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

OUTPUT DATA ANALYSIS FOR SIMULATIONS

Christos Alexopoulos
Seong-Hee Kim

School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332–0205, U.S.A.

ABSTRACT

This paper reviews statistical methods for analyzing output data from computer simulations. First, it focuses on the estimation of steady-state system parameters. The estimation techniques include the replication/deletion approach, the regenerative method, the batch means method, and the standardized time series method. Second, it reviews recent statistical procedures to find the best system among a set of competing alternatives.

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 program only result in *estimates* of system performance measures. Unfortunately, a simulation run does not usually produce independent, identically distributed (i.i.d.) 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 paper reviews (a) statistical methods for computing confidence intervals for system performance measures from output data and (b) statistical methods for determining the best system from a set of alternatives.

For output analysis, there are two types of simulations: **Finite-horizon simulations**. In this case the simulation starts in a specific state and runs until a terminating event occurs. The output process is not expected to achieve steady-state behavior and any parameter estimated from the output will be transient in the sense that its value will depend upon the initial conditions. An example is the simulation of a vehicle storage and distribution facility for a week.

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. 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.

Section 2 discusses methods for analyzing output from terminating simulations. Section 3 reviews approaches for removing bias due to initial conditions in steady-state simulations. Section 4 presents techniques for point and interval estimation of steady-state parameters. Section 5 reviews recent methods for identifying the best system within a set of alternatives.

2 FINITE-HORIZON SIMULATIONS

Suppose that we simulate a system until n output data $X_1, X_2, \ldots, X_n$ are collected with the objective of estimating $\mu = E(\bar{X}_n)$, where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean. 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. Clearly, $\bar{X}_n$ is an unbiased estimator for μ. Unfortunately, the X_i's are generally dependent random variables making the estimation of $\mathrm{Var}(\bar{X}_n)$ a nontrivial problem. Let $S_n^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ be the sample variance of the data. In many queueing systems the X_i's are correlated making S_n^2/n a biased estimator of $\mathrm{Var}(\bar{X}_n)$. In particular, if the X_i's are positively correlated, one has $E(S_n^2/n) < \mathrm{Var}(\bar{X}_n)$.

To overcome this problem, we can run k independent replications of the simulation. Assume that run i produces output data $X_{i1}, \ldots, X_{in}$. Then the "within-run" averages

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

are i.i.d. random variables, their sample mean $\bar{Y}_k = \frac{1}{k}\sum_{i=1}^{k} Y_i$ is also an unbiased estimator of μ, and their sample variance $\hat{V}_R = (k-1)^{-1}\sum_{i=1}^{k}(Y_i - \bar{Y}_k)^2$ is an unbiased estimator of $\mathrm{Var}(\bar{X}_n)$. If in addition k is sufficiently large, an approximate $1-\alpha$ confidence interval for μ is

$$\bar{Y}_k \pm t_{k-1,1-\alpha/2}\sqrt{\hat{V}_R/k}\,, \tag{1}$$

where $t_{d,\gamma}$ represents the γ-quantile of the t distribution with d degrees of freedom.

Alexopoulos and Seila (1998, Section 7.2.2) review sequential procedures for determining the number of replications required to estimate μ with a fixed absolute or relative precision. The procedure for constructing a $1-\alpha$ confidence interval for μ with a small absolute error $|\bar{Y}_k - \mu| \le \beta$ is based on Chow and Robbins (1965). It starts with $k_0 \ge 5$ runs and stops when the halfwidth $t_{k-1,1-\alpha/2}\sqrt{\hat{V}_R/k} \le \beta$. Law and Kelton (2000) describe a method for obtaining an estimate whose relative error satisfies $\Pr(|\bar{Y}_k - \mu|/|\mu| \le \gamma) \ge 1-\alpha$, with $\alpha \le 0.15$. The method starts with $k_0 \ge 10$ runs and stops when the relative halfwidth $t_{k-1,1-\alpha/2}|\bar{Y}_k|^{-1}\sqrt{\hat{V}_R/k}$ drops below $\gamma/(1+\gamma)$.

The method of replications can also be used for estimating performance measures other than means. For example, suppose that we want to estimate the p-quantile, say ξ_p, of the average queue size in a single-server queueing system during a fixed time window. We run k independent replications, denote by Y_i the average observed queue length during replication i, and let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(k)}$ be the order statistics corresponding to the Y_i's. Then a point estimate for y_p is $\hat{\xi}_p = Y_{(kp)}$ if kp is an integer or $\hat{\xi}_p = Y_{(\lfloor kp+1 \rfloor)}$ otherwise ($\lfloor\cdot\rfloor$ is the floor function). A confidence interval for ξ_p is described in Alexopoulos and Seila (1998, Section 7.3.2).

3 INITIALIZATION PROBLEMS FOR STEADY-STATE SIMULATIONS

One of the hardest problems in steady-state simulations is the removal of the *initialization bias*. Suppose that $\{X_i : i \ge 1\}$ is a discrete-time output stochastic process from a single run of a steady-state simulation with initial conditions (system state) I and assume that, as $n \to \infty$, $\Pr(X_n \le x|I) \to \Pr(X \le x)$, where X is the corresponding steady-state random variable. The steady-state mean of $\{X_i\}$ is $\mu = \lim_{n\to\infty} E(X_n|I)$. The problem with the use of $\bar{X}_n$ for a finite n is that $E(\bar{X}_n|I) \ne \mu$.

