*Proceedings of the 2005 Winter Simulation Conference*
*M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.*

# REVIEW OF ADVANCED METHODS FOR SIMULATION OUTPUT ANALYSIS

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 methods based on standardized time series. 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 (IID) 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.

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

**Finite-horizon simulations**. In this case the simulation starts in a specific 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 vehicle storage and distribution facility over 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 finite-horizon simulations. Section 3 presents techniques for point and interval estimation of steady-state parameters. Section 4 reviews recent methods for identifying the best system among 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 $\mu \equiv \mathrm{E}(\bar{X}_n)$, where $\bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean 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 $i$th hour. Clearly, $\bar{X}_n$ is an unbiased estimator for $\mu$. Unfortunately, the $X_i$ are generally dependent random variables making the estimation of the variance $\mathrm{Var}(\bar{X}_n)$ a nontrivial problem. Let $S_n^2(X) \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ be the sample variance of the data. The presence of autocorrelation makes the familiar estimator $S_n^2(X)/n$ a biased estimator of $\mathrm{Var}(\bar{X}_n)$. In particular, if the $X_i$'s are positively correlated, one has $\mathrm{E}(S_n^2(X)/n) < \mathrm{Var}(\bar{X}_n)$ (see Section 3).

To overcome this problem, one can run $k$ independent replications of the system simulation. Assume that run $i$ produces the output data $X_{i1}, X_{i2}, \ldots, X_{in}$. Then the replicate averages $Y_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ are IID random variables, their sample mean $\bar{Y}_k = \frac{1}{k}\sum_{i=1}^{k} Y_i$ is also an unbiased estimator of $\mu$, and their sample variance $\hat{V}_R = S^2(Y)$ is an unbiased estimator of $\mathrm{Var}(\bar{X}_n)$. If in addition $k$ is sufficiently large,

an approximate $1-\alpha$ confidence interval (CI) for $\mu$ is

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

where $t_{d,\delta}$ is the $\delta$-quantile of Student's $t$ 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 $\mu$ with a fixed absolute or relative precision. The procedure for constructing a $1-\alpha$ CI for $\mu$ with a small absolute error $|\bar{Y}_k - \mu| \le \beta$ is based on Chow and Robbins (1965). It starts with $k \ge 5$ 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 $k \ge 10$ 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 independent replications can also be used for estimating performance measures other than means. Let $Y$ be the total cost incurred in an inventory syetem during a certain time window, and let $y_p \equiv \inf\{y : \Pr(Y \le y) \ge p\}$ denote the $p$-quantile of $Y$. Suppose that we run $k$ replications. Let $Y_i$ be the cost observed during replication $i$, and let $Y_{(1)} < Y_{(2)} < \cdots < Y_{(k)}$ be the order statistics corresponding to the $Y_i$. Then a point estimate for $y_p$ is $\hat{\xi}_p = Y_{(kp)}$ if $kp$ is an integer or $\hat{y}_p = Y_{(\lfloor kp+1 \rfloor)}$ otherwise ($\lfloor\cdot\rfloor$ is the floor function). A CI for $y_p$ is described in Alexopoulos and Seila (1998, Section 7.3.2).

## 3 STEADY-STATE ANALYSIS

We focus on methods for computing point and interval estimators for the mean of a discrete-time stationary process. Analogous methods for analyzing continuous-time output data are described in a variety of texts (Fishman 2001; Law and Kelton 2000). The process $X = \{X_i : i \in \mathbb{Z}\}$ 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 $\mathrm{E}(X_i) = \mu$, $\mathrm{Var}(X_i) \equiv \sigma_X^2 < \infty$ for all $i$, and the $\mathrm{Cov}(X_i, X_{i+j})$ is independent of $i$, then $X$ is called *weakly stationary*. We denote the autocovariance function of $X$ by $R_j \equiv \mathrm{Cov}(X_1, X_{1+j})$ ($j = 0, \pm1, \pm2, \ldots$). Notice that $R_0 = \sigma_X^2$.

Clearly, the sample mean $\bar{X}_n$ is not only unbiased for $\mu$, but also strongly consistent by the ergodic theorem (see Durrett 2005). Under the assumption that $\bar{X}_n$ is approximately normally distributed (which is reasonable for sufficiently large $n$), the usual construction of a CI for $\mu$ requires the derivation of an estimator for $\mathrm{Var}(\bar{X}_n)$. A little algebra yields (Anderson 1984),

$$\mathrm{E}\left[\frac{S_n^2(X)}{n}\right] = \frac{\frac{n}{a_n} - 1}{n-1}\mathrm{Var}(\bar{X}_n), \tag{2}$$

where $a_n = 1 + (2/\sigma_X^2)\sum_{j=1}^{n-1}(1 - j/n)R_j$. Then for processes that are positively correlated ($R_i > 0$), equation (2) implies that $\mathrm{E}[S_n^2(X)/n] < \mathrm{Var}(\bar{X}_n)$. Hence the "classical" $1-\alpha$ CI for IID data $\bar{X}_n \pm t_{n-1,1-\alpha/2}\frac{S_n(X)}{\sqrt{n}}$ can have coverage probability that can be considerably below the nominal value $1-\alpha$.

A common assumption facilitating the derivation of a CI for $\mu$ is as follows:

**Functional Central Limit Theorem (FCLT) Assumption**. Suppose that the series

$$\sigma^2 \equiv \sigma_X^2 + 2\sum_{j=1}^{\infty} R_j \tag{3}$$

is absolutely convergent and $\sigma^2 > 0$. Let

$$X_n(t) \equiv \frac{\lfloor nt \rfloor(\bar{X}_{\lfloor nt \rfloor} - \mu)}{\sigma\sqrt{n}}, \quad t \ge 0.$$

Then $X_n(\cdot) \Longrightarrow \mathcal{W}$, where $\{\mathcal{W}(t) : t \ge 0\}$ is a standard Brownian motion process. We call $\sigma^2$ the *(asymptotic) variance parameter* of $X$.

This assumption holds under several conditions (see Durrett 2005). Examples are a condition involving conditional second moments of $X$ and the stronger $\varphi$-mixing condition: $X$ is *$\varphi$-mixing* if there are $\varphi_k \downarrow 0$ such that, for each $k \ge 0$, $A \in \mathcal{F}_{-\infty}^{j}$, and $B \in \mathcal{F}_{j+k}^{\infty}$,

$$|\Pr(A \cap B) - \Pr(A)\Pr(B)| \le \varphi_k \Pr(A).$$

Here $\mathcal{F}_i^j$ ($i \le j$) denotes the $\sigma$-field generated by $X_i, X_{i+1}, \ldots, X_j$.

