Tests for Transient Means in Simulated Time Series

David Goldsman
School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332

Lee W. Schruben
School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York 14853

James J. Swain
Department of Industrial and Systems Engineering, University of Alabama in Huntsville, Huntsville, Alabama 35899

We present a family of tests to detect the presence of a transient mean in a simulation process. These tests compare variance estimators from different parts of a simulation run, and are based on the methods of batch means and standardized time series. Our tests can be viewed as natural generalizations of some previously published work. We also include a power analysis of the new tests, as well as some illustrative examples.

Il termine "batch mean" si riferisce alla media calcolata su un gruppo di dati (chiamato "batch") selezionati casualmente dall'insieme di dati completo.

1. INTRODUCTION

Simulation is frequently used to investigate the long-term behavior of complicated processes. An important problem in simulation concerns the estimation of the steady-state mean of a process (if it exists)—one usually uses the sample mean as the point estimator for the steady-state mean. Unfortunately, if we have little knowledge of the process, we might initialize the system in a "convenient" state that has a very low probability of actually occurring and is far from the steady-state mean. Since simulation processes are almost always serially correlated, an ill-considered initialization can affect a great deal of the output, biasing any estimate of the steady-state mean.

A common way of dealing with such *initialization bias* is to take a large (possibly wasteful) number of observations in an attempt to overwhelm the initialization effects. Another method is simply to delete (truncate) a portion of the output from the beginning of the simulation run. This presumably allows the simulation to warm up before data are retained for analysis. The experimenter would then hope that the biasing observations had been eliminated or their effects ameliorated. Of course, if the output is truncated too early, then significant bias might still be present. If it is truncated too late, then good observations are lost, and the sample mean's variance increases.

Naval Research Logistics, Vol. 41, pp. 171–187 (1994)
 CCC 0894-069X/94/020171-17

Although many of the rules in the simulation literature for detecting and dealing with initialization bias are easy to understand and implement, they are heuristic and often do not perform well in practice; see, e.g., the surveys by Chance [3], Schruben and Goldsman [21], and Wilson and Pritsker [23, 24]. In the current article we develop a class of statistical tests to detect the presence of bias in a simulated series. We henceforth operate under the assumption that the simulation output process $X_1, X_2, \ldots, X_n$ has a *transient mean function* defined by $\mu_i \equiv E[X_i] = \mu(1 - a_i)$, $i = 1, 2, \ldots, n$, where the a_i's are constants. Without loss of generality, we shall say that no transient-mean bias occurs if $\mu_i = \mu$ for all i (i.e., if $a_i = 0$ for all i); otherwise, bias is present. Initialization bias might also exist for higher-order moments, but this is not considered here; see [18].

Figure 1 presents some transient-mean functions that typically arise in simulations. The μ_i's given in Figure 1(a) might be encountered in a queue-length process when initializing a system empty and idle. Figure 1(b) could illustrate inventory level after starting a system fully stocked. Figure 1(c) is a transient-mean function whose oscillations damp out. Functions 1(a), 1(b), and 1(c) each appear to be approaching steady state. The μ_i's in Figure 1(d) have not yet approached steady state (and may never)—this is common in simulated time