The most commonly used method for eliminating the bias of $\bar{X}_n$ identifies an index l and *truncates* the observations $X_1, \ldots, X_l$. Then the estimator $\bar{X}_{n,l} = n^{-1}\sum_{i=l+1}^{n+l} X_i$ is generally less biased than $\bar{X}_n$ 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 Chance and Schruben 1992, Fishman 2001, Gafarian et al. 1978, Goldsman et al. 1994, Kelton 1989, Ockerman 1995, Schruben 1982, Schruben et al. 1983, Wilson and Pritsker 1978ab).

The graphical procedure of Welch (1983) is popular due to its simplicity and generality. This method uses k independent runs with the ith run producing observations $X_{i1}, X_{i2}, \ldots, X_{in}$ and computes the "across-runs" averages

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

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

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

against j. 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 may be a difficult problem for congested systems with output time series having autocorrelation functions with very long tails (see Alexopoulos and Seila 1998, Example 7).

4 STEADY-STATE ANALYSIS

We focus on estimation methods for the steady-state mean μ of a discrete-time output process. Analogous methods for analyzing continuous-time processes are described in several texts (Bratley, Fox, and Schrage 1987, Fishman 2001, Law and Kelton 2000). The process $\{X_i\}$ is called *stationary* if the joint distribution of $X_{i+j_1}, X_{i+j_2}, \ldots, X_{i+j_k}$ is independent of i for all indices $j_1, j_2, \ldots, j_k$ and all $k \ge 1$. If $E(X_i) = \mu$, $\mathrm{Var}(X_i) < \infty$ for all i, and the $\mathrm{Cov}(X_i, X_{i+j})$ is independent of i, then $\{X_i\}$ is called *weakly stationary*.

4.1 The Replication/Deletion Approach

This approach runs k independent replications, each of length $l+n$ observations, and discards the first l observations from each run. One then uses the i.i.d. sample means

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

to compute the point estimate

$$\bar{Y}_k(l,n) = \frac{1}{k}\sum_{i=1}^{k} Y_i(l,n)$$

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

$$\bar{Y}_k(l,n) \pm t_{k-1,1-\alpha/2}\sqrt{\hat{V}_R(n,l)/k}\,, \tag{2}$$

where $\hat{V}_R(l,n)$ is the sample variance of the $Y_i(l,n)$'s.

The method is simple and general, but involves the choice of three parameters, l, n and k. Here are a few points the user should aware of: (a) As l increases for fixed n, the "systematic" error in each $Y_i(l,n)$ due to the initial conditions decreases. (b) As n increases for fixed l, the systematic and sampling errors in $Y_i(l,n)$ decrease. (c) The systematic error in the sample means $Y_i(l,n)$ *cannot* be reduced by increasing the number of replications k. (d) For fixed n and under some mild moment conditions that are satisfied by a variety of simulation output processes, the confidence interval (2) is asymptotically valid only if $l/\ln k \to \infty$ as $k \to \infty$ (Fishman 2001). This means that as one makes more runs in an attempt to compute a narrower confidence interval, the truncation index l must increase faster than $\ln k$ for the confidence interval to achieve the nominal coverage. This requirement is hard to implement in practice. (d) This method is also potentially wasteful of data as the truncated portion is removed from each replication.

The regenerative method (Section 4.2) and the batch means method (Section 4.3) seek to overcome the aforementioned issues. Alexopoulos and Goldsman (2002) present a thorough comparison between the replication/deletion approach and the batch means method.

4.2 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 which can be used for computing point and interval estimates for the mean μ. The method was proposed by Crane and Iglehart (1974ab, 1975) and Fishman (1973, 1974). More precisely, assume that there are (random) time indices $1 \le T_1 < T_2 < \cdots$ such that the portion $\{X_{T_i}+j,\, j \ge 0\}$ 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 $\mu = E(Y_1)/E(Z_1)$.

Now suppose that one simulates the process $\{X_i\}$ over n cycles and collects the observations $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_n$. Then $\hat{\mu} = \bar{Y}_n/\bar{Z}_n$ is a strongly consistent estimator of μ. Furthermore, confidence intervals for μ can be constructed by using the random variables $Y_i - \mu Z_i, i = 1, \ldots, n$ and the central limit theorem (see Iglehart 1975).

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

4.3 The Batch Means Method

The method of batch means is frequently used to estimate the steady-state mean μ or the $\mathrm{Var}(\bar{X}_n)$ (for finite n) and owes its popularity to its simplicity and effectiveness.

To motivate the method, suppose temporarily that the data $X_1, \ldots, X_n$ are from a weakly stationary process with $\lim_{n\to\infty} n\mathrm{Var}(\bar{X}_n) = \sigma_\infty^2 < \infty$ (σ_∞^2 is called the variance parameter of the process $\{X_i\}$). Then 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 $\sigma_m^2 = \mathrm{Var}(\bar{X}_m)$. Since the batch means process $\{Y_i(b), i \ge 1\}$ is also weakly stationary, some algebra yields

$$\sigma_n^2 = \frac{\sigma_b^2}{k}\left(1 + \frac{n\sigma_n^2 - b\sigma_b^2}{b\sigma_b^2}\right). \tag{3}$$

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

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

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

estimates σ_b^2 by the sample variance of the batch means

$$\hat{V}_B(n,k) = \frac{1}{k-1}\sum_{i=1}^{k}(Y_i(b) - \bar{X}_n)^2,$$

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

$$\bar{X}_n \pm t_{k-1,1-\alpha/2}\sqrt{\hat{V}_B(n,k)/k}\,. \tag{4}$$

The main problem with the application of the batch means method in practice is the choice of the batch size b. The literature contains several batch selection approaches for fixed sample size; see Conway (1963), Law and Carson (1979), Mechanic and McKay (1966), and Schriber and Andrews (1979). 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.

