Output Data Analysis

Christos Alexopoulos Andrew F. Seila

June 1998

Chapter 7 in *Handbook of Simulation*
ISBN 0-471-13403-1

1 INTRODUCTION

The primary purpose of most simulation studies is the approximation of prescribed system parameters with the objective of identifying parameter values that optimize some system performance measures. If some of the input processes driving a simulation are random, then the output data are also random and runs of the simulation result in *estimates* of performance measures. Unfortunately, a simulation run does not usually produce independent, identically distributed observations; therefore "classical" statistical techniques are not directly applicable to the analysis of simulation output.

A simulation study consists of several steps such as data collection, coding and verification, model validation, experimental design, output data analysis, and implementation. This chapter focuses on statistical methods for computing confidence intervals for system performance measures from output data. Several aspects of output analysis, such as comparison of systems, design of simulation experiments, and variance reduction methods, will not be discussed. These subjects are treated in other chapters of this handbook and in several texts including Bratley, Fox, and Schrage (1987), Fishman (1978b, 1996), Kleijnen (1974, 1975), and Law and Kelton (1991).

The reader is assumed to be comfortable with probability theory and statistics at the level of Hogg and Craig (1978), and stochastic processes at the level of Ross (1993). A reader who is only interested in computational methods can skip the technical parts of this chapter. Sections 1.1 and 1.2 review definitions and results that are essential for the study of this chapter. Section 2 discusses methods for analyzing output from finite-horizon simulations. Sections 3 and 4 present techniques for point and interval estimation of steady-state parameters.

1.1 Limit Theorems and Their Statistical Implications

This section reviews the tools needed to establish asymptotic (as the sample size increases) properties of estimators and to obtain confidence intervals. Consider the following three forms of convergence for sequences of random variables on the same probability space: The first form is the strongest while the last form is the weakest (and easiest to establish). For

additional forms of convergence as well as the relationships between the forms, see Chapter 5 of Karr (1993).

Almost sure convergence. The sequence $X_1, X_2, \ldots$ converges to the random variable X *almost surely* (or *with probability* 1) (we write $X_n \xrightarrow{a.s.} X$) if $P(X_n \to X \text{ as } n \to \infty) = 1$.
Convergence in probability. The sequence $X_1, X_2, \ldots$ converges to X *in probability* (we write $X_n \xrightarrow{\mathcal{P}} X$) if for every $\epsilon > 0$,

$$P\left(|X_n - X| \le \epsilon\right) \to 1 \quad \text{as } n \to \infty.$$

Convergence in distribution. The sequence $X_1, X_2, \ldots$ converges to X *in distribution* (we write $X_n \xrightarrow{\mathcal{D}} X$) if

$$P(X_n \le x) \to P(X \le x) \quad \text{as } n \to \infty$$

at all points x where the cumulative distribution function $P(X \le x)$ is continuous.

Now suppose that the random variables $X_1, X_2, \ldots$ are from some distribution with an unknown parameter θ and the objective is to estimate a function $g(\theta)$. For fixed n, let $\delta_n = \delta_n(X_1, \ldots, X_n)$ be an estimator of $g(\theta)$. If $E(\delta_n) = g(\theta)$, δ_n is called an *unbiased* estimator. Furthermore, δ_n is said to be a *consistent* (respectively, *strongly consistent*) estimator of $g(\theta)$ if $\delta_n \xrightarrow{\mathcal{P}} g(\theta)$ (respectively, $\delta_n \xrightarrow{a.s.} g(\theta)$). If δ_n is unbiased for each n and $\text{Var}(\delta_n) \to 0$ as $n \to \infty$, then δ_n is also consistent (Lehmann 1991, pp. 331–333).

The remainder of this subsection illustrates the above concepts with a few classical results. Suppose that $X_1, X_2, \ldots, X_n$ are independent, identically distributed (i.i.d.) random variables with finite mean μ and variance σ^2. Let

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

be the sample mean of the X_i's. Since $E(\overline{X}_n) = \mu$, $\overline{X}_n$ is an unbiased estimator of μ. $\overline{X}_n$ is also a strongly consistent estimator of μ by the *strong law of large numbers*

$$\overline{X}_n \xrightarrow{a.s.} \mu \quad \text{as } n \to \infty$$

(Karr 1993, pp. 188–189).

If $0 < \sigma^2 < \infty$, the *central limit theorem* (Karr 1993, p. 174) states that

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1) \quad \text{as } n \to \infty,$$

where $N(0,1)$ denotes a normal random variable with mean 0 and variance 1. In other words,

$$P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \le z\right) \to \Phi(z) \quad \text{as } n \to \infty, \tag{1}$$

where Φ is the distribution function of the standard normal random variable.

The central limit theorem remains valid when the potentially unknown parameter σ^2 is replaced by its unbiased and consistent estimator

$$S_n^2(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Therefore, for sufficiently large n,

$$P\left(\frac{|\overline{X}_n - \mu|}{S_n(X)/\sqrt{n}} \leq z_{1-\alpha/2}\right) \approx 1-\alpha, \tag{2}$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of $N(0,1)$.

Now suppose that the mean μ is unknown. Solving the inequality on the l.h.s. of (2) for μ, one has the well-known approximate (two-sided) $1-\alpha$ confidence interval

$$\overline{X}_n \pm z_{1-\alpha/2}\frac{S_n(X)}{\sqrt{n}}. \tag{3}$$

The l.h.s. of (2) is the probability that the confidence interval (3) contains the true mean μ. Denote this probability by $p_{n,\alpha}$ and call it the "coverage probability" of (3). One interprets this confidence interval as follows: Suppose that a large number of independent trials is performed; in each trial, n observations are collected and a confidence interval for μ is computed using (3). As the number of trials grows, the proportion of confidence intervals that contain μ approaches $1-\alpha$.

The number of observations n required for $p_{n,\alpha} \approx 1-\alpha$ depends on the symmetry of the distribution of X_i. The more skewed (asymmetric) the density/probability function of X_i, the larger n required. To reduce undercoverage problems ($p_{n,\alpha} < 1-\alpha$) for small n, one may replace the normal quantile $z_{1-\alpha/2}$ by the larger quantile $t_{n-1,1-\alpha/2}$ of the t distribution with $n-1$ degrees of freedom. This choice for degrees of freedom is due to the fact that for i.i.d. normally distributed X_i,

$$\frac{\overline{X}_n - \mu}{S_n(X)/\sqrt{n}} \sim t_{n-1},$$

where the notation $X \sim Y$ is used to indicate that the random variables X and Y have the same distribution.

1.2 Stochastic Processes

Simulation output data are realizations (or *sample paths*) of *stochastic processes*. A stochastic process is a probabilistic model of a system that evolves randomly. More formally, a stochastic process is a collection $X = \{X(u), u \in T\}$ of random variables indexed by a parameter u taking values in the set T. The random variables $X(u)$ take values in a set S, called the *state space* of the process X. Throughout this chapter, u will represent time, and we will encounter the following two cases: (a) $T = \{0, 1, 2, \ldots\}$, for which the notation $X = \{X_i, i \geq 0\}$ will be used. For example, X_i may represent the price of a stock at the end of day i or the time in queue of the ith customer at a post office. (b) $T = [0, \infty)$. In

this case, the notation $X = \{X(t), t \geq 0\}$ will be used. Some examples of $X(t)$ would be the number of failed machines in a shop at time t, the throughput of a shop at time t, or the price of a stock at time t.

One way to describe a stochastic process is to specify the joint distribution of $X(t_1), X(t_2), \ldots, X(t_n)$ for each set of times $t_1 < t_2 < \cdots < t_n$ and each n. This approach is typically too complicated to be attempted in practice. An alternative, and simpler approach, is to specify the first and second *moment* functions of the process. These functions are the *mean* function $\mu(t) = E[X(t)]$, the *variance* function $\sigma^2(t) = \text{Var}[X(t)]$, and the *autocovariance* function

$$C(t_1, t_2) = \text{Cov}[X(t_1), X(t_2)], \quad t_1 \leq t_2.$$

Notice that $C(t_1, t_2) = C(t_2, t_1)$ and $C(t, t) = \sigma^2(t)$. (For a discrete-time process, the notation μ_t, σ_t^2, and C_{t_1,t_2} will be used.)

In order to analyze a simulation output process, one must make some structural assumptions. The following are the two most frequently used assumptions.

Strict Stationarity

The process X is called *(strictly) stationary* if the joint distribution of $X(t_1), X(t_2), \ldots, X(t_n)$ is the same as the joint distribution of $X(t_1 + s), X(t_2 + s), \ldots, X(t_n + s)$ for all $t_1, t_2, \ldots, t_n$, and s. In simpler terms, shifting the time origin from zero to any other value s, has no effect on the joint distributions. An immediate result is that the joint distribution of $X(t_1), X(t_2), \ldots, X(t_n)$ depends only on the intervals between $t_1, t_2, \ldots, t_n$.

Example 1 (The M/M/1 queue) Consider an M/M/1 queueing system with i.i.d. interarrival times A_i, $i \geq 1$, from the exponential distribution with rate τ and i.i.d. service times S_i, $i \geq 1$, from the exponential distribution with rate ω ($\tau < \omega$). The ratio $\nu = \tau/\omega$ is called traffic intensity or (long-run) server utilization. Suppose that the service discipline is first-come, first-served. Let D_i be the delay time in queue of the ith customer and assume that the system starts empty. The first of Lindley's recursive equations (Lindley 1952)

$$\begin{aligned} D_1 &= 0 \\ D_{i+1} &= \max\{D_i + S_i - A_{i+1}, 0\}, \quad i \geq 1 \end{aligned} \tag{4}$$

implies $E(D_1) = 0$ whereas $P(D_2 > 0) = P(S_1 > A_2) = \tau/(\tau + \omega) > 0$ implies $E(D_2) > 0$. Therefore the delay process $\{D_i, i \geq 1\}$ is not stationary. Using queueing theory (Ross 1993, Chapter 8) one has

$$\lim_{i \to \infty} P(D_i \leq x) = 1 - \nu + \nu(1 - e^{-(\omega - \tau)x}), \quad x \geq 0, \tag{5}$$

$$\mu = \lim_{i \to \infty} E(D_i) = \frac{\nu}{(1 - \nu)\omega}, \quad \text{and} \quad \sigma^2 = \lim_{i \to \infty} \text{Var}(D_i) = \frac{\nu(2 - \nu)}{\omega^2(1 - \nu)^2}.$$

Equation (5) suggests that the delay process becomes asymptotically stationary. Indeed, if D_1 has the distribution on the r.h.s. of (5), equations (4) imply (after some work) that all D_i have the same distribution and the delay process is stationary.

Weak Stationarity

In practice it is often necessary to consider a less restricted form of stationarity. The process X is said to be *weakly stationary* if its mean and variance functions are constant (equal to μ and σ^2 respectively) and its autocovariance function satisfies

$$\text{Cov}[X(t), X(t+s)] = C(s), \quad t \geq 0, s \geq 0,$$

that is, it depends only on the *lag* s. In this case, the *autocorrelation* function is defined by

$$\rho(s) = \text{Corr}[X(t), X(t+s)] = C(s)/\sigma^2, \quad s \geq 0.$$

Example 2 (A stationary M/M/1 queue) The autocorrelation function of the delay process $\{D_i\}$ in a stationary M/M/1 queueing system is given by (Blomqvist 1967)

$$\rho_j = \frac{(1-\nu)^3(1+\nu)}{(2-\nu)\nu^3} \sum_{k=j+3}^{\infty} \left[\frac{\nu}{(\nu+2)^2}\right]^k \frac{(2k-3)!}{k!(k-2)!}(k-j-1)(k-j-2), \quad j = 0, 1, \ldots$$

This function is monotone decreasing with a very long tail that increases as the server utilization ν increases (for instance, $\rho_{200} \approx 0.30$ when $\nu = 0.9$). This makes the M/M/1 system a good test bed for evaluating simulation methodologies.

Example 3 (Moving average process) A well-studied example of a discrete-time weakly stationary process is the moving average process of order q (often abbreviated to MA(q))

$$X_i = \beta_0 Z_i + \beta_1 Z_{i-1} + \cdots + \beta_q Z_{i-q}, \quad i \geq 0,$$

where the coefficients β_i's are constants and $\{Z_i, i = 0, \pm 1, \pm 2, \ldots\}$ are i.i.d. random variables with mean zero and finite variance a^2. MA processes have applications in several areas, particularly econometrics (Chatfield 1989).

Clearly

$$E(X_i) = 0, \quad \text{Var}(X_i) = a^2 \sum_{i=0}^{q} \beta_i^2$$

while some algebra yields the autocovariace function

$$C_j = \begin{cases} a^2 \sum_{i=0}^{q-j} \beta_i \beta_{i+j} & j = 0, 1, \ldots, q \\ 0 & j > q \end{cases}$$

which "cuts off" at lag q. If in addition the Z_i's are normally distributed, then the MA(q) process is stationary.

Now suppose one observes the portion $X_1, \ldots, X_n$ of a discrete-time weakly stationary process for the purpose of estimating the mean μ. Clearly, $\overline{X}_n$ is an unbiased estimator of μ while some algebra yields

$$\text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}\left[1 + 2\sum_{j=1}^{n-1}(1 - j/n)\rho_j\right] \equiv \frac{\sigma^2}{n}(1 + \gamma_n). \tag{6}$$

In order for $\overline{X}_n$ to be a consistent estimator of μ, we require that $\lim_{n\to\infty} \mathrm{Var}(\overline{X}_n) = 0$. The last condition holds if $\lim_{n\to\infty} n\mathrm{Var}(\overline{X}_n) < \infty$ or, equivalently,

$$\lim_{n\to\infty} \gamma_n < \infty. \tag{7}$$

For (7) to hold, $\lim_{j\to\infty} C_j = 0$ is necessary but not sufficient. A necessary and sufficient condition is

$$\lim_{n\to\infty} \sum_{j=-(n-1)}^{n-1} C_j = \sum_{j=-\infty}^{\infty} C_j < \infty. \tag{8}$$

In simple terms, the covariance between X_i and X_{i+j} must dissipate sufficiently fast so that the summation in (8) remains bounded.

Example 4 (First-order autoregressive process) Another well-known stationary process is the autoregressive process of order one, denoted by AR(1), and often called the Markov process in the time series literature,

$$X_i = \mu + \rho(X_{i-1} - \mu) + Z_i, \quad i \geq 1,$$

where $|\rho| < 1$, $X_0 \sim N(\mu, 1)$, and the Z_i's are i.i.d. $N(0, 1-\rho^2)$ (see Figure 1).

The autocorrelation function of this process

$$\rho_j = \rho^j, \quad j \geq 0$$

is monotone decreasing if $\rho > 0$ with a tail that becomes longer as ρ increases, and exhibits damped harmonic behavior around the zero axis if $\rho < 0$.

Applying equation (6) one has

$$n\mathrm{Var}(\overline{X}_n) = 1 + 2\sum_{j=1}^{n-1}(1 - j/n)\rho^j \to \frac{1+\rho}{1-\rho} \quad \text{as } n \to \infty.$$

Hence $\overline{X}_n$ is a consistent estimator of the mean $\mu = E(X_i)$. The limit $(1+\rho)/(1-\rho)$ is often called the *time-average process variance*.

Brownian Motion and Brownian Bridge

A continuous-time stochastic process with frequent use in simulation output analysis (see Section 3.4) is the standard Brownian motion $\{W(t), t \geq 0\}$. This process has the following properties: (i) $W(0) = 0$; (ii) W has independent *increments*, that is, for $0 \leq t_0 \leq t_1 \cdots \leq t_n$,

$$P[W(t_j) - W(t_{j-1}) \leq w_j,\ 1 \leq j \leq n] = \prod_{j=1}^{n} P[W(t_j) - W(t_{j-1}) \leq w_j];$$

(iii) For $0 \leq s < t$, the increment $W(t) - W(s)$ has the $N(0, t-s)$ distribution.

A well-known function of the Brownian motion is the (standard) Brownian bridge process defined by

$$B(t) = W(t) - tW(1), \quad 0 \leq t \leq 1.$$

Figures 2 and 3 depict sample paths of $W(t)$ and $B(t)$ in the interval $[0, 1]$. Notice that $B(0) = B(1) = 0$.