**Remark 1** Contrary to popular belief, many stochastic processes encountered in simulation output analysis are *not* $\varphi$-mixing. Examples are autoregressive processes, regenerative processes (see Section 3.3) with regenerations not occurring uniformly fast over the state space, and virtually all open queueing networks (Glynn and Iglehart 1985).

The variance of the sample mean in terms of the autocovariance function is

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n}\left[\sigma_X^2 + 2\sum_{j=1}^{n-1}(1 - j/n)R_j\right]. \tag{4}$$

Assumption $0 < \sigma^2 < \infty$ along with equation (4) imply $\lim_{n\to\infty} n\mathrm{Var}(\bar{X}_n) = \sigma^2$ and $\lim_{n\to\infty}\mathrm{Var}(\bar{X}_n) = 0$; hence $\bar{X}_n$ is also consistent (in mean square error). Our focus

will be on methods for obtaining CIs for $\mu$, which involve estimating $\sigma^2$.

Finally, the "little-oh" notation $f(m) = o(g(m))$ means that $f(m)/g(m) \to 0$ as $m \to \infty$; and the "big-oh" notation $f(m) = O(g(m))$ means that there is a positive integer $m_0$ such that $|f(m)/g(m)| \leq C$ for some constant $C$ and all $m \geq m_0$.

### 3.1 Dealing with the Initial Conditions

Several problems arise when the process $X$ does not start in steady-state. For example, $\bar{X}_n$ is not an unbiased estimator of the mean $\mu$. The removal of the effect of the initial conditions is a challenging problem.

The most commonly used method for eliminating the bias of $\bar{X}_n$ identifies an index $l$ and *truncates* the observations $X_1, \ldots, X_l$. Several procedures have been proposed for the detection of a cutoff index $l$ (see Fishman 2001; Law and Kelton 2000; Ockerman 1995; Wilson and Pritsker 1978ab).

The graphical procedure of Welch (1983) uses $k$ independent replications, with the $i$th replication producing observations $X_{i1}, X_{i2}, \ldots, X_{in}$, and computes the "across-runs" averages $\bar{X}_j = \frac{1}{k}\sum_{i=1}^{k} X_{ij}$, $j = 1, \ldots, n$. Then for a given time window $w$, the procedure plots the *moving averages*

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

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 long tails (see Alexopoulos and Seila 1998, Example 7).

### 3.2 The Replication/Deletion Approach

This intuitive approach runs $k$ independent replications, each of length $l + n$ observations, and discards the first $l$ observations from each run. (For example, $l$ can be chosen using Welch's method from Section 3.1.) One then uses the IID sample means $Y_i(l, n) = \frac{1}{n}\sum_{j=l+1}^{l+n} X_{ij}$ from the $k$ runs to compute the point estimate $\bar{Y}_k(l, n) = \frac{1}{k}\sum_{i=1}^{k} Y_i(l, n)$ and the following approximate $1 - \alpha$ CI for $\mu$:

$$\bar{Y}_k(l, n) \pm t_{k-1,1-\alpha/2}\sqrt{\hat{V}_R(l, n)/k}, \tag{5}$$

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

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 be 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 CI (5) is asymptotically valid 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 CI, the truncation index $l$ must increase faster than the logarithm of $k$ for the CI to achieve the nominal coverage. This requirement is hard to implement in practice. (e) This method 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 the aforementioned issues. Alexopoulos and Goldsman (2004) present a thorough comparison between the methods of independent replications and batch means and identify additional shortcomings of the method of independent replications.

### 3.3 The Regenerative Method

This method assumes the identification of time indices at which the process $X$ probabilistically *starts over* and uses these regeneration epochs for obtaining IID random variables which can be used for computing point and interval estimates for the mean $\mu$. The method was proposed by Crane and Iglehart (1975) and Fishman (1973, 1974). More precisely, assume that there are (random) time indices $T_1 < T_2 < \cdots$ 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 $\mathrm{E}(Z_i) < \infty$. Then the steady-state mean $\mu$ is given by $\mu = \mathrm{E}(Y_1)/\mathrm{E}(Z_1)$.

Now suppose that one simulates the process $X$ 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 $\mu$. Furthermore, CIs for $\mu$ can be constructed by using the IID 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.

### 3.4 The Batch Means Method

The method of nonoverlapping batch means (NBM) is a popular approach for computing point and CI estimators for the mean $\mu$ of a stationary process. Original accounts on the method were given by Conway (1963), Fishman (1978), and Law and Carson (1979); see Alexopoulos and Goldsman (2004) and Fishman (2001) for detailed coverage.

Suppose that the sample $X_1, \ldots, X_n$ is divided into $k$ contiguous batches, each consisting of $m$ observations (for simplicity, we assume $n = km$). For $i = 1, \ldots, k$, the $i$th batch consists of the observations $X_{(i-1)m+1}, X_{(i-1)m+2}, \ldots, X_{im}$ and the $i$th batch mean $Y_{i,m} = \frac{1}{m}\sum_{j=1}^{m} X_{(i-1)m+j}$ is the sample average from batch $i$. The NBM-based estimator of the mean is the grand sample mean

$$\bar{X}_n = \frac{1}{k}\sum_{i=1}^{k} Y_{i,m} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Clearly, the stationarity of $X$ implies $\mathrm{E}(\bar{X}_n) = \mu$ and the stationarity of the batch means sequence $\{Y_{i,m} : i \geq 1\}$.

The motivation behind the NBM method is simple. First, under the FCLT, one can show that as $m \to \infty$, the batch means become uncorrelated (Law and Carson 1979) and normally distributed. Since the grand mean $\bar{X}_n$ is the sample average of the batch means, one has the approximation

$$n\mathrm{Var}(\bar{X}_n) \doteq n\mathrm{Var}(Y_{1,m})/k = m\mathrm{Var}(Y_{1,m}).$$

Hence the NBM estimator for $\sigma^2$ is

$$\hat{V}_B(k,m) \equiv \frac{m}{k-1}\sum_{i=1}^{k}(Y_{i,m} - \bar{X}_n)^2, \tag{6}$$

which is $m$ times the sample variance of the batch means. An approximate $1-\alpha$ CI for $\mu$ is

$$\bar{X}_n \pm t_{k-1,1-\alpha/2}\sqrt{\frac{\hat{V}_B(k,m)}{n}}. \tag{7}$$

Of course, the fundamental issue is the choice of the batch size and the number of batches. Several early studies (e.g., Fishman 1978; Schmeiser 1982) addressed this issue, but without the rigor of recent studies.