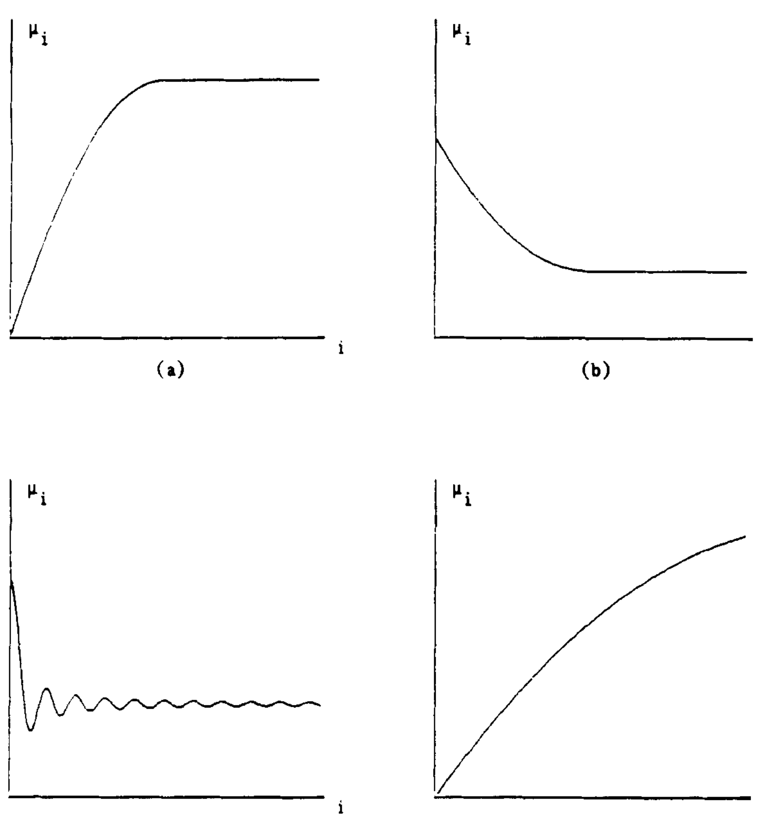

Figure 1. Various transient mean processes (drawn as continuous functions for clarity).

series that have been initialized in very atypical states and have not been run long enough to recover.

The tests to be described here are easily motivated. We partition the process $X_1, X_2, \ldots, X_n$ into two contiguous, nonoverlapping portions. For a particular realization of the process, we calculate an estimate for the variance of the sample mean based solely on the first portion of the output, and then an estimate based only on the latter portion of the output. A large difference between these two estimates is unlikely if the X_i's are stationary; if the two estimates are deemed to be *significantly different*, then we reject H_0: $\mu_i = \mu$ for all i. For the transient-mean functions illustrated in Figures 1(a), 1(b), and 1(c), we would expect an estimate formed from the first portion of the output to be greater than an estimate from the latter portion. Such a conclusion might not be drawn concerning Figure 1(d), since its transient-mean function does not level off.

The organization of the remainder of this article is as follows. We first give some background material in Section 2, reviewing a number of variance estimation methods as well as some previously studied transient-mean tests. In Section 3, we present our new tests. Section 4 is concerned with power calculations for these tests, Section 5 demonstrates the performance of the tests with some examples, and Section 6 provides conclusions and suggestions regarding the use of the tests.

2. BACKGROUND

In Subsection 2.1, we discuss the variance estimation methods of batch means and standardized time series. Subsection 2.2 reviews the transient-mean tests of Schruben [19] and Schruben, Singh, and Tierney [22], since our tests are natural generalizations of their work.

2.1. Some Variance Estimators

The goal here is to estimate the variance of the sample mean of stationary observations. To this end, divide $X_1, X_2, \ldots, X_n$ into b contiguous batches of length m (assume $n = bm$); the observations $X_{(i-1)m+1}, X_{(i-1)m+2}, \ldots, X_{im}$ comprise the ith batch, $i = 1, 2, \ldots, b$. We denote the *grand mean* by

$$\bar{X}_n \equiv \frac{1}{n} \sum_{p=1}^{n} X_p.$$

For $i = 1, 2, \ldots, b$ and $j = 1, 2, \ldots, m$, the jth *cumulative mean* from batch i is

$$\bar{X}_{i,j} \equiv \frac{1}{j} \sum_{p=1}^{j} X_{(i-1)m+p}.$$

(The quantity $\bar{X}_{i,m}$ is called the ith *batch mean*.) For $i = 1, 2, \ldots, b$ and $0 \leq t \leq 1$, the *standardized time series* from batch i is given by

$$T_{i,m}(t) \equiv \frac{\lfloor mt \rfloor(\bar{X}_{i,m} - \bar{X}_{i,\lfloor mt \rfloor})}{\sigma\sqrt{m}},$$

where $\lfloor \cdot \rfloor$ is the greatest integer function, and $\sigma^2 \equiv \lim_{n\to\infty} n \operatorname{var}(\overline{X}_n)$ is the *variance parameter*. Instead of directly estimating $\operatorname{var}(\overline{X}_n)$, it is sometimes more convenient to estimate σ^2.

Schruben [20] shows that if $X_1, X_2, \ldots, X_n$ is a stationary sequence satisfying certain mild moment and mixing conditions, then as $m \to \infty$, $T_{i,m}(t) \xrightarrow{\mathcal{D}} \mathcal{B}(t)$, $0 \le t \le 1$, a standard Brownian bridge process. (The symbol $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.) All finite-dimensional joint distributions of $\mathcal{B}$ are normal and $\operatorname{cov}(\mathcal{B}(s), \mathcal{B}(t)) = \min(s, t)(1 - \max(s, t))$, $0 < s, t < 1$. Schruben also shows that $T_{i,m}(t)$ and $\overline{X}_{i,m}$ are asymptotically independent.