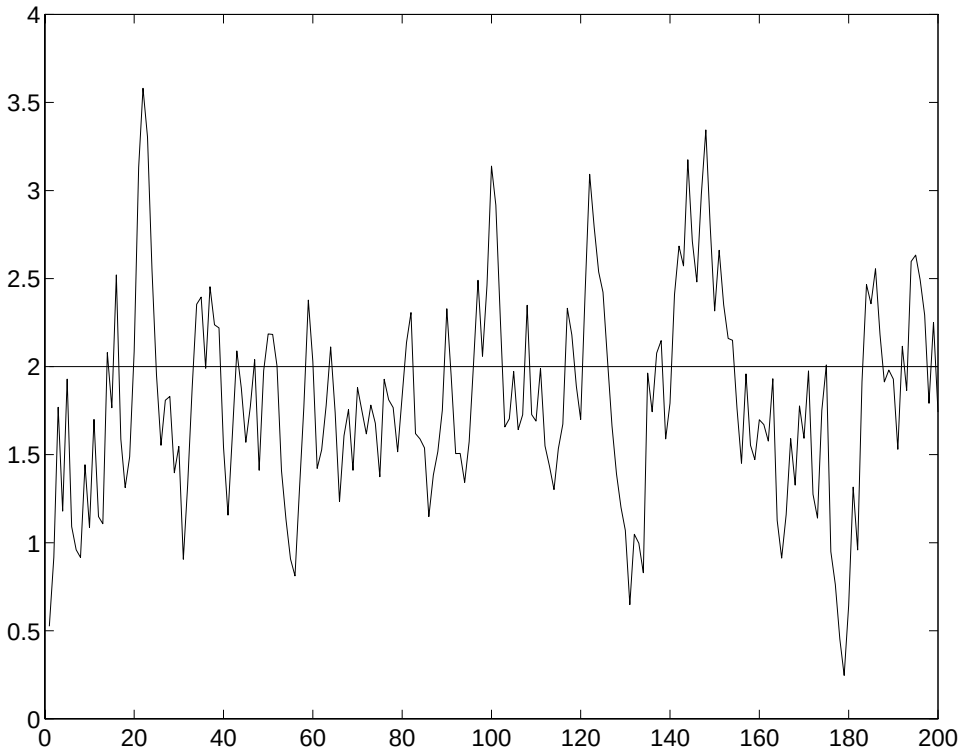

Figure 1: Sample path of the stationary AR(1) process $X_i = 2 + 0.8(X_{i-1} - 2) + Z_i$

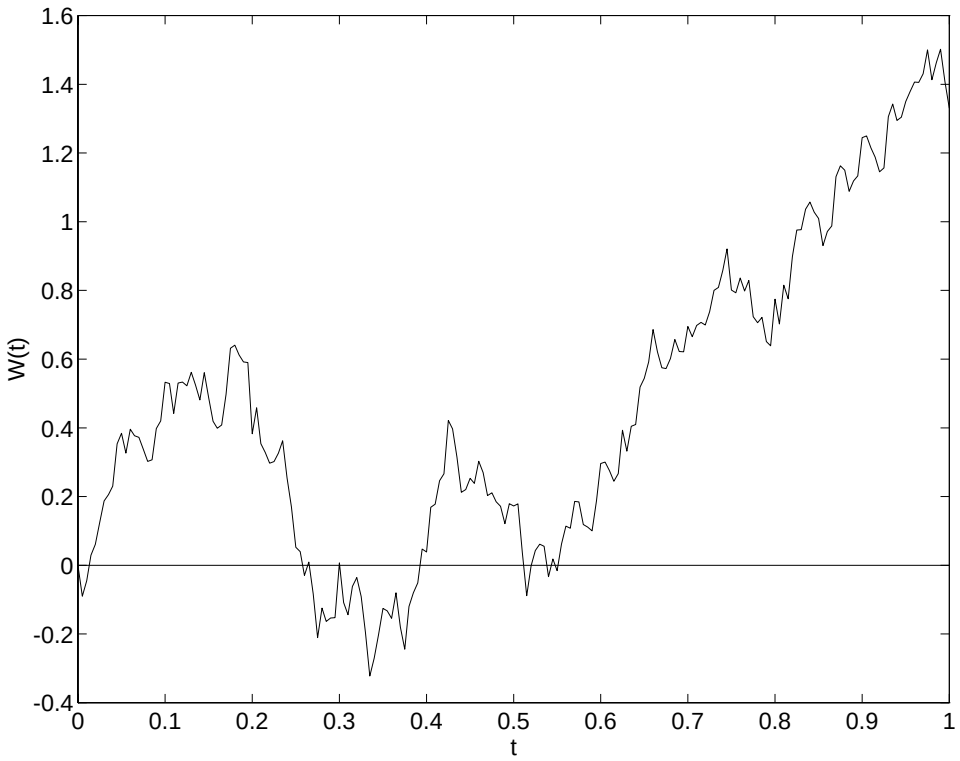

Figure 2: Sample path of a standard Brownian motion in $[0, 1]$

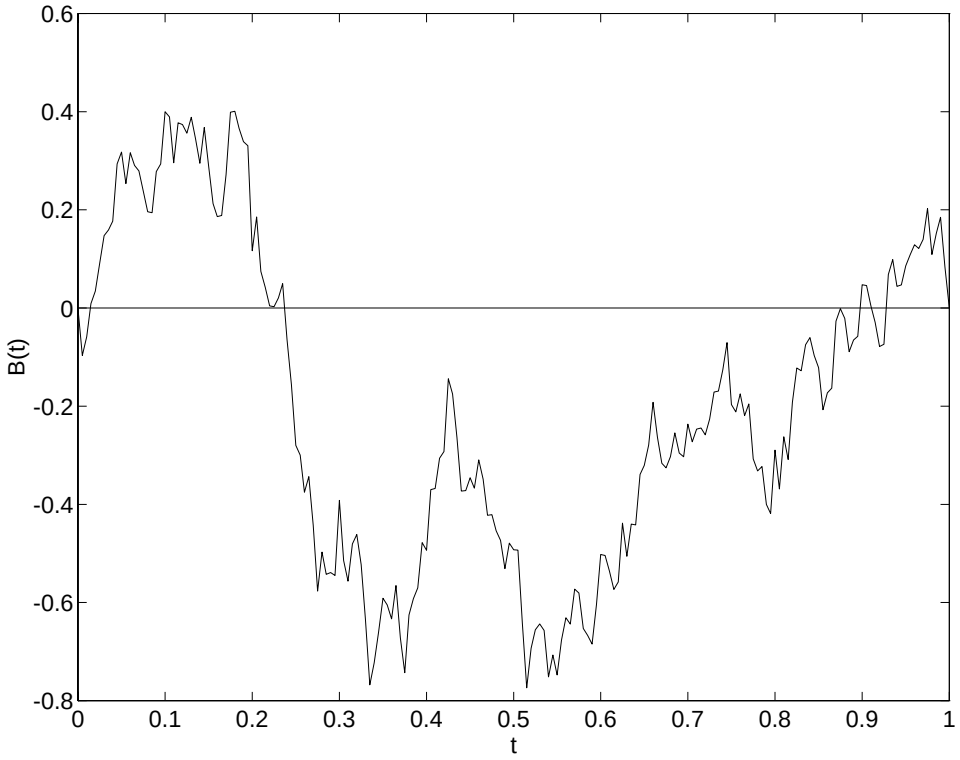

Figure 3: The respective Brownian bridge for the Brownian motion in Figure 2

1.3 Types of Simulations

There are two types of simulations with regard to output analysis:

1. Finite-horizon simulations. In this case the simulation starts in a specific state, such as the empty and idle state, and is run until some terminating event occurs. The output process is not expected to achieve any steady-state behavior and any parameter estimated from the output data will be transient in the sense that its value will depend upon the initial conditions. An example is the simulation of a computer network, starting empty, until n jobs are completed. One might wish to estimate the mean time to complete n jobs, or the mean of the average waiting time for the n jobs.

2. Steady-state simulations. The purpose of a steady-state simulation is the study of the long-run behavior of the system of interest. A performance measure of a system is called a *steady-state parameter* if it is a characteristic of the equilibrium distribution of an output stochastic process (Law and Kelton 1991). An example is the simulation of a continuously operating communication system where the objective is the computation of the mean delay of a data packet.

2 FINITE-HORIZON SIMULATIONS

Suppose that one starts in a specific state and simulates a system until n output data $X_1, X_2, \ldots, X_n$ are collected with the objective of estimating $f(X_1, \ldots, X_n)$, where f is a "nice"[1] function of the data. For example, X_i may be the transit time of unit i through a network of queues or the total time station i is busy during the ith hour and $f(X_1, \ldots, X_n) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the average transit time for the n jobs.

2.1 Estimation of the Mean via Independent Replications

This subsection focuses on the estimation of $\mu = E(\overline{X}_n)$. By definition, $\overline{X}_n$ is an unbiased estimator for μ. Unfortunately, the X_i's are generally dependent random variables which makes the estimation of the variance $\mathrm{Var}(\overline{X}_n)$ a nontrivial problem. In many queueing systems the X_i's are positively correlated. When this is the case, the familiar estimator

$$\frac{S_n^2(X)}{n} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

is a highly biased estimator of $\mathrm{Var}(\overline{X}_n)$.

Example 5 Consider a stationary M/M/1 queueing system (see Examples 1 and 2) with service rate $\omega = 1$ and server utilization $\nu = 0.9$. Using the formulas for ρ_j, one can show that

$$E\left[\frac{S_{10}^2(D)}{10}\right] = 0.033\sigma^2,$$

[1] Formally, f must be a measurable function. In practice, all functions encountered in simulation output analysis are measurable.

where $\sigma^2 = \text{Var}(D_i) = 99$. As a result, the $1-\alpha$ confidence interval $\overline{D}_{10} \pm t_{9,1-\alpha/2} S_{10}(D)/\sqrt{10}$ for the mean delay $\mu = 0.9/(1-0.9) = 9$ computed from 10 consecutive delay observations from a single replication will likely be unacceptably narrow.

To overcome this problem, one can run k independent replications of the system simulation. Each replication starts in the same state and uses a portion of the random number stream that is different from the portions used to run the other replications. Assume that replication i produces the output data $X_{i1}, X_{i2}, \ldots, X_{in}$. Then the sample means

$$Y_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}, \quad i = 1, \ldots, k,$$

are i.i.d. random variables,

$$\overline{Y}_k = \frac{1}{k}\sum_{i=1}^{k} Y_i$$

is also an unbiased estimator of μ, and the sample variance of the Y_i's

$$S_k^2(Y) = \frac{1}{k-1}\sum_{i=1}^{k}(Y_i - \overline{Y}_k)^2$$

is an unbiased estimator of $\text{Var}(\overline{X}_n)$. If, in addition, n and k are sufficiently large, an approximate $1-\alpha$ confidence interval for μ is

$$\overline{Y}_k \pm t_{k-1,1-\alpha/2}\frac{S_k(Y)}{\sqrt{k}}. \tag{9}$$

Denote the halfwidth of the interval (9) by $\delta(k,\alpha) = t_{k-1,1-\alpha/2} S_k(Y)/\sqrt{k}$.

2.2 Sequential Estimation

A fundamental problem is the estimation of μ within a tolerance $\pm d$, where d is user-specified. More formally, one would like to make k runs so that

$$P(\overline{Y}_k - d \le \mu \le \overline{Y}_k + d) \ge 1-\alpha, \tag{10}$$

where $\alpha \in (0,1)$. The sequential procedure of Chow and Robbins (1965) (see also Nadas 1969) is to run one replication at a time and stop at run k^* such that

$$k^* = \min\left[k : k \ge 2, \delta(k,\alpha) \le \sqrt{\frac{k}{k-1}d^2 - \frac{t_{k-1,1-\alpha/2}^2}{k(k-1)}}\right]. \tag{11}$$

The *stopping rule* (11) is based on the limiting result

$$\lim_{d\to 0} P(\overline{Y}_{k^*} - d \le \mu \le \overline{Y}_{k^*} + d) = 1-\alpha. \tag{12}$$

Equation (12) indicates that as the tolerance d decreases, the probability that the interval $\overline{Y}_{k^*} \pm d$ contains μ converges to $1-\alpha$. Notice that as k increases, the r.h.s. of the last inequality in (11) approaches d.

Now suppose that $Y_1, \ldots, Y_k$ are normally distributed. Starr (1966) showed that the stopping rule

$$k^* = \min\left[k : k \geq 3, k \text{ odd}, \delta(k,\alpha) \leq d\right]$$

yields

$$P(\overline{Y}_{k^*} - d \leq \mu \leq \overline{Y}_{k^*} + d) \geq \begin{cases} 0.928 & \text{if } \alpha = 0.05 \\ 0.985 & \text{if } \alpha = 0.01. \end{cases}$$

The last inequalities indicate little loss in the confidence level for arbitrary d. Based on Starr's result and (12), Fishman (1978b) recommended the simpler and more intuitive stopping rule

$$k^* = \min\left[k : k \geq 2, \delta(k,\alpha) \leq d\right].$$

An alternative *two-stage* approach for computing a confidence interval for μ with halfwidth at most d works as follows: The first stage uses k_0 replications to compute a variance estimate $S^2_{k_0}(Y)$ and a confidence interval with halfwidth $\delta(k_0,\alpha)$. Assume that the estimate $S^2_{k_0}(Y)$ does not change significantly as k_0 increases. If $\delta(k_0,\alpha) \leq d$, the procedure terminates. Otherwise, an estimate of the total number of replications required to obtain a halfwidth of at most d is computed from

$$\hat{k} = \min\left[k : k \geq k_0, t_{k-1,1-\alpha/2} S_{k_0}(Y)/\sqrt{k} \leq d\right].$$

The efficacy of this method depends on the closeness of $S^2_{k_0}(Y)$ to the unknown $\mathrm{Var}(Y_i)$. If $S^2_{k_0}(Y)$ underestimates $\mathrm{Var}(Y_i)$, then $\hat{k}$ will be smaller than actually needed. Conversely, if $S^2_{k_0}(Y)$ overestimates $\mathrm{Var}(Y_i)$, then unnecessary replications will have to be made.

Example 6 Table 1 summarizes the results of experiments that were run to estimate the mean number of customers that complete service during the first hour in an M/M/1 queueing system with arrival rate 0.9 per hour, service rate 1, and empty initial state. The sequential procedure was implemented with the stopping rule

$$k^* = \min\left[k : k \geq k_0, \delta(k,\alpha) \leq d\right]$$

and initial sample sizes $k_0 = 2, 3, 4, 5$. The two-stage procedure used initial samples of size 4, 5 and 10. For each experiment, 100 independent replications were run.

Based on Table 1, the sequential procedure with an initial sample of at least 5 replications appears to outperform the two-stage procedure. The advantages of the sequential procedure are: (a) The resulting confidence interval halfwidth is always less than or equal to the target value; (b) The variation in the final sample sizes and confidence interval halfwidths is substantially smaller.

An alternative problem is the computation of an estimate for μ with relative error $|\overline{Y}_k - \mu|/|\mu| \leq c$, where c is a positive constant. Formally, one requests

$$P(|\overline{Y}_k - \mu|/|\mu| \leq c) \geq 1 - \alpha.$$

Using some algebra, one can show

$$P\left(\frac{|\overline{Y}_k - \mu|}{|\mu|} \le c'\right) \ge P\left(\frac{|\overline{Y}_k - \mu|}{|\mu|} \le \frac{\delta(k,\alpha)}{|\overline{Y}_k|}\right),$$

where $c' = c/(1+c)$. Based on these observations, one can use the following stopping rule:

$$k^* = \min\left[k : k \ge k_0, \frac{\delta(k,\alpha)}{|\overline{Y}_k|} \le c'\right]. \tag{13}$$

Law, Kelton, and Koenig (1981) showed that when c is close to 0, the coverage of the confidence interval $\overline{Y}_k \pm \delta(k,\alpha)$ can be arbitrarily close to $1-\alpha$. They recommend that (13) be used with $c \le 0.15$ and $k_0 \ge 10$.

2.3 Quantile Estimation

The method of replications can also be used to implement nonparametric methods for estimating performance measures other than means. For example, suppose that we want to estimate the p-quantile ($0 < p < 1$), say ξ_p, of the maximum queue size Y in a single-server queueing system during a fixed time window. Let $F(y) = P(Y \le y)$ be the cumulative distribution function of Y. Then ξ_p is defined as

$$\xi_p = \inf[y : F(y) \ge p].$$

If the distribution F of Y is monotone increasing, then ξ_p is the unique solution to the equation $P(Y \le y) = p$. Let $Y_1, \ldots, Y_k$ be a random sample from F obtained by performing k independent replications, and let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(k)}$ be the order statistics corresponding to the Y_i's. Then a point estimator for ξ_p is

$$\hat{\xi}_p = \begin{cases} Y_{(kp)} & \text{if } kp \text{ is integer} \\ Y_{(\lfloor kp+1 \rfloor)} & \text{otherwise,} \end{cases}$$

where $\lfloor x \rfloor$ is the greatest integer that is less than or equal to x.