To motivate the description of the modern procedures, we focus on the mean square error (MSE) of $\hat{V}_B(k,m)$ and the coverage of the CI (7). Here we let $\sigma_n^2 \equiv n\mathrm{Var}(\bar{X}_n)$, and define the "center of gravity" constant

$$\gamma \equiv -\sum_{j=-\infty}^{\infty} jR_j = -2\sum_{j=1}^{\infty} jR_j. \tag{8}$$

One can show that

$$\mathrm{E}[\hat{V}_B(k,m)] = \frac{1}{k-1}(k\sigma_m^2 - \sigma_n^2).$$

If in addition $\mathrm{E}(X_1^4) < \infty$, and the process $X$ is $\varphi$-mixing with $\varphi_j = O(j^{-4-\epsilon})$ for some $\epsilon > 0$, then $\gamma$ exists and

$$\sigma_n^2 = \sigma^2 + \gamma/n + o(1/n). \tag{9}$$

Combining the last two equations we obtain

$$\mathrm{E}[\hat{V}_B(k,m)] = \sigma^2 + (k+1)\gamma/n + o(1/n). \tag{10}$$

Hence, $\hat{V}_B(k,m)$, usually has negative first-order bias for positively autocorrelated processes.

Also, the additional assumptions $\mathrm{E}(X_1^{12}) < \infty$ and $\varphi_j = O(j^{-9})$ allow one to write

$$\begin{aligned}\mathrm{Var}[\hat{V}_B(k,m)] &= \frac{2\sigma^4(k+1)}{(k-1)^2} + O(1/(km^{1/4})) \\ &\quad + O(1/k^2)\end{aligned} \tag{11}$$

(Chien, et al. 1997). Then the MSE of the variance estimator $\hat{V}_B(k,m)$ has the form

$$\mathrm{MSE}[\hat{V}_B(k,m)] = O(1/(km^{1/4})) + O(1/k^2) \to 0, \tag{12}$$

as $m, k \to \infty$. Property (12) implies weak consistency for the estimator $\hat{V}_B(k,m)$, but does not guarantee the asymptotic validity of the CI in equation (7). Before we discuss batching rules that yield the last property, we briefly examine how the variance estimator $\hat{V}_B(k,m)$ approaches $\sigma^2$. As in Fishman (2001, p. 251), equation (10) allows us to write

$$\begin{aligned}\hat{V}_B(k,m) - \sigma^2 = {} & \underbrace{\sigma_n^2 - \sigma^2}_{\substack{\text{error due to} \\ \text{finite } n}} \\ & - \underbrace{\sigma_n^2\frac{1-\sigma_m^2/\sigma_n^2}{1-m/n}}_{\substack{\text{error due to} \\ \text{ignoring correlations} \\ \text{between batch means}}} + \underbrace{\epsilon_n}_{\substack{\text{error due to} \\ \text{random sampling}}},\end{aligned} \tag{13}$$

where the error $\epsilon_n$ has mean zero and variance given by equation (11). We call the first two terms on the right-

hand side of equation (13) a systematic error; by equation (9) this error behaves as $O(1/m)$. On the other hand, equation (11) implies that the standard deviation of $\epsilon_n$ behaves as $O(k^{-1/2})$. These growth rates reveal the tradeoff between the two types of error induced by $k$ and $m$. Since $\sigma_n^2$ approaches $\sigma^2$ from below for a variety of systems with positive autocorrelation functions, the systematic error induces a negative bias in $\hat{V}_B(k,m)$ that dissipates as the batch size increases. Then the error due to random sampling fluctuates around zero and decreases at rate $O(k^{-1/2})$.

The recent literature contains a variety of rules for selecting batch sizes $\{m_\ell\}$ and batch counts $\{k_\ell\}$ as the sample size increases. The most intuitive rule fixes the number of batches and doubles the batch size at each iteration. This assignment is computationally attractive because at every iteration, pairs of existing batch means are averaged to compute the new batch means.

**Fixed Number of Batches (FNB) Rule.** Start with $k$ batches of size $m_1$. At stage $\ell \geq 2$, use batch size $m_\ell = 2m_{\ell-1}$ and sample size $n_\ell = km_\ell$.

Under the FCLT assumption, one can show that for fixed $k$ and $m \to \infty$, $\hat{V}_B(k,m) \xrightarrow{d} \sigma^2 \chi^2_{k-1}/(k-1)$, where $\chi^2_d$ denotes a chi-square random variable with $d$ degrees of freedom; and the CI in equation (7) is asymptotically valid (Glynn and Whitt 1991). If we assume uniform integrability for $\hat{V}_B^2(k,m)$ (see Billingsley 1968), we have $\lim_{m\to\infty} \mathrm{E}[\hat{V}_B(k,m)] = \sigma^2$ and $\lim_{m\to\infty} \mathrm{Var}[\hat{V}_B(k,m)] = 2\sigma^4/(k-1)$; hence the FNB rule does not yield a consistent variance estimator. This is in agreement with equation (13) as the error $O(k^{-1/2})$ due to random sampling does not diminish. Therefore the CI in equation (7) tends to be wider than CIs based on consistent variance estimators.

### 3.5 Consistent Batch Means Estimation Methods

Alternative rules that yield strongly consistent estimators for $\hat{V}_B(k,m)$ are based on the following assumption:

**Assumption of Strong Approximation (ASA).** There exists a constant $\lambda \in (0, 1/2]$ and a finite random variable $C$ such that, as $n \to \infty$,

$$|\sqrt{n}(\bar{X}_n - \mu)/\sigma - \mathcal{W}(n)/\sqrt{n}| \leq Cn^{-\lambda}, \quad \text{w.p.1},$$

where $\mathcal{W}$ is a standard Brownian motion process defined on the same space as the standardized process $\{\bar{X}_n\}$.