We denote the weighted area under the ith standardized time series, $i = 1, 2, \ldots, b$, by

$$\hat{A}_i \equiv \frac{\sigma}{m} \sum_{j=1}^{m} w_i(j/m) T_{i,m}(j/m),$$

where the $w_i(\cdot)$'s are prespecified weighting functions that are continuous on $[0, 1]$, not dependent on m, and normalized so that

$$\operatorname{var}\left(\int_0^1 w_i(t)\mathcal{B}(t)\, dt\right) = 2 \int_0^1 \int_0^u w_i(u) w_i(t) t(1 - u)\, dt\, du = 1.$$

(This expression can be simplified considerably; see [10] for details.) We will say more about these weights later. Finally, define the location on $[0, 1]$ of the maximum of the ith standardized time series, $i = 1, 2, \ldots, b$, as

$$\hat{t}_i \equiv \frac{\operatorname{argmax}_{1 \le j \le m}\{T_{i,m}(j/m)\}}{m}.$$

We then have a collection of estimators for σ^2 (cf. Glynn and Iglehart [7] and Goldsman and Schruben [11]): the batch means (BM) estimator:

$$V_{0,b,m} \equiv \frac{Q_{0,b,m}}{b-1} \equiv \frac{m}{b-1} \sum_{i=1}^{b} \left(\overline{X}_{i,m} - \frac{\sum_{j=1}^{b} \overline{X}_{j,m}}{b}\right)^2 \xrightarrow{\mathcal{D}} \frac{\sigma^2 \chi^2(b-1)}{b-1}, \qquad b > 1;$$

the (weighted) area estimator:

$$V_{1,b,m} \equiv \frac{Q_{1,b,m}}{b} \equiv \frac{1}{b} \sum_{i=1}^{b} \hat{A}_i^2 \xrightarrow{\mathcal{D}} \frac{\sigma^2 \chi^2(b)}{b}, \qquad b \ge 1;$$

the combined BM + area estimator:

$$V_{2,b,m} \equiv \frac{Q_{2,b,m}}{2b-1} \equiv \frac{Q_{0,b,m} + Q_{1,b,m}}{2b-1} \xrightarrow{\mathcal{D}} \frac{\sigma^2 \chi^2(2b-1)}{2b-1}, \qquad b > 1;$$

the maximum estimator:

$$V_{3,b,m} \equiv \frac{Q_{3,b,m}}{3b} \equiv \frac{\sigma^2}{3b} \sum_{i=1}^{b} \frac{T_{i,m}^2(\hat{t}_i)}{\hat{t}_i(1-\hat{t}_i)} \xrightarrow{\mathcal{D}} \frac{\sigma^2 \chi^2(3b)}{3b}, \qquad b \ge 1;$$

and the combined BM + maximum estimator:

$$V_{4,b,m} \equiv \frac{Q_{4,b,m}}{4b - 1} \equiv \frac{Q_{0,b,m} + Q_{3,b,m}}{4b - 1} \xrightarrow{\mathscr{D}} \frac{\sigma^2\chi^2(4b - 1)}{4b - 1}, \qquad b > 1.$$

We will write $V_{k,b} \equiv V_{k,b,m}$ and $Q_{k,b} \equiv Q_{k,b,m}$ when the meaning is clear. All of these variance estimators have limiting χ^2 distributions. We shall see in the sequel that certain ratios of these estimators will conveniently have F distributions; this fact will allow us to test $X_1, X_2, \ldots, X_n$ for transient means.

REMARK 1: A number of other estimators for σ^2 arise from standardized time series (see, e.g., [9]), spectral analysis [12], times series modeling [5, 6, 17], regeneration theory [4], and overlapping batch means methods [15]. The limiting distributions of these estimators are sometimes tractable but not necessarily χ^2; for this reason, we will not use these estimators subsequently.

2.2. Some Previous Transient-Mean Tests

This subsection reviews two transient-mean tests from the literature. The reader should be aware that both tests are asymptotic, so they may not perform adequately if the sample size n is small. Also, the tests may do poorly if the transient-mean bias pervades the entire process [as in Figure 1(d)]; but in such cases, one should be able to detect the lack of steady state visually. A final *caveat* is that the numerator and denominator in the appropriate F statistics may not be independent.

2.2.1. *A Test Based on the Maximum Estimator*

Under the stationarity hypothesis H_0: $\mu_i = \mu$ for all i, Schruben [19] notes that $V_{3,1,n} \approx \sigma^2\chi^2(3)/3$. (The notation $\approx$ reads "is approximately distributed as".) Fishman's time series method produces another variance estimator, $\hat{\sigma}^2 \approx \sigma^2\chi^2(\nu)/\nu$, where ν must be estimated. If $V_{3,1,n}$ and $\hat{\sigma}^2$ are independent, then under H_0,

$$F \equiv V_{3,1,n}/\hat{\sigma}^2 \approx F(3, \nu), \tag{1}$$

the F distribution with 3 and ν degrees of freedom. A one-sided test is to reject H_0 at level α if $F > f_{3,\nu,\alpha}$, the $1 - \alpha$ quantile of the $F(3, \nu)$ distribution. This test works well for the battery of simulated systems studied in [19].