4.4 Consistent Estimation Batch Means Methods

These methods assume that a central limit theorem holds

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \sigma_\infty N(0,1) \quad \text{as } n \to \infty \qquad (5)$$

and aim at constructing a consistent estimator for σ_∞^2 (converging in probability to σ_∞^2 as $n \to \infty$) and an asymptotically valid confidence interval for μ. Here $N(0,1)$ denotes the standard normal distribution.

Chien et al. (1997) considered stationary processes and, under quite general moment and sample path conditions, showed that as both $b, k \to \infty$, $\text{MSE}[b\hat{V}_k(b)] \to 0$. Notice that mean squared error consistency differs from consistency.

The limiting result (5) is implied under the following two assumptions, where $\{W(t), t \geq 0\}$ is the standard Brownian motion process (see Billingsley 1968).

Assumption of Weak Approximation (AWA).

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

Assumption of Strong Approximation (ASA). There exists a constant $\lambda \in (0, 1/2]$ and a finite random variable C such that, with probability one,

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

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 1994). The constant λ is closer to $1/2$ for processes having little autocorrelation, while it is closer to zero for processes with high autocorrelation.

4.5 Batching Rules

Equation (3) 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 Fixed Number of Batches (FNB) rule that sets the number of batches at k and uses batch sizes $b_n = \lfloor n/k \rfloor$ as n increases.

The FNB rule along with AWA imply that, as $n \to \infty$, $\bar{X}_n \xrightarrow{p} \mu$ and

$$\frac{\bar{X}_n - \mu}{\sqrt{\hat{V}_B(n,k)/k}} \xrightarrow{d} t_{k-1}$$

(see Glynn and Iglehart 1990). Hence, (4) is an asymptotically valid confidence interval for μ. Unfortunately, the FNB rule has two major limitations (see Fishman 1996, pp. 544–545): (a) Since $b_n\hat{V}_B(n,k)$ is not a consistent estimator of σ_∞^2, the confidence interval (4) tends to be wider than the interval a consistent estimation method would produce. (b) Statistical fluctuations in the halfwidth of the confidence interval (4) do not diminish relative to statistical fluctuation in the sample mean.

The limitations of the FNB rule can be removed by simultaneously increasing the batch size and the number of batches. Indeed, assume that ASA holds and consider batch sizes of the form $b_n = \lfloor n^\theta \rfloor$, $\theta \in (1 - 2\lambda, 1)$. Then as $n \to \infty$, $\bar{X}_n \xrightarrow{a.s.} \mu$, $b_n\hat{V}_B(n,k_n) \xrightarrow{a.s.} \sigma_\infty^2$, and

$$Z_{k_n} = \frac{\bar{X}_n - \mu}{\sqrt{\hat{V}_B(n,k_n)/k_n}} \xrightarrow{d} N(0,1) \qquad (6)$$

(see Damerdji 1994). The last display implies that

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

(z_γ is the γ-quantile of the standard normal distribution) is an asymptotically valid $1-\alpha$ confidence interval for μ. In particular, the Square Root (SQRT) rule (Chien 1989) that uses $\theta = 1/2$ ($b_n = \lfloor\sqrt{n}\rfloor$, $k_n = \lfloor\sqrt{n}\rfloor$) is valid if $1/4 < \lambda < 1/2$. Notice that the last inequality is violated by processes having high autocorrelation (λ is close to zero). Unfortunately, in practice the SQRT rule tends to seriously underestimate the $\text{Var}(\bar{X}_n)$ for small-to-moderate sample sizes n.

With the contrasts between the FNB and SQRT rules in mind, Fishman and Yarberry (1997) proposed two procedures that dynamically shift between the two rules. Both procedures perform "interim reviews" and compute confidence intervals at times $n_l \approx n_1 2^{l-1}, l = 1, 2, \ldots$. The correlation test for the batch means is based on von Neumann's statistic

$$C(n,k_n) = 1 - \frac{\sum_{i=2}^{k}(Y_i(b_n) - Y_{i-1}(b_n))^2}{2\sum_{i=1}^{k}(Y_i(b_n) - \bar{X}_n)^2}$$

(see von Neumann 1941ab).

The LBATCH Procedure. At time n_l, if the 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 Procedure. If at time n_l von Neumann's 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 procedures yield random sequences of batch sizes. Under relatively mild assumptions, these sequences imply convergence results analogous to (6). The respective algorithms require $O(n)$ time and $O(\log_2 n)$ space, where n is the desired sample size (see Alexopoulos et al. 1997 and Yarberry 1993). Although like complexities are known for static fixed batch size algorithms, the dynamic setting of the LBATCH and ABATCH procedures offers an important additional advantage not present in the static approach. As the analysis evolves with increasing sample path length, it allows a user to assess how well the estimated variance of the sample mean stabilizes. This assessment is essential to gauge the quality of the confidence interval for the sample mean. The LABATCH.2 implementation is the only computer package that automatically generates the data for this assessment. C, FORTRAN and SIMSCRIPT II.5 codes of LABATCH.2 can be downloaded via anonymous ftp from the site <http://www.or.unc.edu/~gfish/labatch.2.html>.

