to the X_j's from n independent replications, then a point estimator for x_q is the sample q-quantile $\hat{x}_q$, which is given by

$$\hat{x}_q = \begin{cases} X_{(nq)} & \text{if } nq \text{ is an integer} \\ X_{(\lfloor nq+1 \rfloor)} & \text{otherwise} \end{cases}$$

Let r and s be positive integers that satisfy $1 \le r < s \le n$. If n is "sufficiently large," then a $100(1 - \alpha)$ percent confidence interval for x_q is given by [see Conover (1999, pp. 143–148)]

$$P(X_{(r)} \le x_q \le X_{(s)}) \ge 1 - \alpha$$

where

$$r = \lceil nq + z_{\alpha/2}\sqrt{nq(1-q)} \rceil$$

and

$$s = \lceil nq + z_{1-\alpha/2}\sqrt{nq(1-q)} \rceil$$

The greater than or equal to sign in the confidence-interval expression becomes an equal sign if X is a continuous random variable.

> **EXAMPLE 9.22.** For the bank of Example 9.14, suppose that we would like to decide how large a lobby is needed to accommodate customers waiting in the queue. If we let X be the maximum queue length as defined in Example 9.21, then we might want to build a lobby large enough to hold $x_{0.95}$ customers, the 0.95-quantile of X. From the 100 replications in the previous example, we obtained $\hat{x}_{0.95} = X_{(95)} = 20$. Thus, if the lobby has room for 20 customers waiting in queue, this will be sufficient for approximately 95 out of every 100 days. Furthermore, an approximate 90 percent confidence interval for $x_{0.95}$ is $[X_{(91)}, X_{(99)}] = [19, 23]$. (For this problem, X is a discrete random variable, so that the confidence level is approximate.)

The interested reader may also want to consult Conover (1999, pp. 150–155) for a discussion of a *tolerance interval*, which is an interval that contains a specified proportion of the *values* of the random variable X (and does so with a certain prescribed confidence level).

9.4.3 Choosing Initial Conditions

As stated in Sec. 9.3, the measures of performance for a terminating simulation depend explicitly on the state of the system at time 0; thus, care must be taken in choosing appropriate initial conditions. Let us illustrate this potential problem by means of an example. Suppose that we would like to estimate the expected average delay of all customers who arrive and complete their delays between 12 noon and 1 P.M. (the busiest period) in a bank. Since the bank will probably be quite congested at noon, starting the simulation then with no customers present (the usual initial conditions for a queueing simulation) will cause our estimate of expected average delay to be biased low. We now discuss two heuristic approaches to this problem, the first of which appears to be used widely (see Sec. 9.5.1).

For the first approach, let us assume that the bank opens at 9 A.M. with no customers present. Then we can start the simulation at 9 A.M. with no customers present and run it for 4 simulated hours. In estimating the desired expected average delay, we use only the delays of those customers who arrive and complete their delays between noon and 1 P.M. The evolution of the simulation between 9 A.M. and noon (the "warmup period") determines the appropriate conditions for the simulation at noon. A disadvantage of this approach is that 3 hours of simulated time are not used directly in the estimate. As a result, one might compromise and start the simulation at some other time, say 11 A.M., with no customers present. However, there is no guarantee that the conditions in the simulation at noon will be representative of the actual conditions in the bank at noon.

An alternative approach is to collect data on the number of customers present in the bank at noon for several different days. Let $\hat{p}_i$ be the proportion of these days that i customers ($i = 0, 1, \ldots$) are present at noon. Then we simulate the bank from noon to 1 P.M. with the number of customers present at noon being randomly chosen from the distribution $\{\hat{p}_i\}$. (All customers who are being served at noon might be assumed to be just beginning their services. Starting all services fresh at noon results in an approximation to the actual situation in the bank, since the customers who are in the process of being served at noon would have partially completed their services. However, the effect of this approximation should be negligible for a simulation of length 1 hour.)

If more than one simulation run from noon to 1 P.M. is desired, then a different sample from $\{\hat{p}_i\}$ is drawn for each run. The X_j's that result from these runs are still IID, since the initial conditions for each run are chosen independently from the same distribution.

9.5 STATISTICAL ANALYSIS FOR STEADY-STATE PARAMETERS

Let $Y_1, Y_2, \ldots$ be an output stochastic process from a single run of a nonterminating simulation. Suppose that $P(Y_i \le y) = F_i(y) \to F(y) = P(Y \le y)$ as $i \to \infty$, where Y is the steady-state random variable of interest with distribution function F. (We have suppressed in our notation the dependence of F_i on the initial conditions I.) Then ϕ is a steady-state parameter if it is a characteristic of Y such as $E(Y)$, $P(Y \le y)$, or a quantile of Y. One difficulty in estimating ϕ is that the distribution function of Y_i (for $i = 1, 2, \ldots$) is different from F, since it will generally not be possible to choose I to be representative of "steady-state behavior." This causes an estimator of ϕ based on the observations $Y_1, Y_2, \ldots, Y_m$ not to be "representative." For example, the sample mean $\overline{Y}(m)$ will be a biased estimator of $\nu = E(Y)$ for all finite values of m. The problem we have just described is called the *problem of the initial transient* or the *startup problem* in the simulation literature.

EXAMPLE 9.23. To illustrate the startup problem more succinctly, consider the process of delays $D_1, D_2, \ldots$ for the $M/M/1$ queue with $\rho < 1$ (see Example 9.2). From queueing theory, it is possible to show that

$$P(D_i \le y) \to P(D \le y) = (1 - \rho) + \rho\left[1 - e^{-(\omega-\lambda)y}\right] \qquad \text{as } i \to \infty$$

If the number of customers s present at time 0 is 0, then $D_1 = 0$ and $E(D_i) \ne E(D) = d$ for any i. On the other hand, if s is chosen in accordance with the steady-state number in system distribution [see, for example, Gross et al. (2009)], then for all i, $P(D_i \le y) = P(D \le y)$ and $E(D_i) = d$ (see Prob. 9.11). Thus, there is no initial transient in this case.

In practice, the steady-state distribution will not be known and the above initialization technique will not be possible. Techniques for dealing with the startup problem in practice are discussed in the next section.

9.5.1 The Problem of the Initial Transient

Suppose that we want to estimate the steady-state mean $\nu = E(Y)$, which is also generally defined by

$$\nu = \lim_{i\to\infty} E(Y_i)$$

Thus, the transient means converge to the steady-state mean. The most serious consequence of the problem of the initial transient is probably that $E[\overline{Y}(m)] \neq \nu$ for any m [see Law (1983, pp. 1010–1012) for further discussion]. The technique most often suggested for dealing with this problem is called *warming up the model* or *initial-data deletion*. The idea is to delete some number of observations from the beginning of a run and to use only the remaining observations to estimate ν. For example, given the observations $Y_1, Y_2, \ldots, Y_m$, it is often suggested to use

$$\overline{Y}(m, l) = \frac{\sum_{i=l+1}^{m} Y_i}{m - l}$$

($1 \leq l \leq m - 1$) rather than $\overline{Y}(m)$ as an estimator of ν. In general, one would expect $\overline{Y}(m, l)$ to be less biased than $\overline{Y}(m)$, since the observations near the "beginning" of the simulation may not be very representative of steady-state behavior due to the choice of initial conditions. [If $\hat{\theta}$ is an estimator for a parameter θ, then the bias in $\hat{\theta}$ is $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$.] For example, this is true for the process $D_1, D_2, \ldots$ in the case of an $M/M/1$ queue with $s = 0$, since $E(D_i)$ increases monotonically to d as $i \to \infty$ (see Fig. 9.2). Fishman (1972) showed that this is also the case for a first-order autoregressive [AR(1)] process (see Sec. 6.10.3).

Some authors, however, have questioned the efficacy of initial-data deletion [see, for example, Grassmann (2011)], so we first look at the point estimator $\overline{Y}(m, l)$ more carefully. Although $\overline{Y}(m, l)$ will generally be less biased than $\overline{Y}(m)$, it will often have a larger variance, as was shown by Fishman (1972) for an AR(1) process. However, probably the most commonly used overall measure of point-estimator quality is mean-squared error [see Pasupathy and Schmeiser (2010)]. Blomqvist (1970) showed for the $M/M/1$ queue (and certain other queueing systems) with m sufficiently large, that zero is the value of l that minimizes the mean-squared error of $\overline{D}(m, l)$, which is defined by

$$\text{MSE}[\overline{D}(m, l)] = E\{[\overline{D}(m, l) - d]^2\} = \{\text{Bias}[\overline{D}(m, l)]\}^2 + \text{Var}[\overline{D}(m, l)]$$

On the other hand, Snell and Schruben (1979) and Kelton (1980) showed for a AR(1) process that deletion may either increase or decrease mean-squared error, depending on m, l, and the values of the process parameters. As one might suspect, deletion most significantly reduced mean-squared error when the initialization bias was high and the autocorrelation (see Sec. 5.6) was heavy, causing the bias to dissipate slowly. In these cases, the value of l that minimized $\text{MSE}[\overline{Y}(m, l)]$ decreased as m increased. This observation is consistent with Blomqvist's result that for very large values of m, deletion is not advisable for the mean-squared-error performance measure.

Another criterion that is used for evaluating the efficacy of deletion is confidence-interval quality, which is the one that we prefer. [We believe that one should always construct a confidence interval for ν; otherwise, we have no explicit way of knowing how close $\overline{Y}(m, l)$ is to ν.] The replication/deletion approach for constructing a confidence interval for ν, which is discussed in Sec. 9.5.2, is based on making n independent "short" replications of the process $Y_1, Y_2, \ldots$ of length m observations and deleting the first l observations from each replication. Let $\overline{Y}_j(m, l)$ be the sample mean of the final $m - l$ observations in the jth replication, for $j = 1, 2, \ldots, n$. Then the $\overline{Y}_j(m, l)$'s play the same role as the X_j's in the confidence interval given by (4.12). In order for the replication/deletion approach to produce confidence intervals with acceptable coverage, it is *critical* that m and l be chosen so that $E[\overline{Y}_j(m, l)] \approx \nu$, i.e., that $\overline{Y}_j(m, l)$ is an approximately *unbiased estimator* for ν [see Law (1977)].

Suppose now that we make one "long" run of length m observations, resulting in the observations $Y_1, Y_2, \ldots, Y_m$. There are a number of methods for constructing a confidence interval for ν based on this one-replication scenario (see Sec. 9.5.3). For example, the method of batch means, which is the most widely used of these approaches, divides the m observations into n batches of size k ($m = nk$). Let $\overline{Y}_j(k)$ be the sample (or batch) mean of the k observations in the jth batch, for $j = 1, 2, \ldots, n$. Then the $\overline{Y}_j(k)$'s play the same role as the X_j's in the confidence interval given by (4.12). In order for the batch-means method to produce confidence intervals with acceptable coverage, k must be chosen large enough so that the $\overline{Y}_j(k)$'s are approximately *uncorrelated*. Results in Law (1977), Law and Carson (1979), and Law and Kelton (1984), suggest that batch means (and the other methods of Sec. 9.5.3) will produce confidence intervals with acceptable coverage *without a warmup period* (i.e., $l = 0$), provided that m is "moderate" in value. For instance, in the case of an $M/M/1$ queue with $\rho = 0.9$ and $s = 0$, batch means achieved a coverage of 0.865 for a nominal 90 percent confidence interval for d based on $m = 12{,}800$ and $n = 5$ batches of size 2560 [Law (1977)]. Apparently, if m is large enough, then the observations from the initial transient get "washed out" by the remaining "steady-state" observations.

Thus, in terms of confidence-interval coverage, the need for an effective warmup period appears to be much more important when using the replication/deletion approach, which is based on multiple "short" replications. Thus, we will focus on the replication/deletion approach and its need for an unbiased point estimator in our discussion of the problem of the initial transient that follows.

The question naturally arises as to how to choose the *warmup period* (or deletion amount) l. We would like to pick l (and m) such that $E[\overline{Y}(m, l)] \approx \nu$. If l and m are chosen too small, then $E[\overline{Y}(m, l)]$ may be significantly different from ν. On the other hand, if l is chosen larger than necessary, then $\overline{Y}(m, l)$ will probably have an unnecessarily large variance. There have been a number of methods suggested in the literature for choosing l. However, Gafarian, Ancker, and Morisaku (1978) found that none of the methods available at that time performed well in practice. Kelton and Law (1983) developed an algorithm for choosing l (and m) that worked well {that is, $E[\overline{Y}(m, l)] \approx \nu$} for a wide variety of stochastic models. However, a theoretical limitation of the procedure is that it basically makes the assumption that $E(Y_i)$ is a monotone function of i.

A simple and general technique for determining l is a graphical procedure due to Welch (1981, 1983). Its specific goal is to determine a time index l such that $E(Y_i) \approx \nu$ for $i > l$, where l is the warmup period. [This is equivalent to determining when the transient mean curve $E(Y_i)$ (for $i = 1, 2, \ldots$) "flattens out" at level ν; see Fig. 9.1.] In general, it is very difficult to determine l from a single replication due to the inherent variability of the process $Y_1, Y_2, \ldots$ (see Fig. 9.10 below).

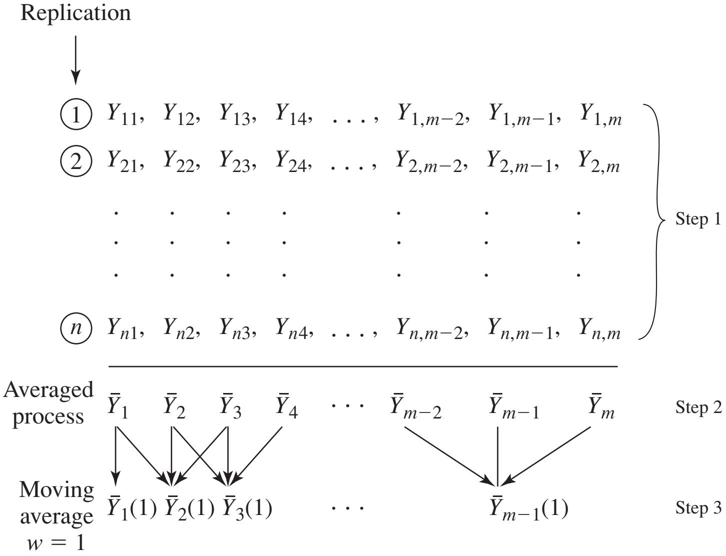

Replication
1 Y_{11}, Y_{12}, Y_{13}, Y_{14}, . . . , $Y_{1,m-2}$, $Y_{1,m-1}$, $Y_{1,m}$
2 Y_{21}, Y_{22}, Y_{23}, Y_{24}, . . . , $Y_{2,m-2}$, $Y_{2,m-1}$, $Y_{2,m}$
n Y_{n1}, Y_{n2}, Y_{n3}, Y_{n4}, . . . , $Y_{n,m-2}$, $Y_{n,m-1}$, $Y_{n,m}$
Step 1
Averaged process
$\bar{Y}_1$ $\bar{Y}_2$ $\bar{Y}_3$ $\bar{Y}_4$ . . . $\bar{Y}_{m-2}$ $\bar{Y}_{m-1}$ $\bar{Y}_m$
Step 2
Moving average $w = 1$
$\bar{Y}_1(1)$ $\bar{Y}_2(1)$ $\bar{Y}_3(1)$ . . . $\bar{Y}_{m-1}(1)$
Step 3

FIGURE 9.8
Averaged process and moving average with $w = 1$ based on n replications of length m.

As a result, Welch's procedure is based on making n independent replications of the simulation and employing the following four steps:

1. Make n replications of the simulation ($n \geq 5$), each of length m (where m is large). Let Y_{ji} be the ith observation from the jth replication ($j = 1, 2, \ldots, n$; $i = 1, 2, \ldots, m$), as shown in Fig. 9.8.
2. Let $\bar{Y}_i = \sum_{j=1}^{n} Y_{ji}/n$ for $i = 1, 2, \ldots, m$ (see Fig. 9.8). The averaged process $\bar{Y}_1, \bar{Y}_2, \ldots$ has means $E(\bar{Y}_i) = E(Y_i)$ and variances $\text{Var}(\bar{Y}_i) = \text{Var}(Y_i)/n$ (see Prob. 9.12). Thus, the averaged process has the same transient mean curve as the original process, but its plot has only $(1/n)$th the variance.
3. To smooth out the high-frequency oscillations in $\bar{Y}_1, \bar{Y}_2, \ldots$ (but leave the low-frequency oscillations or long-run trend of interest), we further define the *moving average* $\bar{Y}_i(w)$ (where w is the *window* and is a positive integer such that $w \leq \lfloor m/4 \rfloor$) as follows:

$$\bar{Y}_i(w) = \begin{cases} \dfrac{\sum\limits_{s=-w}^{w} \bar{Y}_{i+s}}{2w+1} & \text{if } i = w+1, \ldots, m-w \\[2ex] \dfrac{\sum\limits_{s=-(i-1)}^{i-1} \bar{Y}_{i+s}}{2i-1} & \text{if } i = 1, \ldots, w \end{cases}$$

Thus, if i is not too close to the beginning of the replications, then $\bar{Y}_i(w)$ is just the simple average of $2w + 1$ observations of the averaged process centered at observation i (see Fig. 9.8). It is called a moving average since i moves through time.

4. Plot $\overline{Y}_i(w)$ for $i = 1, 2, \ldots, m - w$ and choose l to be that value of i beyond which $\overline{Y}_1(w), \overline{Y}_2(w), \ldots$ appears to have converged. See Welch (1983, p. 292) for an aid in determining convergence.