Now the event $Y_{(i)} < \xi_p < Y_{(j)}$ has the binomial probability

$$\begin{aligned} P(Y_{(i)} < \xi_p < Y_{(j)}) &= \sum_{\ell=i}^{j-1} \binom{k}{\ell} p^\ell (1-p)^{k-\ell} \\ &\approx \Phi\left(\frac{j-1-kp}{\sqrt{kp(1-p)}}\right) - \Phi\left(\frac{i-1-kp}{\sqrt{kp(1-p)}}\right), \end{aligned}$$

where the normal approximation is recommended for $kp \ge 5$ (see Hogg and Craig 1978, pp. 196–198). To compute a $1-\alpha$ confidence interval for ξ_p, one identifies indices $i < j$ such that $P(Y_{(i)} < \xi_p < Y_{(j)}) \ge 1-\alpha$. Then $(Y_{(i)}, Y_{(j)})$ is the required interval. Notice that several index pairs can satisfy the last inequality. Normally, one would choose a symmetric range of indices. In this case, the indices would be

$$i = \left\lfloor kp + 1 - \Phi^{-1}(1-\alpha/2)\sqrt{kp(1-p)} \right\rfloor$$

and

$$j = \left\lfloor kp + 1 + \Phi^{-1}(1-\alpha/2)\sqrt{kp(1-p)} \right\rfloor .$$

It should be noted that quantile estimation is more difficult than estimation of the mean because point estimates for quantiles are biased and significantly larger sample sizes are required to obtain reasonably tight confidence intervals. These problems are much more severe for more extreme quantiles, i.e., for p closer to 1. An introduction to nonparametric interval estimation methods is given in Hogg and Craig (1978, pp. 304–311).

3 STEADY-STATE ANALYSIS

Several methods have been developed for the estimation of steady-state system parameters. In this section we review these methods and provide the interested reader with an extensive list of references. We primarily consider the estimation of the steady-state mean μ of a discrete-time output process $\{X_i : i \geq 1\}$. Analogous methods for analyzing continuous-time output data are described in a variety of texts (Bratley, Fox, and Schrage 1987; Fishman 1978b; Law and Kelton 1991).

3.1 Removal of Initialization Bias

One of the hardest problems in steady-state simulations is the removal of the *initialization bias*. Let I be the set of initial conditions for the simulation model and assume that, as $n \to \infty$, $P(X_n \leq x|I) \to P(X \leq x)$, where X is the corresponding steady-state random variable. The steady-state mean of the process $\{X_i\}$ is $\mu = \lim_{n\to\infty} E(X_n|I)$. The problem with the use of the estimator $\overline{X}_n$ for a finite n is that $E(X_n|I) \neq \mu$ (and thus $E(\overline{X}_n|I) \neq \mu$).

The most commonly used method for reducing the bias of $\overline{X}_n$ involves identifying an index l, $1 \leq l \leq n-1$, and *truncating* the observations $X_1, \ldots, X_l$. Then the estimator

$$\overline{X}_{n,l} = \frac{1}{n-l} \sum_{i=l+1}^{n} X_i$$

is generally less biased than $\overline{X}_n$ because the initial conditions primarily affect data at the beginning of a run. Several procedures have been proposed for the detection of a cutoff index l (see Fishman 1972; Gafarian, Ancker, and Morisaku 1978; Kelton and Law 1983; Schruben 1982; Schruben, Singh, and Tierney 1983; Wilson and Pritsker 1978a,b). The procedure of Kelton (1989) uses a *pilot* run to estimate the steady-state distribution and starts a production run by sampling from the estimated distribution. More sophisticated truncation rules and initialization bias tests have recently been proposed by Chance and Schruben (1992), Goldsman, Schruben, and Swain (1994), and Ockerman (1995).

The graphical procedure of Welch (1981, 1983) is popular due to its generality and ease of implementation. Another graphical method has been proposed by Fishman (1978a,b) in conjunction with the batch means method (see Remark 1 in Section 3.4). Welch's method uses k independent replications with the ith replication producing observations $X_{i1}, X_{i2}, \ldots, X_{in}$ and computes the averages

$$\overline{X}_j = \frac{1}{k} \sum_{i=1}^{k} X_{ij}, \quad j = 1, \ldots, n. \tag{14}$$

Then for a given *time window* w, the procedure plots the moving averages

$$\overline{X}_j(w) = \begin{cases} \frac{1}{2w+1}\sum_{m=-w}^{w} \overline{X}_{j+m} & w+1 \le j \le n-w \\ \frac{1}{2j-1}\sum_{m=-j+1}^{j-1} \overline{X}_{j+m} & 1 \le j \le w \end{cases}$$

against j. For example, when $w = 2$,

$$\begin{aligned} \overline{X}_1(2) &= \overline{X}_1 \\ \overline{X}_2(2) &= \frac{1}{3}(\overline{X}_1 + \overline{X}_2 + \overline{X}_3) \\ \overline{X}_3(2) &= \frac{1}{5}(\overline{X}_1 + \overline{X}_2 + \overline{X}_3 + \overline{X}_4 + \overline{X}_5) \\ &\vdots \\ \overline{X}_{n-2}(2) &= \frac{1}{5}(\overline{X}_{n-4} + \overline{X}_{n-3} + \overline{X}_{n-2} + \overline{X}_{n-1} + \overline{X}_n). \end{aligned}$$

If the plot is reasonably smooth, then l is chosen to be the value of j beyond which the sequence of moving averages converges. Otherwise, a different time window is chosen and a new plot is drawn. The choice of w is similar to the choice of an interval width for a histogram. Since the truncation index is selected visually, the user will generally have to try several window sizes.

Example 7 (The M/M/1 queue revisited) Consider an M/M/1 queueing system with interarrival rate $\tau = 0.09$ and service rate $\omega = 0.1$. The limiting mean customer delay is $\mu = 90$. Assume that the system starts empty. 50 independent replications of the first 5000 delays were run by using equation (4). Figure 4 depicts the plot of the averages $\overline{D}_j$, $1 \le j \le 5000$, computed as in (14).

Figures 5 and 6 show the plots of the moving averages $\overline{D}_j(w)$, $1 \le w \le 5000 - w$, for window sizes $w = 100$ and 500. The transient period is long as the plots of $\overline{D}_j(w)$ first exceed μ for $j \approx 250$. Notice that a large window is required to get a reasonably smooth moving average plot for this system. In the absence of the horizontal line $\mu = 90$, one would hesitate to choose a truncation index $l \ge 2000$ as all $\overline{D}_j(500)$, $j \ge 2500$, are smaller than $\overline{D}_{2000}(500)$ giving the impression that the actual mean is less than 90. Similarly, the plot of $\overline{D}_j(100)$ is asymmetric with respect to μ with more observations smaller than 90. It should be noted that the method of Welch may be difficult to apply in congested systems with output time series having autocorrelation functions with very long tails.

3.2 The Replication/Deletion Approach

This approach runs k independent replications, each of length n observations, and uses the method of Welch (1981, 1983) or some other method to discard the first l observations from each run. One then uses the i.i.d. sample means

$$Y_i = \frac{1}{n-l}\sum_{j=l+1}^{n} X_{ij}$$

Table 1: Comparisons between sequential and two-stage confidence interval procedures

		Final Sample Size		Interval Halfwidth	
Procedure	Initial Sample Size	Mean	Standard Deviation	Mean	Standard Deviation
Sequential	2	91.7	35.0	0.945	0.198
Sequential	3	94.7	26.5	0.981	0.077
Sequential	4	99.0	16.8	0.996	0.005
Sequential	5	97.2	19.6	0.995	0.006
Two-stage	4	88.0	83.9	1.362	0.685
Two-stage	5	92.1	57.7	1.200	0.425
Two-stage	10	101.9	48.5	1.060	0.226

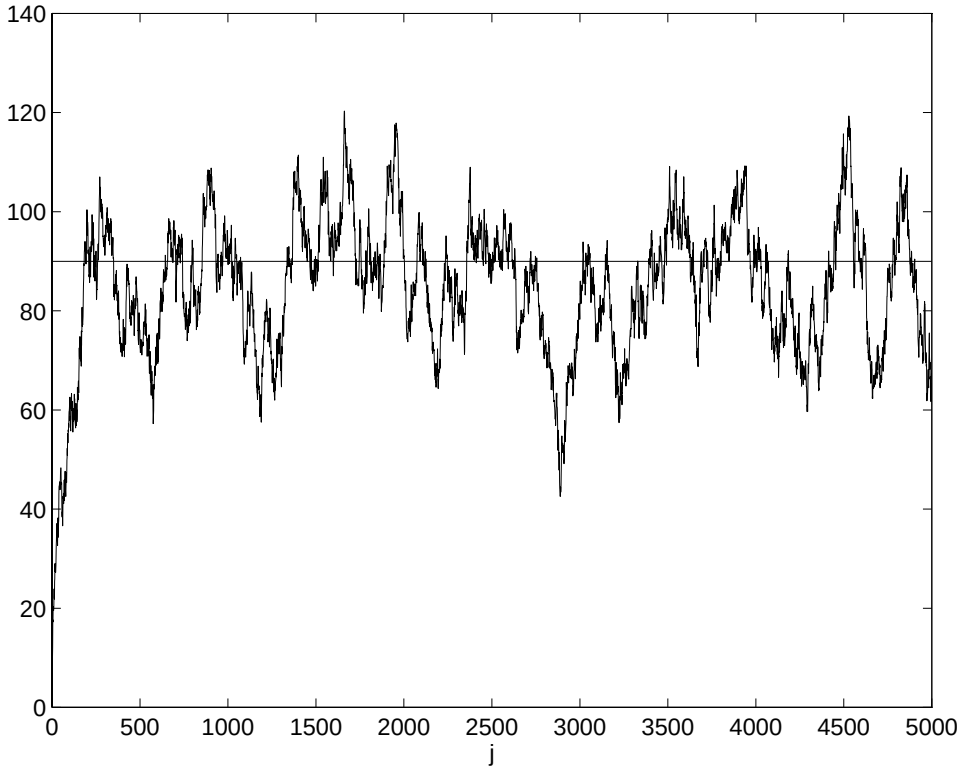

Figure 4: Average delay times $\overline{D}_j$ for the first 5000 customers in an M/M/1 queue from 50 independent replications

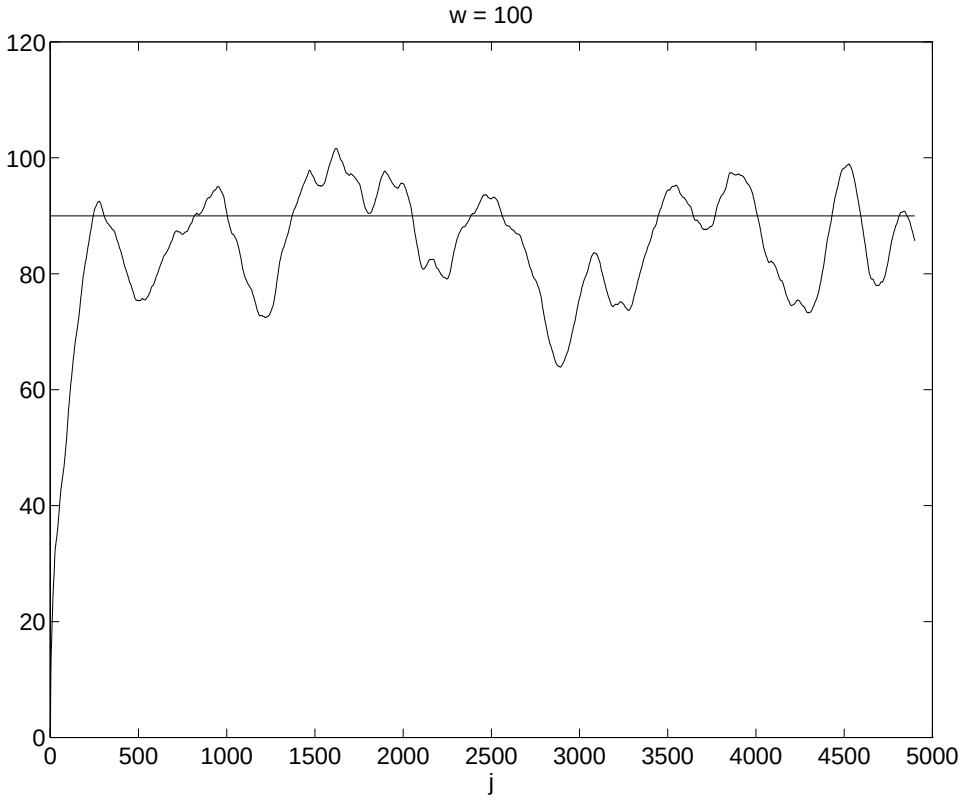

Figure 5: Moving averages with window $w = 100$

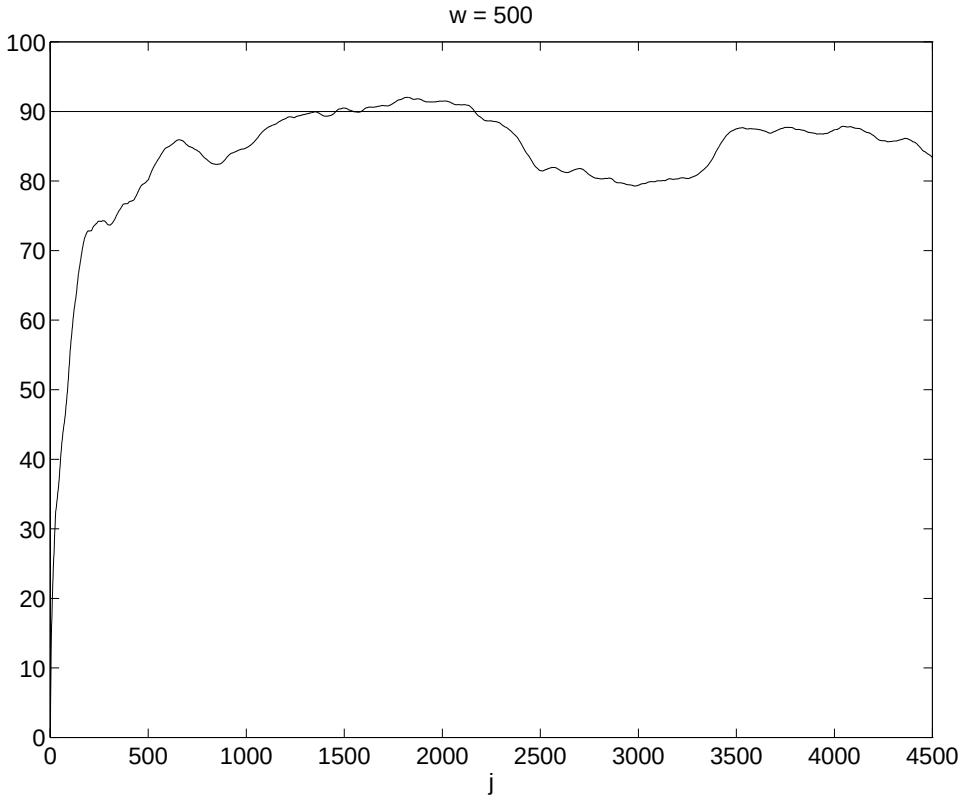

Figure 6: Moving averages with window $w = 500$

to compute point and interval estimators for the steady-state mean μ (see Section 2). The method is characterized by its simplicity and generality. The following list contains important observations about l, n and k.

1. As l increases for fixed n, the "systematic" error in each Y_i due to the initial conditions decreases. However, the sampling error increases because of the smaller number of observations (the variance of Y_i is proportional to $1/(n-l)$).
2. As n increases for fixed l, the systematic and sampling errors in Y_i decrease.
3. The systematic error in the sample means Y_i cannot be reduced by increasing the number of replications k.

Overall, one must be aware that the replication/deletion approach can require a substantial amount of effort to find a "good" truncation index l (as evidenced by Example 7) as well as a large sample size n and a large number of replications to obtain a confidence interval with the required coverage. This approach is also potentially wasteful of data as the truncated portion is removed from each replication. The regenerative method (Section 3.3) and the batch means method (Section 3.4) seek to overcome these disadvantages. The graph of the batch means (see Remark 1) provides an easy means to assess the effect of the initial conditions at a small incremental cost.

3.3 The Regenerative Method

This method assumes the identification of time indices at which the process $\{X_i\}$ probabilistically *starts over* and uses these regeneration epochs for obtaining i.i.d. random variables that can be used to compute point and interval estimates for the mean μ. As a result, it eliminates the need to detect the length of the initial transient period. The method was proposed by Crane and Iglehart (1974a, 1974b, 1975, 1978) and Fishman (1973, 1974). (For a complete treatment, see Crane and Lemoine 1977.)