An alternative sequential method has been proposed by Steiger and Wilson (2001, 2002a) and Steiger et al. (2002b). The associated ASAP and ASAP2 software packages (accessible from the site <http://www.ie.ncsu.edu/jwilson>) can perform sequential sampling subject to absolute or relative precision criteria. ASAP starts with 96 batches, discards the first two batches, and progressively increases the batch size (by a factor of roughly $\sqrt{2}$) until either the last 94 batch means pass von Neumann's test for independence or vectors of spaced batch means pass the Shapiro-Wilk test for multivariate normality (Malkovich and Afifi 1973). In the latter case, the procedure delivers a correlation adjusted confidence interval based on an inverted Cornish-Fisher expansion whose terms are estimated via an ARMA time series model of the batch means. If the resulting confidence interval meets the underlying precision requirement, the method ends; otherwise, it estimates the additional number of batches that the user must collect. The ASAP2 algorithm differs from ASAP in that it starts with 256 batches and disregards the von Neumann test. A comparison between the LABATCH.2 and ASAP/ASAP2 methods is in order: The LABATCH.2 methodology aims at computing a (strongly) consistent estimator for σ_∞^2 along with an asymptotically valid confidence interval for the process mean. On the other hand, ASAP/ASAP2 do not yield consistent estimators for σ_∞^2; instead they aim at computing valid confidence intervals for the mean with specified precision. At small sample sizes, they take advantage of the fact that the batch means achieve multivariate normality prior to becoming independent to compute confidence intervals with substantially better coverage at the expense of larger and more variable halfwidths.

4.5.1 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, this method uses all $n-b+1$ overlapping batches to estimate μ and $\mathrm{Var}(\bar{X}_n)$. 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

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

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. Let $\hat{V}_O$ be the sample variance of the $Y_i(b)$'s. The following list contains properties of the estimators $\bar{Y}_O$ and $\hat{V}_O$: (a) The OBM estimator is a weighted average of non-overlapping batch means estimators. (b) 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, but $\mathrm{Var}(\hat{V}_O)/\mathrm{Var}(\hat{V}_B) \to 2/3$ (Meketon and Schmeiser 1984). (c) The behavior of $\mathrm{Var}(\hat{V}_O)$ appears to be less sensitive to the choice of the batch size than the behavior of $\mathrm{Var}(\hat{V}_B)$ (Song and Schmeiser 1995, Table 1). (d) If $\{X_i\}$ satisfies ASA and $\{b_n\}$ is a sequence of batches with $b_n = \lfloor n^\theta \rfloor$, $\theta \in (1-2\lambda, 1)$ and $b_n^2/n \to 0$ as $n \to \infty$, then (Damerdji 1994) $b_n \hat{V}_O \xrightarrow{a.s.} \sigma_\infty^2$.

Welch (1987) noted that both traditional batch means and overlapping batch means are special cases of spectral estimation 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.

4.6 The Standardized Time Series Method

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

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

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

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

where $\{B(t) : t \geq 0\}$ is the standard Brownian bridge process defined by $B(t) = W(t) - tW(1)$, $0 \leq t \leq 1$.

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. Schruben's method splits the data $X_1, \ldots, X_n$ into k (contiguous) batches, each of size b. Then for sufficiently large n the random variables

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

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

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

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

$$\bar{Y}_k \pm t_{k,1-\alpha/2} \sqrt{\hat{V}_T / n}; \quad \hat{V}_T = 12\widehat{E(A^2)}.$$

The standardized time series method 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, et al. (1992). Some useful theoretical foundations of the method are given in Glynn and Iglehart (1990). Also, Damerdji (1994) shows that under ASA in Section 4.3, batching sequences with $b_n = \lfloor n^\theta \rfloor$, $\theta \in (1 - 2\lambda, 1)$, yield asymptotically consistent estimators for the process variance σ_∞^2. Additional developments on the method, as well as other estimators based on the standardized time series, are contained in Alexopoulos et al. (2001), Goldsman et al. (1990) and Goldsman and Schruben (1984, 1990).

4.7 Quantile Estimation

A variety of methods have been proposed for estimating quantiles of steady-state data (see Iglehart 1976, Moore 1980, Seila 1982ab, Heidelberger and Lewis 1984). The methods differ in the way the variance of the sample quantile is estimated. It should be mentioned that quantile estimation is a harder problem than the estimation of steady-state means.