The following example illustrates the calculation of the moving average.

EXAMPLE 9.24. For simplicity, assume that $m = 10$, $w = 2$, $\overline{Y}_i = i$ for $i = 1, 2, \ldots, 5$, and $\overline{Y}_i = 6$ for $i = 6, 7, \ldots, 10$. Then

$$\overline{Y}_1(2) = 1 \qquad \overline{Y}_2(2) = 2 \qquad \overline{Y}_3(2) = 3$$
$$\overline{Y}_4(2) = 4 \qquad \overline{Y}_5(2) = 4.8 \qquad \overline{Y}_6(2) = 5.4$$
$$\overline{Y}_7(2) = 5.8 \qquad \overline{Y}_8(2) = 6$$

Before giving examples of applying Welch's procedure to actual stochastic models, we make the following recommendations on choosing the parameters n, m, and w:

- Initially, make $n = 5$ or 10 replications (depending on model execution time), with m as large as practical. In particular, m should be much larger than the anticipated value of l (see Sec. 9.5.2) and also large enough to allow infrequent events (e.g., machine breakdowns) to occur a reasonable number of times.
- Plot $\overline{Y}_i(w)$ for several values of the window w and choose the smallest value of w (if any) for which the corresponding plot is "reasonably smooth." Use this plot to determine the length of the warmup period l. [Choosing w is like choosing the interval width Δb for a histogram (see Sec. 6.4.2). If w is too small, the plot of $\overline{Y}_i(w)$ will be "ragged." If w is too large, then the $\overline{Y}_i$ observations will be overaggregated and we will not have a good idea of the shape of the transient mean curve, $E(Y_i)$ for $i = 1, 2, \ldots$.]
- If no value of w in step 3 is satisfactory, make 5 or 10 additional replications of length m. Repeat step 2 using all available replications. [For a fixed value of w, the plot of $\overline{Y}_i(w)$ will get "smoother" as the number of replications increases. Why?]

The major difficulty in applying Welch's procedure in practice is that the required number of replications, n, may be relatively large if the process $Y_1, Y_2, \ldots$ is highly variable [see Alexopoulos and Seila (1998, p. 240)]. Also, the choice of l is somewhat subjective.

EXAMPLE 9.25. A small factory consists of a machining center and inspection station in series, as shown in Fig. 9.9. Unfinished parts arrive to the factory with exponential interarrival times having a mean of 1 minute. Processing times at the machine are uniform on the interval [0.65, 0.70] minute, and subsequent inspection times at the inspection

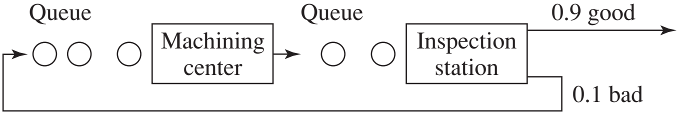


FIGURE 9.9
Small factory consisting of a machining center and an inspection station.

station are uniform on the interval [0.75, 0.80] minute. Ninety percent of inspected parts are "good" and are sent to shipping; 10 percent of the parts are "bad" and are sent back to the machine for rework. (Both queues are assumed to have infinite capacity.) The machining center is subject to randomly occurring breakdowns. In particular, a new (or freshly repaired) machine will break down after an exponential amount of *calendar* time with a mean of 6 hours (see Sec. 14.4.2). Repair times are uniform on the interval [8, 12] minutes. If a part is being processed when the machine breaks down, then the machine continues where it left off upon the completion of repair. Assume that the factory is initially empty and idle, and is open 8 hours per day.

Consider the stochastic process $N_1, N_2, \ldots,$ where N_i is the number of parts produced in the ith hour. Suppose that we want to determine the warmup period l so that we can eventually estimate the steady-state mean hourly throughput $\nu = E(N)$ (see Example 9.30). We made $n = 10$ independent replications of the simulation each of length $m = 160$ hours (or 20 days). In Fig. 9.10 we plot the averaged process $\overline{N}_i$ for $i = 1, 2, \ldots, 160$. It is clear that further smoothing of the plot is necessary, and that one replication, in general, is not sufficient to estimate l. In Figs. 9.11*a* and 9.11*b* we plot the moving average $\overline{N}_i(w)$ for both $w = 20$ and $w = 30$. From the plot for $w = 30$ (which is smoother), we chose a warmup period of $l = 24$ hours. Note that it is better to choose l too large rather than too small, since our goal is to have $E(Y_i)$ close to ν for $i > l$. (We choose to tolerate slightly higher variance in order to be more certain that our point estimator for ν will have a small bias.)

EXAMPLE 9.26. Consider a simple model of a Signaling System Number 7 (SS7) network that is used for setting up and tearing down of telephone calls, and for processing of "800" calls. (The actual calls are transmitted on an associated circuit-switched

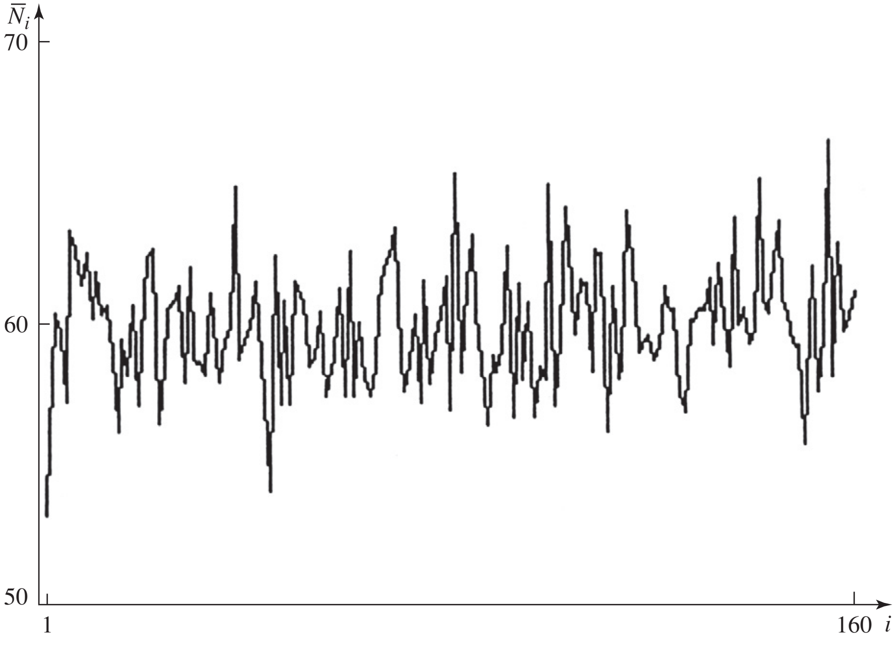


FIGURE 9.10
Averaged process for hourly throughputs, small factory.

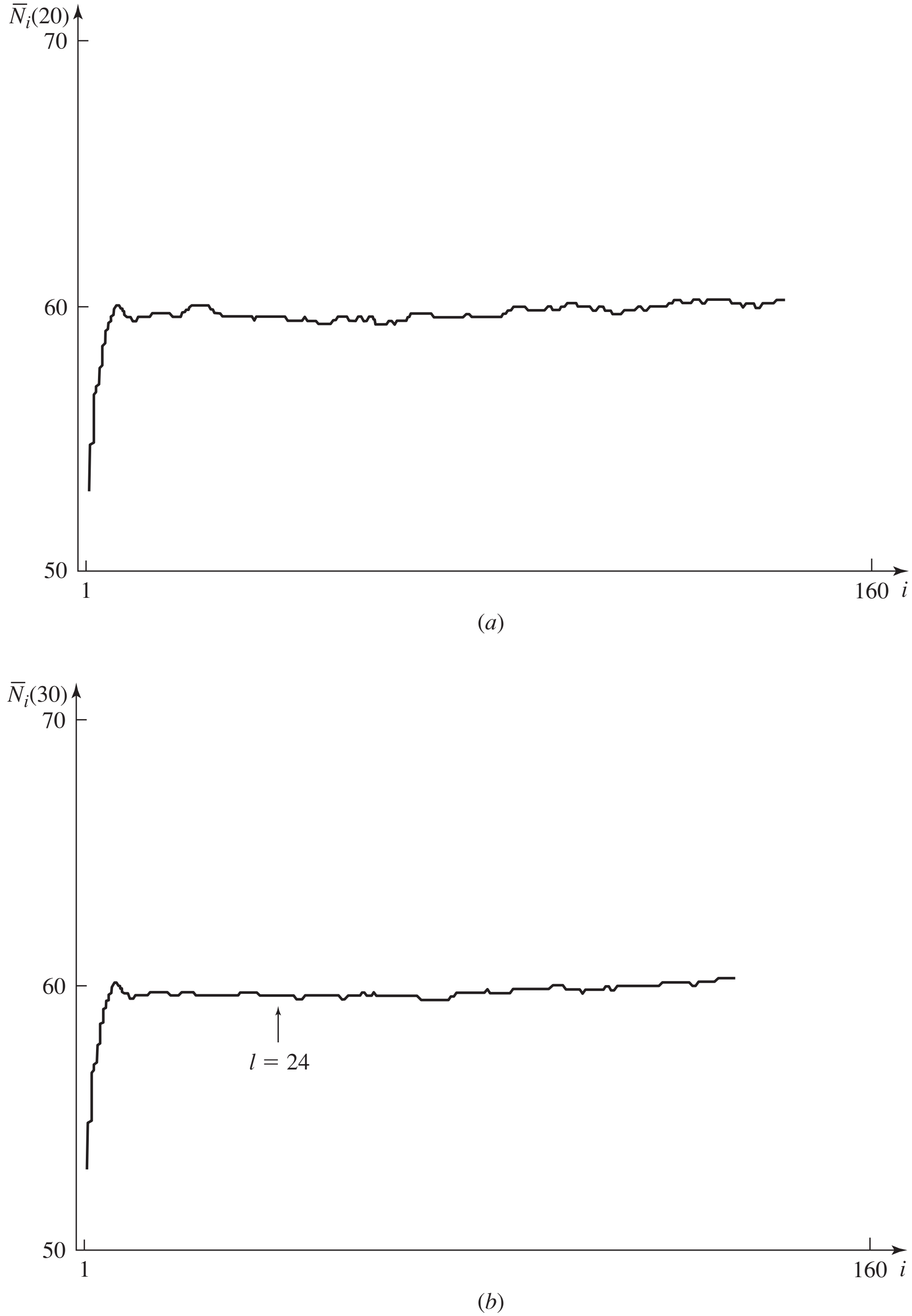
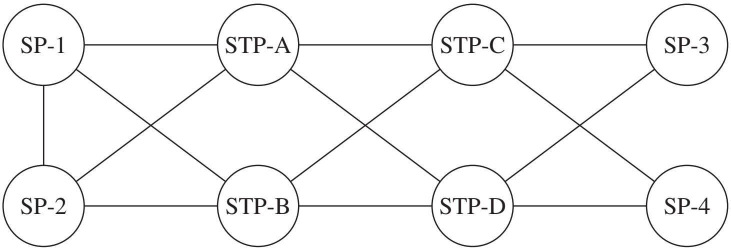


FIGURE 9.11
Moving averages for hourly throughputs, small factory: (a) $w = 20$; (b) $w = 30$.



Note: Each line segment represents two links.

FIGURE 9.12
Topology of the SS7 network.

network.) The network consists of four Signaling Points (denoted SP-1, . . . , SP-4), two Signal Transfer Point pairs (STP-A/STP-B and STP-C/STP-D), and pairs (see Prob. 9.33) of 56 kilobits per second, full-duplex (bidirectional) links as shown in Fig. 9.12. (A line segment in Fig. 9.12 corresponds to two links.) The links from node SP-1 to node STP-A are denoted by 1-A, the links from STP-A to STP-C are denoted by A-C, etc. There is a system requirement that the utilization of each STP and link cannot exceed 0.4. (This requirement is necessary in the actual network to allow extra capacity in case a resource breaks down; however, we do *not* model breakdowns here.)

Each SP sends messages (signals) to each of the other SPs in accordance with a Poisson process (i.e., exponential interarrival times) with rates given in Table 9.6. The length of a message is a discrete uniform random variable in the range 23 to 29 bytes. Each message also includes a 7-byte routing label (containing the source and destination nodes) when it is sent over a link.

Each STP (SP) contains three (two) parallel processors (see Prob. 9.34) that are fed by a single input queue, and there is an output queue for each link that emanates from the node. A message must be processed on one of the processors in a node, and processing times are a constant 2.5 milliseconds.

The initial links used to send a message from one node to another are given in Table 9.7. When two links are available, each one is chosen with a probability of 0.5.

Consider the stochastic process $E_1, E_2, \ldots$, where E_i is the end-to-end delay (i.e., the time to go from a source SP to a destination SP) of the ith completed message. Suppose that we want to determine the warmup period l so that we can eventually estimate the steady-state mean $\nu = E(E)$ (see Example 9.31). [The symbol $E(E)$ is the expected value of the steady-state random variable E.] We made $n = 5$ independent replications of the simulation, each of length $m = 10$ seconds. In Fig. 9.13 we plot the end-to-end delay moving average $\overline{E}_i(w)$ for $w = 600$. [Note that the number of E_i observations in a 10-second simulation run is a random variable with approximate mean 12,400, since the

TABLE 9.6
Traffic rates (in messages per minute) from one SP to another SP

Node	SP-1	SP-2	SP-3	SP-4
SP-1		9600	7200	4800
SP-2	8000		4800	7200
SP-3	6400	4800		6400
SP-4	4800	5600	4800	

TABLE 9.7
Initial links used (see Fig. 9.12) in going from one node (row) to another node (column)

Node	SP-1	SP-2	SP-3	SP-4
SP-1		1-2	1-A, 1-B	1-A, 1-B
SP-2	1-2		2-A, 2-B	2-A, 2-B
SP-3	3-C, 3-D	3-C, 3-D		3-C, 3-D
SP-4	4-C, 4-D	4-C, 4-D	4-C, 4-D	
Node	**SP-1**	**SP-2**	**SP-3**	**SP-4**
STP-A	1-A	2-A	A-C, A-D	A-C, A-D
STP-B	1-B	2-B	B-C, B-D	B-C, B-D
STP-C	A-C, B-C	A-C, B-C	3-C	4-C
STP-D	A-D, B-D	A-D, B-D	3-D	4-D

overall arrival rate is 1240 messages per second and the system is stable (see Prob. 9.35). Therefore, for our analysis we used the minimum number of observations for any one of the 5 runs, which was 12,306. However, $\overline{E}_i(w)$ is only plotted for $i = 1, 2, \ldots, 9920$ (a multiple of 1240) in Fig. 9.13.] From the plot, we conservatively chose a warmup period of $l = 6200$ (5×1240) end-to-end delays. However, for the construction of a confidence interval in Example 9.31, we will actually use a warmup period of 5 seconds, since the run length m is in units of seconds.

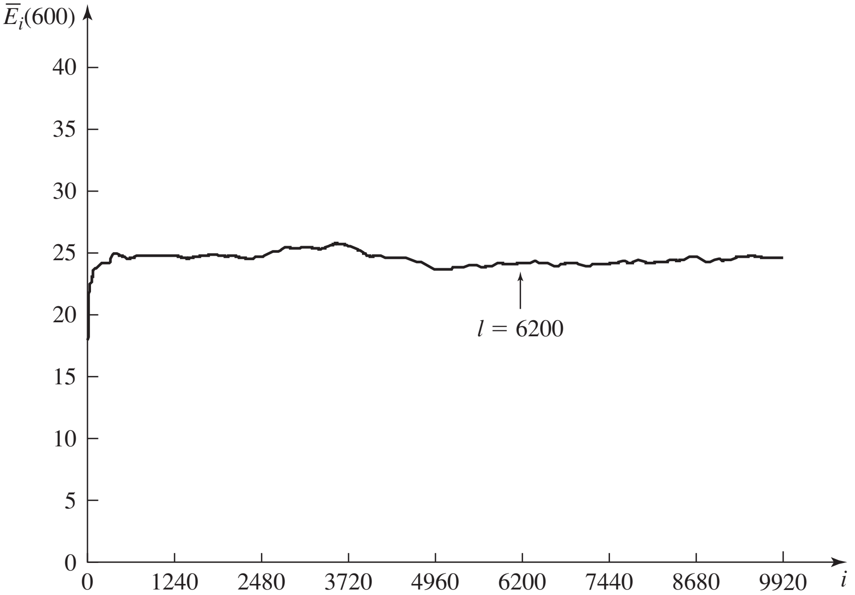


FIGURE 9.13
Moving average with $w = 600$ for end-to-end delays, SS7 network.

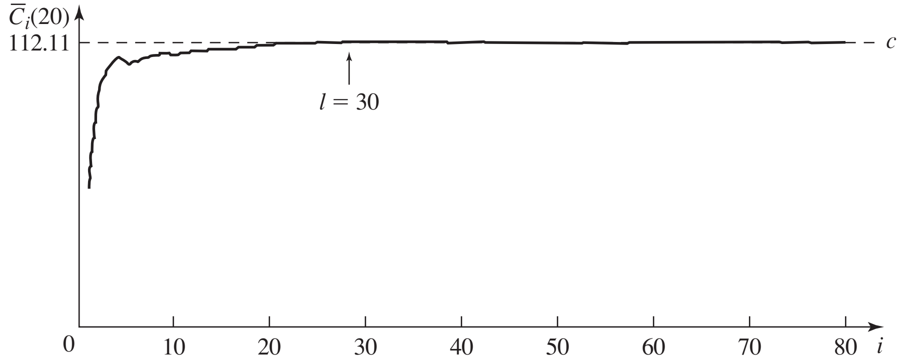


FIGURE 9.14
Moving average with $w = 20$ for monthly costs, inventory system.