Schruben gives an alternative procedure in which the process is divided into two batches, each with $m = n/2$. Since the standardized time series from each half of the run are asymptotically independent, we have, under H_0,

$$\frac{Q_{3,1}}{Q_{3,2} - Q_{3,1}} \approx F(3, 3). \tag{2}$$

The family of tests to be presented in our Section 3 represents extensions of the above idea in two directions. First, $b > 2$ is permitted. Second, variance estimators other than the maximum estimator are considered.

2.2.2. A Test Based on the Area Estimator

Schruben, Singh, and Tierney [22] (SST) use a weighted area estimator to test H_0 versus H_1: $\mu_i = \mu(1 - a_i)$, $i = 1, 2, \ldots, n$, for some given sequence of a_i's. As in Subsection 2.2.1, SST assume $V_{1,1,n}$ and $\hat{\sigma}^2$ are independent so that

$$F' \equiv V_{1,1,n}/\hat{\sigma}^2 \approx F(1, \nu). \tag{3}$$

Under certain strong assumptions, the authors find the asymptotically most powerful test against H_1; viz., use $w_1(i/n) \equiv a_i - a_{i+1}$ for all i, and reject H_0 when F' is large. They offer several examples that show the experimenter can make reasonable choices for the a_i's. The SST procedure works well for the examples given in their article (even though the test is not always as powerful as the maximum test of [19] against alternatives *other than* H_1). We feel that the SST approach is a valuable one, and we extend it in the next section.

L'algoritmo è questo in linea di massima, nella prossima sezione, però, andrà a dire come calcolare k, b e b' in modo tale che ci siano tutti gli elementi per implementarlo
Quindi per l'algoritmo serve
-La sezione 2 per le convergenze e i modi per calcolare le medie
-La sezione 4 per le stime dei "coefficienti"

3. A NEW CLASS OF TESTS FOR TRANSIENT MEANS

Once more, divide $X_1, X_2, \ldots, X_n$ into b adjacent batches, each of size m. Variance estimators based on the first b' batches will be compared to the corresponding estimators from the remaining $b - b'$ batches. Using the notation and results from Subsection 2.1, we have that under H_0: $\mu_i = \mu$ for all i,

$$\boxed{Q_{k,b'} \xrightarrow{\mathscr{D}} \sigma^2\chi^2(d_{k,b'}),}$$

where

$$\boxed{d_{k,p} \equiv \begin{cases} p - 1, & \text{for } k = 0, \\ p, & \text{for } k = 1, \\ 2p - 1, & \text{for } k = 2, \\ 3p, & \text{for } k = 3, \\ 4p - 1, & \text{for } k = 4. \end{cases}}$$

Similarly,

$$\boxed{\begin{aligned} Q^*_{0,b-b'} &\equiv m \sum_{i=b'+1}^{b} \left(\bar{X}_{i,m} - \frac{\sum_{j=b'+1}^{b} \bar{X}_{j,m}}{b - b'} \right)^2 \xrightarrow{\mathscr{D}} \sigma^2\chi^2(b - b' - 1), \\ Q^*_{1,b-b'} &\equiv Q_{1,b} - Q_{1,b'} \xrightarrow{\mathscr{D}} \sigma^2\chi^2(b - b'), \\ Q^*_{2,b-b'} &\equiv Q^*_{0,b-b'} + Q^*_{1,b-b'} \xrightarrow{\mathscr{D}} \sigma^2\chi^2(2b - 2b' - 1), \\ Q^*_{3,b-b'} &\equiv Q_{3,b} - Q_{3,b'} \xrightarrow{\mathscr{D}} \sigma^2\chi^2(3b - 3b'), \\ Q^*_{4,b-b'} &\equiv Q^*_{0,b-b'} + Q^*_{3,b-b'} \xrightarrow{\mathscr{D}} \sigma^2\chi^2(4b - 4b' - 1). \end{aligned}}$$

So for all k,

$$\boxed{Q^*_{k,b-b'} \xrightarrow{\mathscr{D}} \sigma^2\chi^2(d_{k,b-b'}).}$$

Further, define for all k, b', and b,

$$\boxed{V_{k,b'} \equiv \frac{Q_{k,b'}}{d_{k,b'}} \quad \text{and} \quad V^*_{k,b-b'} \equiv \frac{Q^*_{k,b-b'}}{d_{k,b-b'}}.}$$