More precisely, assume that there are (random) time indices $1 \leq T_1 < T_2 < \cdots$ such that the portion $\{X_{T_i+j}, j \geq 0\}$ has the same distribution for each i and is independent of the portion prior to time T_i. The portion of the process between two successive regeneration epochs is called a *cycle*. Let $Y_i = \sum_{j=T_i}^{T_{i+1}-1} X_j$ and $Z_i = T_{i+1} - T_i$ for $i = 1, 2, \ldots$ and assume that $E(Z_i) < \infty$. Then the mean μ is given by

$$\mu = \frac{E(Y_1)}{E(Z_1)}.$$

In addition, the long-run fraction of time the process spends in a set of states E is equal to

$$\lim_{n\to\infty} P(X_n \in E) = \frac{E(\text{total time the process } X \text{ spends in } E \text{ during a cycle})}{E(Z_1)}.$$

Example 8 (An (s, S) inventory system) The demand for an item on day i at a store is a nonnegative integer random variable L_i with positive mean. The store uses the following inventory management policy: If the inventory at the end of the day (after the demand is met) is at least s ($s \geq 0$), the store takes no action. If, however, the inventory is less than

s, the store orders enough to bring the inventory at the beginning of the next day to level S ($S > s$).

Assume that the inventory is replenished instantaneously (overnight) and let X_i be the level of inventory at the start of day i (before demands occur but after inventory replenishment). Then $\{X_i, i \geq 1\}$ is a regenerative process with return state S. If T_i is the day of the ith return to state S ($T_1 = 1$), the steady-state probability of stockout can be computed by

$$\lim_{n\to\infty} P(X_n < 0) = \frac{E(\text{total time stockout occurs during a cycle})}{E(Z_1)} = \frac{P\left(X_{T_2-1} < L_{T_2-1}\right)}{E(Z_1)}.$$

3.3.1 Estimation of the Mean

Suppose that one simulates the process $\{X_i\}$ over k cycles and collects the observations $Y_1, \ldots, Y_k$ and $Z_1, \ldots, Z_k$. Then

$$\hat{\mu} = \frac{\overline{Y}_k}{\overline{Z}_k}$$

is a strongly consistent, although typically biased for finite k, estimator of μ.

Confidence intervals for μ can be constructed by using the random variables $V_i = Y_i - \mu Z_i, i = 1, \ldots, k$, and the central limit theorem. Indeed, $E(V_i) = 0$ and

$$\sigma^2 = \text{Var}(V_i) = \text{Var}(Y_i) - 2\mu\text{Cov}(Y_i, Z_i) + \mu^2\text{Var}(Z_i).$$

By the central limit theorem $\overline{V}_k / \sqrt{\text{Var}(\overline{V}_k)}$ asymptotically has the standard normal distribution and for large k

$$P\left(\frac{\sqrt{k}|\overline{V}_k|}{\sigma} \leq z_{1-\alpha/2}\right) \approx 1 - \alpha\,.$$

The classical, and most commonly used, approach estimates σ^2 by

$$S_k^2(V) = S_k^2(Y) - 2\hat{\mu}S_k(Y, Z) + \hat{\mu}^2 S_k^2(Z),$$

where

$$S_k(Y, Z) = \frac{1}{k-1}\sum_{i=1}^{k}(Y_i - \overline{Y}_k)(Z_i - \overline{Z}_k)$$

is the sample covariance of Y_i and Z_i, to produce the approximate $1-\alpha$ confidence interval

$$\hat{\mu} \pm z_{1-\alpha/2}\frac{S_k(V)}{\overline{Z}_k\sqrt{k}}\,.$$

For a small sample size k, Iglehart (1975) showed that the approximate confidence interval

$$\hat{\mu}_J \pm z_{1-\alpha/2}\frac{S_J}{\sqrt{k}}\,,$$

where

$$\hat{\mu}_J = \frac{1}{k}\sum_{i=1}^{k}\theta_i\,,$$

$$\theta_i = k\frac{\overline{Y}_k}{\overline{Z}_k} - (k-1)\frac{\sum_{j\neq i} Y_j}{\sum_{j\neq i} Z_j},$$

is a *jackknife* estimator of μ (with smaller bias than $\hat{\mu}$) and

$$S_J^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\theta_i - \hat{\mu}_J)^2,$$

often provides better coverage than the classical regenerative confidence interval. However, its evaluation requires substantial bookkeeping in addition to $O(n^2)$ operations, making its use costly for large sample sizes. The jackknife method also generally increases the width of the confidence interval. A comprehensive review of the jackknife method is given by Efron (1982).

For small sample sizes and bounded Y_i and Z_i, one can also compute the confidence interval in Alexopoulos (1993) which provides superior coverage over confidence intervals based on the central limit theorem at the expense of increased width.

The regenerative method is difficult to apply in practice in simulations that have either no regenerative points or very long cycle lengths. Two classes of systems the regenerative method has successfully been applied to are inventory systems and highly reliable communications systems with repairs.

3.3.2 Quantile Estimation

Iglehart (1976), Moore (1980) and Seila (1982a,b) have proposed methods for computing confidence intervals for quantiles when the output process is regenerative. Seila's method will be presented because it is somewhat simpler to apply than the other methods and has been shown to produce intervals that are as reliable. The method will be illustrated by describing the estimation of the p-quantile, say ξ, of the steady-state customer delay distribution in an M/M/1 queueing system that starts empty. In this case, the regeneration epochs are the times the system returns to the empty state.

The method begins by simulating a total of $r \cdot m$ cycles and then partitions the data into r contiguous "batches" with m cycles per batch. For example, the ith batch contains the delay times from the cycles $(i-1)m+1, \ldots, im$. Denote the delay times in this cycle by $D_{i1}, \ldots, D_{iM_i}$ and notice that M_i is generally a random variable.

Now

$$\hat{\xi}_i = \begin{cases} D_{i,(M_i p)} & \text{if } M_i p \text{ is integer} \\ D_{i,(\lfloor M_i p+1 \rfloor)} & \text{otherwise} \end{cases}$$

is the quantile estimator for this batch and the overall estimator for q is

$$\overline{\xi} = \frac{1}{r}\sum_{i=1}^{r}\hat{\xi}_i.$$

To reduce the bias of the estimators $\hat{\xi}_i$, one can use the jackknife method by forming the estimators

$$\hat{\xi}_{J,i} = 2\hat{\xi}_i - \frac{\hat{\xi}_i^{(1)} + \hat{\xi}_i^{(2)}}{2},$$

where $\hat{\xi}_i^{(1)}$ (respectively, $\hat{\xi}_i^{(2)}$) is the sample quantile computed from the first (respectively, second) half of the m cycles in the ith batch. Then the overall jackknife estimator for ξ is

$$\overline{\xi}_J = \frac{1}{r}\sum_{i=1}^{r} \hat{\xi}_{J,i}$$

and, for large m and r,

$$\frac{\overline{\xi}_J - q}{S_r(\hat{\xi}_J)/\sqrt{r}} \approx t_{r-1},$$

where $S_r^2(\hat{\xi}_J)$ is the sample variance of the sample quantiles $\hat{\xi}_{J,i}$. The resulting approximate $1-\alpha$ confidence interval for ξ is

$$\overline{\xi}_J \pm t_{r-1,1-\alpha/2}\frac{S_r(\hat{\xi}_J)}{\sqrt{r}}\,. \tag{15}$$

Experiments in Seila (1982a,b) indicate that the confidence interval (15) has better coverage than the confidence interval resulting from the estimator $\overline{\xi}$.

The selection of m and r is an open research problem. Based on practical experience, m should be large enough so that $E(M_i) \geq 100$ and that $r \geq 10$. The mean $E(M_i)$ can be estimated by the sample mean $\overline{M}_r$ of the M_i's.

3.4 The Batch Means Method

The method of batch means is frequently used to estimate the steady-state mean μ or the $\mathrm{Var}(\overline{X}_n)$ and owes its popularity to its simplicity and effectiveness. Original accounts on the method were given by Conway (1963), Fishman (1978a,b), and Law and Carson (1979).

The classical approach divides the output $X_1, \ldots, X_n$ of a long simulation run into a number of contiguous *batches* and uses the sample means of these batches (or *batch means*) to produce point and interval estimators.

To motivate the method, assume temporarily that the process $\{X_i\}$ is weakly stationary with $\lim_{n\to\infty} n\mathrm{Var}(\overline{X}_n) < \infty$ and split the data into k batches, each consisting of b observations. (Assume $n = kb$.) The ith batch consists of the observations

$$X_{(i-1)b+1}, X_{(i-1)b+2}, \ldots, X_{ib}$$

for $i = 1, 2, \ldots, k$ and the ith batch mean is given by

$$Y_i(b) = \frac{1}{b}\sum_{j=1}^{b} X_{(i-1)b+j}.$$

For fixed m, let $\sigma_m^2 = \mathrm{Var}(\overline{X}_m)$. Since the batch means process $\{Y_i(b)\}$ is also weakly stationary, some algebra yields

$$\sigma_n^2 = \mathrm{Var}(\overline{X}_n) = \frac{\sigma_b^2}{k} + \frac{1}{k^2}\sum_{i\neq j}\mathrm{Cov}[Y_i(b), Y_j(b)] = \frac{\sigma_b^2}{k}\left(1 + \frac{n\sigma_n^2 - b\sigma_b^2}{b\sigma_b^2}\right).$$

Since $n \geq b$, $(n\sigma_n^2 - b\sigma_b^2)/(n\sigma_b^2) \to 0$ as first $n \to \infty$ and then $b \to \infty$. As a result, σ_b^2/k approximates σ_n^2 with error that diminishes as b and n approach infinity. Equivalently, the correlation among the batch means diminishes as b and n approach infinity.

To use the last limiting property, one forms the grand batch mean

$$\overline{Y}_k = \overline{X}_n = \frac{1}{k}\sum_{i=1}^{k} Y_i(b),$$

estimates the σ_b^2 by

$$\hat{V}_B(n,k) = \frac{1}{k-1}\sum_{i=1}^{k}(Y_i(b) - \overline{Y}_k)^2, \tag{16}$$

and computes the approximate $1-\alpha$ confidence interval for μ

$$\overline{Y}_k \pm t_{k-1,1-\alpha/2}\sqrt{\hat{V}_B(n,k)/k}\,. \tag{17}$$

The main problem with the application of the batch means method in practice is the choice of the batch size b. If b is small, the batch means $Y_i(b)$ can be highly correlated and the resulting confidence interval will frequently have coverage below the user-specified nominal coverage $1-\alpha$. Alternatively, a large batch size can result in very few batches and potential problems with the application of the central limit theorem to obtain (17).

The method of Fishman (1978) selects the smallest batch size from the set $\{1, 2, 4, \ldots, n/8\}$ that passes the test of independence based on von Neumann's statistic (see Section 3.4.3). A variant of this method was proposed by Schriber and Andrews (1979). Mechanic and McKay (1966) choose a batch size from the set $\{16b_1, 64b_1, 256b_1, \ldots, n/25\}$ (usually $b_1 = 1$) and select the batch size that passes an alternative test for independence. The procedure of Law and Carson (1979) starts with 400 batches of size 2. Then it considers sample sizes that double every two iterations until an estimate for lag-1 correlation among 400 batch means becomes smaller than 0.4 and larger than the estimated lag-1 correlation among 200 batch means. The procedure stops when the confidence interval (17) computed with 40 batches satisfies a relative width criterion. Schmeiser (1982) reviews the above procedures and concludes that selecting between 10 and 30 batches should suffice for most simulation experiments. The major drawback of these methods is their inability to yield a consistent variance estimator.

Remark 1 For fixed sample size, a plot of the batch means is a very useful tool for checking the effects of initial conditions, non-normality of batch means, and existence of correlation between batch means. For example, consider the M/M/1 queueing system in Example 7. A sample of 100,000 customer delays was generated by means of (4) starting with an empty system. Figure 7 shows the plot of the batch means $Y_1(2000), \ldots, Y_{50}(2000)$ for batch size $b = 2000$. The first batch mean is small but not the smallest, relaxing one's worries about the effect of the initial transient period. This also hints that $l = 2000$ is a reasonable truncation index for Welch's method. Had the first batch mean been smaller than the other batch means, one can assess the effect of the initial conditions by removing the first batch and comparing the new grand batch mean with the old. Although the plot does not indicate the presence of serious autocorrelation among the batch means, the asymmetric dispersion of the means about the actual mean should make the experimenter concerned about the coverage of the confidence interval (17).

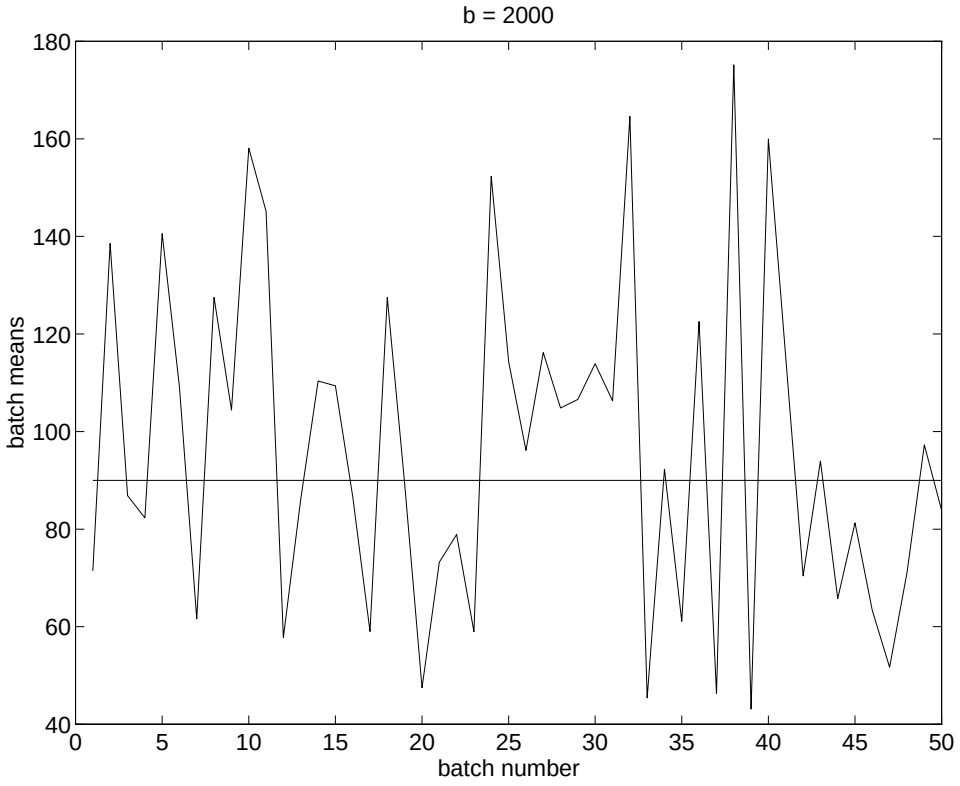

Figure 7: Batch means for delay times in an M/M/1 queue

Example 9 shows how an asymptotically optimal batch size can be obtained in special cases.

Example 9 For the AR(1) process in Example 4, Carlstein (1986) showed that

$$\text{Bias}[\hat{V}_B(n,k)] = -\frac{2\rho}{(1-\rho)^3(1+\rho)}\frac{1}{b} + o\left(\frac{1}{b}\right) \tag{18}$$

and

$$\text{Var}[\hat{V}_B(n,k)] = \frac{2}{(1-\rho)^4}\frac{b}{n} + o\left(\frac{b}{n}\right),$$

where $o(h)$ is a function such that $\lim_{h\to 0} o(h)/h = 0$. Then the batch size that minimizes the asymptotic (as $n \to \infty$ and $k \to \infty$) mean squared error $\text{MSE}[\hat{V}_B(n,k)] = \text{Bias}^2[\hat{V}_B(n,k)] + \text{Var}[\hat{V}_B(n,k)]$ is

$$b_0 = \left(\frac{2|\rho|}{1-\rho^2}\right)^{2/3} n^{1/3}. \tag{19}$$

Clearly, the optimal batch size increases with the absolute value of the correlation ρ between successive observations.

The reader should keep in mind that the optimal batch size may differ substantially from (19) for a finite sample size (e.g., Song and Schmeiser 1995), and the model generally does not apply to the analysis of data from queueing systems. Furthermore, it is not evident that this strategy for batch size selection allows the space and time complexities achievable by the LBATCH and ABATCH strategies in Sections 3.4.2–3.4.4 for generating an assessment of the stability of the variance of the sample mean.