EXAMPLE 9.27. Consider the process $C_1, C_2, \ldots$ for the inventory system of Example 9.3. Suppose that we want to determine the warmup period l in order to estimate the steady-state mean cost per month $c = E(C) = 112.11$. We made $n = 10$ independent replications of the simulation of length $m = 100$ months. In Fig. 9.14 we plot the moving average $\overline{C}_i(w)$ for $w = 20$, from which we chose a warmup period of $l = 30$ months.

Additional applications of Welch's procedure are given in Chaps. 10 through 12 and 14. Note also that a version of Welch's procedure is available in the manufacturing-oriented simulation package AutoMod [see Banks (2004)].

White (1997) introduced a procedure for determining a warmup period or deletion amount called MSER (*M*arginal *S*tandard *E*rror *R*ules), which is based on minimizing mean-squared error. Let $\overline{Y}(m, l)$ be as defined above and let

$$S^2(m, l) = \frac{\sum_{i=l+1}^{m} [Y_i - \overline{Y}(m, l)]^2}{m - l}$$

be the sample variance of $Y_{l+1}, Y_{l+2}, \ldots, Y_m$ if we divide by $m - l$ rather than the usual $m - l - 1$. Define the MSER(m, l) statistic as

$$\text{MSER}(m, l) = \frac{S^2(m, l)}{m - l}$$

If the Y_i's *were* IID and $S^2(m, l)$ had an $m - l - 1$ in the denominator, then the MSER(m, l) statistic would be an unbiased estimator of the variance of $\overline{Y}(m, l)$ [the square of the *standard error* of $\overline{Y}(m, l)$]. Then the optimal deletion amount, l^*, is that value of l that minimizes MSER(m, l) over the values $l = 0, 1, \ldots, m - 1$, which is often written as

$$l^* = \underset{l=0,1,\ldots,m-1}{\arg\min} \ \text{MSER}(m, l) \tag{9.5}$$

where "arg" is an abbreviation for argument. Pasupathy and Schmeiser (2010) show that MSER(m, l) is asymptotically (as m goes to infinity) proportional to the mean-squared error MSE[$\overline{Y}(m, l)$] for every l. Therefore, for large m the value of l that minimizes MSER(m, l) will tend to lie close to the value of l that minimizes MSE[$\overline{Y}(m, l)$]. Although MSER is explicitly designed to minimize mean-squared error, it is also stated by its proponents to be a procedure for reducing the bias in $\overline{Y}(m, l)$ [see Hoad et al. (2009, p. 9) and Franklin and White (2008, p. 545)].

White et al. (2000) discuss a variant of MSER called MSER-k. Let

$$Z_j = \frac{\sum_{i=1}^{k} Y_{k(j-1)+i}}{k} \qquad \text{for } j = 1, 2, \ldots, \lfloor m/k \rfloor$$

Then MSER-k applies the rule given by Eq. (9.5) to the batch averages Z_j's rather than to the Y_i's, where MSER-1 is, of course, the same as MSER. In practice, MSER-5 is typically used rather than MSER, since the former operates on "smoother" data. Also, if $l^* > \lfloor m/k \rfloor/2$ (half the number of batches), then l^* is rejected as a valid warmup period. In this case, m should be increased and MSER-5 applied to the new set of batch averages, etc. Note that the choices $k = 5$ and $\lfloor m/k \rfloor/2$ are somewhat arbitrary, but have been found to work fairly well in practice. Finally, Hoad et al. (2009) recommend that MSER-5 be applied to data averaged over five replications rather than to data from one replication (see Example 9.28). Additional papers that discuss MSER-5 are Mokashi et al. (2010) and Sanchez and White (2011).

EXAMPLE 9.28. Consider the delay-in-queue process $D_1, D_2, \ldots$ for the $M/M/1$ queue with $\rho = 0.9$ ($\lambda = 1$, $\omega = 10/9$) and initial conditions $s = 0$. We made $n = 5$ independent replications of length $m = 65{,}000$ observations and applied MSER-5 to the averaged process $\overline{D}_1, \overline{D}_2, \ldots, \overline{D}_{65{,}000}$ (see Sec. 9.5.1 for the definition of $\overline{D}_i$). MSER-5 chose a warmup period of $l^* = 3$ batches, resulting in a total deletion amount of 15 observations. (All calculations were performed using an Excel macro graciously provided by Professor Katy Hoad of the University of Warwick; it allows m to have a maximum value of 65,536.) The fact that l^* is close to 0 is not surprising given the result of Blomqvist discussed earlier in this section. A plot of the MSER-5 statistic as a function of the batch number, l, is given Fig. 9.15, where it can be seen that $l = 3$ does, in fact, minimize the MSER-5 statistic. It can also be seen from Fig. 9.2 that MSER-5 failed to delete a large amount of biased data.

In order to see how sensitive the optimal deletion amount l^* might be to the value of m, we repeated the above analysis using subsets of the existing 65,000 observations of size 1000, 5000, 10,000, 20,000, and 40,000. For each of the five subset sizes, l^* was equal to 3 or 4.

To see the effect of the batch size k on the selected warmup period, we applied MSER-1 (i.e., $k = 1$) to the $n = 5$ replications of length $m = 65{,}000$ and obtained $l^* = 16$, resulting in a total deletion amount of 16 observations (compared to a deletion amount of 15 above).

EXAMPLE 9.29 Consider again the SS7 network of Example 9.26, where we had $n = 5$ replications of length $m = 12{,}306$ observations. We applied MSER-5 to the averaged process $\overline{E}_1, \overline{E}_2, \ldots, \overline{E}_{12{,}306}$ and obtained a warmup period of $l^* = 2$ batches, resulting in a total deletion amount of 10 observations. (Recall from Example 9.26 that Welch's procedure called for a warmup period of 6200 observations.) A plot of the MSER-5

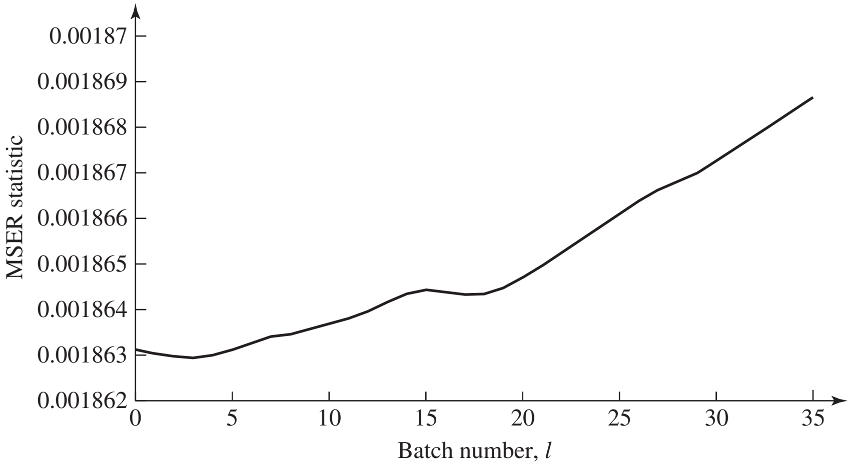


FIGURE 9.15
Plot of MSER-5 statistic as a function of the batch number, l, $M/M/1$ queue with $\rho = 0.9$.

statistic as a function of the batch number, l, is given in Fig. 9.16, where it can be seen that $l = 2$ minimizes the MSER-5 statistic. It appears from Fig. 9.13 that MSER-5 failed to delete a large amount of biased data.

In order to see how sensitive the optimal deletion amount l^* might be to the value of m, we repeated the above analysis using subsets of the existing 12,306 observations of size 1000, 2000, 4000, and 8000. For each of the four subset sizes, l^* was equal to 2.

Based on the two examples presented above (and several others not shown), it appears that MSER may fail to delete a significant amount of highly biased data for some simulation models.

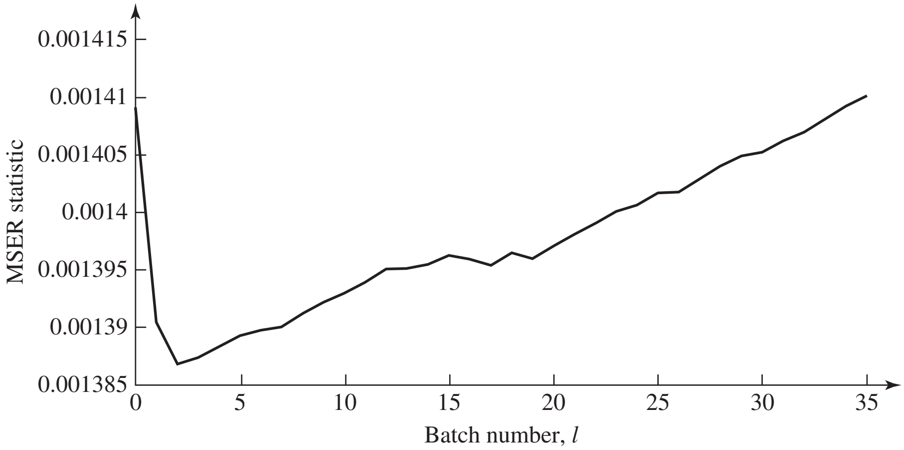


FIGURE 9.16
Plot of MSER-5 statistic as a function of the batch number, l, SS7 network.

Schruben (1982) developed a very general procedure based on standardized time series (see Sec. 9.5.3) for determining whether the observations $Y_{s+1}, Y_{s+2}, \ldots, Y_{s+t}$ (where s need not be zero) contain initialization bias with respect to the steady-state mean $\nu = E(Y)$, that is, whether $E(Y_i) \neq \nu$ for at least one i (where $s + 1 \leq i \leq s + t$). As the procedure is constituted, it is not an algorithm for determining a deletion amount l, but rather a test to determine whether a set of observations contains initialization bias. For example, it could be applied to the truncated averaged process $\overline{Y}_{l+1}, \overline{Y}_{l+2}, \ldots, \overline{Y}_m$ resulting from applying Welch's procedure, in order to determine if there is significant remaining bias. Schruben tested his procedure on several stochastic models with a known value of ν, and found that it had high power in detecting initialization bias [see also Glynn (1995)]. Variations of this initialization-bias test are given in Schruben, Singh, and Tierney (1983) and in Goldsman, Schruben, and Swain (1994). Additionally, Vassilacopoulos (1989) proposed a rank test for accessing the presence of initialization bias. Limited testing on the *M*/*M*/*s* queue produced encouraging results. Finally, comprehensive lists of references on the problem of the initial transient can be found in Hoad et al. (2009) and Pasupathy and Schmeiser (2010).

In Example 9.23 we saw that initializing the *M*/*M*/1 queue with the steady-state number in system distribution resulted in the process $D_1, D_2, \ldots$ not having an initial transient. This suggests trying to estimate the steady-state distribution from a "pilot" run, and then independently sampling from this estimated distribution in order to determine the initial conditions for each production run. Kelton (1989) applied this idea to several queueing systems and also a computer model, where in each case the state of the system is an integer-valued random variable. He found that random initialization reduced the severity and duration of the initial transient period as compared with starting the simulation in a fixed state (e.g., no one present in a queueing system). This technique would be harder to apply, however, in the case of many real-world simulations, where the state of the system has a multivariate distribution [see Murray (1988) and Law (1983, p. 1016) for further discussion]. Glynn (1988) discusses a related method where a one-time pass through the "transient period" is used to specify the starting conditions for subsequent replications.

9.5.2 Replication/Deletion Approach for Means

Suppose that we want to estimate the steady-state mean $\nu = E(Y)$ of the process $Y_1, Y_2, \ldots$. There are six fundamental approaches for addressing this problem, which are discussed in this and the next section. We will for the most part, however, concentrate on one of these, the replication/deletion approach, for the following reasons:

1. If properly applied, this approach should give reasonably good statistical performance.
2. It is the easiest approach to understand and implement. (This is very important in practice due to the time constraints of many simulation projects and because many analysts do not have the statistical background necessary to use some of the more complicated analysis approaches.)

3. This approach applies to all types of output parameters (i.e., Secs. 9.4 through 9.6).
4. It can easily be used to estimate several different parameters for the same simulation model (see Sec. 9.7).
5. This approach can be used to compare different system configurations, as discussed in Chap. 10.
6. Multiple replications can be made simultaneously on different cores within a single computer or on different computers on a network, provided that the software being used for simulation supports this.

We now present the *replication/deletion approach* for obtaining a point estimate and confidence interval for ν. The analysis is similar to that for terminating simulations except that now only those observations beyond the warmup period l in each replication are used to form the estimates. Specifically, suppose that we make n' replications of the simulation each of length m' observations, where m' is much larger than the warmup period l determined by Welch's graphical method (see Sec. 9.5.1). Let Y_{ji} be as defined before and let X_j be given by

$$X_j = \frac{\sum_{i=l+1}^{m'} Y_{ji}}{m' - l} \qquad \text{for } j = 1, 2, \ldots, n'$$

(Note that X_j uses only those observations from the jth replication corresponding to "steady state," namely, $Y_{j,l+1}, Y_{j,l+2}, \ldots, Y_{j,m'}$.) Then the X_j's are IID random variables with $E(X_j) \approx \nu$ (see Prob. 9.15), $\bar{X}(n')$ is an approximately unbiased point estimator for ν, and an approximate $100(1 - \alpha)$ percent confidence interval for ν is given by

$$\bar{X}(n') \pm t_{n'-1,1-\alpha/2}\sqrt{\frac{S^2(n')}{n'}} \tag{9.6}$$

where $\bar{X}(n')$ and $S^2(n')$ are computed from Eqs. (4.3) and (4.4), respectively.

One legitimate objection that might be levied against the replication/deletion approach is that it uses one set of n replications (the pilot runs) to determine the warmup period l, and then uses *only* the last $m' - l$ observations from a different set of n' replications (production runs) to perform the actual analyses. However, this is usually not a problem due to the relatively low cost of computer time.

In some situations, it should be possible to use the initial n pilot runs of length m observations both to determine l and to construct a confidence interval. In particular, if m is substantially larger than the selected value of the warmup period l, then it is probably safe to use the "initial" runs for both purposes. Since Welch's graphical method is only approximate, a "small" number of observations beyond the warmup period l might contain significant bias relative to ν. However, if m is much larger than l, these biased observations will have little effect on the overall quality (i.e., lack of bias) of X_j (based on $m - l$ observations) or $\bar{X}(n)$. Strictly speaking, however, it is more correct statistically to base the replication/deletion approach on two independent sets of replications (see Prob. 9.16 and Example 9.31).

EXAMPLE 9.30. For the manufacturing system of Example 9.25, suppose that we would like to obtain a point estimate and 90 percent confidence interval for the steady-state mean hourly throughput $\nu = E(N)$. From the $n = 10$ replications of length $m = 160$ hours used there, we specified a warmup period of $l = 24$ hours. Since $m = 160$ is much larger than $l = 24$, we will use these same replications to construct a confidence interval. Let

$$X_j = \frac{\sum_{i=25}^{160} N_{ji}}{136} \qquad \text{for } j = 1, 2, \ldots, 10$$

Then a point estimate and 90 percent confidence interval for ν are given by

$$\hat{\nu} = \overline{X}(10) = 59.97$$

and $$\overline{X}(10) \pm t_{9,0.95}\sqrt{\frac{0.62}{10}} = 59.97 \pm 0.46$$

Thus, in the long run we would expect the small factory to produce an average of about 60 parts per hour. Does this throughput seem reasonable? (See Prob. 9.17.)

EXAMPLE 9.31. For the SS7 network of Example 9.26, suppose that we would like to obtain a point estimate and 95 percent confidence interval for the steady-state mean end-to-end delay $\nu = E(E)$. For this example, we made $n' = 5$ *new* independent replications of the simulation of length $m' = 65$ seconds and used the previously determined warmup period of $l = 5$ seconds. Let X_j be the average end-to-end delay of all messages that are completed in the interval [5, 65] seconds for replication j. Then a point estimate and 95 percent confidence interval for ν (in milliseconds) are given by

$$\hat{\nu} = \overline{X}(5) = 24.11$$

and

$$\overline{X}(5) \pm t_{4,0.975}\sqrt{\frac{0.0114}{5}} = 24.11 \pm 0.13$$

From these five replications, we also found that the utilization of each STP and link was less than 0.4, as expected. In particular, the utilization of STP-A was 0.316, and the utilization of link 1-2 was 0.377. Do these values seem reasonable? (See Prob. 9.36.)

The half-length of the replication/deletion confidence interval given by (9.6) depends on the variance of X_j, $\text{Var}(X_j)$, which will be unknown when the first n replications are made. Therefore, if we make a fixed number of replications of the simulation, the resulting confidence-interval half-length may or may not be small enough for a particular purpose. We know, however, that the half-length can be decreased by a factor of approximately 2 by making 4 times as many replications. See also the discussion of "Obtaining a Specified Precision" in Sec. 9.4.1.

A criticism that is sometimes made about the replication/deletion approach is that a $100(1 - \alpha)$ percent confidence interval is actually being constructed for $E(X_j)$ rather than for ν [i.e., $\overline{X}(n')$ is a biased estimator of ν]. As a result, if we make a large number of replications n' in an effort to make the confidence-interval half-length small, then the coverage of the confidence interval might be much less than the desired $1 - \alpha$. However, since a simulation model is only an approximation to the corresponding real-world system, we feel that for many, if not most, models it is sufficient to estimate $E(X_j)$, provided that it is "close" to ν. This should be the case if we choose the run length m' sufficiently large and use Welch's procedure to choose a *conservative* warmup period l.