4.8 Multivariate Estimation

Frequently, the output from a single simulation run is used for estimating several system parameters. The estimators of these parameters are typically correlated. As an example, consider the average customer delays at two stations on a path of a queueing network. In general, Bonferroni's inequality can be used for computing a conservative confidence coefficient for a set of confidence intervals. Indeed, suppose that D_i is a $1 - \alpha$ confidence interval for the parameter μ_i, $i = 1, \ldots, m$. Then

$$\Pr\left[\cap_{i=1}^{m} \{\mu_i \in D_i\}\right] \geq 1 - \sum_{i=1}^{m} \alpha_i\,.$$

This result can have serious implications as for $m = 10$ and $\alpha_i = 0.10$ the r.h.s. of the above inequality is equal to 0. If the overall confidence level must be at least $1 - \alpha$, then the α_i's can be chosen so that $\sum_{i=1}^{m} \alpha_i = \alpha$. Multivariate estimation methods are described in Charnes (1989, 1990, 1991) and Chen and Seila (1987).

5 RANKING AND SELECTION METHODS

Simulation is performed not only to assess the performance of a single system, but also to compare a number of alternatives. One can argue that simulation almost always involves comparisons because even when a system is simulated to access its feasibility, the performance will be compared to a minimum standard. There exist at least four classes of comparison problems that arise in simulation: determining which alternative configurations have similar performance, comparing all systems to a standard, comparing alternatives to a default, and selecting the best system. By "best" we mean the system with a maximum performance measure, assumed to be the expectation of a random variable, such as throughput or delay time. Goldsman and Nelson (1998) reviewed procedures for each class of the above comparison problems. In this section, we review recent procedures to find the best system among a relatively large number of simulated systems, say more than 20 systems.

5.1 Background

Many classical procedures assume that the output data generated by each system are i.i.d. and normally distributed. However, raw data from within a single run are generally dependent and not quite normal. As suggested in Sections 2, 4.1, and 4.3, this can be overcome by using either within-run sample means or batch means from sufficiently large batches as basic observations. Issues related to comparisons with regard to steady-state measures will be discussed in Section 5.2.

5.1.1 Notation

The goal is to compare m systems via simulation and find the best system (with the largest expected performance), guaranteeing a correct selection with probability at least $1-\alpha$. Let Y_{ij} denote the jth observation from system i ($i = 1, 2, \ldots, m$). We assume that Y_{ij}'s are either within-run averages for system i or batch means from a single sufficiently long run after accounting for the elimination of the initialization bias. Thus it will always be assumed that the outputs from system i ($Y_{i1}, Y_{i2}, \ldots$) are i.i.d. and normally distributed. See Alexopoulos and Seila (1998) and Goldsman and Nelson (1998) for the plausibility of such assumptions.

5.1.2 Indifference Zone Approaches

In stochastic simulation, it is impossible to find the true best with certainty; so many procedures employ the Indifference Zone (IZ) approach as a good compromise. The IZ approach attempts to find a system whose mean is at least a user-specified amount better than the means for the other systems while guaranteeing a "correct selection" with high probability. The user-specified amount δ is called the indifference zone parameter and it is interpreted as a practically significant difference worth detecting. Goldsman and Nelson (1998) and Law and Kelton (2000) present IZ-based procedures that have been proven to be useful. The text of Bechhofer et al. (1995) provides a comprehensive review of IZ procedures. The problem is that IZ procedures become inefficient when the number of alternatives is large. This is because these procedures are developed under the Least Favorable Configuration (LFC) condition. LFC is the configuration considered as the most difficult to resolve. Therefore, if a procedure guarantees at least $1-\alpha$ probability of correct selection under LFC, it will do so for all other configurations. The Slippage Configuration (SC) is known to be the LFC in most procedures. Under SC, the expected performances of all other systems are assumed to be exactly δ smaller than that of the best so all inferior systems are equally close to the best. As IZ procedures are developed under the assumption that all inferior systems are close to the best, they become conservative when the number of systems is larger than 20.

To overcome this inefficiency of IZ approaches, one can introduce screening. When the number of systems is large, the performance measures for the systems are likely to spread out. Thus, after obtaining some observations, we may identify clearly inferior systems with high probability and then stop sampling from those inferior systems. If many systems can be eliminated early, one can save a lot of observations making the procedures more efficient. Gupta (1965) and Gupta and Huang (1976) proposed single-stage subset selection procedures that divide systems into a "maybe-best" group and a "clearly-not-best" group. Nelson et al. (2001) proposed a subset selection procedure that handles unknown and unequal variances and new procedures that combine subset selection algorithms with two-stage IZ algorithms. Also, Kim and Nelson (2001) proposed a fully sequential procedure where the systems in contention are compared after every observation until only one system survives. Procedures presented in Nelson et al. (2001) and Kim and Nelson (2001) have been shown to be efficient when hundreds of systems are compared.

5.1.3 Common Random Numbers

Many procedures require the assumption that systems are simulated independently. This can be achieved by assigning different streams of pseudo-random numbers to the simulation of each system. Alternatively, if the same random number streams are assigned to each simulation, then under fairly general conditions this induces positive correlation among the competitors and decreases the variance of the difference in observed avarage performances between two systems. This technique is called Common Random Numbers (CRN). Although CRN makes statistical procedures more complicated when there are more than two systems, under appropriate statistical procedures CRN makes comparison sharper, meaning spending less number of observations to find the best. The procedure of Nelson and Matejcik (1995) extended Rinott's procedure (1978), which is known as one of the simplest and most popular IZ procedures. Their procedure works in conjunction with CRN under a special structure of the variance-covariance matrix, called sphericity. However, it is not appropriate for a large number of systems as the sphericity assumption is often violated.