The V's are variance estimators calculated from the first b' batches, and the V^*'s are the analogous estimators from the remaining $b - b'$ batches. Thus, the V's are asymptotically independent of the V^*'s. To keep things simple, we will work with statistics of the form (under H_0),

$$\boxed{\frac{V_{k,b'}}{V^*_{k,b-b'}} \xrightarrow{\mathscr{D}} F(d_{k,b'}, d_{k,b-b'}).} \tag{4}$$

The test statistics from Section 2.2 are similar to (or special cases of) (4): in (1), $V_{3,1,n}/\hat{\sigma}^2 \approx F(3, \nu)$; in (2), $V_{3,1,n/2}/V^*_{3,1,n/2} \approx F(3, 3)$; and in (3), $V_{1,1,n}/\hat{\sigma}^2 \approx F(1, \nu)$.

The tests based on the statistics given by (4) are all asymptotically *valid* as $m \to \infty$; i.e., the tests achieve the desired level α as the batch size becomes large. Given a choice of valid tests, it is natural to ask which is the best. The goal here is to find (for fixed n) combinations of k, b', and b that yield test statistics having high *power*.

La potenza (power) di un test è la probabilità di rifiutare l'ipotesi nulla quando essa è falsa. In altre parole, è la probabilità di non commettere un errore di "falso negativo" (concludere che non c'è un effetto quando in realtà c'è).

4. POWER CALCULATIONS FOR THE NEW TESTS

The most common criterion for comparison among tests with fixed level α is power. In Subsections 4.1–4.3, we derive analytical results for $k = 0, 1$, and 2 (BM, area, and BM + area), respectively. An analytical comparison is carried out in Subsection 4.4. Some Monte Carlo results for $k = 3$ and 4 are given in the examples of Section 5.

Let $X_{i,j} \equiv X_{(i-1)m+j}$ denote the jth observation from the ith batch, $i = 1, 2, \ldots, b$ and $j = 1, 2, \ldots, m$, and consider the test

$$H_0\colon E[X_{i,j}] = \mu, \quad \text{for all } i, j \quad \text{versus } H_1\colon E[X_{i,j}] = \mu(1 - a_{i,j}), \text{ for all } i, j,$$

where the $a_{i,j}$'s are prespecified. In order to calculate the power, we shall henceforth use the model $X_{i,j} = Y_{i,j} - \mu a_{i,j}$, where the $Y_{i,j}$'s are stationary with mean μ. For ease of exposition, suppose we suspect that the transient-mean function is of the form of Figure 1(a). [Although the BM, area, and BM + area tests are appropriate for use with transient-mean functions of the forms of Figures 1(a), 1(b), or 1(c), the maximum and BM + max tests are only designed for use with transient-mean functions of the form of Figure 1(a).] Then by (4), a level-α test is to reject H_0 if

$$\frac{V_{k,b'}}{V^*_{k,b-b'}} > f_{d_{k,b'},d_{k,b-b'},\alpha}.$$

4.1. Batch Means Tests

Suppose H_1 is true so that we can compute the power, i.e., $E[X_{i,j}] = \mu(1 - a_{i,j})$, for all i, j. For $i = 1, 2, \ldots, b$ and $j = 1, 2, \ldots, m$, let

$$\overline{Y}_{i,j} \equiv \frac{1}{j} \sum_{p=1}^{j} Y_{i,p},$$

$$\overline{a}_{i,j} \equiv \frac{1}{j} \sum_{p=1}^{j} a_{i,p},$$

and

$$\overline{\overline{a}}_b \equiv \frac{1}{b} \sum_{i=1}^{b} \overline{a}_{i,m}.$$

If the $Y_{i,j}$'s are stationary and satisfy certain additional mild assumptions, then Billingsley [1, Theorem 21.1], implies

$$\sqrt{m}[\overline{X}_{i,m} - \mu(1 - \overline{a}_{i,m})] = \sqrt{m}(\overline{Y}_{i,m} - \mu) \xrightarrow{\mathscr{D}} \mathrm{Nor}(0, \sigma^2).$$