9.5.3 Other Approaches for Means

In this section we present a more comprehensive discussion of procedures for constructing a point estimate and a confidence interval for the steady-state mean $\nu = E(Y)$ of a simulation output process $Y_1, Y_2, \ldots$. The following definitions of ν are usually equivalent:

$$\nu = \lim_{i\to\infty} E(Y_i)$$

and

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

General references on this subject include Alexopoulos, Goldsman, and Serfozo (2006), Banks et al. (2010), Bratley, Fox, and Schrage (1987), Fishman (1978, 2001), Law (1983), and Welch (1983).

Two general strategies have been suggested in the simulation literature for constructing a point estimate and confidence interval for ν:

1. *Fixed-sample-size procedures.* A single simulation run of an *arbitrary* fixed length is made, and then one of a number of available procedures is used to construct a confidence interval from the available data.
2. *Sequential procedures.* The length of a single simulation run is sequentially increased until an "acceptable" confidence interval can be constructed. There are several techniques for deciding when to stop the simulation run.

Fixed-Sample-Size Procedures

There have been six fixed-sample-size procedures suggested in the literature [see Law (1983) and Law and Kelton (1984) for surveys]. The replication/deletion approach, which was discussed in Sec. 9.5.2, is based on n independent "short" replications of length m observations. It tends to suffer from bias in the point estimator $\hat{\nu}$ (see Sec. 9.1). The five other approaches are based on one "long" replication, and tend to have a problem with bias in the estimator $\widehat{\text{Var}}(\hat{\nu})$ of the variance of the point estimator $\hat{\nu}$. Properties of the six approaches are given in Table 9.8, and details of the five new approaches are now presented.

The method of *batch means*, like the replication/deletion approach, seeks to obtain independent observations so that the formulas of Chap. 4 can be used to obtain a confidence interval. However, since the batch-means method is based on a

TABLE 9.8
Properties of steady-state estimation procedures

Approach	Number of replications	Most serious bias problem	Potential difficulties
Replication/deletion	n $(n \geq 2)$	$\hat{\nu}$	Choice of warmup period, l
Batch means	1	$\widehat{\text{Var}}(\hat{\nu})$	Choice of batch size, k, to obtain uncorrelated batch means
Autoregressive	1	$\widehat{\text{Var}}(\hat{\nu})$	Quality of autoregressive model
Spectral	1	$\widehat{\text{Var}}(\hat{\nu})$	Choice of number of covariance lags, q
Regenerative	1	$\widehat{\text{Var}}(\hat{\nu})$	Existence of cycles with "small" mean length
Standardized time series	1	$\widehat{\text{Var}}(\hat{\nu})$	Choice of batch size, k

single long run, it has to go through the "transient period" only once. Assume that $Y_1, Y_2, \ldots$ is a covariance-stationary process (see Sec. 4.3) with $E(Y_i) = \nu$ for all i. (Alternatively, suppose that the first l observations have been deleted and we are dealing with $Y_{l+1}, Y_{l+2}, \ldots$. If ν exists, in general $Y_{l+1}, Y_{l+2}, \ldots$ will be approximately covariance-stationary if l is large enough.) Suppose that 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 k. (Assume that $m = nk$.) Thus, batch 1 consists of observations $Y_1, \ldots, Y_k$, batch 2 consists of observations $Y_{k+1}, \ldots, Y_{2k}$, etc. Let $\overline{Y}_j(k)$ (where $j = 1, 2, \ldots, n$) be the sample (or batch) mean of the k observations in the jth batch, and let $\overline{\overline{Y}}(n, k) = \sum_{j=1}^{n} \overline{Y}_j(k)/n = \sum_{i=1}^{m} Y_i/m$ be the grand sample mean. We shall use $\overline{\overline{Y}}(n, k)$ as our point estimator for ν. [The $\overline{Y}_j(k)$'s will eventually play the same role for batch means as the X_j's did for the replication/deletion approach in Sec. 9.5.2.]

If the process $Y_1, Y_2, \ldots$ satisfies some additional conditions in addition to being covariance-stationary, then, for a fixed number of batches n, Steiger and Wilson (2001) show that the $\overline{Y}_j(k)$'s are asymptotically (as $k \rightarrow \infty$) distributed as independent normal random variables with mean ν. Therefore, if the batch size k is large enough, it is reasonable to treat the $\overline{Y}_j(k)$'s as if they were IID normal random variables with mean ν. Then a point estimate and approximate $100(1 - \alpha)$ percent confidence interval for ν are obtained by substituting $X_j = \overline{Y}_j(k)$ into (4.3), (4.4), and (4.12).

The major source of error for batch means lies in choosing the batch size k too small, which results in the $\overline{Y}_j(k)$'s possibly being highly correlated and $S^2(n)/n$ being a severely biased estimator of $\text{Var}[\overline{X}(n)] = \text{Var}[\overline{\overline{Y}}(n, k)]$; see Sec. 4.4. In particular, if the Y_i's are positively correlated (as is often the case in practice), the $\overline{Y}_j(k)$'s will be too, giving a variance estimator that is biased low and a confidence interval that is too small. Thus, the confidence interval will cover ν with a probability that is lower than the desired $1 - \alpha$.

There have been several variations of batch means proposed in the literature. Meketon and Schmeiser (1984) introduced the method of *overlapping batch means* (OBM), where $\overline{\overline{Y}}(n, k)$ is once again the point estimator for ν but the expression for $\widehat{\mathrm{Var}}[\overline{\overline{Y}}(n, k)]$ involves all $m - k + 1$ batch means of size k. In particular, batch 1 consists of observations $Y_1, \ldots, Y_k$, batch 2 consists of observations $Y_2, \ldots, Y_{k+1}$, etc. Let $\overline{Y}_j(k)$ (where $j = 1, 2, \ldots, m - k + 1$) be the sample (or batch) mean of the k observations in the jth batch; the $\overline{Y}_j(k)$'s will, in general, be highly correlated. Then the OBM-based estimator of $\mathrm{Var}[\overline{\overline{Y}}(n, k)]$ is given by

$$\widehat{\mathrm{Var}}_O[\overline{\overline{Y}}(n, k)] = \frac{k \sum_{j=1}^{m-k+1} [\overline{Y}_j(k) - \overline{\overline{Y}}(n, k)]^2}{(m - k + 1)(m - k)}$$

and an approximate $100(1 - \alpha)$ percent confidence interval for ν is

$$\overline{\overline{Y}}(n, k) \pm t_{f,1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}_O[\overline{\overline{Y}}(n, k)]}$$

where the degrees of freedom, f, for the t distribution is discussed in Alexopoulos, Goldsman, and Serfozo (2006). Empirical results for the OBM confidence interval can be found in Sargent, Kang, and Goldsman (1992).

Bischak, Kelton, and Pollak (1993) studied the idea of *weighted batch means*, where a weight of w_i is assigned to the ith observation in a batch and the w_i's sum to 1. In the usual batch-means approach, $w_i = 1/k$ for all i. Fox, Goldsman, and Swain (1991) consider the idea of *spaced batch means*, where a spacer of size s is inserted between the batches used for the actual analysis to reduce the correlations among the $\overline{Y}_j(k)$'s.

Tafazzoli et al. (2011) propose N-Skart. It accounts for skewness in the batch means by using the Willink confidence interval given by (4.13) and it makes a correlation adjustment to the half-length of the confidence interval based on the estimated lag-one correlation between the batch means.

Argon and Andradóttir (2006) introduced the method of *replicated batch means*, which is based on making a "small" number, r, of replications of length m, and then breaking each replication into n batches of length k ($m = nk$). The sample mean of the r replication averages is used as a point estimator for ν, and the rn batch means are used to construct a variance estimator. This method includes replication as a special case when $n = 1$, and it includes batch means as a special case when $r = 1$. Other papers that discuss batch means in general are by Alexopoulos and Goldsman (2004); Alexopoulos, Goldsman, and Serfozo (2006); Damerdji (1994); Fishman and Yarberry (1997); Sargent, Kang, and Goldsman (1992); Schmeiser (1982); Schmeiser and Song (1996); and Song and Schmeiser (1995). Sequential procedures based on batch means are discussed at the end of this section.

Rather than attempt to achieve independence, the two methods we discuss next use estimates of the autocorrelation structure of the underlying stochastic process to obtain an estimate of the variance of the sample mean and ultimately to construct a confidence interval for ν. Assume that we have the observations $Y_1, Y_2, \ldots, Y_m$ from a single replication of the simulation and let $\overline{Y}(m) = \sum_{i=1}^{m} Y_i/m$ be our point estimator for ν. The *autoregressive method*, developed by Fishman (1971, 1973a, 1978), assumes that the process $Y_1, Y_2, \ldots$ is covariance-stationary with $E(Y_i) = \nu$ and can be represented by the pth-order autoregressive model

$$\sum_{j=0}^{p} b_j(Y_{i-j} - \nu) = \epsilon_i \tag{9.7}$$

where $b_0 = 1$ and $\{\epsilon_i\}$ is a sequence of uncorrelated random variables with common mean 0 and variance $\sigma_\epsilon{}^2$. For known autoregressive order p and

$$\sum_{j=-\infty}^{\infty} |C_j| < \infty \tag{9.8}$$

it is possible to show that $m\,\text{Var}[\overline{Y}(m)] \rightarrow \sigma_\epsilon{}^2/(\sum_{j=0}^{p} b_j)^2$ as $m \rightarrow \infty$. Based on estimating the covariances C_j from the observations $Y_1, \ldots, Y_m$, Fishman (1973a) gives a procedure for determining the order p and obtaining estimates $\hat{b}_j$ (where $j = 1, 2, \ldots, \hat{p}$) and $\hat{\sigma}_\epsilon{}^2$, where $\hat{p}$ is the estimated order. Let $\hat{b} = 1 + \sum_{j=1}^{\hat{p}} \hat{b}_j$. Then, for large m, an estimate of $\text{Var}[\overline{Y}(m)]$ and an approximate $100(1 - \alpha)$ percent confidence interval for ν are given by

$$\widehat{\text{Var}}[\overline{Y}(m)] = \frac{\hat{\sigma}_\epsilon{}^2}{m(\hat{b})^2}$$

and

$$\overline{Y}(m) \pm t_{\hat{f},1-\alpha/2} \sqrt{\widehat{\text{Var}}[\overline{Y}(m)]}$$

where an expression for the estimated df $\hat{f}$ is given by

$$\hat{f} = \frac{m\hat{b}}{2\sum_{j=0}^{\hat{p}} (\hat{p} - 2j)\hat{b}_j}$$

Yuan and Nelson (1994) give an alternative approach for estimating the autoregressive order p and the df f. Their approach gives better coverage than Fishman's approach for the *M*/*M*/1 queue with $\rho = 0.9$.

A major concern in using these approaches is whether the autoregressive model provides a good representation for an arbitrary stochastic process. Schriber and Andrews (1984) give a generalization of the autoregressive method that allows for moving-average components as well.

The method of *spectrum analysis* also assumes that the process $Y_1, Y_2, \ldots$ is covariance-stationary with $E(Y_i) = \nu$, but does not make any further assumptions such as that given by Eq. (9.7). Under this stationarity assumption, it is possible to show that

$$\text{Var}[\overline{Y}(m)] = \frac{C_0 + 2\sum_{j=1}^{m-1} (1 - j/m)C_j}{m} \tag{9.9}$$

[which is essentially the same as Eq. (4.7)], and the method of spectrum analysis uses this relationship as a starting point for estimating $\text{Var}[\overline{Y}(m)]$. The name of this method is based on the fact that, provided (9.8) holds, we have $m\,\mathrm{Var}[\overline{Y}(m)] \rightarrow 2\pi g(0)$ as $m \rightarrow \infty$, where $g(\tau)$ is called the *spectrum* of the process at frequency τ, and is defined by the Fourier transform $g(\tau) = (2\pi)^{-1}\sum_{j=-\infty}^{\infty} C_j \exp(-i\tau j)$ for $|\tau| \le \pi$ and $i = \sqrt{-1}$. Thus, for large m, $\mathrm{Var}[\overline{Y}(m)] \approx 2\pi g(0)/m$ and the problem of estimating $\mathrm{Var}[\overline{Y}(m)]$ can be viewed as that of estimating the spectrum at zero frequency.

An estimator of $\mathrm{Var}[\overline{Y}(m)]$ that immediately presents itself is obtained by simply replacing C_j in Eq. (9.9) by an estimate $\hat{C}_j$ computed from $Y_1, Y_2, \ldots, Y_m$ and Eq. (4.9). However, for large m and j near m, C_j will generally be nearly zero, but $\hat{C}_j$ will have a large variance since it will be based on only a few observations. As a result, several authors have suggested estimators of the following form:

$$\widehat{\mathrm{Var}}[\overline{Y}(m)] = \frac{\hat{C}_0 + 2\sum_{j=1}^{q-1} W_q(j)\hat{C}_j}{m}$$

where q (which determines the number of $\hat{C}_j$'s in the estimator) must be specified and the weighting function $W_q(j)$ is designed to improve the sampling properties of $\widehat{\mathrm{Var}}[\overline{Y}(m)]$. Then an approximate $100(1 - \alpha)$ percent confidence interval for ν is given by

$$\overline{Y}(m) \pm t_{f,1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\overline{Y}(m)]}$$

where f depends on m, q, and the choice of weighting function [see Fishman (1969, 1973a) and Law and Kelton (1984)]. Welch (1987) discusses the relationships among batch means, overlapping batch means, and spectrum analysis.

This technique is complicated, requiring a fairly sophisticated background on the part of the analyst. Moreover, there is no definitive procedure for choosing the value of q. Additional discussions of spectral methods may be found in Damerdji (1991), Heidelberger and Welch (1981a, 1981b, 1983), Lada and Wilson (2006a), and Lada et al. (2007).

The *regenerative method* is an altogether different approach to simulation and thus leads to different approaches to constructing a confidence interval for ν. The idea is to identify random times at which the process probabilistically "starts over," i.e., regenerates, and to use these regeneration points to obtain independent random variables to which classical statistical analysis can be applied to form point and interval estimates for ν. This method was developed simultaneously by Crane and Iglehart (1974a, 1974b, 1975) and by Fishman (1973b, 1974); we follow the presentation of the former authors.

Assume for the output process $Y_1, Y_2, \ldots$ that there is a sequence of random indices $1 \le B_1 < B_2 < \cdots$, called *regeneration points*, at which the process starts over probabilistically; i.e., the distribution of the process $\{Y_{B_j+i-1}, i = 1, 2, \ldots\}$ is the same for each $j = 1, 2, \ldots$, and the process from each B_j on is assumed to be independent of the process prior to B_j. The portion of the process between two successive B_j's is called a *regeneration cycle*, and it can be shown that successive cycles are IID replicas of each other. In particular, comparable random variables defined over the successive cycles are IID. Let $N_j = B_{j+1} - B_j$ for $j = 1, 2, \ldots$ and assume that $E(N_j) < \infty$. If $Z_j = \sum_{i=B_j}^{B_{j+1}-1} Y_i$, the random vectors $\mathbf{U}_j = (Z_j, N_j)^T$

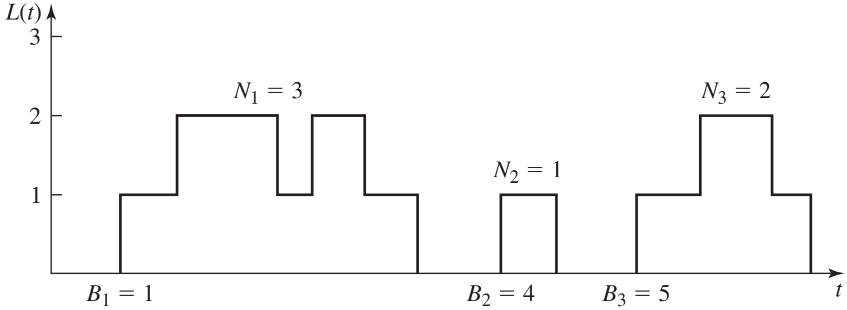


FIGURE 9.17
A realization of the number-in-system process $\{L(t), t \geq 0\}$ for a single-server queue.

(where $\mathbf{A}^T$ is the transpose of the vector $\mathbf{A}$) are IID, and provided that $E(|Z_j|) < \infty$, the steady-state mean ν is given (see Prob. 9.21) by

$$\nu = \frac{E(Z)}{E(N)}$$

EXAMPLE 9.32. Consider the output process of delays $D_1, D_2, \ldots$ for a single-server queue with IID interarrival times, IID service times, customers served in a FIFO manner, and $\rho < 1$. The indices of those customers who arrive to find the system completely empty are regeneration points (see Fig. 9.17). Let N_j be the total number of customers served in the jth cycle and let $Z_j = \sum_{i=B_j}^{B_{j+1}-1} D_i$ be the total delay of all customers served in the jth cycle. Then the steady-state mean delay d is given by $d = E(Z)/E(N)$.

Note that the indices of customers who arrive to find l customers present ($l \geq 1$ and fixed) will not, in general, be regeneration points for the process $D_1, D_2, \ldots$. This is because the distribution of the remaining service time of the customer in service will be different for successive customers who arrive to find l customers present. However, if service times are exponential random variables, these indices *are* regeneration points due to the memoryless property of the exponential distribution (see Probs. 4.26 and 9.22).