Below we review two recent procedures: the Subset+Rinott procedure due to Nelson et al. (2001) and the $\mathcal{KN}$ procedure due to Kim and Nelson (2001). Both procedures employ the IZ approach where basic observations are either within-run sample means or batch means, and utilize screening to gain efficiency in the case of many systems. The Subset+Rinott procedure does not use CRN (thus it requires systems to be simulated independently) while the $\mathcal{KN}$ procedure allows CRN.

Procedure Subset + Rinott

1. Specify the overall desired probability of correct selection $1-\alpha$, the IZ parameter δ, a common initial sample size from each system $k_0 \geq 2$, and the initial number of competing systems m.
   Further, set

$$t = t_{k_0-1,1-(1-\alpha/2)^{\frac{1}{m-1}}}$$

and obtain Rinott's constant $h = h(k_0, m, 1-\alpha)$ from the tables in Wilcox (1984) or Bechhoffer et al. (1995).

2. Take k_0 observations from each system. Calculate the first-stage sample means $\bar{Y}_i^{(1)} = \sum_{j=1}^{k_0} Y_{ij}/k_0$ and marginal sample variances

$$S_i^2 = \frac{1}{k_0 - 1} \sum_{j=1}^{k_0} (Y_{ij} - \bar{Y}_i^{(1)})^2,$$

for $i = 1, 2, \ldots, m$.

3. *Subset Selection.* Calculate the quantity

$$W_{i\ell} = t \left(\frac{S_i^2 + S_\ell^2}{k_0} \right)^{1/2}$$

for all $i \neq \ell$. Form the screening subset I, containing every alternative i such that $1 \leq i \leq m$ and

$$\bar{Y}_i^{(1)} \geq \bar{Y}_\ell^{(1)} - (W_{i\ell} - \delta)^+ \quad \text{for all } \ell \neq i.$$

4. If $|I| = 1$, then stop and return the system in I as the best. Otherwise, for all $i \in I$, compute the second-stage sample sizes

$$K_i = \max \left\{ k_0, \lceil (hS_i/\delta)^2 \rceil \right\},$$

where $\lceil \cdot \rceil$ is the ceiling function.

5. Take $K_i - k_0$ additional observations from all systems $i \in I$.

6. Compute the overall sample means $\bar{\bar{Y}}_i = \sum_{j=1}^{K_i} Y_{ij}/K_i$ for all $i \in I$. Select the system with the largest $\bar{\bar{Y}}_i$ as best.

Nelson et al. (2001) showed that any screening procedure and any two-stage procedure that satisfy certain conditions can be combined while guaranteeing the overall probability of correct selection. The Subset+Rinott procedure can handle a relatively large number of systems as several systems can be screened out after the first stage. This procedure requires running all systems at the same time, but this is sometimes difficult. Nelson et al. (2001) provide a revised version of the Subset+Rinott procedure, the Group-Screening procedure in which one can avoid simulating all the systems simultaneously. Under the Group-Screening procedure,

- the set of systems is divided into several groups of any size, and
- the groups are visited one at a time with the procedure being performed completely for each group.

If the group size is 1, then we can simulate systems one by one as we would under Rinott's procedure. When a group containing a good system is visited early in the procedures, a lot of savings can occur as the good system can eliminate many inferior systems. On the other hand, the Group-Screening procedure could be no better than Rinott's procedure if a good system is visited at the end. Thus the efficacy of the Group-Screening procedure depends on the ability to find a good system early in the procedures. Boesel et al. (2002) extended the Group-Screening procedure to the general optimization area. Now we present a procedure that allows CRN.

The $\mathcal{KN}$ procedure is *fully sequential* because it takes only a single basic observation from each alternative still in contention at each stage. Also, if there exists a clear evidence that a system is inferior, then it will be eliminated from our consideration immediately — unlike the Subset+Rinott procedure, where elimination occurs only after the first stage. As the $\mathcal{KN}$ procedure accounts for CRN and has more chances to detect inferior systems, it is expected to be more efficient than the Subset+Rinott procedure. Kim and Nelson (2001) showed that the $\mathcal{KN}$ procedure is uniformly superior to two-stage procedures with or without screening and its superiority is more noticeable as the number of systems increases.

Procedure $\mathcal{KN}$

1. *Setup.* Select confidence level $1-\alpha$, IZ parameter δ and first stage sample size $k_0 \geq 2$. Set

$$\eta = \frac{1}{2} \left[\left(\frac{2\alpha}{m-1} \right)^{-2/(k_0-1)} - 1 \right].$$

2. *Initialization.* Let $I = \{1, 2, \ldots, m\}$ be the set of systems still in contention, and let $h^2 = 2\eta(k_0 - 1)$. Obtain k_0 observations Y_{ij} $(j = 1, 2, \ldots, k_0)$ from each system i $(i = 1, 2, \ldots, m)$ and let $\bar{Y}_i(k) = k^{-1} \sum_{j=1}^{k} Y_{ij}$ denote the sample mean of the first k observations from system i.
For all $i \neq \ell$ compute

$$S_{i\ell}^2 = \frac{1}{k_0 - 1} \sum_{j=1}^{k_0} \left(Y_{ij} - Y_{\ell j} - \left[\bar{Y}_i(k_0) - \bar{Y}_\ell(k_0) \right] \right)^2,$$

the sample variance of the difference between the sample means for systems i and ℓ. Let

$$K_{i\ell} = \left\lfloor \frac{h^2 S_{i\ell}^2}{\delta^2} \right\rfloor \text{ and } K_i = \max_{\ell \neq i} K_{i\ell}.$$

Here $K_i + 1$ is the maximum number of observations that can be taken from system i.