3.4.1 Consistent Estimation Batch Means Methods

Consistent estimation batch means methods assume the existence of a parameter σ_∞^2 (the time-average variance of the process $\{X_i\}$), such that a central limit theorem holds

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \sigma_\infty N(0,1) \quad \text{as } n \to \infty \tag{20}$$

and aim at constructing a consistent estimator for σ_∞^2 and an asymptotically valid confidence interval for μ. [Notice that the X_i's in (20) need not be i.i.d.] Consistent estimation methods are often preferable to methods that "cancel" σ_∞^2 (see Glynn and Iglehart 1990) because: (a) The expectation and variance of the halfwidth of the confidence interval resulting from (20) is asymptotically smaller for consistent estimation methods; and (b) Under reasonable assumptions $n\text{Var}(\overline{X}_n) \to \sigma_\infty^2$ as $n \to \infty$.

Example 10 The delay process $\{D_i\}$ of a stationary M/M/1 system has

$$\sigma_\infty^2 = \frac{\nu}{\omega^2(1-\nu)^4}(\nu^3 - 4\nu^2 + 5\nu + 2)$$

(Blomqvist 1967) whereas a stationary AR(1) process has $\sigma_\infty^2 = (1+\rho)/(1-\rho)$.

Chien, Goldsman, and Melamed (1997) considered stationary processes and, under quite general moment and sample path conditions, showed that as both $b \to \infty$ and $k \to \infty$, $E[b\hat{V}_B(n,k)] \to \sigma_\infty^2$, $k\text{Var}[\hat{V}_B(n,k)] \to 2\sigma_\infty^2$, and $\text{MSE}[b\hat{V}_B(n,k)] \to 0$. Notice that the last limiting property of $b\hat{V}_B(n,k)$ differs from consistency.

The limiting result (20) is implied under the following two assumptions.

Assumption of Weak Approximation (AWA). There exist finite constants μ and $\sigma_\infty > 0$ such that

$$\frac{n(\overline{X}_n - \mu)}{\sigma_\infty} \xrightarrow{\mathcal{D}} W(n) \quad \text{as } n \to \infty.$$

Assumption of Strong Approximation (ASA). There exist finite constants μ, $\sigma_\infty > 0$, $\lambda \in (0, 1/2]$, and a finite random variable C such that, with probability one,

$$|n(\overline{X}_n - \mu) - \sigma_\infty W(n)| \le Cn^{1/2-\lambda} \quad \text{as } n \to \infty.$$

Both AWA and ASA state that the process $\{n(\overline{X}_n - \mu)/\sigma_\infty\}$ is close to a standard Brownian motion. However the stronger ASA addresses the convergence rate of (20).

The ASA is not restrictive as it holds under relatively weak assumptions for a variety of stochastic processes including Markov chains, regenerative processes and certain queueing systems (see Damerdji 1994a for details). The constant λ is closer to 1/2 for processes having little autocorrelation while it is closer to zero for processes with high autocorrelation. In the former case the "distance" between the processes $\{n(\overline{X}_n - \mu)/\sigma_\infty\}$ and $\{W(n)\}$ "does not grow" as n increases.

3.4.2 Batching Rules

Fishman and Yarberry (1997) and Fishman (1996, Chapter 6) present a thorough discussion of batching rules. Our account in this section and in Section 3.4.4 parallels their development. Both references contain detailed instructions for obtaining FORTRAN, C, and SIMSCRIPT II.5 implementations for various platforms via anonymous ftp from the site ftp.or.unc.edu.

The discussion prior to the derivation of (17) suggests that fixing the number of batches and letting the batch size grow as $n \to \infty$ ensures that $\sigma_b^2/k \to \sigma_n^2$. This motivates the following rule.

The Fixed Number of Batches (FNB) Rule. Fix the number of batches at k. For sample size n, use batch size $b_n = \lfloor n/k \rfloor$.

The FNB rule along with AWA lead to the following result.

Theorem 1 (Glynn and Iglehart 1990) If $\{X_i\}$ satisfies AWA, then as $n \to \infty$, $\overline{X}_n \xrightarrow{\mathcal{P}} \mu$ and (20) holds. Furthermore, if k is constant and $\{b_n, n \ge 1\}$ is a sequence of batch sizes such that $b_n \to \infty$ as $n \to \infty$, then

$$\frac{\overline{X}_n - \mu}{\sqrt{\hat{V}_B(n,k)/k}} \xrightarrow{\mathcal{D}} t_{k-1} \quad \text{as } n \to \infty.$$

∎

The primary implication of Theorem 1 is that (17) is an asymptotically valid confidence interval for μ. Unfortunately, the FNB rule has two major limitations: (a) $b_n\hat{V}_B(n,k)$ is not a consistent estimator of σ_∞^2. Therefore the confidence interval (17) tends to be wider than the interval a consistent estimation method would produce; (b) After some algebra (see Fishman 1996, Chapter 6) one has

$$\lim_{n\to\infty}\frac{\mathrm{Var}\left[\sqrt{\hat{V}_B(n,k)/k}\right]}{\mathrm{Var}(\overline{X}_n)} = \frac{1}{2(k-1)} - \frac{1}{8(k-1)^2} - \frac{1}{16(k-1)^3} - \cdots$$

so that statistical fluctuations in the halfwidth of the confidence interval (17) do not diminish relative to statistical fluctuation in the sample mean.

The following theorem proposes batching assumptions which along with ASA yield a strongly consistent estimator for σ_∞^2.

Theorem 2 (Damerdji 1994a) If $\{X_i\}$ satisfies ASA, then $\overline{X}_n \xrightarrow{a.s.} \mu$ as $n\to\infty$. Furthermore suppose that $\{(b_n,k_n), n\geq 1\}$ is a batching sequence satisfying

(A.1) $b_n\to\infty$ and $k_n\to\infty$ monotonically as $n\to\infty$

(A.2) $b_n^{-1}n^{1-2\lambda}\ln n\to 0$ as $n\to\infty$

(A.3) there exists a finite positive integer m such that

$$\sum_{n=1}^{\infty}(b_n/n)^m < \infty.$$

Then, as $n\to\infty$,

$$b_n\hat{V}_B(n,k_n) \xrightarrow{a.s.} \sigma_\infty^2 \tag{21}$$

and

$$Z_{k_n} = \frac{\overline{X}_n - \mu}{\sqrt{\hat{V}_B(n,k_n)/k_n}} \xrightarrow{\mathcal{D}} N(0,1). \tag{22}$$

∎

The last display implies that

$$\overline{X}_n \pm t_{k_n-1,1-\alpha/2}\sqrt{\hat{V}_B(n,k_n)/k_n}$$

is an asymptotically valid $1-\alpha$ confidence interval for μ.

Theorem 2 motivates the consideration of batch sizes of the form $b_n = \lfloor n^\theta\rfloor$, $0<\theta<1$. In this case one can show that the conditions (A.1)–(A.3) are met if $\theta\in(1-2\lambda,1)$. In particular, the assignment $\theta=1/2$ and the SQRT rule below are valid if $1/4<\lambda<1/2$. Notice that the last inequality is violated by processes having high autocorrelation ($\lambda\approx 0$).

The Square Root (SQRT) Rule. For sample size n, use batch size $b_n = \lfloor\sqrt{n}\rfloor$ and number of batches $k_n = \lfloor\sqrt{n}\rfloor$.

Under some additional moment conditions, Chien (1989) showed that the convergence of Z_{k_n} to the $N(0,1)$ distribution is fastest if b_n and k_n grow proportionally to $\sqrt{n}$. Unfortunately, in practice the SQRT rule tends to seriously underestimate the $\mathrm{Var}(\overline{X}_n)$ for fixed n.

Example 11 (The M/M/1 queue revisited) Consider an M/M/1 queueing system with interarrival rate $\tau = 0.9$ and service rate $\omega = 1$, and assume that the system starts empty. Table 2 contains performance statistics for 0.95 confidence intervals for the steady-state mean customer delay $\mu = 0.9/[1 \times (1 - 0.9)] = 9$. The confidence intervals resulted from 500 independent replications. Within each replication, the delays were generated by means of (4). The FNB rule used 16 batches and batch sizes 2^m, $m \geq 0$. The SQRT rule started with batch size $b_1 = 1$ and number of batches $k_1 = 8$, and computed confidence intervals with batch sizes

$$b_l = 2^{(l-1)/2} \times \begin{cases} b_1 & \text{if } l \text{ is odd} \\ 3/(2\sqrt{2}) & \text{otherwise} \end{cases}$$

and numbers of batches

$$k_l = 2^{(l-1)/2} \times \begin{cases} k_1 & \text{if } l \text{ is odd} \\ 11/\sqrt{2} & \text{otherwise.} \end{cases}$$

The resulting sample sizes $n_l = k_l b_l$ are roughly powers of 2 (see Section 3.4.4 for details).

Columns 2 and 4 contain the estimated coverage probabilities of the confidence intervals produced by the FNB rule and the SQRT rule, respectively. Columns 3 and 5 display the respective average interval halfwidths. Specifically, for sample size $n \approx 2^{17} = 131{,}072$, roughly 94 percent of the confidence intervals resulting from the FNB rule contained μ whereas only 78 percent of the confidence intervals resulting from the SQRT rule contained μ. However, the latter intervals were 43 percent narrower. Experiments by Fishman and Yarberry showed that the disparity in coverage between the two rules grows with increasing traffic intensity $\nu = \tau/\omega$.

Table 2: Performance statistics for the FNB and SQRT rules on 0.95 confidence intervals for the mean customer delay in an M/M/1 queue with utilization $\nu = 0.9$

	FNB Rule		SQRT Rule	
$\log_2 n$	Coverage	Average Halfwidth	Coverage	Average Halfwidth
10	0.544	3.244	0.326	1.694
11	0.640	3.506	0.366	1.665
12	0.746	3.304	0.414	1.437
13	0.798	2.963	0.466	1.271
14	0.838	2.435	0.498	1.063
15	0.880	1.901	0.604	0.904
16	0.912	1.437	0.664	0.738
17	0.944	1.053	0.778	0.599
18	0.934	0.756	0.810	0.471
19	0.950	0.541	0.854	0.369
20	0.940	0.385	0.858	0.283

With the contrasts between the FNB and SQRT rules in mind, Fishman and Yarberry proposed two strategies that dynamically shift between the two rules. Both strategies perform "interim reviews" and compute confidence intervals at times $n_l \approx n_1 2^{l-1}, l = 1, 2, \ldots$.

The LBATCH Strategy. At time n_l, if a hypothesis test detects autocorrelation between the batch means, the batching for the next review is determined by the FNB rule. If the test fails to detect correlation, all future reviews omit the test and employ the SQRT rule.

The ABATCH Strategy. If at time n_l the hypothesis test detects correlation between the batch means, the next review employs the FNB rule. If the test fails to detect correlation, the next review employs the SQRT rule.

Both strategies LBATCH and ABATCH yield random sequences of batch sizes. Under relatively mild assumptions, these sequences imply convergence results analogous to (21) and (22) (see Fishman and Yarberry 1997 and Fishman 1996).

3.4.3 Test for Correlation

This subsection reviews a test for the hypothesis

$$H_0 : \text{the batch means } Y_1(b), \ldots, Y_k(b) \text{ are uncorrelated.}$$

The test is due to von Neumann (1941) and is very effective when the number of batches k is small.

Assume that the process $\{X_i\}$ is weakly stationary and let

$$\rho_l(b) = \text{Corr}[Y_i(b), Y_{i+l}(b)], \quad l = 0, 1, \ldots$$

be the autocorrelation function of the batch means process. The von Neumann test statistic for H_0 is

$$\Gamma_k = \sqrt{\frac{k^2 - 1}{k - 2}} \left[\hat{\rho}_1(b) + \frac{(Y_1(b) - \overline{X}_n)^2 + (Y_k(b) - \overline{X}_n)^2}{2 \sum_{i=1}^{k} (Y_i(b) - \overline{X}_n)^2} \right], \tag{23}$$

where

$$\hat{\rho}_1(b) = \frac{\sum_{i=1}^{k-1} (Y_i(b) - \overline{X}_n)(Y_{i+1}(b) - \overline{X}_n)}{\sum_{i=1}^{k} (Y_i(b) - \overline{X}_n)^2}$$

is an estimator for the lag-1 autocorrelation $\rho_1(b)$. The rightmost ratio in equation (23) carries more weight when k is small but it approaches zero as $k \to \infty$.

Suppose that H_0 is true. If the batch means are i.i.d. normal, then the distribution of Γ_k is very close to $N(0, 1)$ for as few as $k = 8$ batches (von Neumann 1941; Young 1941, Table 1). On the other hand, if the batch means are i.i.d. but non-normal, the first four cumulants of Γ_k converge to the respective cumulants of the $N(0, 1)$ distribution as $k \to \infty$. This discussion suggests the approximation

$$\Gamma_k \approx N(0, 1)$$

for large b (the batch means become approximately normal) or large k (by the central limit theorem).

If $\{X_i\}$ has a monotone decreasing autocorrelation function (e.g., the delay process for an M/M/1 queueing system), the batch means process also has a monotone decreasing autocorrelation function. As a result, one rejects H_0 at level β if

$$\Gamma_k > z_{1-\beta}.$$

Alternatively, if $\{X_i\}$ has an autocorrelation function with damped harmonic behavior around the zero axis (e.g., the AR(1) process with $\rho < 0$), the test can lead to erroneous conclusions. In this case, repeated testing under the ABATCH strategy reduces this possibility.

The p–value of the test, $1 - \Phi(\Gamma_k)$, is the largest value of the type I error $\beta = P(\text{reject } H_0 \mid H_0 \text{ is true})$ for which H_0 is rejected given the observed value of Γ_k. Equivalently, H_0 is accepted if the p–value is larger than β. Hence, a p–value close to zero implies low credibility for H_0. The plot of the p–value versus the batch size is a useful graphical device.

3.4.4 Implementing the ABATCH Strategy

This subsection presents a pseudo-code for implementing the ABATCH strategy. The implementation of the LBATCH strategy is discussed in short after the pseudo-code.

To understand the role of the hypothesis test in the LBATCH and ABATCH algorithms, define the random variables

$$R_l = \begin{cases} 1 & \text{if } H_0 \text{ is rejected on review } l \\ 0 & \text{otherwise} \end{cases}$$

and

$$\begin{aligned} \overline{R}_l &= (R_1 + \cdots + R_l)/l \\ &= \text{fraction of rejected tests for } H_0 \text{ on reviews } 1, \ldots, l. \end{aligned}$$

A sufficient condition for strong consistency (equation 21) and asymptotic normality (equation 22) is $\beta_0 > 1 - 4\lambda$ (or $\lambda > (1 - \beta_0)/4$), where $\beta_0 = \lim_{l\to\infty} \overline{R}_l$ is the long-run fraction of rejections. In practice, β_0 differs from but is expected to be close to the type I error β. Clearly, $\lambda > 1/4$ guarantees (21) and (22) regardless of β_0. However, β_0 plays a role when $\lambda \leq 1/4$. Specifically, for β_0 equal to 0.05 or 0.10, the lower bound $(1 - \beta_0)/4$ becomes 0.2375 or 0.2225, respectively, a small reduction from 1/4.

On review l, the ABATCH strategy induces batch size

$$b_l = 2^{(l-1)(1+\overline{R}_{l-1})/2} \times \begin{cases} b_1 & \text{if } (l-1)(1+\overline{R}_{l-1}) \text{ is even} \\ \tilde{b}_1/\sqrt{2} & \text{otherwise,} \end{cases}$$

where

$$\tilde{b}_1 = \begin{cases} 3/2 & \text{if } b_1 = 1 \\ \lfloor \sqrt{2}\, b_1 + 0.5 \rfloor & \text{if } b_1 > 1, \end{cases}$$

and number of batches

$$k_l = 2^{(l-1)(1-\overline{R}_{l-1})/2} \times \begin{cases} k_1 & \text{if } (l-1)(1-\overline{R}_{l-1}) \text{ is even} \\ \tilde{k}_1/\sqrt{2} & \text{otherwise,} \end{cases}$$

where $\tilde{k}_1 = \lfloor \sqrt{2}\, k_1 + 0.5 \rfloor$.

The resulting sample sizes are

$$n_l = k_l b_l = \begin{cases} 2^{l-1} k_1 b_1 & \text{if } (l-1)(1+\overline{R}_{l-1}) \text{ is even} \\ 2^{l-2} \tilde{k}_1 \tilde{b}_1 & \text{otherwise} \end{cases}$$

and the definitions for $\tilde{b}_1$ and $\tilde{k}_1$ guarantee that if H_0 is never rejected, then both b_l and k_l grow approximately as $\sqrt{2}$ with l (i.e., they follow the SQRT rule).