We now discuss how to obtain a point estimator and a confidence interval for ν using the regenerative method. Suppose that we simulate the process $Y_1, Y_2, \ldots$ for exactly n' regeneration cycles, resulting in the following data:

$$Z_1, Z_2, \ldots, Z_{n'}$$

$$N_1, N_2, \ldots, N_{n'}$$

Each of these sequences consists of IID random variables. In general, however, Z_j and N_j are not independent. A point estimator for ν is then given by

$$\hat{\nu}(n') = \frac{\overline{Z}(n')}{\overline{N}(n')}$$

Although $\overline{Z}(n')$ and $\overline{N}(n')$ are unbiased estimators of $E(Z)$ and $E(N)$, respectively, $\hat{\nu}(n')$ is *not* an unbiased estimator of ν (see App. 9A). It is true, however, that $\hat{\nu}(n')$ is a *strongly consistent estimator* of ν, that is, $\hat{\nu}(n') \to \nu$ as $n' \to \infty$ (w.p. 1); see Prob. 9.21.

Let the covariance matrix of the vector $\mathbf{U}_j = (Z_j, N_j)^T$ be

$$\sum = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

for example, $\sigma_{12} = E\{[Z_j - E(Z_j)][N_j - E(N_j)]\}$, and let $V_j = Z_j - \nu N_j$. Then the V_j's are IID random variables with mean 0 and variance $\sigma_V^2 = \sigma_{11} - 2\nu\sigma_{12} + \nu^2\sigma_{22}$ (see Prob. 4.13). Therefore, if $0 < \sigma_V^2 < \infty$, it follows from the classical central limit theorem (see Theorem 4.1 in Sec. 4.5) that

$$\frac{\overline{V}(n')}{\sqrt{\sigma_V^2/n'}} \xrightarrow{\mathcal{D}} \text{N}(0, 1) \qquad \text{as } n' \to \infty \tag{9.10}$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution. Let

$$\hat{\sum}(n') = \begin{bmatrix} \hat{\sigma}_{11}(n') & \hat{\sigma}_{12}(n') \\ \hat{\sigma}_{12}(n') & \hat{\sigma}_{22}(n') \end{bmatrix} = \frac{\sum_{j=1}^{n'} [\mathbf{U}_j - \overline{\mathbf{U}}(n')][\mathbf{U}_j - \overline{\mathbf{U}}(n')]^T}{n' - 1}$$

be the estimated covariance matrix and let

$$\hat{\sigma}_V^2(n') = \hat{\sigma}_{11}(n') - 2\hat{\nu}(n')\hat{\sigma}_{12}(n') + [\hat{\nu}(n')]^2\hat{\sigma}_{22}(n')$$

be the estimate of σ_V^2 based on n' regeneration cycles. It can be shown that $\hat{\sigma}_V^2(n') \to \sigma_V^2$ as $n' \to \infty$ (w.p. 1). Consequently, we can replace σ_V^2 in (9.10) by $\hat{\sigma}_V^2(n')$ [see Chung (1974, p. 93)], and dividing through the ratio by $\overline{N}(n')$ yields

$$\frac{\hat{\nu}(n') - \nu}{\sqrt{\hat{\sigma}_V^2(n')/\{n'[\overline{N}(n')]^2\}}} \xrightarrow{\mathcal{D}} \text{N}(0, 1) \qquad \text{as } n' \to \infty$$

Therefore, if the number of cycles n' is sufficiently large, an approximate (in terms of coverage) $100(1 - \alpha)$ percent confidence interval for ν is given by

$$\hat{\nu}(n') \pm \frac{z_{1-\alpha/2}\sqrt{\hat{\sigma}_V^2(n')/n'}}{\overline{N}(n')} \tag{9.11}$$

We call this regenerative approach to constructing a confidence interval for ν the *classical approach* (C). For an alternative regenerative approach to constructing a confidence interval for ν, known as the *jackknife approach* (J), see App. 9A.

The difficulty with using the regenerative method in practice is that real-world simulations may not have regeneration points, or (even if they do) the expected cycle length may be so large that only a very few cycles can be simulated [in which case the confidence interval given by (9.11) will not be valid]. For example, suppose one wants to estimate by simulation the steady-state mean total delay in queue for a network consisting of k queueing systems in series. [A customer departing from queueing system i (where $i = 1, 2, \ldots, k - 1$) proceeds to queueing system $i + 1$.] Then regeneration points for the process $D_1, D_2, \ldots$ (where D_i is the total delay of the ith customer to arrive) are the indices of those customers who arrive at the first queueing system to find the *entire* network empty. If the queueing systems composing the network are highly utilized, as is typical, regeneration points for the network will be few and far between. A more complete discussion of the regenerative method may be found in Crane and Lemoine (1977), Henderson and Glynn (2001), and Shedler (1993).

The *standardized time series method* [see Schruben (1983a)] assumes that the process $Y_1, Y_2, \ldots$ is strictly stationary with $E(Y_i) = \nu$ for all i and is also phi-mixing. *Strictly stationary* means that the joint distribution of $Y_{i_1+j}, Y_{i_2+j}, \ldots, Y_{i_n+j}$ is independent of j for all time indices $i_1, i_2, \ldots, i_n$. (If ν exists, then, in general, $Y_{l+1}, Y_{l+2}, \ldots$ should be approximately strictly stationary if l is large enough.) Roughly speaking, $Y_1, Y_2, \ldots$ is *phi-mixing* if Y_i and Y_{i+j} become essentially independent as j becomes large [see Billingsley (1999) for a precise definition]. Suppose that we make one simulation run of length m and divide $Y_1, Y_2, \ldots, Y_m$ into n batches of size k (where $m = nk$). Let $\overline{Y}_j(k)$ be the sample mean of the k observations in the jth batch. The grand sample mean $\overline{Y}(m)$ is the point estimator for ν. Furthermore, if m is large, then $\overline{Y}(m)$ will be approximately normally distributed with mean ν and variance τ^2/m, where

$$\tau^2 = \lim_{m\to\infty} m \operatorname{Var}[\overline{Y}(m)]$$

and is called the *variance parameter.* [See Alexopoulos et al. (2007a, 2007b), Antonini et al. (2009), Alexopoulos et al. (2010), and Meterelliyoz et al. (2012) for recent papers on estimating τ^2.] Let

$$A = \left(\frac{12}{k^3 - k}\right) \sum_{j=1}^{n} \left\{ \sum_{s=1}^{k} \sum_{i=1}^{s} [\overline{Y}_j(k) - Y_{i+(j-1)k}] \right\}^2$$

For a fixed number of batches n, A will be asymptotically (as $k \to \infty$) distributed as τ^2 times a chi-square random variable with n df and asymptotically independent of $\overline{Y}(m)$. Therefore, for k large, we can treat

$$\frac{[\overline{Y}(m) - \nu]/\sqrt{\tau^2/m}}{\sqrt{(A/\tau^2)/n}} = \frac{\overline{Y}(m) - \nu}{\sqrt{A/(mn)}}$$

as having a t distribution with n df, and an approximate $100(1 - \alpha)$ percent confidence interval for ν is given by

$$\overline{Y}(m) \pm t_{n,1-\alpha/2} \sqrt{A/(mn)}$$

The major source of error for standardized time series is choosing the batch size k too small [see Schruben (1983a) for details]. It should be noted that this approach is based on the same underlying theory as Schruben's test for initialization bias discussed in Sec. 9.5.1. Additional references for standardized time series, including alternative confidence-interval formulations, are Glynn and Iglehart (1990), Goldsman, Meketon, and Schruben (1990), Goldsman and Schruben (1984, 1990), and Sargent, Kang, and Goldsman (1992).

Since the five fixed-sample-size confidence-interval approaches presented in this section depend on assumptions that will not be strictly satisfied in an actual simulation, it is of interest to see how these approaches perform in practice. We first present the results from 400 independent simulation experiments for the *M*/*M*/1 queue with $\rho = 0.8$ ($\lambda = 1$ and $\omega = \frac{5}{4}$), where in each experiment our goal was to construct a 90 percent confidence interval for the steady-state mean delay $d = 3.2$ using all five procedures. Not knowing how to select definitively the total sample size m for batch means (B), the autoregressive method (A), spectrum analysis (SA), and standardized times series (STS), we arbitrarily chose $m = 320$, 640, 1280, and 2560. For the regenerative method (R), it can be shown that $E(N) = 1/(1 - \rho) = 5$ for the $M/M/1$ queue with $\rho = 0.8$ (see Prob. 9.25). We therefore chose the number of regeneration cycles $n' = 64$, 128, 256, and 512 so that, on the average, all procedures used the same number of observations, that is, $m = n'E(N)$. Furthermore, we considered both the classical and jackknifed regenerative confidence intervals. For batch means and standardized time series, we chose the number of batches $n = 5$, 10, and 20. The df f for spectrum analysis was chosen so that $f + 1 = n$, where f is related to the number of covariance estimates q in the variance expression by $q = 1.33m/f$ [see Law and Kelton (1984) for details]. Table 9.9 gives the proportion of the 400 confidence intervals that covered d for each of the 48 cases discussed above. [All results are taken from Law and Kelton (1984), except those for standardized time series, which were graciously provided by Professor David Goldsman of Georgia Tech.] For example, in the case of $m = 320$ and $n = 5$ for batch means (i.e., each confidence interval was based on five batches of size 64), 69 percent of the 400 confidence intervals covered d, falling considerably short of the desired 90 percent. (Note that for fixed m, the estimated coverage for batch means decreases as n increases. This is because as n increases, the batch means become more correlated, resulting in a more biased estimate of the variance of the sample mean.)

We next present the results from 200 independent simulation experiments for the time-shared computer model with 35 terminals [see Law and Kelton (1984)], which was discussed in Sec. 2.5. Our objective was to construct 90 percent confidence intervals for the steady-state mean response time $r = 8.25$ [see Adiri and Avi-Itzhak (1969)]. We chose m and n as above and, since $E(N) \approx 32$ for the computer model, we took $n' = 10$, 20, 40, and 80. Table 9.10 gives the proportion of the 200 confidence intervals that covered r for each of 36 cases (results for standardized time series were not available). Even though the computer model is physically much more complex than the $M/M/1$ queue, it can be seen from Table 9.10 that batch means with $n = 5$ produces an estimated coverage very close to 0.90 for m as small as 640. Thus, the $M/M/1$ queue with $\rho = 0.8$ is much more difficult statistically, despite its very simple structure. These two examples illustrate that one cannot infer anything about the statistical behavior of the output data by looking at how "complex" the model's structure might be.

From the empirical results presented in Tables 9.9 and 9.10 and also those in Law (1977), Law and Kelton (1984), and Sargent, Kang, and Goldsman (1992), we came to the following conclusions with regard to fixed-sample-size procedures:

1. If the total sample size m (or n') is chosen too small, the actual coverages of *all* existing fixed-sample-size procedures (including replication/deletion) may be considerably lower than desired. This is really not surprising, since a steady-state parameter is defined as a limit as the length of the simulation (total number of observations) goes to infinity.

TABLE 9.9

Estimated coverages based on 400 experiments, *M/M/*1 queue with $\rho = 0.8$

	B			STS			SA			A	R	
	n			n			$f + 1$				Method	
$m(n')$	5	10	20	5	10	20	5	10	20		C	J
320 (64)	0.690	0.598	0.490	0.520	0.340	0.208	0.713	0.625	0.538	0.688	0.560	0.670
640 (128)	0.723	0.708	0.588	0.628	0.485	0.318	0.760	0.735	0.645	0.723	0.683	0.728
1280 (256)	0.780	0.740	0.705	0.730	0.645	0.485	0.783	0.770	0.745	0.753	0.705	0.748
2560 (512)	0.798	0.803	0.753	0.798	0.725	0.598	0.833	0.808	0.773	0.755	0.745	0.763

TABLE 9.10

Estimated coverages based on 200 experiments, time-shared computer model

	B			SA			A	R	
	n			$f + 1$				Method	
$m(n')$	5	10	20	5	10	20		C	J
320 (10)	0.860	0.780	0.670	0.880	0.815	0.720	0.680	0.545	0.725
640 (20)	0.890	0.855	0.790	0.870	0.870	0.820	0.805	0.730	0.830
1280 (40)	0.910	0.885	0.880	0.910	0.910	0.905	0.890	0.830	0.865
2560 (80)	0.905	0.875	0.895	0.910	0.885	0.900	0.885	0.870	0.915

2. The "appropriate" choice of m (or n') would appear to be extremely model-dependent and thus impossible to choose arbitrarily. For the method of batch means with $n = 5$, $m = 640$ gave good results for the computer model; however, even for m as large as 2560, we did not obtain good results for the *M*/*M*/1 queue.
3. For m fixed, the methods of batch means, standardized time series, and spectrum analysis will achieve the best coverage for n and f small.

Sequential Procedures

We now discuss procedures that sequentially determine the length of a single simulation run needed to construct an acceptable confidence interval for the steady-state mean ν. The need for such sequential procedures is evident from the fixed-sample-size results reported above. Specifically, no procedure in which the run length is fixed before the simulation begins can generally be relied upon to produce a confidence interval that covers ν with the desired probability $1 - \alpha$, if the fixed run length is too small for the system being simulated.

In addition to the problem of coverage, an analyst might want to determine a run length large enough to obtain an estimate of ν with a specified absolute error β or relative error γ (see Sec. 9.4.1). It will seldom be possible to know in advance even the order of magnitude of the run length needed to meet these goals for a given simulation problem, so some sort of procedure to increase the run length iteratively would seem to be in order.

Law and Kelton (1982) and Law (1983) surveyed the sequential procedures available at those times and found three that appeared to perform well in terms of achieved coverage if the specified absolute or relative error was small enough. In particular, Fishman (1977) developed a procedure based on the regenerative method and an absolute-error stopping rule. Law and Kelton found that it achieved acceptable coverage for 9 out of 10 stochastic models tested if $\beta = 0.075\ \nu$. Fishman's procedure has the disadvantage of being based on the regenerative method, which we feel limits its application to real-world problems. Also, specifying an appropriate value for β in practice may be troublesome, since ν will, of course, be unknown.

Law and Carson (1979) developed a procedure based on batch means and a relative-error stopping rule. For a fixed number of batches $n = 40$, the batch size k is increased until the resulting batch means are approximately uncorrelated and the corresponding confidence interval satisfies the specified relative error. However, at a particular iteration, 400 batch means each based on $k/10$ observations are actually used to determine whether the corresponding 40 batch means, each based on k observations, are uncorrelated. This scheme was necessary because correlation estimators are generally biased and for small n have a large variance. The Law and Carson (L&C) procedure does not test the batch means to see whether they are approximately normally distributed. However, since $n = 40$ and each batch mean is an average of many individual observations, this is probably not a major issue in general (see Table 4.1).

They applied their procedure to 14 stochastic models for which ν can be computed analytically. For each model, they tried to construct a 90 percent confidence

TABLE 9.11

Estimated coverage, 90 percent confidence interval for true coverage, and average sample size when constructing nominal 90 percent confidence intervals with $\gamma = 0.075$, batch-means procedures

Model	L&C	ASAP3	SBatch	Skart
$M/M/1$, $\rho = 0.8$	0.87 ± 0.06 75,648	0.868 ± 0.028 72,060	0.903 ± 0.024 89,434	0.912 ± 0.014 82,508
$M/M/1$ LIFO, $\rho = 0.8$	0.84 ± 0.06 74,624	0.875 ± 0.027 68,325	0.888 ± 0.016 97,172	0.916 ± 0.014 81,441
$M/H_2/1$, $\rho = 0.8$	0.90 ± 0.05 229,632	0.900 ± 0.025 228,482	0.896 ± 0.025 254,400	0.913 ± 0.015 255,363
$M/M/1/M/1$, $\rho = 0.8$	0.87 ± 0.06 49,920	0.913 ± 0.023 58,844	0.931 ± 0.013 55,398	0.909 ± 0.016 58,573
Central-server model 3	0.88 ± 0.05 3740	0.870 ± 0.026 18,447	0.921 ± 0.014 85,842	0.873 ± 0.017 12,231
Experiments	100	400	1000	1000

interval with a relative error of $\gamma = 0.075$, and they carried out 100 independent experiments. (In general, we believe that if someone is going to use a sequential procedure, then they will choose $\gamma \leq 0.1$. Otherwise, they will use a fixed-sample-size procedure.) Note that the L&C procedure does not explicitly address the startup problem (e.g., no warmup period is used).

In Table 9.11 we give for the L&C procedure the proportion, $\hat{p}$, of the 100 confidence intervals that contained ν, a 90 percent confidence interval for the true coverage p, and the average run length (sample size) at termination, respectively, for five of the tested models. The 90 percent confidence interval for the true coverage p was computed from $\hat{p} \pm z_{0.95}\sqrt{\hat{p}(1 - \hat{p})/100}$ (see Sec. 9.4.2). The results in the first four rows of Table 9.11 are for the delay-in-queue process for the $M/M/1$ queue, the $M/M/1$ LIFO queue, the $M/H_2/1$ queue [hyperexponential service times with cv = 2; see Law (1974) for the exact definition], and the $M/M/1/M/1$ queue (two $M/M/1$ queues in series), respectively; for each model, $\rho = 0.8$. The fifth row of Table 9.11 is for the response-time process for a simple model of a computer system, which Law and Carson call central-server model 3. The last row in Table 9.11 gives the number of independent experiments used in the evaluation of each procedure.