If $k_0 > \max_i K_i$, then stop and select the system with the largest $\bar{Y}_i(k_0)$ as the best.
Otherwise, set $r = k_0$ and go to *Screening*.

3. *Screening.* Set $I^{\text{old}} = I$. Let

$$I = \left\{ i : i \in I^{\text{old}} \text{ and } \bar{Y}_i(r) \geq \bar{Y}_\ell(r) - W_{i\ell}(r), \forall \ell \in I^{\text{old}}, \ell \neq i \right\},$$

where

$$W_{i\ell}(r) = \max\left\{0, \frac{\delta}{2r}\left(\frac{h^2 S_{i\ell}^2}{\delta^2} - r\right)\right\}.$$

Notice that the *continuation region* $W_{i\ell}(r)$, shrinks monotonically as the number of replications r increases.

4. *Stopping Rule.* If $|I| = 1$, then stop and select the system whose index is in I as the best.
Otherwise, take one additional observation $Y_{i,r+1}$ from each system $i \in I$ and set $r = r + 1$.
If $r = \max_i K_i + 1$, then stop and select the system whose index is in I and has the largest $\bar{Y}_i(r)$ as the best. Otherwise, go to *Screening*.
(Notice that the stopping rule can also be $|I| = s > 1$ if it is desired to find a subset containing the best, rather than the single best.)

The $\mathcal{KN}$ procedure requires simulation of all systems simultaneously. This requirement has become less restrictive due to the development of faster computers and parallel computing environments.

5.2 Finding the Best in Steady-State Simulations

The procedures presented in the previous subsections can be applied to steady-state simulation as is if one is willing to use within-run sample means or batch means. However, as discussed in Section 4.1, the replication/deletion approach is usually inefficient. Nakayama (1995) presented single-stage multiple-comparison procedures, and Damerdji and Nakayama (1999) developed two-stage multiple-comparison procedures to select the best system for steady-state simulation. These procedures use batch means from a single sufficiently long run for each system as basic observations. The batch means method lessens the loss of data compared to the replication/deletion approach, but it can still be inefficient in fully sequential type procedures as elimination only occurs between large batches; see Goldsman et al. (2000, 2001). The last two references proposed three procedures that take a single replication from each system and use a single observation as a basic observation. One is a two-stage procedure based on Rinott's procedure and the others are the extension of $\mathcal{KN}$ procedure to steady-state simulation.

5.3 Closing Comments

Some researchers have considered completely different approaches from than IZ approach. Chen et al. (1997) proposed a procedure to find a system that maximizes the probability of correct selection under a budget constraint. Chick (1997) and Chick and Inoue (2001ab) approached this problem from a decision-theoretic point of view. Chick and Inoue (2001ab) and Inoue et al. (1999) showed that their Bayes procedures work fairly well when hundreds of systems are compared.

REFERENCES

Alexopoulos, C., and D. Goldsman. 2002. To batch or not to batch? Technical Report, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia.

Alexopoulos, C., D. Goldsman, and G. Tokol. 2001. Properties of batched quadratic-form variance parameter estimators for simulations. *INFORMS Journal on Computing* 13:149–156.

Alexopoulos, C., and A. F. Seila. 1998. Output data analysis. In *Handbook of Simulation*, ed. J. Banks, Chapter 7. New York: John Wiley & Sons.

Alexopoulos, C., A. F. Seila, and G. S. Fishman. 1997. Computational experience with the batch means method. In *Proceedings of the 1997 Winter Simulation Conference*, 194–201. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Bechhofer, R. E., T. J. Santner, and D. Goldsman. 1995. *Design and analysis of experiments for statistical selection, screening and multiple comparisons*. New York: John Wiley & Sons.

Billingsley, P. 1968. *Convergence of probability measures*. New York: John Wiley & Sons.

Boesel, J., B. L. Nelson, and S.-H. Kim. 2002. Using ranking and selection to "clean up" after simulation optimization. To appear in *Operations Research*.

Bratley, P., B. L. Fox, and L. E. Schrage. 1987. *A guide to simulation*, 2d Ed. New York: Springer-Verlag.

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. 1990. Power comparisons for the multivariate batch-means method. In *Proceedings of the 1990*

Winter Simulation Conference, 281–287. Piscataway, New Jersey: IEEE.

Charnes, J. M. 1991. Multivariate simulation output analysis. In *Proceedings of the 1991 Winter Simulation Conference*, 187–193. Piscataway, New Jersey: IEEE.

Chen, H.-C., C.-H. Chen, L. Dai, and E. Yücesan. 1997. New development of optimal computing budget allocation for discrete event simulation, In *Proceedings of the 1997 Winter Simulation Conference*, 334–341. Piscataway, New Jersey: IEEE.

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

Chick, S. E. 1997. Selecting the best system: A decision-theoretic approach. In *Proceedings of the 1997 Winter Simulation Conference*, 326–333. Piscataway, New Jersey: IEEE.

Chick, S. E., and K. Inoue. 2001a. New procedures for identifying the best simulated system using common random numbers. *Management Science* 47:1133–1149.