Suppose $L+1$ reviews are performed. The final implementation issue for the ABATCH strategy is the relative difference between the potential terminal sample sizes

$$\Delta(b_1, k_1) = \frac{|2^L k_1 b_1 - 2^{L-1} \tilde{k}_1 \tilde{b}_1|}{2^L k_1 b_1} = \frac{|2 k_1 b_1 - \tilde{k}_1 \tilde{b}_1|}{2 k_1 b_1}.$$

This quantity is minimized (i.e., the final sample size is deterministic) when $2k_1 b_1 = \tilde{k}_1 \tilde{b}_1$. Pairs (b_1, k_1), with small b_1, satisfying the last equality are $(1,3)$, $(1,6)$, $(2,3)$, and $(2,6)$. Unfortunately, the condition $2k_1 b_1 = \tilde{k}_1 \tilde{b}_1$ excludes several practical choices for b_1 and k_1, such as $b_1 = 1$ (to test the original sample for independence) and $8 \le k_1 \le 10^5$. Overall, $\Delta(b_1, k_1)$ remains small for numerous choices of b_1 and k_1. For instance, $b_1 = 1$ and $8 \le k_1 \le 32$ ensure $\Delta(b_1, k_1) \le 0.078$.

Algorithm ABATCH
Source: Fishman and Yarberry (1997) and Fishman (1996, Chapter 6). Minor notational changes have been made.
Input: Minimal number of batches k_1, minimal batch size b_1, desired sample size $n = 2^L k_1 b_1$ (L is a positive integer), and confidence level $1 - \alpha$.
Output: Sequences of point estimates and confidence intervals for sample sizes $N \le n$.
Method:

1. $b \leftarrow b_1$ and $k \leftarrow k_1$.
2. If $b_1 = 1$, $\tilde{b}_1 \leftarrow 3/2$; otherwise $\tilde{b}_1 \leftarrow \lfloor \sqrt{2}\, b_1 + 0.5 \rfloor$.
3. $\tilde{k}_1 \leftarrow \lfloor \sqrt{2}\, k_1 + 0.5 \rfloor$.
4. $g \leftarrow \tilde{b}_1 / b_1$ and $f \leftarrow \tilde{k}_1 / k_1$.
5. $i \leftarrow 0$.
6. $\tilde{n} \leftarrow 2^{L-1} \tilde{k}_1 \tilde{b}_1$.
   Until $N = n$ or $N = \tilde{n}$:
7. $N \leftarrow kb$.
8. Randomly generate $X_{i+1}, \ldots, X_N$.
   Compute:
9. The batch means $Y_1(b), \ldots, Y_k(b)$.
10. $\overline{X}_N$ as a point estimate of μ.
11. The sample variance $\hat{V}_B$ of the batch means.
12. The halfwidth $\delta = t_{k-1,1-\alpha/2} \sqrt{\hat{V}_B / k}$ of the confidence interval (17).
13. Print N, k, b, $\overline{X}_N$, $\overline{X}_N - \delta$, $\overline{X}_N + \delta$, $\hat{V}_B$.
14. $i \leftarrow N$.

15. Test $H_0 : Y_1(b), \ldots, Y_k(b)$ are uncorrelated. Print the p-value of this test.
16. If H_0 is rejected, $b \leftarrow 2b$. (FNB rule)
    If H_0 is accepted:
17. If $b = 1$, $b \leftarrow 2$. (FNB rule)
    Otherwise: (SQRT rule)
18. $b \leftarrow bg$ and $k \leftarrow kf$.
19. If $g = \tilde{b}_1/b_1$, $g \leftarrow 2b_1/\tilde{b}_1$ and $f \leftarrow 2k_1/\tilde{k}_1$; otherwise $g \leftarrow \tilde{b}_1/b_1$ and $f \leftarrow \tilde{k}_1/k_1$.

Remark 2 Algorithm ABATCH requires $O(n)$ time and $O(\log_2 n)$ space. For details, see Chapter 5 of Yarberry (1993).

Remark 3 The implementation of strategy LBATCH is simpler. Once H_0 is accepted in step 15, the steps 17–19 are ignored for the remainder of the execution.

3.4.5 Tests for the Batching Rules

The experiments in Examples 12–14 compare the FNB rule and the LBATCH and ABATCH strategies by means of three queueing systems with traffic intensity $\nu = 0.9$. Each system starts empty and has a first-come, first-served discipline. Each experiment computed 0.95 confidence intervals for the long-run mean customer delay from 500 independent replications. The FNB rule relied on 16 batches whereas the LBATCH and ABATCH strategies started with $k_1 = 8$ batches of size $b_1 = 1$ and used type I error $\beta = 0.1$ for H_0.

Example 12 (Example 11 continued) The entries of Tables 2 and 3 indicate that the ABATCH strategy comes closer to the FNB rule's superior coverage with shorter confidence intervals.

Example 13 (An M/G/1 queue) Consider an M/G/1 queueing system with i.i.d. interarrival times from the exponential distribution with parameter $\tau = 0.9$ and i.i.d. service times S_i from the hyperexponential distribution with density function

$$f(x) = 0.9\left(\frac{1}{0.5}e^{-x/0.5}\right) + 0.1\left(\frac{1}{5.5}e^{-x/5.5}\right), \quad x \geq 0.$$

This distribution applies when customers are classified into two types, 1 and 2, with respective probabilities 0.9 and 0.1; type 1 customers have exponential service times with mean 0.5, and type 2 customers have exponential service times with mean 5.5. The service times have mean $E(S) = 0.9(0.5) + 0.1(5.5) = 1$, second moment $E(S^2) = 0.9 \times 2(0.5^2) + 0.1 \times 2(5.5^2) = 6.5$, and coefficient of variation

$$\frac{\sqrt{\mathrm{Var}(S)}}{E(S)} = 2.739,$$

which is larger than 1, the coefficient of variation of the exponential distribution. Then the traffic intensity is $\nu = \tau E(S) = 0.9$.

The long-run mean delay time in queue is given by the Pollaczek-Khintchine formula (Ross 1993, Chapter 8)

$$\mu = \lim_{i \to \infty} E(D_i) = \frac{\tau E(S^2)}{2(1-\nu)} = 29.25. \tag{24}$$

Notice that the M/M/1 system in Example 11 with the same arrival rate and traffic intensity has much smaller long-run mean delay time.

Table 4 displays the results of this experiment. As n increases, the conservative ABATCH strategy produces 0.95 confidence intervals for μ that are roughly 50 to 100 percent wider than the respective confidence intervals produced by the LBATCH strategy but have coverage rates that are acceptably close to 0.95 for substantially smaller sample sizes (as small as $2^{17} = 131{,}072$).

Example 14 (An M/D/1 queue) Consider an M/G/1 queueing system with i.i.d. interarrival times from the exponential distribution with parameter $\tau = 0.9$ and fixed unit service times. Then, by (24), the long-run mean delay time in queue is $\mu = 4.5$.

The results of this experiment are contained in Table 5. As in Examples 12 and 13, the performance of the ABATCH strategy makes it an attractive compromise between the "extreme" FNB and SQRT rules.

Example 15 tests the FNB, LBATCH and ABATCH methods by means of an AR(1) process.

Example 15 (The AR(1) process revisited) Consider the stationary AR(1) process $X_i = -0.9X_{i-1} + Z_i$ with mean 0 (see Example 4). The autocorrelation function $\rho_j = (-0.9)^j$, $j \geq 0$, of this process oscillates around the zero axis and the time-average process variance is $\sigma_\infty^2 = (1 - 0.9)/(1 + 0.9) = 0.053$.

The entries of Table 6 were obtained from 500 independent replications. The FNB rule used 16 batches and the type I error for H_0 was $\beta = 0.1$. The 0.95 confidence intervals for μ produced by the three methods have roughly equal halfwidths and coverages. In fact, almost all coverage estimates are greater than the nominal coverage 0.95. This behavior is due to the fact that $b\text{Var}(\hat{V}_B(n,k))$ tends to overestimate σ_∞^2 (the coefficient of $1/b$ in equation (18) is $2.624 > 0$).

From equation (19), the batch size that minimizes $\text{MSE}(\hat{V}_B(n,k))$ is $b_0 = 113.71$. 500 independent replications with 144 batches of size 114 (sample size 16416) produced 0.95 confidence intervals with estimated coverage 0.958 and average halfwidth 0.0016 — not a substantial improvement over the statistics in the last row of Table 6 (for sample size roughly equal to $2^{14} = 16384$).

Based on Examples 12–14 (Tables 3–5), the ABATCH strategy appears to provide approximately 10 percent reduction in confidence interval width for sample sizes large enough to achieve the nominal coverage probability.

3.4.6 Overlapping Batch Means

An interesting variation of the traditional batch means method is the method of *overlapping* batch means (OBM) proposed by Meketon and Schmeiser (1984). For given batch size b,

Table 3: Performance statistics for the LBATCH and ABATCH strategies on 0.95 confidence intervals for the mean customer delay in an M/M/1 queue with utilization $\nu = 0.9$

	LBATCH Strategy		ABATCH Strategy		
$\log_2 n$	Coverage	Average Halfwidth	Rejection Proportion	Coverage	Average Halfwidth
10	0.398	2.085	0.622	0.562	3.384
11	0.420	1.992	0.552	0.632	3.450
12	0.464	1.693	0.458	0.712	3.100
13	0.518	1.477	0.394	0.760	2.686
14	0.562	1.227	0.340	0.816	2.168
15	0.652	1.029	0.266	0.850	1.708
16	0.714	0.834	0.206	0.902	1.296
17	0.808	0.663	0.200	0.932	0.955
18	0.852	0.513	0.176	0.938	0.688
19	0.866	0.395	0.154	0.930	0.493
20	0.876	0.298	0.156	0.936	0.353

Table 4: Performance statistics for the FNB, LBATCH and ABATCH algorithms on 0.95 confidence intervals for the mean customer delay in an M/G/1 queue with hyperexponential service times and utilization $\nu = 0.9$

	FNB Rule		LBATCH Strategy		ABATCH Strategy		
$\log_2 n$	Coverage	Average Halfwidth	Coverage	Average Halfwidth	Rejection Proportion	Coverage	Average Halfwidth
10	0.324	9.251	0.204	5.865	0.742	0.356	10.305
11	0.420	10.756	0.254	5.962	0.724	0.436	11.426
12	0.560	11.482	0.294	5.552	0.614	0.566	11.635
13	0.688	11.711	0.354	5.083	0.584	0.652	11.166
14	0.774	11.021	0.392	4.418	0.502	0.746	10.147
15	0.832	9.715	0.452	3.863	0.386	0.794	8.658
16	0.886	8.053	0.540	3.215	0.344	0.856	7.057
17	0.908	6.208	0.620	2.678	0.300	0.898	5.483
18	0.900	4.542	0.632	2.178	0.202	0.896	4.090
19	0.914	3.262	0.694	1.761	0.148	0.900	2.997
20	0.934	2.339	0.748	1.387	0.134	0.924	2.145
21	0.942	1.669	0.806	1.083	0.118	0.926	1.525

Table 5: Performance statistics for the FNB, LBATCH and ABATCH algorithms on 0.95 confidence intervals for the mean customer delay in an M/D/1 queue with unit service times and utilization $\nu = 0.9$

	FNB Rule		LBATCH Strategy		ABATCH Strategy		
$\log_2 n$	Coverage	Average Halfwidth	Coverage	Average Halfwidth	Rejection Proportion	Coverage	Average Halfwidth
10	0.618	1.626	0.460	1.062	0.552	0.616	1.631
11	0.740	1.620	0.548	0.962	0.504	0.720	1.538
12	0.826	1.529	0.598	0.842	0.396	0.788	1.391
13	0.858	1.222	0.648	0.686	0.334	0.842	1.101
14	0.878	0.946	0.696	0.556	0.274	0.858	0.852
15	0.908	0.707	0.794	0.445	0.220	0.884	0.632
16	0.920	0.517	0.808	0.351	0.222	0.924	0.472
17	0.946	0.375	0.862	0.271	0.140	0.942	0.343

Table 6: Performance statistics for the FNB, LBATCH and ABATCH algorithms on 0.95 confidence intervals for the mean $\mu = 0$ of the stationary AR(1) process $X_i = -0.9X_{i-1} + Z_i$

	FNB Rule		LBATCH Strategy		ABATCH Strategy		
$\log_2 n$	Coverage	Average Halfwidth	Coverage	Average Halfwidth	Rejection Proportion	Coverage	Average Halfwidth
5	0.954	0.0515	1.000	0.1217	0.020	1.000	0.1153
6	0.978	0.0364	0.978	0.0364	0.102	0.980	0.0367
7	0.980	0.0248	0.980	0.0244	0.060	0.980	0.0246
8	0.974	0.0164	0.982	0.0166	0.038	0.980	0.0167
9	0.960	0.0108	0.972	0.0111	0.018	0.966	0.0111
10	0.966	0.0071	0.984	0.0076	0.022	0.980	0.0075
11	0.944	0.0048	0.976	0.0051	0.012	0.978	0.0050
12	0.962	0.0034	0.982	0.0035	0.014	0.984	0.0035
13	0.938	0.0023	0.964	0.0024	0.012	0.962	0.0023
14	0.942	0.0017	0.960	0.0016	0.020	0.962	0.0016

this method uses all $n - b + 1$ overlapping batches to estimate μ and $\text{Var}(\overline{X}_n)$. The first batch consists of observations $X_1, \ldots, X_b$, the second batch consists of $X_2, \ldots, X_{b+1}$, etc. The OBM estimator of μ is

$$\overline{Y}_O = \frac{1}{n-b+1} \sum_{i=1}^{n-b+1} Y_i(b),$$

where

$$Y_i(b) = \frac{1}{b} \sum_{j=i}^{i+b-1} X_j, \quad i = 1, \ldots, n-b+1$$

are the respective batch means, and has sample variance

$$\hat{V}_O = \frac{1}{n-b} \sum_{i=1}^{n-b+1} (Y_i(b) - \overline{Y}_O)^2.$$

The following list contains properties of the estimators $\overline{Y}_O$ and $\hat{V}_O$:

(i) The OBM estimator is a weighted average of non-overlapping batch means estimators.

(ii) Asymptotically (as $n, b \to \infty$ and $b/n \to 0$), the OBM variance estimator $\hat{V}_O$ and the non-overlapping batch means variance estimator $\hat{V}_B \equiv \hat{V}_B(n,k)$ have the same expectation. Furthermore,

$$\frac{\text{Var}(\hat{V}_O)}{\text{Var}(\hat{V}_B)} \to \frac{2}{3}.$$

In words, the asymptotic ratio of the mean squared error of $\text{Var}(\hat{V}_O)$ to the mean squared error of $\text{Var}(\hat{V}_B)$ is equal to 2/3 (Meketon and Schmeiser 1984).

(iii) The behavior of $\text{Var}(\hat{V}_O)$ appears to be less sensitive to the choice of the batch size than the behavior of $\text{Var}(\hat{V}_B)$ (Song and Schmeiser 1995, Table 1).

(iv) If $\{X_i\}$ satisfies ASA and $\{(b_n, k_n), n \geq 1\}$ is a sequence that satisfies the assumptions (A.1)–(A.3) in Theorem 2 and

$$\frac{b_n^2}{n} \to 0 \quad \text{as } n \to \infty,$$

then (Damerdji 1994a)

$$b_n \hat{V}_O \xrightarrow{a.s.} \sigma_\infty^2.$$

Pedrosa and Schmeiser (1994) and Song (1996) considered weakly stationary processes with $\gamma_m = \sum_{j=-\infty}^{\infty} j^m C_j < \infty$ for $m = 0, 1$ and studied batch means variance estimators with

$$\text{Bias}(\hat{V}) = -c_b \gamma_1 \frac{1}{b} + o\left(\frac{1}{b}\right)$$

and

$$\mathrm{Var}(\hat{V}) = c_v \gamma_0^2 \frac{b}{n} + o\left(\frac{b}{n}\right).$$

The constants c_b and c_v depend on the amount of overlapping between the batches. In particular, the estimator $\hat{V}_B$ has $c_b = 1$ and $c_v = 2$, while $\hat{V}_O$ has $c_b = 1$ and $c_v = 4/3$. Then the asymptotic batch size that minimizes $\mathrm{MSE}(\hat{V}) = \mathrm{Bias}^2(\hat{V}) + \mathrm{Var}(\hat{V})$ is

$$b^* = \left(\frac{2c_b^2\gamma_1^2}{c_v\gamma_0^2}\right)^{1/3} n^{1/3}. \tag{25}$$

Song (1996) developed methods for estimating the ratio $(\gamma_1/\gamma_0)^2$ for a variety of processes, including moving average processes and autoregressive processes. Then one can obtain an estimator for b^* by plugging the ratio estimator into equation (25). Sherman (1995) proposed a method that does not rely on the estimation of $(\gamma_1/\gamma_0)^2$.