Heidelberger and Welch (1983) developed a procedure (denoted H&W) based on spectral methods and a relative-error stopping rule, which uses regression techniques to estimate the spectrum at zero frequency. Their procedure requires the user to specify a maximum sample size, so some of the confidence intervals produced may not satisfy the relative-error requirement. For the response-time process for two models of computer systems, they tried to construct a 90 percent confidence interval for ν, and they carried out 50 independent experiments [see Heidelberger and Welch (1981b)]. Let $\hat{p}$ be the proportion of confidence intervals *satisfying the precision requirement* that cover ν. For $\gamma = 0.05$ they obtained $\hat{p}$ equal to 0.88 and

TABLE 9.12
Estimated coverage, 90 percent confidence interval for true coverage, and average sample size when constructing nominal 90 percent confidence intervals with $\gamma = 0.075$, spectral procedures

Model	H&W	WASSP
$M/M/1$, $\rho = 0.8$	0.790 ± 0.034 77,971	0.885 ± 0.026 117,540
$M/M/1$ LIFO, $\rho = 0.8$	0.795 ± 0.033 80,098	0.902 ± 0.024 152,355
$M/H_2/1$, $\rho = 0.8$	0.780 ± 0.034 233,430	0.910 ± 0.024 330,580
$M/M/1/M/1$, $\rho = 0.8$	0.803 ± 0.033 52,700	0.890 ± 0.026 82,680
Central-server model 3	0.880 ± 0.027 12,562	0.930 ± 0.021 79,188
Experiments	400	400

0.84 for the two models. Additional empirical results for the H&W procedure are given in Table 9.12 above.

Steiger et al. (2005) proposed a modified version of the *A*utomated *S*imulation *A*nalysis *P*rocedure [Steiger and Wilson (2002)], called ASAP3, which is based on the method of batch means. It operates as follows: The batch size is incrementally increased until spaced groups of four adjacent batch means pass a test for multivariate normality (see Sec. 6.10.1), where the spacer preceding each group also consists of four adjacent batch means. Then after skipping the first spacer as the warmup period, ASAP3 fits an AR(1) process to the *nonspaced* batch means. If necessary, the batch size is further increased until the autoregressive parameter ϕ in the AR(1) model does not exceed 0.8. Next, ASAP3 uses a modified t confidence interval based on the AR(1) parameter estimates to account for the remaining correlations in the batch means. Note that unlike the L&C procedure, ASAP3 does *not* try to obtain uncorrelated batch means. For each model they tried to construct a 90 percent confidence interval for ν with a relative error of $\gamma = 0.075$, and they carried out 400 independent experiments, with the results given in Table 9.11. Additional performance results for ASAP3 are given in Tafazzoli et al. (2011).

Lada et al. (2008) introduced SBatch (*S*paced *Batch* means) for constructing a confidence interval for a steady-state mean. SBatch uses a randomness test and a normality test to iteratively determine the size s of a spacer proceeding each batch and the batch size k so that the resulting spaced batch means are approximately IID normal random variables. To check for any residual correlation between the spaced batch means, SBatch tests the lag-one correlation of the spaced batch means to make sure that it does not exceed 0.8. Each time the correlation test is failed, the batch size is increased, additional observations are obtained from the simulation, a new set of spaced batch means is computed, and the correlation test is repeated for the new set of spaced batch means.

Once the correlation test is passed, SBatch constructs a correlation-adjusted confidence interval for ν using the current set of spaced batch means as follows: The center of the confidence interval is the average of *all* observations beyond the first spacer (the warmup period), and the half-length uses the sample variance of the spaced batch means and a correlation adjustment based on the estimated lag-one correlation [see Eq. (4.9)] between the spaced batch means. For each model they tried to construct a 90 percent confidence interval for ν with a relative error of $\gamma = 0.075$, and they carried out 1000 independent experiments, with the results given in Table 9.11. (We would like to thank Dr. Emily Lada, Dr. Ali Tafazzoli, and Professor James Wilson for supplying some of results presented in Table 9.11.) Additional results for SBatch are given in Tafazzoli et al. (2011).

Tafazzoli and Wilson (2011) developed a procedure that they called Skart (*Sk*ewness- and *a*uto*r*egression-adjusted Student's *t* analysis). [See also Tafazzoli et al. (2011).] Skart addresses the startup problem by successively applying a randomness test to spaced batch means with progressively increasing batch sizes and inter-batch spacer sizes. When the test is finally passed with a batch size k and a spacer size dk (d a nonnegative integer), the warmup period is taken to be $l = dk$.

For the data beyond the warmup period l, Skart computes and uses n' *nonspaced* batch means with batch size k. Since the batch means will, in general, be skewed, Skart uses (4.13) [by Willink (2005)] to construct a confidence interval for the steady-state mean ν, with the batch means' playing the role of the X_i's. The "center" of the nonsymmetric confidence interval is the sample mean of the batch means, and the half-length uses the sample variance of the batch means and a correlation adjustment based on the estimated lag-one correlation between the batch means. For each model they tried to construct a 90 percent confidence interval for ν with a relative error of $\gamma = 0.075$, and they carried out 1000 independent experiments, with the results given in Table 9.11.

Lada and Wilson (2006a) developed WASSP (*WA*velet-based *S*equential *S*pectral *P*rocedure) that also uses a relative-error stopping rule. Lada and Wilson (2006b) tested WASSP and H&W on the same five models discussed above. They performed 400 independent experiments and attempted to construct 90 percent confidence intervals for ν with a relative error of $\gamma = 0.075$, with the results given in Table 9.12. Based on the coverage results in Table 9.12 and in Lada et al. (2007), which we believe are more important than average sample sizes, we conclude that WASSP is superior to H&W. Additional results for WASSP can be found in Tafazzoli et al. (2011).

We do, however, believe that the ASAP3, SBatch, and Skart are preferable to WASSP, since the latter procedure generally requires much larger average sample sizes. Also, WASSP is more complicated since it requires 21 steps as compared to, for example, the 13 steps employed by Skart. (If a particular step for one of these procedures requires p substeps, then we consider the overall step as actually consisting of p steps.) Using the results in Table 9.11 and also those in Tables 1 and 6 of Tafazzoli et al. (2011), we believe that Skart provides the best overall results among the ASAP3, SBatch, and Skart procedures using coverage and average sample size as the criteria for comparison.

However, we believe that the L&C procedure is also worthy of some consideration, since it provided satisfactory performance on 14 stochastic models, albeit based on only 100 experiments per model (compared to 1000 experiments for Skart). It is also the simplest of all sequential procedures discussed here, requiring only five steps. This is very important if a procedure has to be programmed from scratch for a particular application. Finally, Chen and Kelton (2009) present an additional sequential procedure based on batch means, but only limited and noncomparable experimental results are given.

If one wants to construct a confidence interval for a steady-state mean ν that is likely to have coverage close to $1 - \alpha$ and to require a "reasonable" sample size, then one might consider the use of the Skart or L&C procedures with a relative error of $\gamma = 0.075$ or smaller. These two procedures have been tested on a large number of stochastic models and generally produced good results in terms of estimated coverage and average sample size. The reader should be aware, however, that these procedures are more complicated to understand and implement than, say, the replication/deletion approach of Sec. 9.5.2. They may also require larger sample sizes and will probably not easily generalize to the common situation of multiple measures of performance [see Sec. 9.7 and Tafazzoli et al. (2011)].

It might be noted that Glynn and Whitt (1992b) give sufficient conditions for a sequential procedure to be asymptotically valid, i.e., produce a coverage of $1 - \alpha$ as the run length goes to infinity.

9.5.4 Estimating Other Measures of Performance

As we saw in Sec. 9.4.2, the mean does not always provide us with an appropriate measure of system performance. We thus consider the estimation of steady-state parameters ϕ other than the mean $\nu = E(Y)$.

Suppose that we would like to estimate the steady-state probability $p = P(Y \in B)$, where B is a set of real numbers. By way of example, for a communications network we might want to determine the steady-state probability that the end-to-end delay of a message is less than or equal to 5 seconds ($B = \{\text{all real numbers} \le 5\}$). Estimating the probability p, as it turns out, is just a special case of estimating the mean ν, as we now see. Let the steady-state random variable Z be defined by

$$Z = \begin{cases} 1 & \text{if } Y \in B \\ 0 & \text{otherwise} \end{cases}$$

Then

$$\begin{aligned} P(Y \in B) &= P(Z = 1) = 1 \cdot P(Z = 1) + 0 \cdot P(Z = 0) \\ &= E(Z) \end{aligned}$$

Thus, estimating p is equivalent to estimating the steady-state mean $E(Z)$, which has been discussed in Secs. 9.5.2 and 9.5.3. In particular, let

$$Z_i = \begin{cases} 1 & \text{if } Y_i \in B \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, 2, \ldots,$ where $Y_1, Y_2, \ldots$ is the original stochastic process of interest. Then, for example, the replication/deletion approach could be applied to the output process $Z_1, Z_2, \ldots$ to obtain a point estimate and confidence interval for $E(Z) = p$. Note that the warmup period for the (binary) process $Z_1, Z_2, \ldots$ may be different from that for the original process $Y_1, Y_2, \ldots.$

Another parameter of the steady-state distribution of considerable interest is the q-quantile, y_q, which was defined in Sec. 6.4.3. That is, y_q is the value of y such that $P(Y \leq y_q) = q$, where Y is the steady-state random variable. For example, in the case of the communications network discussed above, it might be desired to estimate the 0.9-quantile of the steady-state end-to-end delay distribution. Estimating quantiles is both conceptually and computationally (in terms of the number of observations required to obtain a specified precision) a more difficult problem than estimating the steady-state mean. Furthermore, most procedures for estimating quantiles are based on order statistics and require storage and sorting of the observations.

There have been several procedures proposed for estimating quantiles based on batch means (or extensions), spectral, and regenerative methods [see Law (1983) and Heidelberger and Lewis (1984)]. One drawback of these procedures is that they are all based on a fixed sample size, which must be chosen somewhat arbitrarily. If this sample size is chosen too small, the coverage of the resulting confidence interval will be somewhat less than desired.

Raatikainen (1990) proposed a procedure for estimating quantiles based on the P^2 algorithm of Jain and Chlamtac (1985), which does not require storing and sorting the observations. It is a sequential procedure based on a spectral method and a relative-error stopping rule. Raatikainen tested his procedure on several stochastic models of computer systems and appeared to obtain good results in terms of coverage. The procedure is, however, difficult to implement.

Chen and Kelton (2006) proposed two sequential procedures—*zoom in* (ZI) and *quasi-independent* (QI)—for constructing a confidence interval for a quantile. They tested their procedures on the delay-in-queue process for the $M/M/1$ and $M/M/2$ queues, performing 100 independent experiments in each case. By way of example, suppose that the goal is to construct a 95 percent confidence interval for the 0.9-quantile for the $M/M/1$ queue with $\rho = 0.9$. The ZI procedure had an estimated coverage of 1.00 and an average sample size at termination of approximately 12,464,000. On the other hand, the QI procedure had an estimated coverage of 0.95 and an average sample size of approximately 14,793,000.

Alexopoulos et al. (2012) studied the feasibility of developing a sequential procedure to construct point estimators and confidence intervals for quantiles based on nonoverlapping batches. (The batched quantiles become approximately independent.) Their results are encouraging in terms of the sample sizes that will be required.

9.6 STATISTICAL ANALYSIS FOR STEADY-STATE CYCLE PARAMETERS

Suppose that the output process $Y_1, Y_2, \ldots$ does not have a steady-state distribution. Assume, on the other hand, that there is an appropriate cycle definition so that the process $Y_1^C, Y_2^C, \ldots$ has a steady-state distribution F^C, where Y_i^C is the random variable defined on the *i*th cycle (see Sec. 9.3). If $Y^C \sim F^C$, then we are interested in estimating some characteristic of Y^C such as the mean $\nu^C = E(Y^C)$ or the probability $P(Y^C \le y)$. Clearly, estimating a steady-state cycle parameter is just a special case of estimating a steady-state parameter, so all of the techniques of Sec. 9.5 apply, except to the *cycle* random variables Y_i^C's rather than to the original Y_i's. For example, we could use Welch's method to identify a warmup period and then apply the replication/deletion approach to obtain a point estimate and confidence interval for ν^C.

EXAMPLE 9.33. Consider once again the small factory of Example 9.25 but suppose that there is a half-hour lunch break that starts 4 hours into each 8-hour shift. This break stops the inspection process, but unfinished parts continue to arrive and to be processed by the unmanned machine. If N_i is the throughput in the *i*th hour, then the process $N_1, N_2, \ldots$ does not have a steady-state distribution (see Example 9.11). We might, however, expect that it is periodic with a cycle length of 8 hours. To substantiate this, we made $n = 10$ replications of length $m = 160$ hours (20 shifts). From the plot of the averaged process $\overline{N}_i$ (where $i = 1, 2, \ldots, 160$) in Fig. 9.18, we see that the process $N_1, N_2, \ldots$ does indeed appear to have a cycle of length 8 hours.

Let N_i^C be the average production in the *i*th 8-hour cycle and assume that $N_1^C, N_2^C, \ldots$ has a steady-state distribution. Suppose that we want to obtain a point estimate and a 99 percent confidence interval for the steady-state expected average production over a shift, $\nu^C = E(N^C)$, using the replication/deletion approach. Let N_{ji}^C be the average production in the *i*th cycle of our *j*th available replication ($j = 1, 2, \ldots, 10$; $i = 1, 2, \ldots, 20$), and let $\overline{N}_i^C$ for $i = 1, 2, \ldots, 20$ be the corresponding averaged process (that is, $\overline{N}_i^C = \sum_{j=1}^{10} N_{ji}^C/10$), which is plotted in Fig. 9.19. We conclude from this plot that further smoothing is desirable. As a result, we plot the moving average $\overline{N}_i^C(w)$ (from Welch's procedure) for both $w = 3$ and $w = 6$ shifts in Figs. 9.20*a* and 9.20*b*. From the plot for $w = 6$ (which is smoother), we chose a warmup period of $l = 5$ shifts or 40 hours. (Compare this l with that obtained in Example 9.25.)

Let

$$X_j^C = \frac{\sum_{i=6}^{20} N_{ji}^C}{15} \qquad \text{for } j = 1, 2, \ldots, 10$$

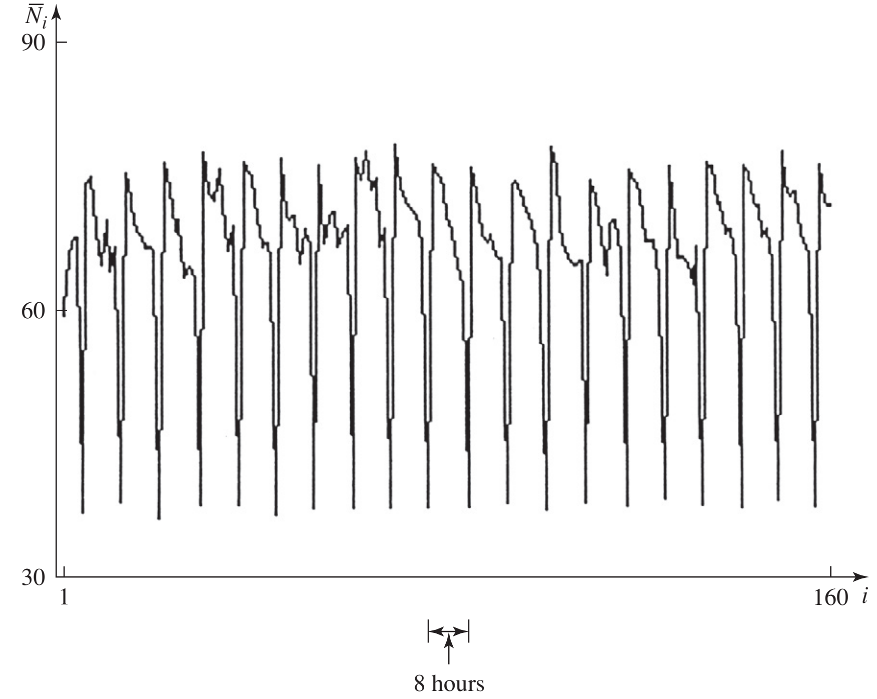


FIGURE 9.18
Averaged process for hourly throughputs, small factory with lunch breaks.