Then from the Appendix, we get

$$m \sum_{i=1}^{b} (\overline{X}_{i,m} - \overline{\overline{X}}_n)^2 \approx \sigma^2\chi^2\left(b - 1, \frac{m\mu^2}{\sigma^2} \sum_{i=1}^{b} (\overline{a}_{i,m} - \overline{\overline{a}}_b)^2\right), \tag{5}$$

where $\chi^2(d, \delta)$ is the noncentral χ^2 distribution with d degrees of freedom and noncentrality parameter δ. We immediately have

$$Q_{0,b'} \approx \sigma^2\chi^2(b' - 1, \delta_{0,b'})$$

and

$$Q^*_{0,b-b'} \approx \sigma^2\chi^2(b - b' - 1, \delta^*_{0,b-b'}),$$

where

$$\delta_{0,b'} \equiv \frac{m\mu^2}{\sigma^2} \sum_{i=1}^{b'} \left(\overline{a}_{i,m} - \frac{\sum_{j=1}^{b'} \overline{a}_{j,m}}{b'}\right)^2$$

and

$$\delta^*_{0,b-b'} \equiv \frac{m\mu^2}{\sigma^2} \sum_{i=b'+1}^{b} \left(\overline{a}_{i,m} - \frac{\sum_{j=b'+1}^{b} \overline{a}_{j,m}}{b - b'}\right)^2.$$

So

$$R_{0,b',b} \equiv V_{0,b'}/V^*_{0,b-b'} \approx F(b' - 1, b - b' - 1, \delta_{0,b'}, \delta^*_{0,b-b'}), \tag{6}$$

where $F(d_1, d_2, \gamma_1, \gamma_2)$ is the doubly noncentral F distribution with d_1 and d_2 degrees of freedom and noncentrality parameters γ_1 and γ_2. The power is

$$P\{\text{Reject } H_0 | H_1 \text{ true}\} = P\{R_{0,b',b} > f_{b'-1,b-b'-1,\alpha}\}.$$

REMARK 2: Consider the interesting special case in which the X_i's are *independent* normal observations with common variance σ^2. Then it is easy to see that the preceding manipulations lead to *equalities in distribution* in (5) and (6).

4.2. Area Tests

For $i = 1, 2, \ldots, b$ and $j = 1, 2, \ldots, m$, let

$$T'_{i,m}(j/m) \equiv \frac{j(\overline{Y}_{i,m} - \overline{Y}_{i,j})}{\sigma\sqrt{m}}$$

and

$$c_{i,j} \equiv -j(\overline{a}_{i,m} - \overline{a}_{i,j}).$$

The standardized time series from the ith batch is given by

$$T_{i,m}(j/m) = \frac{j(\overline{X}_{i,m} - \overline{X}_{i,j})}{\sigma\sqrt{m}} = T'_{i,m}(j/m) + \frac{\mu c_{i,j}}{\sigma\sqrt{m}}, \qquad j = 1, 2, \ldots, m.$$

Hence,

$$\hat{A}_i = \frac{\sigma}{m} \sum_{j=1}^{m} w_i(j/m) T_{i,m}(j/m) = \frac{\sigma}{m} \sum_{j=1}^{m} w_i(j/m) T'_{i,m}(j/m) + \frac{\mu}{m^{3/2}} \sum_{j=1}^{m} w_i(j/m) c_{i,j}.$$

Since the $Y_{i,j}$'s are stationary, $T'_{i,m}(t) \xrightarrow{\mathcal{D}} \mathcal{B}(t)$ as $m \to \infty$. So for large m,

$$E[\hat{A}_i] = \frac{\mu}{m^{3/2}} \sum_{j=1}^{m} w_i(j/m) c_{i,j} \equiv \mu c_i,$$

say, and

$$\text{var}(\hat{A}_i) = \text{var}\left(\frac{\sigma}{m} \sum_{j=1}^{m} w_i(j/m) T'_{i,m}(j/m)\right) \doteq \sigma^2.$$

These results imply that for large m,

$$\hat{A}_i \doteq \text{nor}(\mu c_i, \sigma^2), \qquad i = 1, 2, \ldots, b.$$

Therefore, if the $\hat{A}_i$'s are approximately independent,

$$Q_{1,b'} \approx \sigma^2\chi^2\left(b', \frac{\mu^2}{\sigma^2}\sum_{i=1}^{b'} c_i^2\right), \tag{7}$$

$$Q^*_{1,b-b'} \approx \sigma^2\chi^2\left(b - b', \frac{\mu^2}{\sigma^2}\sum_{i=b'+1}^{b} c_i^2\right),$$

and

$$R_{1,b',b} \equiv V_{1,b'}/V^*_{1,b-b'} \approx F(b', b - b', \delta_{1,b'}, \delta^*_{1,b-b'}), \tag{8}$$

where $\delta_{1,b'}$ and $\delta^*_{1,b-b'}$ are the noncentrality parameters in $Q_{1,b'}$ and $Q^*_{1,b-b'}$, respectively. The power is $P\{R_{1,b',b} > f_{b',b-b',\alpha}\}$.