Welch (1987) noted that both traditional batch means and overlapping batch means are special cases of spectral estimation (see Section 3.6) at frequency 0 and, more importantly, suggested that overlapping batch means yield near-optimal variance reduction when one forms sub-batches within each batch and applies the method to the sub-batches. For example, a batch of size 64 is split into 4 sub-batches and the first (overlapping) batch consists of observations $X_1, \ldots, X_{64}$, the second consists of observations $X_{17}, \ldots, X_{80}$, etc.

3.5 The Standardized Time Series Method

This method was proposed by Schruben (1983). The standardized time series is defined by

$$T_n(t) = \frac{\lfloor nt \rfloor (\overline{X}_n - \overline{X}_{\lfloor nt \rfloor})}{\sigma_\infty \sqrt{n}}, \quad 0 \le t \le 1$$

and, under some mild assumptions (e.g., stationarity and ϕ-mixing),

$$(\sqrt{n}(\overline{X}_n - \mu), \sigma_\infty T_n) \xrightarrow{\mathcal{D}} (\sigma_\infty W(1), \sigma_\infty B),$$

where $\{B(t) : t \ge 0\}$ is the Brownian bridge process (see Billingsley 1968). Informally, $\{X_i\}$ is ϕ-mixing if X_i and X_{i+j} are approximately independent for large j. Figure 8 shows the standardized time series for the AR(1) sample path in Figure 1.

If $A = \int_0^1 \sigma_\infty B(t)\, dt$ is the area under B, then the identity

$$E(A^2) = \sigma_\infty^2/12$$

implies that σ_∞^2 can be estimated by multiplying an estimator of $E(A^2)$ by 12. Suppose that the data $X_1, \ldots, X_n$ are divided into k (contiguous) batches, each of size b. Then for sufficiently large n the random variables

$$A_i = \sum_{j=1}^{b} [(n+1)/2 - j] X_{(i-1)b+j}, \quad i = 1, \ldots, k$$

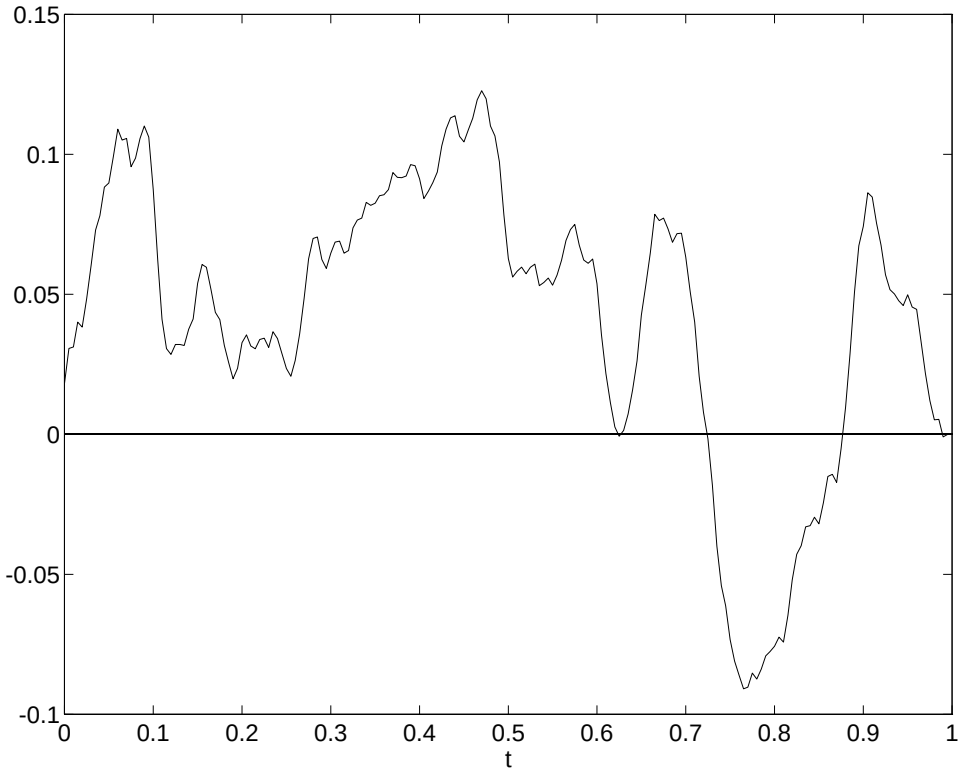

Figure 8: The standardized time series for the AR(1) sample path in Figure 1

become approximately i.i.d. normal and an estimator of $E(A^2)$ is

$$\hat{E}(A^2) = \frac{1}{(b^3 - b)k} \sum_{i=1}^{k} A_i^2.$$

Hence an (approximate) $1 - \alpha$ confidence interval for μ is

$$\overline{Y}_k \pm t_{k,1-\alpha/2} \sqrt{\hat{V}_T / n},$$

where

$$\hat{V}_T = 12\hat{E}(A^2).$$

The standardized time series method is easy to implement and has asymptotic advantages over the batch means method (see Goldsman and Schruben 1984). However, in practice it can require prohibitively long runs as noted by Sargent, Kang, and Goldsman (1992). Some useful theoretical foundations of the method are given in Glynn and Iglehart (1990). Additional developments on the method, as well as other standardized time series estimators, are contained in Goldsman, Meketon, and Schruben (1990) and Goldsman and Schruben (1984, 1990). Finally, Damerdji (1994a) shows that under the assumption of

strong approximation in Section 3.4, batching sequences satisfying assumptions (A.1)–(A.3) yield consistent estimators for the process variance σ_∞^2.

3.6 The Spectral Estimation Method

This method also assumes that the process $\{X_i\}$ is weakly stationary. Under this assumption, the variance of $\overline{X}_n$ is given by (6). The name of the method is due to the fact that if $\sum_{j=-\infty}^{\infty} |C_j| < \infty$, then $n\text{Var}(\overline{X}_n) \to 2\pi g(0)$ as $n \to \infty$, where $g(\lambda)$ is the *spectrum* of the process at frequency λ and is defined by

$$g(\lambda) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} C_j e^{-i\lambda j}, \quad |\lambda| \leq \pi,$$

where $i = \sqrt{-1}$. Therefore, for large n the estimation of $\text{Var}(\overline{X}_n)$ can be viewed as that of estimating $g(0)$. Estimators of this variance have the form

$$\hat{V}_S = \frac{1}{n} \left(\hat{C}_0 + 2 \sum_{j=1}^{p-1} w_j \hat{C}_j \right),$$

where p and the *weights* w_j are chosen to improve the properties of the variance estimator $\hat{V}_S$. The selection of these parameters is discussed in Fishman (1978b) and Law and Kelton (1984). Further discussions of spectral methods are given in Heidelberger and Welch (1981a,b, 1983) and Damerdji (1991).

3.7 Quantile Estimation from Stationary Output Data

Heidelberger and Lewis (1984) proposed three methods for computing confidence intervals for quantiles when the output process is stationary but not regenerative. Only the average group quantile method will be presented because it is simpler to implement than the competitors and has performed as well or better than the others in terms of the width of confidence intervals and the coverage probabilities. This method makes use of the *maximum transform*.

3.7.1 The Maximum Transform

The purpose of the maximum transform is to convert the problem of computing an extreme quantile to one of computing a quantile close to the median. The transform works as follows: Let $X_1, X_2, \ldots, X_v$ be i.i.d. random variables with p-quantile equal to ξ, and let $Y = \max\{X_1, X_2, \ldots, X_v\}$. Then

$$P(Y \leq \xi) = F_Y(\xi) = P(X_1 \leq \xi, \ldots, X_v \leq \xi) = [F_X(\xi)]^v = p^v \equiv p'.$$

Thus, the p'-quantile of Y is the p^v-quantile of X. The idea is to choose v such that $p^v \approx 0.5$ since estimators for the median will generally have smaller bias and variance than estimators for more extreme quantiles. For example, if $p = 0.99$, then $v \approx \ln(0.5)/\ln(0.99) = 6.58$. So, choosing $v = 7$ gives $p' = 0.99^7 = 0.48$. Notice that by choosing groups of 7 observations and

applying the maximum transform, the amount of data that must be processed is reduced by a factor of 7.

Applying the maximum transform generally results in an inflation of the variance by a factor of approximately 1.4 (see Heidelberger and Lewis 1984). It is also possible to use other schemes such as the next-to-maximum. The maximum transform is clearly applicable to quantile estimation via independent replications (see Section 3.3.2). For processes that are stationary but not i.i.d., Heidelberger and Lewis apply the maximum transform to observations at least m positions apart, where $m = n/v$, n is the sample size of the output, and v is an integer such that $p^v \approx 0.5$.

3.7.2 The Average Group Quantile Method

This method works as follows: First, determine a v so that $p^v \approx 0.5$, i.e., $v \approx \lfloor \ln(0.5)/\ln(p) \rfloor$. Then form k contiguous batches of $m \cdot v$ observations each. Within each batch, form m sub-batches of v observations. The first sub-batch consists of observations $1, m+1, 2m+1, \ldots, (v-1)m+1$, the second of observations $2, m+2, 2m+2, \ldots, (v-1)m+2$, and so on. All of the observations in each sub-batch are m positions apart. The maximum transform is applied within each sub-batch, producing m maximum transformed observations within the sub-batch. The p'-quantile is then computed from these observations, producing a quantile from each batch. Denote these batch quantiles by $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_k$. Then the overall quantile estimate is the sample mean of these batch quantiles, and an approximate $1-\alpha$ confidence interval for ξ is computed by treating $\hat{\xi}_1, \hat{\xi}_2, \ldots, \hat{\xi}_k$ as a set of i.i.d. observations and applying the usual confidence interval estimator for the mean

$$\overline{\xi} \pm t_{k-1,\alpha/2} \frac{S_k(\hat{\xi})}{\sqrt{k}} .$$

Heidelberger and Lewis compared this estimator to two competitors, one based on estimation of the spectral density of a binary process and another based on nested group quantiles. The average group quantile method performed well relative to the other methods, was not dominated by any of the other methods, and has the advantage that it is the easiest method to implement. The performance of the method depends upon the choice of the quantities m and k. While Heidelberger and Lewis do not provide a specific method or specific guidelines for choosing these parameters, they do recommend making m as large as possible to assure that the observations used in the maximum transform have a maximum distance between them and make the spaced observations approximately independent.

4 Multivariate Estimation

Frequently the output from a single simulation run is used to estimate several system parameters. The estimators of these parameters are typically correlated. As an example, one might wish to simultaneously estimate the average delays for customers at three stations in a queueing network.

Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_h)$ be a vector of h parameters that will be estimated using simulation output data. Two types of multivariate interval estimation procedures are generally used: *simultaneous confidence intervals* and *confidence regions*. The set of intervals

$\{I_i = (\hat{\theta}_{il}, \hat{\theta}_{iu}), i = 1, 2, \ldots, h\}$ is said to be a set of $1-\alpha$ simultaneous confidence intervals for $\boldsymbol{\theta}$ if

$$P\left(\cap_{i=1}^{h}\{\theta_i \in I_i\}\right) = 1-\alpha.$$

A region $\boldsymbol{\Theta} \subset \mathbb{R}^h$ is said to be a $1-\alpha$ confidence region for $\boldsymbol{\theta}$ if $P(\boldsymbol{\theta} \in \boldsymbol{\Theta}) = 1-\alpha$. Note that simultaneous confidence intervals form a rectangular region in $\mathbb{R}^h$. In general, a confidence region will not be rectangular but will have smaller volume.

There are many articles in the literature concerning multivariate estimation in general and multivariate methods for simulation in particular. For a general introduction to multivariate statistical methods, see Anderson (1984). Seila (1984) and Charnes (1995) survey multivariate methods for simulation, primarily methods for estimating the mean, and provide extensive lists of references.

4.1 Bonferroni Intervals

Bonferroni's inequality provides a means for computing a lower bound on the simultaneous confidence coefficient for any set of confidence intervals. Let $E_1, E_2, \ldots, E_h$ be any set of events. Bonferroni's inequality states that

$$P(E_1 \cap E_2 \cap \cdots \cap E_h) \geq 1 - \sum_{j=1}^{h}(1 - P(E_j)).$$

To apply this inequality to a set of confidence intervals, let I_j be a $1-\alpha_j$ confidence interval for $\theta_j, j = 1, 2, \ldots, h$, and let E_j represent the event $\theta_j \in I_j$. Then, $P(E_j) = 1-\alpha_j$. By Bonferroni's inequality, the simultaneous confidence coefficient is

$$1-\alpha = P(E_1 \cap E_2 \cap \cdots \cap E_h) \geq 1 - \sum_{j=1}^{h}\alpha_j. \tag{26}$$

Bonferroni's inequality applies in very general circumstances. No conditions or restrictions are placed on the population, the parameters, or the methods of computing the intervals $I_1, I_2, \ldots, I_h$. Normally, to apply this approach one would compute a $1-\alpha/h$ confidence interval for each parameter θ_i. Then by (26) the simultaneous confidence coefficient is at least $1-\alpha$. The correctness of this simultaneous confidence coefficient depends upon the correctness of the individual confidence coefficients for the individual intervals, however. See the last paragraph in Section 4.2.

4.2 Multivariate Inference for the Mean Using Independent Replications

Suppose that the simulation run consists of n identical, independent replications, and that replication i produces output data vector $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ih})$, where X_{ij} is an observation that will be used to estimate θ_j. Thus, the output of the entire simulation experiment consists of n i.i.d. vectors of observations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$. Let $\boldsymbol{\mu} = E(\mathbf{X}_i)$ be the vector of population means, and $\boldsymbol{\Sigma} = E[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})']$ be the variance-covariance matrix for each $\mathbf{X}_i$ with components $\Sigma_{jk} = \mathrm{Cov}(X_{ij}, X_{ik})$. The point estimator for $\boldsymbol{\mu}$ is the multivariate sample mean

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i,$$

with components

$$\hat{\mu}_j = \overline{X}_{nj} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \quad j = 1, 2, \ldots, h$$

and the estimator of $\boldsymbol{\Sigma}$ is

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}}_n)(\mathbf{X}_i - \overline{\mathbf{X}}_n)'.$$

Here $\overline{\mathbf{X}}_n$ and $\boldsymbol{\Sigma}$ are the basic sample statistics that are used for multivariate inference for the mean $\boldsymbol{\mu}$. If $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ have a multivariate normal distribution, then a $1-\alpha$ confidence region for $\boldsymbol{\mu}$ is given by all vectors $\mathbf{x}$ for which

$$n(\overline{\mathbf{X}}_n - \mathbf{x})'\mathbf{S}^{-1}(\overline{\mathbf{X}}_n - \mathbf{x}) \leq \frac{(n-1)h}{n-h}F_{h,n-h,1-\alpha},$$

where $F_{h,n-h,1-\alpha}$ is the $1-\alpha$ quantile of the F distribution with h and $n-h$ degrees of freedom in the numerator and denominator, respectively. More generally, if $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_d$ are h-dimensional non-null vectors of constants, then a $1-\alpha$ confidence region for $(\phi_1, \ldots, \phi_d)$ with

$$\phi_l = \boldsymbol{\pi}_l'\boldsymbol{\mu} = \sum_{j=1}^{l}\pi_{lj}\mu_j, \quad l = 1, 2, \ldots, d,$$

is given by all vectors $\mathbf{x} \in \mathbb{R}^d$ such that

$$n(\hat{\boldsymbol{\phi}} - \mathbf{x})'[\boldsymbol{\pi}'\mathbf{S}\boldsymbol{\pi}]^{-1}(\hat{\boldsymbol{\phi}} - \mathbf{x}) \leq \frac{(n-1)r}{n-r}F_{r,n-r,1-\alpha},$$

where $\hat{\phi}_l = \boldsymbol{\pi}_l'\hat{\boldsymbol{\mu}}$ is the estimator for ϕ_l. If the data $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are not multivariate normal but are approximately multivariate normal, the above regions may be used as approximate $1-\alpha$ confidence regions for $\boldsymbol{\mu}$. This would be the case, for example, if $\mathbf{X}_i$ were the sample mean for a sequence of observations and conditions for a central limit theorem were met.