A $\lambda$ close to $1/2$ indicates a marginal normal distribution and low correlation among the $X_i$. Conversely, a $\lambda$ close to zero indicates the absence of at least one of these properties (Philipp and Stout 1975). The following theorem proposes batching assumptions which along with ASA yield a strongly consistent estimator for $\sigma^2$. (Notice that the batching sequences are indexed by the sample size.)

**Theorem 1** *(Damerdji 1994a) Suppose that the ASA holds and that $\{m_n\}$ and $\{k_n\}$ are deterministic sequences of batch sizes and batch counts, respectively, such that $m_n \to \infty$, $k_n \to \infty$, $n^{1-2\lambda}\ln(n)/m_n \to 0$ (as $n \to \infty$), and $\sum_{n=1}^{\infty} k_n^{-q} < \infty$ for some finite integer $q \geq 1$. Then, as $n \to \infty$, $\hat{V}_B(k_n, m_n) \to \sigma^2$, w.p.1 and*

$$Z(k_n, m_n) \equiv \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\hat{V}_B(k_n, m_n)}} \xrightarrow{d} N(0,1), \qquad (14)$$

*where $N(0,1)$ is a standard normal random variable.*

Suppose that $m_n \doteq n^\theta$, for some $\theta \in (0,1)$. One can verify that the conditions of Theorem 1 are satisfied if $\theta \in (1-2\lambda, 1)$. In particular, the square root (SQRT) rule that uses $m_n \doteq k_n \doteq \sqrt{n}$ ($\theta = 1/2$) yields a strongly consistent variance estimator when $1/4 < \lambda < 1/2$. In addition to the derivation of a strongly consistent estimator for $\sigma^2$, the SQRT rule induces an optimal property: Assuming that $\mathrm{E}(X_1^{20}) < \infty$ and that $X$ is $\varphi$-mixing with $\varphi_j = O(j^{-13})$, Chien (1989) showed that the CDF of the standardized statistic $Z(k,m)$ converges to the standard normal CDF at the fastest possible rate. Unfortunately, the CIs for $\mu$ that result from an implementation of the SQRT rule often exhibit low coverage for small sample sizes (see Example 11 in Alexopoulos and Seila 1998).

Although both the FNB and SQRT rules yield asymptotically valid CIs for $\mu$, each has desirable properties and limitations. To close the "gap," Fishman and Yarberry (1997) proposed the LABATCH.2 suite of algorithms. Among the two recommended algorithms, LBATCH and ABATCH, we present the latter because it is more conservative with regard to the coverage of the resulting CI (7). This method uses von Neumann's test (von Neumann 1941ab; Young 1941) to assess the hypothesis $H_0$: "the batch means are independent." The associated test statistic is

$$\Gamma(k,m) \equiv \sqrt{\frac{k^2-1}{k-2}}\left[1 - \frac{\sum_{i=1}^{k-1}(Y_{i,m} - Y_{i+1,m})^2}{2\sum_{i=1}^{k}(Y_{i,m} - \bar{X}_n)^2}\right].$$

Assume that the hypothesis $H_0$ is true. If the batch means are normally distributed, the distribution of $\Gamma(k,m)$ is very close to $N(0,1)$ for $k$ as small as 8. On the other hand, if the batch means are nonnormal, the first four cumulants of $\Gamma(k,m)$ converge to the respective cumulants of the $N(0,1)$ distribution as $k \to \infty$. Hence, under $H_0$, $\Gamma(k,m) \approx N(0,1)$ for large $m$ (the batch means become approximately normal) or large $k$. To guard against positive correlation, one can use a one-sided test and reject $H_0$ at level $\beta$ when $\Gamma(k,m) > z_{1-\beta}$, where $z_\delta$ is the $\delta$-quantile of the standard normal distribution.

The ABATCH algorithm evolves as follows. For a complete description, see Fishman (2001).

**Algorithm ABATCH**

- Select initial batch size $m_1$, initial batch count $k_1$, confidence level $1-\alpha$, and type I error $\beta$ for von Neumann's test.
- On iteration $\ell \geq 1$: Compute von Neumann's statistic $\Gamma(k_\ell, m_\ell)$. If $\Gamma(k_\ell, m_\ell) > z_{1-\beta}$, reject $H_0$ and use the FNB rule on iteration $\ell+1$. Otherwise, use the SQRT rule on iteration $\ell+1$.

Since the ABATCH algorithm uses random $m_\ell$ and $k_\ell$, Theorem 6.6 of Fishman (2001) lists conditions that imply strong consistency for $\hat{V}_B(k_\ell, m_\ell)$ and asymptotic validity for the CI $\bar{X}_n \pm z_{1-\alpha/2}\sqrt{\hat{V}_B(k_\ell, m_\ell)/n_\ell}$. The FNB and SQRT rules can be implemented easily within the ABATCH algorithm by setting $\beta = 0$ or $\beta = 1$, respectively. Two features of the LABATCH.2 suite that are often overlooked are algorithm efficiency and low space requirements: each algorithm requires $O(n)$ *total* time and $O(\log_2 n)$ space. Although similar complexities are known for static fixed-batch-size algorithms (e.g., all the methods in the remainder of this paper have a linear time complexity per iteration), the dynamic setting of ABATCH 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, in linear total time. This assessment is essential to gauge the quality of the variance parameter estimates and the CI for the mean. 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> .

Steiger et al. (2004) proposed an alternative sequential NBM approach, ASAP3, that delivers a CI for $\mu$ that satisfies user-specified requirements on absolute or relative precision as well as coverage probability. This approach takes advantage of the fact that the batch means often become approximately multivariate normal random variables before achieving independence. ASAP3 operates as follows: the batch size is progressively increased until the batch means pass the Shapiro-Wilk test for multivariate normality; and then ASAP3 fits a first-order autoregressive (AR(1)) time series model to the batch means. If necessary, the batch size is further increased until the autoregressive parameter in the AR(1) model does not significantly exceed 0.8. Next ASAP3 computes the terms of an inverted Cornish-Fisher expansion for the classical batch means $t$-ratio based on the AR(1) parameter estimates; and finally ASAP3 delivers a correlation-adjusted CI based on this expansion. Although ASAP3 does not possess the computational efficiency of the LABATCH.2 algorithms, it performs very well with regard to conformance to the precision and coverage probability requirements as well as with regard to the mean and variance of the half-length of the delivered CI. Related papers, experimental results, and the ASAP3 software are accessible from the site <http://www.ie.ncsu.edu/jwilson>.