REMARK 3: Consider $V_{1,1,mb'}$ (the area estimator based on one batch of size mb') and $V^*_{1,b-b',m}$ (the area estimator based on the following $b - b'$ batches of size m), where $1 \le b' < b$. Then $V_{1,1,mb'}/V^*_{1,b-b',m}$ is very similar to the SST statistic given by (3). Perhaps the assumption of independence between the numerator and denominator is more justified here, since the two are based on disjoint batches. Although the ratio is not strictly of the form (4), we can still derive an expression for the power. By (7),

$$V_{1,1,mb'} = Q_{1,1,mb'} = \hat{A}_1^2 \approx \sigma^2\chi^2\left(1, \frac{\mu^2}{\sigma^2} c_1^2\right),$$

where

$$c_1 = (mb')^{-3/2}\sum_{j=1}^{mb'} w_1(j/mb')c_{1,j} = (mb')^{-3/2}\sum_{j=1}^{mb'} w_1(j/mb')j(\bar{a}_{1,j} - \bar{a}_{1,mb'}).$$

The distribution of $V^*_{1,b-b',m}$ is immediate from (7); an expression for the power then follows.

REMARK 4: Suppose we operate in the special case in which the X_i's are *independent* normal observations with common variance σ^2. If we define

$$g_i(j) \equiv \sum_{k=j}^{m} w_i(k/m) - \sum_{k=1}^{m} (k/m)w_i(k/m),\ i = 1, 2, \ldots, b,\ j = 1, 2, \ldots, m,$$

then we can write (see [10])

$$\hat{A}_i = \frac{1}{m^{3/2}}\sum_{j=1}^{m} g_i(j)X_{i,j}, \qquad i = 1, 2, \ldots, b,$$

so that $\hat{A}_i$ is normal with mean $(\mu/m^{3/2}) \sum_{j=1}^{m} g_i(j)(1 - a_{i,j})$ and variance $(\sigma^2/m^3) \sum_{j=1}^{m} g_i^2(j)$, $i = 1, 2, \ldots, b$. Manipulations similar to those yielding (7) and (8) will produce an *exact* expression for power.

4.3. Combined BM + Area Tests

We can add the asymptotically independent noncentral χ^2 random variables to obtain other noncentral χ^2's:

$$Q_{2,b'} = Q_{0,b'} + Q_{1,b'} \approx \sigma^2\chi^2(2b' - 1, \delta_{2,b'})$$

and

$$Q^*_{2,b-b'} = Q^*_{0,b-b'} + Q^*_{1,b-b'} \approx \sigma^2\chi^2(2(b - b') - 1, \delta^*_{2,b-b'}),$$

where

$$\delta_{2,b'} \equiv \delta_{0,b'} + \delta_{1,b'} \quad \text{and} \quad \delta^*_{2,b-b'} \equiv \delta^*_{0,b-b'} + \delta^*_{1,b-b'}.$$

So

$$R_{2,b',b} \equiv V_{2,b'}/V^*_{2,b-b'} \approx F(2b' - 1, 2(b - b') - 1, \delta_{2,b'}, \delta^*_{2,b-b'}),$$

which has power $P\{R_{2,b',b} > f_{2b'-1,2(b-b')-1,\alpha}\}$.

4.4. Analytical Comparison of Tests

We now compare the power of the BM, area, and BM + area tests ($k = 0$, 1, and 2). As shown earlier,

$$R_{k,b'} \equiv R_{k,b',b} \approx F(d_{k,b'}, d_{k,b-b'}, \delta_{k,b'}, \delta^*_{k,b-b'})$$

with power

$$P\{R_{k,b'} > f_{d_{k,b'},d_{k,b-b'},\alpha}\}.$$

Since tables for doubly noncentral F distributions are not readily available, we approximate the distribution of $R_{k,b'}$ by using the familiar result (cf. Johnson and Kotz [13, p. 197]) that

$$F(\nu_1, \nu_2, \gamma_1, \gamma_2) \approx \beta F(\tilde{\nu}_1, \tilde{\nu}_2),$$

where

$$\beta \equiv \frac{(\nu_1 + \gamma_1)\nu_2}{(\nu_2 + \gamma_2)\nu_1} \quad \text{and} \quad \tilde{\nu}_i \equiv \frac{(\nu_i + \gamma_i)^2}{\nu_i + 2\gamma_i}, \qquad i = 1, 2.$$

This gives

$$R_{k,b'} \approx \frac{(d_{k,b'} + \delta_{k,b'})d_{k,b-b'}}{(d_{k,b-b'} + \delta^*_{k,b-b'})d_{k,b'}} F\left(\frac{(d_{k,b'} + \delta_{k,b'})^2}{d_{k,b'} + 2\delta_{k,b'}}, \frac{(d_{k,b-b'} + \delta^*_{k,b-b'})^2}{d_{k,b-b'} + 2\delta^*_{k,b-b'}}\right).$$