Two methods are available for computing simultaneous confidence intervals for the mean. The approach using Bonferroni's inequality has already been discussed. A second method, which was originally proposed by Roy and Bose (1953) and Scheffé (1953), computes the intervals

$$\overline{X}_{nj} \pm T_{h,n-h,1-\alpha/2}\frac{S_{jj}}{\sqrt{n}}, \quad j = 1, 2, \ldots, h,$$

where $T_{h,n-h,1-\alpha/2}$ is the $1-\alpha/2$ quantile of Hotelling's T^2 distribution with h and $n-h$ degrees of freedom in the numerator and denominator, respectively (see Hotelling 1931 or Anderson 1984, Chapter 5).

Bonferroni's inequality is rather tight; Scheffé intervals are very conservative. Therefore, Bonferroni intervals will normally be shorter than Scheffé intervals. However, if the true confidence coefficients of the individual intervals are less than the nominal values, Scheffé intervals may be preferred in order to protect against an unacceptably small simultaneous confidence coefficient. For example, suppose that $h = 5$ parameters are to be estimated using simultaneous confidence intervals with simultaneous confidence coefficient $1-\alpha = 0.95$.

To use Bonferroni's inequality, one would compute a 0.99 confidence interval for each parameter. If, in fact, the parameter estimators were independent, then the true simultaneous confidence coefficient would be $0.99^5 = 0.951$. However, if the true coverage probability for each confidence interval were actually 0.98 instead of 0.99, then the simultaneous confidence coefficient would be only $0.98^5 = 0.904$, considerably below the desired value of 0.95.

4.3 Multivariate Inference for the Mean Using Stationary Data

Methods for computing simultaneous confidence intervals and confidence regions for the mean have been developed when output data are stationary. As in the case with univariate inference, the initial transient portion must be identified and removed, leaving observations that are approximately stationary. One option is to run a series of n independent replications, compute the mean from each replication and use the methods of Section 4.2 to compute either approximate simultaneous confidence intervals or an approximate confidence region for μ. This approach, like the replication/deletion approach for univariate inference in Section 3.2, is wasteful of data and will result in a biased point estimator if the initial transient portion is judged too short. Bias in the point estimator of the mean or variance will reduce the coverage probability of the intervals or regions.

4.3.1 Multivariate Batch Means

An alternative to independent replications is to apply a generalization of the univariate batch means method. As in the univariate batch means procedure, this method divides a long run into batches of multivariate observations. These vectors could be produced because the output process is naturally in the form of a vector. Such a process would result, for example, in a queueing network if an observation is produced each time a customer leaves the system, and X_{ij} is the time required by customer i to travel path j in the network. Vector observations could also be produced by sampling continuous-time processes. More generally, if only means for continuous-time processes are to be estimated, then batches could be formed using continuous data accumulated every t time units.

It should be noted that for certain combinations of parameters, one can encounter synchronization problems. Suppose, for example, that a queueing system has two classes of customers, A and B, and suppose that 90 percent of the customers are of class A while the remaining 10 percent are of class B. The objective is to simultaneously estimate the mean waiting times for each class, say μ_A and μ_B. Then, if the batch size is set to 100, for example, the amount of simulation time required to collect a batch of 100 observations for class B customers will be approximately 9 times that for class A customers. One can easily see that the batches for class A customers will be completed long before those for class B customers, and the relationship between batches for classes A and B customers will change over time. In the following, the observation processes are assumed to be synchronous, in the sense that for any batch size the statistical relationship among batch means does not change.

The multivariate batch means method is applied analogously to the univariate batch means procedure. Suppose that a stationary multivariate output process is divided into k batches of b vectors each and let $\overline{\mathbf{Y}}_1, \ldots, \overline{\mathbf{Y}}_k$ be the sequence of batch means. If

$$\sum_{l=-\infty}^{\infty} \mathrm{Cov}(X_{ij}, X_{i+l,j}) < \infty \quad \text{for all } j = 0, 1, \ldots, h,$$

then the vectors $\overline{\mathbf{Y}}_1, \ldots, \overline{\mathbf{Y}}_k$ are asymptotically uncorrelated and their sample mean vector is a consistent estimator of the steady-state mean vector $\boldsymbol{\mu}$. The multivariate batch means method then treats $\overline{\mathbf{Y}}_1, \ldots, \overline{\mathbf{Y}}_k$ as a sequence of i.i.d. random vectors and applies the methods of Section 4.2 to compute a confidence region or simultaneous confidence intervals. One is left with the problem of determining the appropriate batch size and number of batches. This problem is complicated by the fact that since the batch means are vectors, the autocorrelation function will be a sequence of correlation matrices. Chen and Seila (1987) developed a procedure that is based upon fitting a first-order autoregressive process to the sequence of batch means to test for autocorrelation and determine the batch size. This procedure has been shown to work well in a variety of systems.

ACKNOWLEDGMENTS

The authors would like to thank George Fishman and David Goldsman for their many fruitful discussions. The research of the first author was partially supported by the Air Force Office of Scientific Research under contract 93-0043.

REFERENCES

Alexopoulos, C. 1993. Distribution-free confidence intervals for conditional probabilities and ratios of expectations. *Management Science* 40(12):1748–1763.

Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, New York.

Billingsley, P. 1968. *Convergence of Probability Measures*. Wiley, New York, New York.

Blomqvist, N. The covariance function of the M/G/1 queueing system. *Skandinavisk Aktuarietidskrift* 50: 157–174.

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

Carlstein, E. 1986. The use of subseries for estimating the variance of a general statistic from a stationary sequence. *Annals of Mathematical Statistics* 14:1171–1179.

Chance, F., and L. W. Schruben. 1992. Establishing a truncation point in simulation output. Technical Report, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York.

Charnes, J. M. 1989. Statistical analysis of multivariate discrete-event simulation output. Ph.D. Thesis, Department of Operations and Management Science, University of Minnesota, Minneapolis, Minnesota.

Charnes, J. M. 1995. Analyzing multivariate output. In *Proceedings of the 1995 Winter Simulation Conference*, eds. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman, 201–208. IEEE, Piscataway, New Jersey.

Chatfield, C. 1989. *The Analysis of Time Series: An Introduction*, 4th edition. Chapman and Hall, New York, New York.

Chen, R. D., and A. F. Seila. 1987. Multivariate inference in stationary simulation using batch means. In *Proceedings of the 1987 Winter Simulation Conference*, eds. A. Thesen, H. Grant, and W. D. Kelton, 302–304. IEEE, Piscataway, New Jersey.

Chien, C.-H. 1989. Small sample theory for steady state confidence intervals. Technical Report No. 37, Department of Operations Research, Stanford University, Palo Alto, California.

Chien, C., D. Goldsman, and B. Melamed. 1997. Large-sample results for batch means. *Management Science* 43:1288–1295.

Chow, Y. S., and H. Robbins. 1965. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Annals of Mathematical Statistics* 36:457–462.

Conway, R. W. 1963. Some tactical problems in digital simulation. *Management Science* 10:47–61.

Crane, M. A., and D. L. Iglehart. 1974a. Simulating stable stochastic systems, I: General multiserver queues. *Journal of the ACM* 21:103–113.

Crane, M. A., and D. L. Iglehart. 1974b. Simulating stable stochastic systems, II: Markov chains. *Journal of the ACM* 21:114–123.

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

Crane, M. A., and A. J. Lemoine. 1977. *An Introduction to the Regenerative Method for Simulation Analysis*. Springer-Verlag, New York, New York.

Damerdji, H. 1991. Strong consistency and other properties of the spectral variance estimator. *Management Science* 37:1424–1440.

Damerdji, H. 1994a. Strong consistency of the variance estimator in steady-state simulation output analysis. *Mathematics of Operations Research* 19:494–512.

Damerdji, H. 1994b. On the batch means and area variance estimators. In *Proceedings of the 1994 Winter Simulation Conference*, eds. S. Manivannan, J. D. Tew, D. A. Sadowski, and A. F. Seila, 340–344. IEEE, Piscataway, New Jersey.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia, Pennsylvania.

Fishman, G. S. 1972. Bias considerations in simulation experiments. *Operations Research* 20:785–790.

Fishman, G. S. 1973. Statistical analysis for queueing simulations. *Management Science* 20:363–369.

Fishman, G. S. 1974. Estimation of multiserver queueing simulations. *Operations Research* 22:72–78.

Fishman, G. S. 1978a. Grouping observations in digital simulation. *Management Science* 24:510–521.

Fishman, G. S. 1978b. *Principles of Discrete Event Simulation*. John Wiley, New York.

Fishman, G. S. 1996. *Monte Carlo: Concepts, Algorithms and Applications*. Chapman and Hall, New York.

Fishman, G. S., and L. S. Yarberry. 1997. An implementation of the batch means method. *INFORMS Journal on Computing* 9:296–310.

Gafarian, A.V., C. J. Ancker, and F. Morisaku. 1978. Evaluation of commonly used rules for detecting steady-state in computer simulation. *Naval Research Logistics Quarterly* 25:511–529.

Glynn, P. W., and D. L. Iglehart. 1990. Simulation analysis using standardized time series. *Mathematics of Operations Research* 15:1–16.

Goldsman, D., M. Meketon, and L. W. Schruben. 1990. Properties of standardized time series weighted area variance estimators. *Management Science* 36:602–612.

Goldsman, D., and L. W. Schruben. 1984. Asymptotic properties of some confidence interval estimators for simulation output. *Management Science* 30:1217–1225.

Goldsman, D., and L. W. Schruben. 1990. New confidence interval estimators using standardized time series. *Management Science* 36:393–397.

Goldsman, D., L. W. Schruben, and J. J. Swain. 1994. Tests for transient means in simulated time series. *Naval Research Logistics Quarterly* 41:171–187.

Heidelberger, P., and P. A. W. Lewis. 1984. Quantile estimation in dependent sequences. *Operations Research* 32:185–209.

Heidelberger, P., and P. D. Welch. 1981a. A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM* 24:233–245.

Heidelberger, P., and P. D. Welch. 1981b. Adaptive spectral methods for simulation output analysis. *IBM Journal of Research and Development* 25:860–876.

Heidelberger, P., and P. D. Welch. 1983. Simulation run length control in the presence of an initial transient. *Operations Research* 31:1109–1144.

Hogg, R. V., and A. T. Craig. 1978. *Introduction to Mathematical Statistics*, 4th edition. Macmillan, New York.

Hotelling, H. 1931. The generalization of Student's ratio. *Annals of Mathematical Statistics* 2:360–378.

Iglehart, D. L. 1975. Simulating stable stochastic systems, V: Comparison of ratio estimators. *Naval Research Logistics Quarterly* 22:553–565.

Iglehart, D. L. 1976. Simulating stable stochastic systems, VI: Quantile estimation. *Journal of the ACM* 23:347–360.

Iglehart, D. L. 1978. The regenerative method for simulation analysis. In *Current Trends in Programming Methodology, Vol. III*, eds. K. M. Chandy, and K. M. Yeh, 52–71. Prentice-Hall, Englewood Cliffs, New Jersey.

Karr, A. F. 1993. *Probability*. Springer-Verlag, New York.

Kelton, W. D. 1989. Random initialization methods in simulation. *IIE Transactions* 21:355–367.

Kelton, W. D., and A. M. Law. 1983. A new approach for dealing with the startup problem in discrete event simulation. *Naval Research Logistics Quarterly* 30:641–658.

Kleijnen, J. P. C. 1974. *Statistical Techniques in Simulation, Part I*. Marcel Dekker, New York.

Kleijnen, J. P. C. 1975. *Statistical Techniques in Simulation, Part II*. Marcel Dekker, New York.

Law, A. M., and J. S. Carson. 1979. A sequential procedure for determining the length of a steady-state simulation. *Operations Research* 27:1011–1025.

Law, A. M., and W. D. Kelton. 1984. Confidence intervals for steady-state simulations, I: A survey of fixed sample size procedures. *Operations Research* 32:1221–1239.

Law, A. M., and W. D. Kelton. 1991. *Simulation Modeling and Analysis*, 2d edition. McGraw-Hill, New York.

Law, A. M., W. D. Kelton, and L. W. Koenig. 1981. Relative width sequential confidence intervals for the mean. *Communic. Statist.* B10:29–39.

Lehmann, E. L. 1991. *Theory of Point Estimation*, 2d edition. Wadsworth, Belmont, California.

Lindley, D. V. 1952. The theory of queues with a single server. *Proceedings of the Cambridge Philosophical Society* 48:277–289.

Mechanic, H., and W. McKay. 1966. Confidence intervals for averages of dependent data in simulations II. Technical Report ASDD 17–202, IBM Corporation, Yorktown Heights, New York.

Meketon, M. S., and B. W. Schmeiser. 1984. Overlapping batch means: Something for nothing? In *Proceedings of the 1984 Winter Simulation Conference*, eds. S. Sheppard, U. W. Pooch, and C. D. Pegden, 227–230. IEEE, Piscataway, New Jersey.

Moore, L. W. 1980. Quantile estimation in regenerative processes. Ph.D. Thesis, Curriculum in Operations Research and Systems Analysis, University of North Carolina, Chapel Hill, North Carolina.

Nadas, A. 1969. An extension of the theorem of Chow and Robbins on sequential confidence intervals for the mean. *Annals of Mathematical Statistics* 40:667–671.

Ockerman, D. H. 1995. Initialization bias tests for stationary stochastic processes based upon standardized time series techniques. Ph.D. Thesis, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia.

Pedrosa, A. C., and B. W. Schmeiser. 1994. Estimating the variance of the sample mean: Optimal batch size estimation and 1–2–1 overlapping batch means. Technical Report SMS94–3, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

Ross, S. M. 1993. *Introduction to Probability Models*, 5th edition. Academic Press, San Diego, California.

Roy, S. N., and R. C. Bose. 1953. Simultaneous confidence interval estimation. *Annals of Mathematical Statistics* 24:513–536.

Sargent, R. G., K. Kang, and D. Goldsman. 1992. An investigation of finite-sample behavior of confidence interval estimators. *Operations Research* 40:898–913.

Scheffé, H. 1953. A method of judging all contrasts in analysis of variance. *Biometrica* 40:87–104.

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

Schriber, T. J., and R. W. Andrews. 1979. Interactive analysis of simulation output by the method of batch means. In *Proceedings of the 1979 Winter Simulation Conference*, eds. M. G. Spiegel, N. R. Nielsen, and H. J. Highland, 513–525. IEEE, Piscataway, New Jersey.

Schruben, L. W. 1982. Detecting initialization bias in simulation output. *Operations Research* 30:569–590.

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

Schruben, L. W., H. Singh, and L. Tierney. 1983. Optimal tests for initialization bias in simulation output. *Operations Research* 31:1167–1178.

Seila, A. F. 1982a. A batching approach to quantile estimation in regenerative simulations. *Management Science* 28:573–581.

Seila, A. F. 1982b. Percentile estimation in discrete event simulation. *Simulation* 39:193–200.

Seila, A. F. 1984. Multivariate simulation output analysis. *American Journal of Mathematical and Management Sciences* 4:313–334.

Sherman, M. 1995. On batch means in the simulation and statistical communities. In *Proceedings of the 1995 Winter Simulation Conference*, eds. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman, 297–302. IEEE, Piscataway, New Jersey.

Song, W.-M. T., and B. W. Schmeiser. 1995. Optimal mean-squared-error batch sizes. *Management Science* 41:110–123.

Song, W.-M. T. 1996. On the estimation of optimal batch sizes in the analysis of simulation output analysis. *European Journal of Operations Research* 88:304–309.

Starr, N. 1966. The performance of a statistical procedure for the fixed-width interval estimation for the mean. *Annals of Mathematical Statistics* 37(1):36–50.

von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference and the variance. *Annals of Mathematical Statistics* 12:367–395.

Welch, P. D. 1981. On the problem of the initial transient in steady state simulations. Technical Report, IBM Watson Research Center, Yorktown Heights, New York.

Welch, P. D. 1983. The statistical analysis of simulation results. In *The Computer Performance Modeling Handbook*, ed. S. Lavenberg, 268–328. Academic Press, New York.

Welch, P. D. 1987. On the relationship between batch means, overlapping batch means and spectral estimation, In *Proceedings of the 1987 Winter Simulation Conference*, eds. A. Thesen, H. Grant, and W. D. Kelton, 320–323. IEEE, Piscataway, New Jersey.

Wilson, J. R., and A. A. B. Pritsker. 1978a. A survey of research on the simulation startup problem. *Simulation* 31:55–58.

Wilson, J. R., and A. A. B. Pritsker. 1978b. Evaluation of startup policies in simulation experiments. *Simulation* 31:79–89.

Yarberry, L. S. 1993. Incorporating a dynamic batch size selection mechanism in a fixed-sample-size batch means procedure. Ph.D. thesis, Department of Operations Research, University of North Carolina, Chapel Hill, North Carolina.

Young, L. C. 1941. Randomness in ordered sequences. *Annals of Mathematical Statistics* 12:293–300.