**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 $m$, this method uses all $n-m+1$ overlapping batches to estimate $\mu$ and $\mathrm{Var}(\bar{X}_n)$. The first batch consists of observations $X_1, \ldots, X_m$, the second batch consists of $X_2, \ldots, X_{m+1}$, etc. The OBM estimator of $\mu$ is

$$\bar{Y}_O = \frac{1}{n-m+1}\sum_{i=1}^{n-m+1} Y'_{i,m},$$

where $Y'_{i,m} = \frac{1}{m}\sum_{j=i}^{i+m-1} X_j$ $(i = 1, \ldots, n-m+1)$ are the respective batch means. The OBM-based estimator of $\sigma^2$ is

$$\hat{V}_O(k,m) = \frac{nm}{(n-m+1)(n-k)}\sum_{i=1}^{n-m+1}(Y'_{i,m} - \bar{X}_n)^2,$$

with $k \equiv n/m$. The OBM variance estimator is almost identical to Bartlett's spectral estimator (see Anderson 1984).

Under conditions similar to those required to derive equations (10) and (11) one has (Song and Schmeiser 1995)

$$\mathrm{E}[\hat{V}_O(k,m)] = \sigma^2 + \gamma/m + o(1/m) \qquad (15)$$

and, as $m \to \infty$,

$$\mathrm{Var}[\hat{V}_O(k,m)] \to \frac{2(2k^2-3k-3)\sigma^4}{3(k-1)^2} \doteq \frac{4\sigma^4}{3k}. \qquad (16)$$

Equations (10) and (15) show that the estimators $\hat{V}_B(k,m)$ and $\hat{V}_O(k,m)$ have the same asymptotic means (as $k, m \to \infty$). However a comparison between equations (11) and (16) reveals that

$$\frac{\mathrm{Var}[\hat{V}_O(k,m)]}{\mathrm{Var}[\hat{V}_B(k,m)]} \to \frac{2}{3}, \quad \text{as } k, m \to \infty.$$

Thus, the OBM method gives better (asymptotic) performance than NBM with regard to MSE. Also, the behavior of $\mathrm{Var}[\hat{V}_O(k,m)]$ appears to be less sensitive to the choice of the batch size than does the behavior of $\mathrm{Var}[\hat{V}_B(k,m)]$ (see Song and Schmeiser 1995, Table 1).

An approximate $1-\alpha$ CI for $\mu$ is $\bar{X}_n \pm t_{d,1-\alpha/2}\sqrt{\hat{V}_O(k,m)/n}$, with the degrees of freedom $d$ chosen so that $\hat{V}_O(k,m)$ is asymptotically $\sigma^2\chi^2_d/d$. Meketon and Schmeiser (1984) use the value $d = 1.5(k-1)$ whereas,

based on Monte Carlo studies, Schmeiser (1986) recommends the larger value $d = 1.5(k-1)[1+(k-1)^{-0.5-0.6k}]$.

The OBM method can also yield a consistent variance estimator. If $X$ satisfies ASA, and the deterministic sequences satisfy the assumptions of Theorem 1 and $\lim_{n\to\infty}(k_n^2/n) = 0$, then $\mathrm{Var}[\hat{V}_O(k_n, m_n)] \to \sigma^2$, w.p.1 (Damerdji 1994a).

Using equations (15) and (16), one can show that for a sample size $n$, the batch size that minimizes the $\mathrm{MSE}[\hat{V}_O(k, m)]$ is given by

$$m^* = \left(\frac{3\gamma^2 n}{2\sigma^4}\right)^{1/3}. \tag{17}$$

Song (1996) developed methods for estimating the ratio $\gamma^2/\sigma^4$ for a variety of processes, including moving average processes and autoregressive processes. Then one can obtain an estimator for $m^*$ by plugging the ratio estimator into equation (17).

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.

### 3.6 Methods Based on Standardized Time Series

Now we turn to estimators based on standardized time series (STS). We start with estimators based on the entire sample, and then present estimators based on standardized time series applied to batches.

The STS for the sample $X_1, \ldots, X_n$ is formed as follows (see Schruben 1983): One defines $D_{0,n} \equiv 0$ and $D_{i,n} \equiv \bar{Y}_i - \bar{Y}_n$, for $i = 1, \ldots, n$; scales the sequence $\{D_{i,n}\}$ by $i/(\sigma\sqrt{n})$; and then scales the time index $i$ of the resulting sequence to the unit interval by setting $t = i/n$. The resulting STS is

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

If $X$ satisfies a FCLT, it can be shown that, as $n \to \infty$,

$$(\sqrt{n}(\bar{X}_n - \mu), \sigma T_n) \Longrightarrow (\sigma\mathcal{W}(1), \sigma\mathcal{B}), \tag{18}$$

where $\mathcal{B}$ is a standard Brownian bridge process on $[0, 1]$ defined by $\mathcal{B}(t) = \mathcal{W}(t) - t\mathcal{W}(1)$ (for a set of sufficient conditions, see Glynn and Iglehart 1990). In addition, the STS $T_n(\cdot)$ is asymptotically independent of $\bar{X}_n$. Recall that all finite-dimensional joint distributions of $\mathcal{B}$ are normal with $\mathrm{E}(\mathcal{B}(t)) = 0$ and $\mathrm{Cov}(\mathcal{B}(s), \mathcal{B}(t)) = \min(s, t) - st$, $0 \le s, t \le 1$.

#### The Weighted Area Estimator

We start with the weighted area estimator (Goldsman et al. 1990; Goldsman and Schruben 1990; Schruben 1983). Suppose that the function $f$ is continuous on the interval $[0, 1]$ and normalized so that $\mathrm{Var}(\int_0^1 f(t)\mathcal{B}(t)\,dt) = 1$. Then $\int_0^1 f(t)\mathcal{B}(t) \sim \sigma N(0, 1)$. The square of the weighted area under the STS is defined by

$$A(f; n) \equiv \left[\frac{1}{n}\sum_{i=1}^{n} f(i/n)\sigma T_n(i/n)\right]^2.$$

Under mild conditions, the continuous mapping theorem (see Billingsley 1968, Theorem 5.1) implies

$$A(f; n) \xrightarrow{d} A(f) \equiv \left[\int_0^1 f(t)\sigma\mathcal{B}(t)\,dt\right]^2 \sim \sigma^2\chi_1^2,$$

as $n \to \infty$. For this reason, we call $A(f; n)$ the *weighted area* estimator for $\sigma^2$.