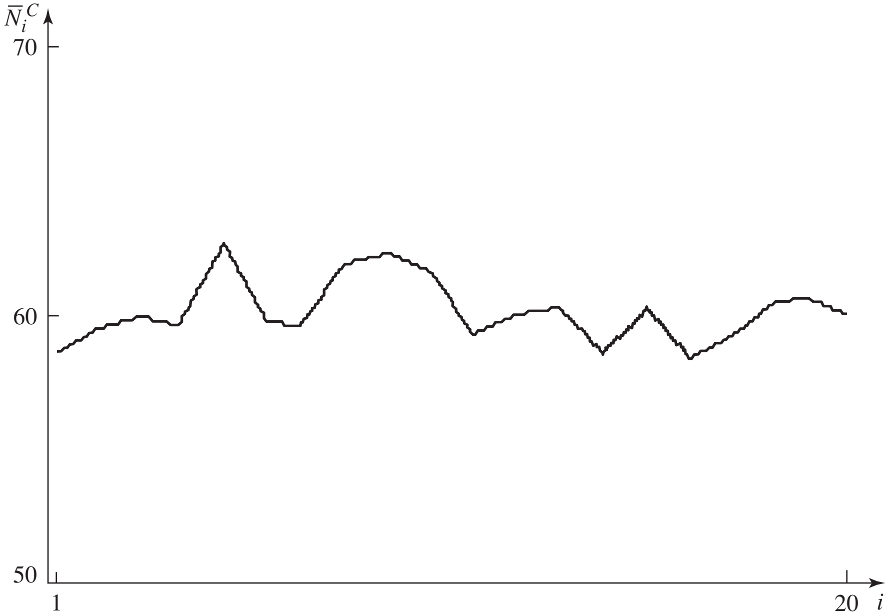
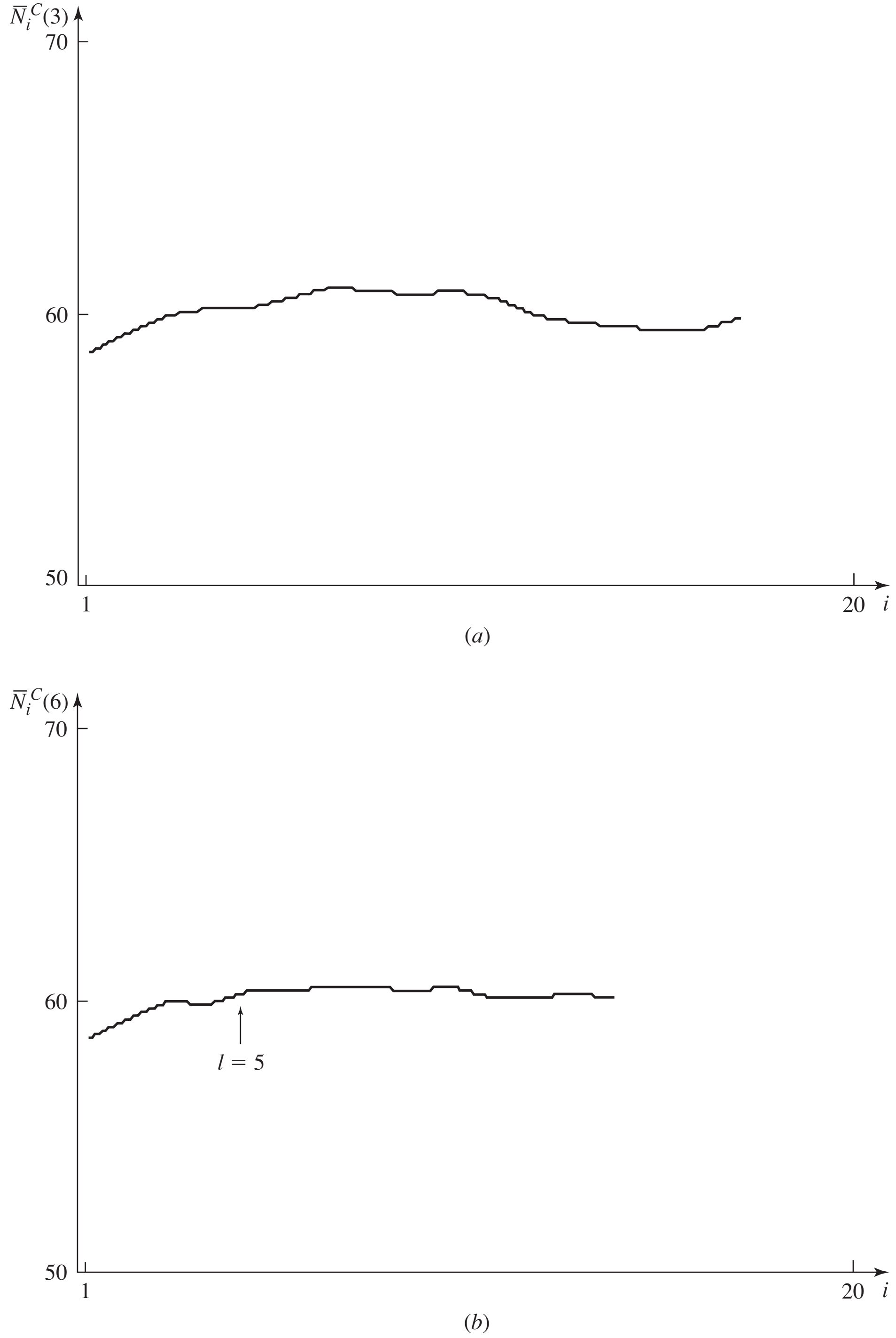


FIGURE 9.19
Averaged process for average hourly throughputs over a shift, small factory with lunch breaks.



(*a*)

(*b*)

FIGURE 9.20
Moving averages for average hourly throughputs over a shift, small factory with lunch breaks: (*a*) $w = 3$; (*b*) $w = 6$.

Then a point estimate and 99 percent confidence interval for ν^C are given by

$$\hat{\nu}^C = \overline{X}^C(10) = 60.24$$

and
$$\overline{X}^C(10) \pm t_{9,0.995}\sqrt{\frac{0.79}{10}} = 60.24 \pm 0.91$$

which also contains 60 (see Prob. 9.27).

9.7 MULTIPLE MEASURES OF PERFORMANCE

In Secs. 9.4 through 9.6 we presented procedures for constructing a confidence interval for a single measure of performance. However, for most real-world simulations several measures of performance are of interest simultaneously. Suppose that I_s is a $100(1 - \alpha_s)$ percent confidence interval for the measure of performance μ_s (where $s = 1, 2, \ldots, k$). (The μ_s's may all be measures of performance for a terminating simulation or may all be measures for a nonterminating simulation.) Then the probability that *all k* confidence intervals *simultaneously* contain their respective true measures satisfies (see Prob. 9.31)

$$P(\mu_s \in I_s \text{ for all } s = 1, 2, \ldots, k) \geq 1 - \sum_{s=1}^{k} \alpha_s \tag{9.12}$$

whether or not the I_s's are independent. This result, known as the *Bonferroni inequality*, has serious implications for a simulation study. For example, suppose that one constructs 90 percent confidence intervals, that is, $\alpha_s = 0.1$ for all s, for 10 different measures of performance. Then the probability that each of the 10 confidence intervals contains its true measure can only be claimed to be greater than or equal to *zero*. Thus, one must be careful in interpreting the results from such a study. The difficulty we have just described is known in the statistics literature as the *multiple-comparisons problem*.

We now describe a practical solution to the above problem when the value of k is small. If one wants the overall confidence level associated with k confidence intervals to be at least $100(1 - \alpha)$ percent, choose the α_s's so that $\sum_{s=1}^{k} \alpha_s = \alpha$. (Note that the α_s's do *not* have to be equal. Thus, α_s's corresponding to more important measures could be chosen smaller.) Therefore, one could construct ten 99 percent confidence intervals and have the overall confidence level be *at least* 90 percent. The difficulty with this solution is that the confidence intervals will be larger than they were originally if a fixed-sample-size procedure is used, or more data will be required for a specified set of k relative errors if a sequential procedure is used. For this reason, we recommend that k be no larger than about 10.

If one has a very large number of measures of performance, the only recourse available is to construct the usual 90 percent or 95 percent confidence intervals but to be aware that one or more of these confidence intervals probably does not contain its true measure.

TABLE 9.13
Results of making 10 replications of the bank model with five tellers and one queue

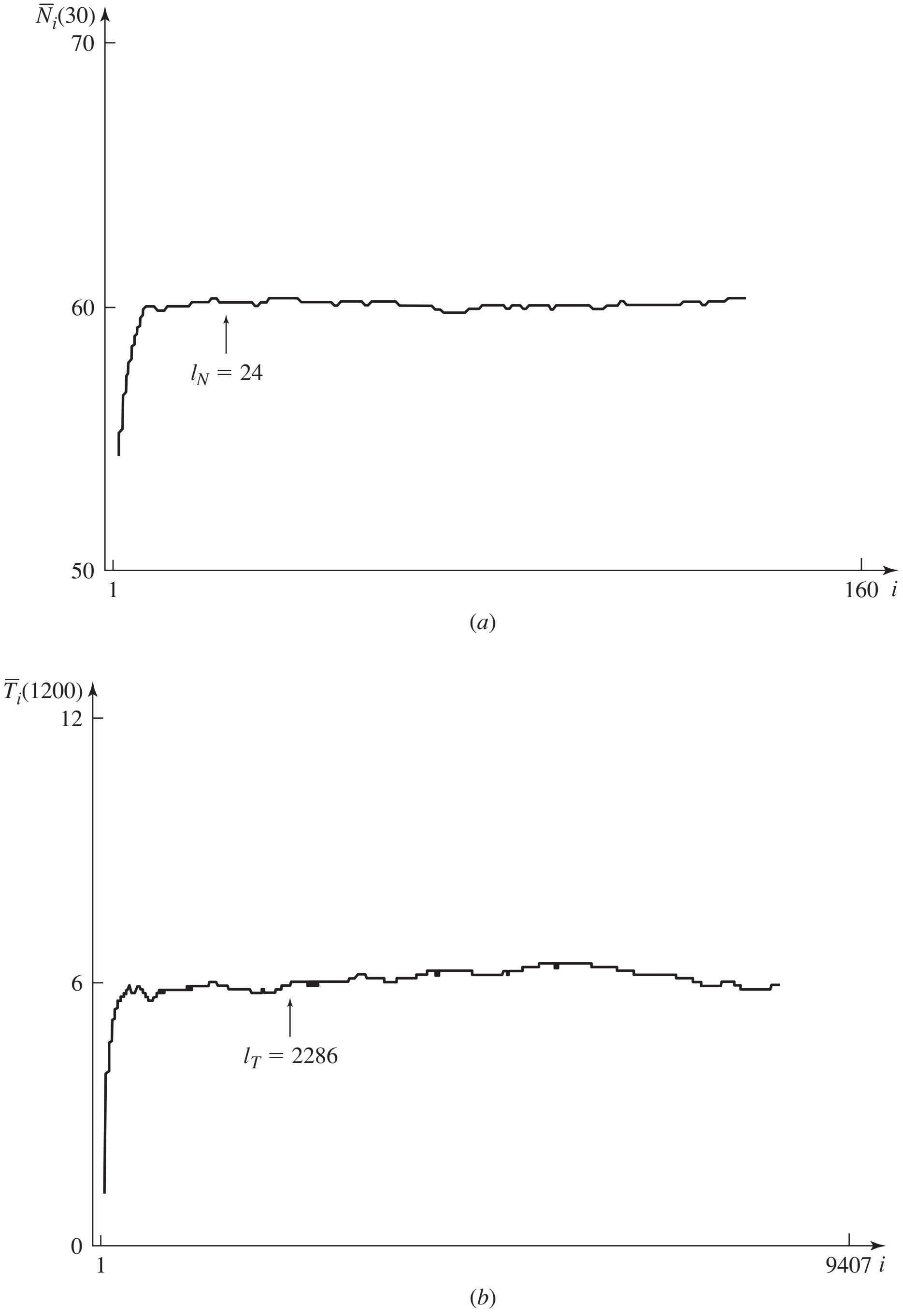
Measure of performance	Point estimate	96.667% confidence interval
$E\left[\frac{\int_0^T Q(t)\,dt}{T}\right]$	1.97	[1.55, 2.40]
$E\left(\frac{\sum_{i=1}^N D_i}{N}\right)$	2.03	[1.59, 2.47]
$E\left[\frac{\sum_{i=1}^N I_i(0, 5)}{N}\right]$	0.85	[0.80, 0.90]

EXAMPLE 9.34. Consider the bank of Example 9.1 with five tellers and one queue. Table 9.13 gives the results of using these 10 replications of the (terminating) simulation and (9.1) to construct 96.667 percent confidence intervals for each of the measures of performance

$$E\left[\frac{\int_0^T Q(t)\,dt}{T}\right], \qquad E\left(\frac{\sum_{i=1}^N D_i}{N}\right), \qquad E\left[\frac{\sum_{i=1}^N I_i(0, 5)}{N}\right]$$

so that the overall confidence level is at least 90 percent (see Prob. 9.38).

EXAMPLE 9.35. Suppose for the small factory of Example 9.25 that we would like to obtain point estimates and confidence intervals for both the steady-state mean hourly throughput ν_N and the steady-state mean time in system of a part ν_T, with the overall confidence level being at least 90 percent. Therefore, we will make the confidence level of each individual interval 95 percent. Using the 10 replications from Example 9.30, we plotted the moving average $\overline{T}_i(w)$ $(i = 1, 2, \ldots)$ for the time-in-system process T_1, $T_2, \ldots$ in order to determine its warmup period. (Here T_i is the time in system of the ith departing part.) Since this plot was highly variable, we made an additional 10 replications of length 160 hours and used the entire 20 replications for our analysis. In Fig. 9.21*a* we plot the hourly throughput moving average $\overline{N}_i(w)$ for $w = 30$, and in Fig. 9.21*b* we plot the time-in-system moving average $\overline{T}_i(w)$ for $w = 1200$. (Note that the number of T_i observations in a 160-hour simulation run is a random variable with approximate mean 9600. Therefore, for our analysis we used the minimum number of observations for any one of the 20 runs, which was 9407.) From Figs. 9.21*a* and 9.21*b*, we decided on warmup periods of $l_N = 24$ hours and $l_T = 2286$ times, respectively. Note, however, that 2286 times corresponds to approximately 38 hours. Since 24 and 2286 are much smaller than 160 and 9407, respectively, we will use these same replications to construct our confidence intervals.



(*a*)

(*b*)

FIGURE 9.21
Moving averages for small factory: (*a*) $w = 30$ for hourly throughputs; (*b*) $w = 1200$ for times in system.

Let

$$X_j = \frac{\sum_{i=25}^{160} N_{ji}}{136} \qquad \text{for } j = 1, 2, \ldots, 20$$

$$Y_j = \frac{\sum_{i=2287}^{9407} T_{ji}}{7121}$$

Then point estimates and 95 percent confidence intervals for ν_N and ν_T are given by

$$\hat{\nu}_N = \overline{X}(20) = 60.03, \qquad \hat{\nu}_T = \overline{Y}(20) = 6.16 \text{ minutes}$$

and

$$\overline{X}(10) \pm t_{19,0.975}\sqrt{\frac{0.70}{20}} = 60.03 \pm 0.39$$

$$\overline{Y}(10) \pm t_{19,0.975}\sqrt{\frac{0.55}{20}} = 6.16 \pm 0.35$$

Thus, we are at least 90 percent confident that ν_N and ν_T are simultaneously in the intervals [59.64, 60.42] and [5.81, 6.51], respectively.

Additional methods for constructing confidence intervals (or regions) for multiple measures of performance are surveyed by Charnes (1995).

9.8 TIME PLOTS OF IMPORTANT VARIABLES

In this chapter we have seen how to construct point estimates and confidence intervals for several different measures of performance, with an emphasis on mean system response. Although these measures are clearly quite useful, there are situations where we need a better indication of how system performance changes dynamically over time. This is particularly true when characteristics of the system (e.g., number of available workers) vary as a function of time. Animation (see Sec. 3.4.3) can provide considerable insight into the short-term dynamic behavior of a system, but it does not give us an easily interpreted record of system performance over the entire length of the simulation. On the other hand, plotting one or more key variables over the duration of the simulation is an easy way to gain an understanding of long-run dynamic system behavior. For example, a graph of queue size over time can provide information on whether the corresponding server (or servers) has sufficient processing capacity and also on the required floor space or capacity for the queue. The following example illustrates the use of time plots, with additional applications being given in Chap. 14 (see also Prob. 9.28).

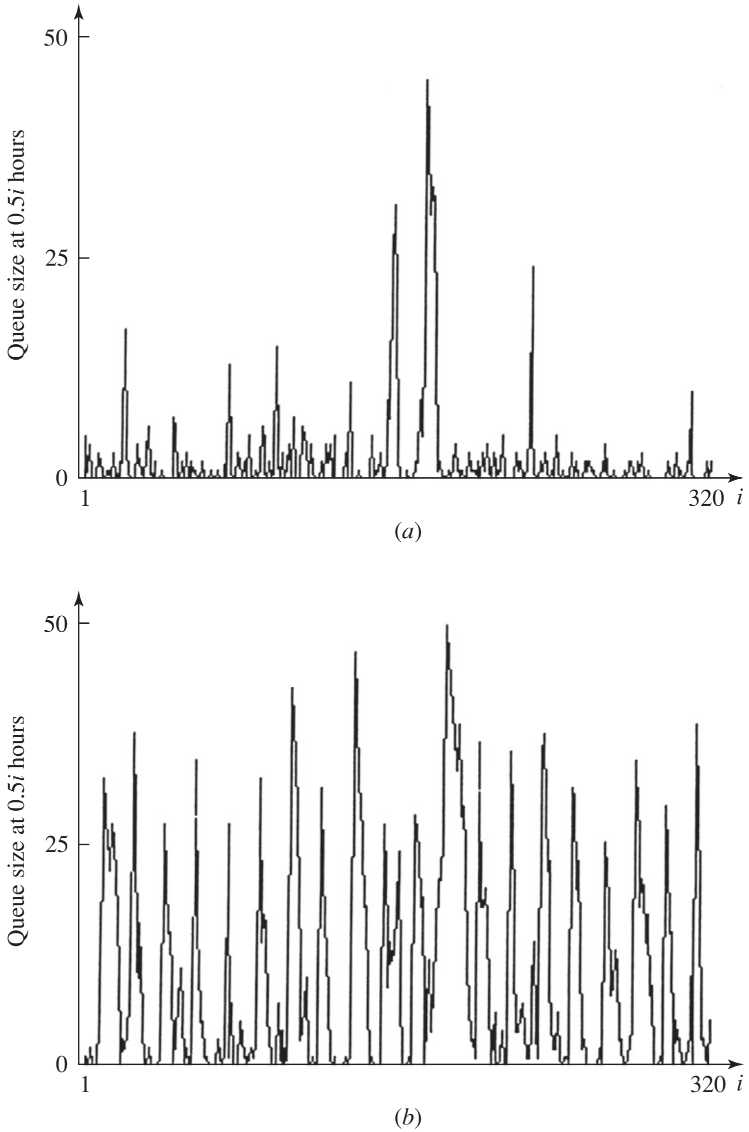


FIGURE 9.22
Time plots for number in queue in time increments of 30 minutes (run 1), small factory: (*a*) machine queue; (*b*) inspector queue.