Chick, S. E., and K. Inoue. 2001b. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research* 49:732–743.

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.-H., 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.

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

Damerdji, H. and M. K. Nakayama. 1999. Two-stage multiple-comparison procedures for steady-state simulation. *ACM TOMACS* 9:1–30.

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. 1996. *Monte Carlo: Concepts, algorithms, and applications*. New York: Springer Verlag.

Fishman, G. S. 2001. *Discrete-event simulation*. New York: Springer Verlag.

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., S.-H. Kim, W. S. Marshall, and B. L. Nelson. 2000. Ranking and selection for steady-state simulation. In *Proceedings of the 2000 Winter Simulation Conference*, 544–553, Piscataway, New Jersey: IEEE.

Goldsman, D., S.-H. Kim, W. Marshall, and B. L. Nelson. 2001. Ranking and selection procedures for steady-state simulation: Perspectives and procedures. *INFORMS Journals on Computing* 14:2–19.

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 B. L. Nelson. 1998. Comparing systems via simulation. In *Handbook of Simulation*, ed. J. Banks, Chapter 8. New York: John Wiley & Sons.

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* 41:171–187.

Gupta, S. S. 1965. On some multiple decision (selection and ranking) rules. *Technometrics* 7:225–245.

Gupta, S. S., and D.-Y. Huang. 1976. Subset selection procedures for the means and variances of normal populations: Unequal sample sizes case, *Sankhyā B*, 38:112–128.

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

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.

Inoue, K., S. E. Chick, and C.-H. Chen. 1999. An empirical evaluation of several methods to select the best system. *ACM TOMACS* 9:381–407.

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

Kim, S-H., and B. L. Nelson. 2001. A fully sequential procedure for indifference-zone selection in simulation. *ACM TOMACS* 11:251–273.

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. 2000. *Simulation modeling and analysis*, 3d Ed. McGraw-Hill, New York.

Malkovich, J. F., and A. A. Afifi. 1973. On tests for multivariate normality. *Journal of the American Statistical Association* 68:176–179.

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*, 227–230. Piscataway, New Jersey: IEEE.

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.

Nakayama, M. K. 1995. Selecting the best system in steady-state simulations using batch means. In *Proceedings of the 1995 Winter Simulation Conference*, 362–366. Piscataway, New Jersey: IEEE.

Nelson, B. L., and F. J. Matejcik. 1995. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Science* 41:1935–1945

Nelson, B. L., J. Swann, D. Goldsman, and W.-M. T. Song. 2001. Simple procedures for selecting the best system when the number of alternatives is large. *Operations Research* 49:950–963.

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.

Rinott, Y. 1978. On two-stage selection procedures and related probability-inequalities. *Communications in Statistics — Theory and Methods A* 7:799–811.

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

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*, 513–525. Piscataway, New Jersey: IEEE.

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.

Song, W.-M. T., and B. W. Schmeiser. 1993. Optimal mean-squared-error batch sizes. *Management Science* 41:110–123.

Steiger, N. M., and J. R. Wilson. 2001. Convergence properties of the batch-means method for simulation output analysis. *INFORMS Journal on Computing* 13(4): 277–293.

Steiger, N. M., and J. R. Wilson. 2002a. An improved batch means procedure for simulation output analysis. Technical Report, Department of Industrial Engineering, North Carolina State University, Raleigh, North Carolina.

Steiger, N. M., E. K. Lada, J. R. Wilson, C. Alexopoulos, D. Goldsman, F. Zouaoui. 2002b. ASAP2: An improved batch means procedure for simulation output analysis. In *Proceedings of the 2002 Winter Simulation Conference*, this volume. Piscataway, New Jersey: IEEE.

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

von Neumann, J. 1941b. The mean square difference. *Annals of Mathematical Statistics* 12:153–162.

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

Welch, P. D. 1987. On the relationship between batch means, overlapping batch means and spectral estima-

tion, In *Proceedings of the 1987 Winter Simulation Conference*, 320–323. Piscataway, New Jersey: IEEE.

Wilcox, R. R. 1984. A table for Rinott's selection procedure. *Journal of Quality Technology* 16:97–100.

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. dissertation, Department of Operations Research, University of North Carolina, Chapel Hill, North Carolina.

AUTHOR BIOGRAPHIES

CHRISTOS ALEXOPOULOS is an Associate Professor in the School of Industrial and Systems Engineering at the Georgia Institute of Technology. He received his Ph.D. in Operations Research from the University of North Carolina at Chapel Hill. His research interests are in the areas of applied probability, statistics, and optimization of stochastic systems. His recent work involves problems related to the optimal design of telecommunications and transportation systems. Dr. Alexopoulos is a member of INFORMS and the INFORMS College of Simulation. He served as a Co-Editor for the *Proceedings of the 1995 Winter Simulation Conference*. His email address is <christos@isye.gatech.edu>, and his web page is <www.isye.gatech.edu/~christos>.

SEONG-HEE KIM is an Assistant Professor in the School of Industrial and Systems Engineering at the Georgia Institute of Technology. She received her Ph.D. in Industrial Engineering and Management Sciences from Northwestern University. Her research interests include ranking and selection procedures and simulation output analysis of stochastic systems. She is a member of INFORMS. Her email address is <skim@isye.gatech.edu>.