The following theorem gives expressions for the mean and variance of the weighted area estimator.

**Theorem 2** *(Foley and Goldsman 2000; Goldsman et al. 1990) Suppose that $X$ is $\varphi$-mixing and satisfies a FCLT, the constant $\gamma$ in equation (8) exists, and $A^2(f; n)$ is uniformly integrable. Then, as $n \to \infty$,*

$$\mathrm{E}[A(f; n)] = \sigma^2 + \frac{[(F(1) - \bar{F}(1))^2 + \bar{F}^2(1)]\gamma}{2n} + o(1/n)$$

*and*

$$\mathrm{Var}[A(f; n)] \to \mathrm{Var}[A(f)] = \mathrm{Var}(\sigma^2\chi_1^2) = 2\sigma^4,$$

*where $F(s) \equiv \int_0^s f(t)\,dt$, $0 \le s \le 1$, and $\bar{F}(u) \equiv \int_0^u F(s)\,ds$, $0 \le u \le 1$.*

Notice that the limiting variance does not depend on the weight function $f$.

**Example 1** Schruben (1983) studied the area estimator with constant weight function $f_0(t) \equiv \sqrt{12}$, for $t \in [0, 1]$; in this case, Theorem 2 implies that $\mathrm{E}[A(f_0; n)] = \sigma^2 + 3\gamma/n + o(1/n)$.

If one chooses weights having $F(1) = \bar{F}(1) = 0$, the resulting estimator is *first-order unbiased* for $\sigma^2$, i.e., its bias is $o(1/n)$. An example of a weight function yielding a first-order unbiased estimator for $\sigma^2$ is $f_2(t) \equiv \sqrt{840}(3t^2 - 3t + 1/2)$ (Goldsman et al. 1990; Goldsman and Schruben 1990).

Other examples of weight functions yielding first-order unbiased estimators for $\sigma^2$ are given by the family $f_{\cos,j}(t) = \sqrt{8}\pi j\cos(2\pi jt)$, $j = 1, 2, \ldots$. Foley and Goldsman (2000)

showed that this "orthonormal" sequence of weights produces area estimators $A(f_{\cos,1}, n)$, $A(f_{\cos,2}, n), \ldots$ that are not only first-order unbiased, but asymptotically independent; that is, $A(f_{\cos,1}), A(f_{\cos,2}), \ldots$ are IID $\sigma^2\chi_1^2$.

**Batched Area Estimators**

Up to now, the STS-based variance estimators have been constructed directly from a single long run of $n$ observations. We now examine what happens if we (a) divide the run into contiguous, nonoverlapping *batches*; (b) form an STS estimator from each batch; and (c) take the average of the estimators.

The STS from batch $i$ $(i = 1, \ldots, k)$ is

$$T_{i,m}(t) \equiv \frac{\lfloor mt \rfloor (Y_{i,m} - Y_{i,\lfloor mt \rfloor})}{\sigma\sqrt{m}}, \quad 0 \le t \le 1,$$

where $Y_{i,j} = \frac{1}{j}\sum_{\ell=1}^{j} X_{(i-1)m+\ell}$. Under the same mild conditions as before, one has

$$\begin{aligned}&(\sqrt{m}(Y_{1,m} - \mu), \sqrt{m}(Y_{2,m} - \mu), \ldots,\\ &\sqrt{m}(Y_{k,m} - \mu); \sigma T_{1,m}, \sigma T_{2,m}, \ldots, \sigma T_{k,m})\\ &\quad \Longrightarrow (\sigma Z_1, \sigma Z_2, \ldots, \sigma Z_k; \sigma\mathcal{B}_0, \sigma\mathcal{B}_1, \ldots, \sigma\mathcal{B}_{k-1}),\end{aligned}$$

where the $Z_i$ are IID standard normal random variables, and $\mathcal{B}_s$ denotes a standard Brownian bridge on $[s, s+1]$, for $s \in [0, k-1]$. That is,

$$\mathcal{B}_s(t) = \mathcal{W}(s+t) - \mathcal{W}(s) - t[\mathcal{W}(s+1) - \mathcal{W}(s)], \quad 0 \le t \le 1.$$

One can easily show that the Brownian bridges $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k$ are independent.

The area estimator from batch $i$ is

$$A_i(f; m) \equiv \left[\frac{1}{m}\sum_{\ell=1}^{m} f(\ell/m)\sigma T_{i,m}(\ell/m)\right]^2, \ i = 1, \ldots, k,$$

and the *batched area* estimator for $\sigma^2$ is

$$\hat{V}_A(f; k, m) \equiv \frac{1}{k}\sum_{i=1}^{k} A_i(f; m). \tag{19}$$

Since the $T_{i,m}$, $i = 1, \ldots, k$, converge to independent Brownian bridges as $m$ becomes large (with fixed $k$), we shall assume that the the $A_i(f; m)$ are asymptotically independent as $m \to \infty$. Then by the discussion above, we have $\hat{V}_A(f; k, m) \xrightarrow{d} \sigma^2\chi_k^2/k$, and an approximate $1-\alpha$ CI for $\mu$ is $\bar{X}_n \pm t_{k-1,1-\alpha/2}\sqrt{\hat{V}_A(k, m)/n}$.

Theorem 2 implies

$$\begin{aligned}\mathrm{E}[\hat{V}_A(f; k, m)] &= \mathrm{E}[A_1(f; m)]\\ &= \sigma^2 + \frac{[(F(1) - \bar{F}(1))^2 + \bar{F}^2(1)]\gamma}{2m}\\ &\quad + o(1/m). \end{aligned}\tag{20}$$

Further, if we assume uniform integrability of $\hat{V}_A^2(f; k, m))$, we can also make an analogous statement concerning the variance of the batched area estimator: As $m \to \infty$,

$$\begin{aligned}\mathrm{Var}[\hat{V}_A(f; k, m)] &= k^{-1}\mathrm{Var}[A_1(f; m)]\\ &\to k^{-1}\mathrm{Var}[A(f)] = 2\sigma^4/k. \end{aligned}\tag{21}$$

Equations (20) and (21) indicate that the batched area estimator has a bit more bias than the area estimator obtained from the entire sample, but smaller asymptotic variance (by a factor of $k$).