EXAMPLE 9.36. Consider the small factory with lunch breaks discussed in Example 9.33. In Figs. 9.22*a* and 9.22*b*, we plot the numbers in the machine and inspector queues sampled in 30-minute increments of time, respectively, based on the first of the 10 available simulation replications. Note the periodic behavior of the inspector plot due to the half-hour lunch break.

APPENDIX 9A
RATIOS OF EXPECTATIONS AND JACKKNIFE ESTIMATORS

Much of this chapter has been concerned with estimating the expectation of a single random variable X, namely, $E(X)$. However, as the following examples show, there are many situations in simulation where it is of interest to estimate the ratio of two expectations, such as $E(Y)/E(X)$:

1. For the regenerative method, we saw in Sec. 9.5.3 that steady-state parameters can be expressed as the ratio of two expectations.
2. For the combat simulation of Example 9.5, it is sometimes of interest to estimate $E(R)/E(B)$, where R and B are the numbers of red losses and blue losses in a battle.
3. For the bank simulation of Example 9.14, let $P = \sum_{i=1}^{N} D_i$ be the total delay of all customers served in a day. Then it is of interest to estimate $E(P/N)$, which can be interpreted as the expectation of the average delay of a customer where the expectation is taken with respect to all possible days. However, it may also be of interest to estimate the long-run average delay of all customers, which can be shown to be equal to $E(P)/E(N)$.

Estimators of ratios of expectations, however, are usually biased. We now discuss a method of obtaining a less biased point estimator, as well as an alternative confidence interval.

Suppose that we want to estimate the ratio $\phi = E(Y)/E(X)$ from the data $Y_1, Y_2, \ldots, Y_n$ and $X_1, X_2, \ldots, X_n$, where the X_i's are IID random variables, the Y_i's are IID random variables, and $\text{Cov}(Y_i, X_j) = 0$ for $i \neq j$. The classical point estimator of ϕ is given by $\hat{\phi}_C(n) = \overline{Y}(n)/\overline{X}(n)$; see the discussion of the regenerative method in Sec. 9.5.3 for the classical confidence interval for ϕ. We now discuss the jackknife approach to point and interval estimation of ϕ [see Iglehart (1975) and Miller (1974)]. First define

$$\theta_g = n\hat{\phi}_C(n) - (n-1)\frac{\sum\limits_{\substack{j=1\\ j\neq g}}^{n} Y_j}{\sum\limits_{\substack{j=1\\ j\neq g}}^{n} X_j} \qquad \text{for } g = 1, 2, \ldots, n$$

Then the jackknife point estimator for ϕ is given by $\hat{\phi}_J(n) = \sum_{g=1}^{n} \theta_g/n$, which is, in general, less biased than $\hat{\phi}_C(n)$. Let

$$\hat{\sigma}_J^2(n) = \frac{\sum\limits_{g=1}^{n} [\theta_g - \hat{\phi}_J(n)]^2}{n-1}$$

Then it can be shown [see Miller (1974)] that

$$\frac{\hat{\phi}_J(n) - \phi}{\sqrt{\hat{\sigma}_J^2(n)/n}} \xrightarrow{\mathcal{D}} N(0, 1) \qquad \text{as } n \to \infty$$

which gives the jackknife $100(1 - \alpha)$ percent confidence interval $\hat{\phi}_J(n) \pm z_{1-\alpha/2}\sqrt{\hat{\sigma}_J^2(n)/n}$ for ϕ. (See Sec. 9.5.3 for some empirical results on the relative performance of the classical and jackknife confidence intervals.)

PROBLEMS

9.1. Argue heuristically that comparable output random variables from replications using different random numbers should be independent.

9.2. Consider a machine that works for an exponential amount of time having mean $1/\lambda$ before breaking down. Suppose that it takes an exponential amount of time having mean $1/\omega$ to repair the machine. Let $Y(t)$ be the state of the machine at time t for $t \geq 0$, where

$$Y(t) = \begin{cases} 1 & \text{if the machine is working at time } t \\ 0 & \text{otherwise} \end{cases}$$

Then $\{Y(t), t \geq 0\}$ is a continuous-time stochastic process. Furthermore, it can be shown that [see Ross (2003, pp. 364–366)]

$$P(Y(t) = 1 | Y(0) = 1) = \frac{\lambda}{\lambda + \omega} e^{-(\lambda+\omega)t} + \frac{\omega}{\lambda + \omega}$$

and
$$P(Y(t) = 1 | Y(0) = 0) = -\frac{\omega}{\lambda + \omega} e^{-(\lambda+\omega)t} + \frac{\omega}{\lambda + \omega}$$

Thus, the distribution of $Y(t)$ depends on both t and $Y(0)$. By letting $t \to \infty$ in these equations, compute the steady-state distribution of $Y(t)$. Does it depend on $Y(0)$?

9.3. In Example 9.9, suppose that condition (*b*) is violated. In particular, suppose that it takes workers 20 minutes to put their tools away at the end of a shift and it takes the new workers 20 minutes to set up their tools at the beginning of the next shift. Does $N_1, N_2, \ldots$ have a steady-state distribution?

9.4. Suppose in Example 9.9 that we would like to estimate the steady-state mean total time in system of a part. Does our approach to simulating the manufacturing system present a problem?

9.5. Why is determining the required number of tellers for a bank different from determining the hardware requirements for a computer or communications system (see Example 9.10)?

9.6. In Example 9.11, why doesn't the process of hourly throughputs $N_1, N_2, \ldots$ have a steady-state distribution?

9.7. For the following systems, state whether you think a terminating or nonterminating simulation would be more appropriate. In the terminating cases, state the terminating event E. In the nonterminating cases, would the parameter of interest be a steady-state parameter or a steady-state cycle parameter?

(*a*) Consider a telephone system for which an arriving call may experience a delay before obtaining a line. Suppose that the goal is to estimate the mean delay of the 100th arriving call, $E(D_{100})$.

(*b*) Consider a military inventory system (see Sec. 1.5) during peacetime, which is assumed to have a long duration. Assume that system parameters (e.g., the interdemand time distribution) do not change over time and we are interested in the output process $C_1, C_2, \ldots$, where C_i is the total cost in the ith month. Suppose further that we want a measure of mean cost.

(*c*) Consider a manufacturing system for food products. A production schedule is issued, the system produces product for 13 days, and then the system is completely cleaned out on the fourteenth day. Then a new production schedule is issued and the 2-week cycle is repeated, etc. The goal is to estimate the mean throughput over a cycle.

(*d*) Consider an air freight company that provides overnight delivery of packages. Aircraft loaded with packages start arriving at the hub operations at approximately 11 P.M. The packages are unloaded and then sorted in a warehouse according to the destination Zip code. Packages with similar Zip codes are placed on one aircraft, and the last plane departs at approximately 5 A.M. It is desired to estimate the mean (across departing planes) amount of time that planes are late in departing.

(*e*) Consider a manufacturing system that operates in a similar manner 7 days a week. Suppose, however, that 6 machines operate during the first two shifts in each day, but only 4 machines operate during the third shift. Let $N_1, N_2, \ldots$ be the output process of interest, where N_i is the number of parts produced in the ith shift. We are interested in a measure of mean throughput. Does your answer depend on the relationship between the arrival rate and the service rate of an individual machine?

9.8. For the small factory of Example 9.25, suppose that the system operates 24 hours a day for 5 days and then is completely cleaned out. Thus, we have a terminating simulation of length 120 hours. Make five independent replications and construct a point estimate and 95 percent confidence interval for the mean weekly throughput. Approximately how many replications would be required to obtain an absolute error of 50? A relative error of 5 percent?

9.9. Let p be a probability of interest for a terminating simulation, as discussed in Sec. 9.4.2. Define IID random variables $Y_1, Y_2, \ldots, Y_n$ such that $\hat{p} = \bar{Y}(n)$ and use these Y_j's in (4.3), (4.4), and (4.12) to derive one possible confidence interval for p. Show that the variance estimate given by Eq. (4.4) can be written as $\hat{p}(1 - \hat{p})/(n - 1)$.

9.10. Consider the bank of Example 9.1. Use the data from the 10 replications in Table 9.1 to construct a point estimate for the median (i.e., 0.5-quantile) of the distribution of the average delay over a day. How does this estimate compare with the sample mean in Example 9.14?

9.11. For the $M/M/1$ queue with $\rho < 1$ of Example 9.23, suppose that the number of customers present when the first customer arrives has the following discrete distribution:

$$p(x) = (1 - \rho)\rho^x \qquad \text{for } x = 0, 1, \ldots$$

which is the steady-state distribution of the number of customers in the system. Compute the distribution function of D_1 and its mean. In this case, it can also be shown that D_i for $i \geq 2$ has this same distribution.

9.12. For Welch's procedure in Sec. 9.5.1, show that $E(\overline{Y}_i) = E(Y_i)$ and $\mathrm{Var}(\overline{Y}_i) = \mathrm{Var}(Y_i)/n$.

9.13. Assume that $Y_1, Y_2, \ldots$ is a covariance-stationary process and that $\rho_i < 1$ for $i \geq 1$. Show for Welch's procedure that $\mathrm{Var}[\overline{Y}_i(w)] < \mathrm{Var}(\overline{Y}_i)$.

9.14. Suppose that $Y_1, Y_2, \ldots$ is an output process with steady-state mean ν and that $\overline{Y}(m)$ is the usual sample mean based on m observations. Consider plotting $\overline{Y}(m)$ as a function of m and let l' be the point beyond which $\overline{Y}(m)$ does not change appreciably. Is l' a good warmup period in the sense that $E(Y_i) \approx \nu$ for $i > l'$ and also that l' is not excessively large? Why?

9.15. Consider the replication/deletion approach of Sec. 9.5.2. Show that $E(X_j) \approx \nu$. Give two reasons why the confidence interval given by (9.6) is only approximate in terms of coverage.

9.16. Consider the replication/deletion approach in Sec. 9.5.2 based on using the same set of replications to determine the warmup period l and to construct a confidence interval. Are the resulting X_j's truly independent?

9.17. For the small factory of Example 9.30, what should the steady-state mean hourly throughput be if the system is well defined in the sense that $\rho < 1$ for both the machine and the inspector?

9.18. Consider a continuous-time stochastic process such as $\{Q(t), t \geq 0\}$, where $Q(t)$ is the number of customers in queue at time t. Suppose that we would like to estimate the steady-state time-average number in queue, Q (see App. 1B for one definition), using the method of batch means based on one simulation run of length m time units. Discuss two approaches for getting *exactly* m basic discrete observations $Q_1, Q_2, \ldots, Q_m$ for use in the method of batch means. The m Q_i's will be batched to form n batch means.

9.19. If $Y_1, Y_2, \ldots$ is a covariance-stationary process, show for the method of batch means that $C_i(k) = \mathrm{Cov}[\overline{Y}_j(k), \overline{Y}_{j+i}(k)]$ is given by

$$C_i(k) = \sum_{l=-(k-1)}^{k-1} \frac{(1 - |l|/k)\, C_{ik+l}}{k} \qquad \text{where } C_l = \mathrm{Cov}(Y_i, Y_{i+l})$$

9.20. Let $Y_1, Y_2, \ldots$ be a covariance-stationary process. For the method of batch means, let $\rho_i(k) = \mathrm{Cor}[\overline{Y}_j(k), \overline{Y}_{j+i}(k)]$ and let $b(n, k)$ be such that $E\{\widehat{\mathrm{Var}}[\overline{\overline{Y}}(n, k)]\} = b(n, k) \cdot \mathrm{Var}[\overline{\overline{Y}}(n, k)]$. Show that $\rho_i(k) \to 0$ (for $i = 1, 2, \ldots, n-1$) as $k \to \infty$ implies that $E\{\widehat{\mathrm{Var}}[\overline{\overline{Y}}(n, k)]\} \to \mathrm{Var}[\overline{\overline{Y}}(n, k)]$ as $k \to \infty$. *Hint:* First show that

$$b(n, k) = \frac{\left\{ n \Big/ \left[1 + 2 \sum_{i=1}^{n-1} (1 - i/n)\, \rho_i(k) \right] \right\} - 1}{n - 1}$$

9.21. For the regenerative method, show that $\nu = E(Z)/E(N)$. [*Hint:* Observe that

$$\frac{\sum_{j=1}^{n'} Z_j}{\sum_{j=1}^{n'} N_j} = \frac{\sum_{i=1}^{M(n')} Y_i}{M(n')}$$

where n' is the number of regeneration cycles and $M(n')$ is the total number of observations (a random variable) in the n' cycles. Let $n' \to \infty$ and apply the strong law of large numbers (see Sec. 4.6) to both sides of the above equation.] Also conclude that $\hat{\nu}(n') = \overline{Z}(n')/\overline{N}(n') \to \nu$ as $n' \to \infty$ (w.p. 1), so that $\hat{\nu}(n')$ is a strongly consistent estimator of ν. (See the definitions of ν in Sec. 9.5.3.)

9.22. For the queueing system considered in Example 9.32, are the indices of those customers who depart and leave exactly l customers behind ($l \geq 0$ and fixed) regeneration points for the process $D_1, D_2, \ldots$? If not, under what circumstances would they be?

9.23. For the inventory example of Sec. 1.5, identify a sequence of regeneration points for the monthly-cost process. Repeat assuming that the interdemand times are not exponential random variables.

9.24. Suppose that $\hat{\nu}(n')$ is the (biased) regenerative point estimator for the steady-state mean ν based on simulating the process $Y_1, Y_2, \ldots$ for n' regeneration cycles. Do you think that it is advisable to have a warmup period of l cycles to reduce the point estimator bias?

9.25. Consider an *M*/*M*/1 queue with $\rho < 1$, and let the number of customers served in a cycle, N, be as defined in Example 9.32. By conditioning on whether the second customer arrives before or after the first customer departs, show that $E(N) = 1/(1 - \rho)$.

9.26. For Example 9.33, compute the utilization factor ρ for both the machine and the inspector. What arrival rate should be used? Is this system well defined in the sense that $\rho < 1$ in both cases?

9.27. In Example 9.33, what should be the value for ν^C if the system is well defined?

9.28. A manufacturing system consists of two machines in parallel and a single queue. Jobs arrive with exponential interarrival times at a rate of 10 per hour, and each machine has exponential processing times at a rate of 8 per hour. During the first 16 hours of each day both machines are operational, but only one machine is used during the final 8 hours.

(*a*) Determine whether the system is well defined by computing the utilization factor ρ and comparing it with l.

(*b*) Let N_i be the throughput for the *i*th hour. Does $N_1, N_2, \ldots$ have a steady-state distribution?

(*c*) Make 10 replications of the simulation of length 480 hours (20 days) each. Plot the averaged process $\overline{N}_1, \overline{N}_2, \ldots, \overline{N}_{480}$.

(*d*) Let M_i be the throughput for the *i*th 24-hour day. Use the data from part (*c*) and the replication/deletion approach to construct a point estimate and 90 percent confidence interval for the steady-state mean daily throughput $\nu = E(M) = 240$.

9.29. For the system in Prob. 9.28, make one replication of length 200 days and let M_i be as previously defined. Use the M_i's and the method of batch means to construct a point estimate and a 90 percent confidence interval for $\nu = 240$ based on $n = 10$ batches and also on $n = 5$ batches.

9.30. Repeat Prob. 9.29 using standardized time series rather than batch means.

9.31. Let E_s be an event that occurs with probability $1 - \alpha_s$ for $s = 1, 2, \ldots, k$. Then prove that

$$P\left(\bigcap_{s=1}^{k} E_s\right) \geq 1 - \sum_{s=1}^{k} \alpha_s$$

where $\bigcap_{s=1}^{k} E_s$ is the intersection of the events $E_1, E_2, \ldots, E_k$. Do not assume that the E_s's are independent. [This result is called the *Bonferroni inequality*; see (9.12).] *Hint:* The proof is by mathematical induction. That is, first show that $P(E_1 \cap E_2) \geq 1 - \alpha_1 - \alpha_2$. Then show that if

$$P\left(\bigcap_{s=1}^{k-1} E_s\right) \geq 1 - \sum_{s=1}^{k-1} \alpha_s$$

is true, the desired result is also true.

9.32. For Example 9.21, compute the approximate number of replications required to reduce the half-length of the confidence interval for p to 0.05.

9.33. For Example 9.26, show that there must be two links between SP-1 and SP-2 so that the utilization of this link group does not exceed 0.4.

9.34. For Example 9.26, show that STP-A must contain three processors so that its utilization does not exceed 0.4. Also show that SP-1 must contain two processors so that its utilization is less than 1.

9.35. For Example 9.26, show that the overall arrival rate of messages is 1240 per second.

9.36. For Example 9.31, does the utilization of 0.316 for STP-A seem reasonable? What about the utilization of 0.377 for link 1-2?

9.37. Suppose that one constructs k $100(1 - \alpha)$ percent confidence intervals from the same set of replications (see Sec. 9.7), so the confidence intervals are *dependent.* Derive an expression for the expected number of confidence intervals that do *not* contain their respective true measures of performance.

9.38. If the 10 replications and subsequent analysis corresponding to Table 9.13 were performed independently by 100 different banks, what could be said about the confidence intervals for approximately 90 of the banks?