It is worth mentioning that, under the assumptions of Theorem 1, Damerdji (1994ab) showed that the batched area estimator $\hat{V}_A(f; k, m)$ is a strongly consistent estimator of $\sigma^2$.

**Remark 2** Another class of estimators is based on the weighted area under the square of the STS (Goldsman et al. 1999). Also, additional benefits result from combining NBM-based and area estimators (Schruben 1983) or by forming area estimators based on overlapping batches (Alexopoulos, et al. 2005ab).

### 3.7 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 often a harder problem than the estimation of steady-state means.

### 3.8 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 CIs. Indeed, suppose that $D_i$ is a $1-\alpha$ CI for the parameter $\mu_i$, $i = 1, \ldots, m$. Then $\Pr[\cap_{i=1}^{m}\{\mu_i \in D_i\}] \ge 1 - \sum_{i=1}^{m}\alpha_i$.

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 $\alpha_i$ 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).

## 4 RANKING AND SELECTION METHODS

Up to this point we reviewed output analysis methods for a single system. In this section we focus on comparing a number of alternative systems. There exist at least four classes of comparison problems that arise in simulation: finding the system with the largest or smallest expected performance measure (finding the best), comparing alternative systems to a standard (comparisons with a standard), finding the system with the largest success probability (Bernoulli selection), and finding the system with the largest probability of being the actual best (multinomial selection). Goldsman and Nelson (1998) give detailed definitions for each class of the above comparison problems and review the relative procedures. Kim and Nelson (2003) provide an advanced tutorial on the construction of ranking and selection procedures for selecting the best simulated system.

This section focuses on the problem of finding the best. 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. We review recent procedures to find the best system among a relatively large number of simulated systems, say more than 20 systems.

Many classical procedures assume that the output data generated by each system are IID and normally distributed. However, as we discussed in the preceding sections, raw data from within a single run are generally dependent and not quite normal. For finite-horizon simulations, the "IID normal" assumption is not a problem since within-run sample means across a number of independent replications are basic observations (see Section 2). For steady-state simulations, the assumption still makes sense if one is willing to use the replication/deletion approach as discussed in Section 3.2. Even when the single-replication design is employed, batch means from sufficiently large batches are approximately IID normal (see Section 3.4). However, Goldsman et al. (2000, 2001) and Kim and Nelson (2005) discuss the inefficiency of using within-run sample means or batch means as basic observations in steady-state simulations and suggest that individual observations (such as individual wait times) should be used as basic observations. Issues related to comparisons with regard to steady-state measures will be discussed more in Section 4.3.

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 $j$th observation from system $i$ ($i = 1, 2, \ldots, m$). We assume that the $Y_{ij}$ 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 IID and normally distributed.

### 4.1 Indifference-Zone Procedures

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. This 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 $\delta$ 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 problem is that IZ procedures become inefficient when the number of alternatives is large. This is due to the fact that these procedures are developed under the Least Favorable Configuration (LFC) condition, which is considered as the most difficult to resolve. Therefore, if a procedure guarantees at least $1 - \alpha$ probability of correct selection under the 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 $\delta$ smaller than that of the best, so that 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 Kim and Nelson (2001) and Nelson et al. (2001) have been shown to be efficient when hundreds of systems are compared.

Below we review two recent procedures: the NSGS 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 NSGS procedure does not use common random numbers (thus it requires systems to be simulated independently) while the $\mathcal{KN}$ procedure allows the use of common random numbers.

**Procedure NSGS**

- Specify the overall desired probability of correct selection $1-\alpha$, the IZ parameter $\delta$, 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).
- 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 $i = 1, 2, \ldots, m$.
- *Subset Selection.* Calculate the quantity

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

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

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

- 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 $K_i = \max\{k_0, \lceil (hS_i/\delta)^2 \rceil\}$, where $\lceil \cdot \rceil$ is the ceiling function.
- Take $K_i - k_0$ additional observations from all systems $i \in I$.
- 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.

**Procedure** $\mathcal{KN}$

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

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

- *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 $i \in I$ and let $\bar{Y}_i(k) = \frac{1}{k}\sum_{j=1}^{k} Y_{ij}$ 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 $\ell$. Set $r = k_0$.
- *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.
- *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 $i \in I$, set $r = r+1$, and go to *Screening*.

NSGS is a two-stage procedure while 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 NSGS procedure, where elimination occurs only after the first stage. As the $\mathcal{KN}$ procedure accounts for common random numbers and has more chances to detect inferior systems, it is expected to be more efficient than the NSGS 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. However, the $\mathcal{KN}$ procedure requires simultaneous simulation of all systems with many stoppages and restarts of simulation due to switching between systems. Hong and Nelson (2005) present a procedure that minimizes the cost of sampling and switching among the simulations of alternative systems.

Table 1: The Five Alternative Inventory Policies

| Policy $i$ | $s$ | $S$ | Expected Cost |
|---|---|---|---|
| 1 | 20 | 40 | 114.176 |
| 2 | 20 | 80 | 112.742 |
| 3 | 40 | 60 | 130.550 |
| 4 | 40 | 100 | 130.699 |
| 5 | 60 | 100 | 147.382 |

Table 2: Simulation Results for the $(s, S)$ Inventory Policy Example

| | NSGS | | $\mathcal{KN}$ | |
|---|---|---|---|---|
| Policy $i$ | # Obs. | Average | # Obs. | Average |
| 1 | 209 | 114.243 | 98 | 114.274 |
| 2 | 349 | 112.761 | 98 | 113.612 |
| 3 | 10 | 130.257 | 16 | 130.331 |
| 4 | 10 | 128.990 | 10 | 128.990 |
| 5 | 10 | 147.133 | 10 | 147.133 |
| Total | 588 | | 232 | |

### 4.2 An Application

This subsection illustrates NSGS and $\mathcal{KN}$ using an $(s, S)$ inventory system with the five inventory policies as described in Koenig and Law (1985). The goal of this study is to compare the five polices given in Table 1 and find the one with the smallest expected average cost per month for the first 30 months of operation. Table 1 also contains the expected cost (in thousands of dollars) of each policy, which can be analytically computed in this case. We set $\delta = \$1000$, $k_0 = 10$ initial replications, and $1 - \alpha = 0.95$.

Table 2 shows the results of the simulation study for each procedure, including the total number of outputs taken and the sample average cost per month for each policy. In NSGS, policies 3, 4, and 5 were eliminated after the first stage of sampling, so only policies 1 and 2 required second-stage samples. Under $\mathcal{KN}$, only policies 4 and 5 were eliminated after the first stage, but the elimination of policies 1 and 3 occurred after they received 16 and 98 observations, respectively. This illustrates the value of the tighter initial screen in NSGS, which takes only one look at the data, and the potential savings from taking many looks, as $\mathcal{KN}$ does. Both procedures chose policy 2 as the best (which is in fact correct). Since $\delta$ is smaller than the true difference, NSGS and $\mathcal{KN}$ will choose the true best with 95% confidence. However, in general we do not have any information about the true differences; therefore, the best we can conclude without prior knowledge is that policy 2 is either the true best, or has expected cost per month within \$1000 of the true best policy, with 95% confidence.

### 4.3 Finding the Best System in Steady-State

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 3.2, the replication/deletion approach is usually inefficient due to the elimination of data during warm-up period at each replication. Nakayama (1995, 1997) presented single-stage multiple-comparison procedures, and Damerdji and Nakayama (1999) developed two-stage multiple-comparison procedures to select the best system using steady-state simulations. 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 using large batches; see Goldsman et al. (2000, 2001) and Kim and Nelson (2005). The last three references propose three procedures that take a single replication from each system and use a single individual observation as a basic observation. One is a two-stage procedure based on a procedure due to Rinott (1978) and the others are the extensions of the $\mathcal{KN}$ procedure to steady-state simulation. Those three procedures require an estimator of the asymptotic variance parameter $\sigma^2$ and their performance is affected a lot by the quality of the variance estimator. Goldsman et al. (2000, 2001) and Kim and Nelson (2005) illustrate the performance of each procedure with NBM, OBM, and first-order unbiased weighted area estimators (Section 3.6) for a number of different batch sizes when the number of raw observations is fixed. Since the OBM estimator has the largest number of degrees of freedom and smallest variance among the three estimators, it appears to work best.

Kim and Nelson (2005) prove the asymptotic validity of the extensions of the $\mathcal{KN}$ procudure (called $\mathcal{KN}$+ and $\mathcal{KN}$++) as $\delta$ goes to zero and show that $\mathcal{KN}$++, which updates the variance estimator as more observations become available, is highly efficient compared to other competitors.

### 4.4 Closing Comments

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

## REFERENCES

Alexopoulos, C., N. T. Argon, D. Goldsman, G. Tokol, and J. R. Wilson. 2005a. Overlapping variance estimators

for simulation: I. Theory. Technical Report, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia.

Alexopoulos, C., N. T. Argon, D. Goldsman, N. M. Steiger, G. Tokol, and J. R. Wilson. 2005b. Overlapping variance estimators for simulation: II. Implementation and evaluation. Technical Report, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia.

Alexopoulos, C., and D. Goldsman. 2004. To batch or not to batch? *ACM Transactions on Modeling and Computer Simulation* 14:76–114.

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

Anderson, T. W. 1984. *An Introduction to multivariate statistical analysis*. New York: John Wiley & Sons.

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.

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 two-stage and sequential procedures for selecting the best simulated system. *Operations Research* 49:732–743.

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

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. 1975. Simulating stable stochastic systems III: Regenerative processes and discrete-event simulations. *Operations Research* 23:33–45.

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*, 340–344. Piscataway, New Jersey: IEEE.

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

Durrett, R. 2005. *Probability: Theory and examples*, 3d Ed. Belmont, California: Brooks/Cole–Thomson Learning.

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. 1978. *Principles of discrete event simulation*. New York: John Wiley & Sons.

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.

Foley, R. D., and D. Goldsman. 2000. Confidence intervals using orthonormally weighted standardized time series. *ACM Transactions on Modeling and Simulation* 9:297–325.

Glynn, P. W., and D. L. Iglehart. 1985. Large-sample theory for standardized time series: An overview. In *Proceedings of the 1985 Winter Simulation Conference*, 129–134. Piscataway, New Jersey: IEEE.

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

Glynn, P. W., and W. Whitt. 1991. Estimating the asymptotic variance with batch means. *Operations Research Letters* 10:431–435.

Goldsman, D., K. Kang, and A. F. Seila. 1999. Cramér-von Mises variance estimators for simulations. *Operations Research* 47, 299–309.

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. 1990. New confidence interval estimators using standardized time series. *Management Science* 36:393–397.

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.

Hong, L. J., and B. L. Nelson. 2005. The tradeoff between sampling and switching: New sequential procedures for indifference-zone selection. *IIE Transactions* 37:623–634.

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.

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.

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

Kim, S.-H. and B. L. Nelson. 2003. Selecting the best system: Theory and methods. In *Proceedings of the 2003 Winter Simulation Conference*, 101–112. Piscataway, New Jersey: IEEE.

Kim, S.-H., and B. L. Nelson. 2005. On the asymptotic validity of fully sequential selection procedures for steady-state simulation. *Operations Research*, forthcoming.

Koenig, L. W., and A. M. Law. 1985. A procedure for selecting a subset of size $m$ containing the $\ell$ best of $k$ independent normal populations, with applications to simulation. *Communications in Statistics—Simulation and Computation* B14:719–734.

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. New York: McGraw-Hill.

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. 1997. Multiple-comparison procedures for steady-state simulations. *Annals of Statistics* 25:2433–2450.

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., 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.

Philipp, W., and W. Stout. 1975. Almost sure invariance principles for partial sums of weakly dependent random variables. *Memoirs of the American Mathematical Society* 2.

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.

Schmeiser, B. W. 1986. Personal communication.

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

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. 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.

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, and D. Goldsman. 2004. ASAP3: An batch means procedure for steady-state simulation analysis. *ACM Transactions on Modeling and Computer Simulation* 15(1):39–73

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

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

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 estimation, 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.

Young, L. C. 1941. On randomness of order sequences. *Annals of Mathematical Statistics* 12:293–300.

## 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. He is a member of INFORMS. He is an active participant in the Winter Simulation Conference, having been Proceedings Co-Editor in 1995. He is also the Simulation Department Editor of *IIE Transactions*. His e-mail address is <christos@isye.gatech.edu>, and his web page is <http://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, quality control, and simulation output analysis. She is a member of INFORMS and IIE, and serves on the editorial board of the Simulation Department of *IIE Transactions*. Her e-mail and web addresses are <skim@isye.gatech.edu> and <http://www.isye.gatech.edu/~skim>.