So the power is approximately

$$P\{F(\tilde{\nu}_1, \tilde{\nu}_2) > \beta^{-1} f_{d_{k,b'}, d_{k,b-b'}, \alpha}\}. \tag{9}$$

Since β, $\tilde{\nu}_1$, and $\tilde{\nu}_2$ are functions of $\delta_{k,b'}$ and $\delta^*_{k,b-b'}$, they are functions of μ^2/σ^2, which is unknown. If we are willing to estimate (or make a rough guess of) μ^2/σ^2, the combination of b, b', and k can be found that maximizes (9).

5. EXAMPLES

We present some examples to illustrate the performance of the new tests. The first example involves the analytic calculation of the power of the tests for the cases $k = 0$, 1, and 2, in the presence of a quadratic transient-mean function.

EXAMPLE 1: Suppose that $m = 100$, $b = 20$, $\alpha = 0.05$, and H_1: $E[X_p] = \mu(1 - a_p)$, where $a_p = [1 - (p/1000)]^2$ for $p = 1, 2, \ldots, 1000$, and $a_p = 0$ for $p > 1000$; in the parlance of Section 4, $a_{i,j} = a_{(i-1)m+j}$. The transient mean function is of the form given by Figure 1(a). Figures 2(a) and 2(b) give plots of the power (9) as a function of μ^2/σ^2 for $b' = 5$ and 8, respectively. Both of the figures illustrate plots for $k = 0$, 1, and 2. We have (arbitrarily) used the constant weights $w_i(\cdot) = \sqrt{12}$ when calculating (9) for $k = 1$ and 2. The figures show that the BM and BM + area tests ($k = 0$ and 2) have power that quickly increases toward unity; unfortunately, the area test ($k = 1$) has power that is barely greater than the size α.

There is a heuristic explanation for this phenomenon. The BM variance estimator $V_{0,b'}$ draws its observations from *all* b' batches of the first portion of the output. In the presence of transient-mean bias of the form given by Figure 1(a), we would expect many of the first portion's batch means to differ significantly from the first portion's grand mean; it follows that a realization of $V_{0,b'}$ would tend to greatly overestimate the steady-state variance parameter σ^2. On the other hand, each individual $\hat{A}_i$ from the first portion of the output draws its observations from a single batch. Although the presence of transient-mean bias suggests that the expected value of a particular $\hat{A}_i^2$ might be greater than σ^2, it stands to reason that $E[\hat{A}_i^2] < E[V_{0,b'}]$, $i = 1, 2, \ldots, b'$; i.e., there is simply not as much variance within a single batch as there is within the entire first portion. Since $V_{1,b'}$ is the average of the $\hat{A}_i^2$'s, we would then have $E[V_{1,b'}] < E[V_{0,b'}]$. In other words, our results make sense since the BM and BM + area variance estimators are more sensitive to *between-batch* changes in mean than is the area estimator.

The results from Example 1 are substantiated by the following empirical example, which also evaluates the power for the maximum and BM + maximum tests.

EXAMPLE 2: We use Monte Carlo simulation to compare the power for the cases $k = 0, 1, 2, 3, 4$. Consider the stationary first-order autoregressive [AR(1)] process $Z_p = \phi Z_{p-1} + \epsilon_p$, where the ϵ_p's are independent nor$(0, 1 - \phi^2)$ random variables, $p = 1, 2, \ldots, n$, and $-1 < \phi < 1$. Let $X_p \equiv Z_p +$

Table 1. Estimated power of tests for shifted AR(1).

	$\phi = 0.0$				$\phi = 0.5$			
k	$b' = 2$	$b' = 5$	$b' = 8$	$b' = 10$	$b' = 2$	$b' = 5$	$b' = 8$	$b' = 10$
0	0.004	0.803	0.997	0.997	0.025	0.391	0.710	0.738
1	0.082	0.088	0.073	0.066	0.067	0.061	0.056	0.054
2	0.007	0.768	0.996	0.997	0.031	0.318	0.633	0.680
3	0.137	0.160	0.158	0.154	0.128	0.144	0.150	0.136
4	0.013	0.781	0.996	0.998	0.056	0.376	0.688	0.737

$\mu(1 - a_p)$, $p = 1, 2, \ldots, n$. We test H_0: $E[X_p] = \mu$, $p = 1, 2, \ldots, n$, versus H_1: not H_0. Suppose $\mu = 1$ with m, b, α, and a_p as in Example 1. For each of various values of b' and ϕ, we ran 2000 independent Monte Carlo experiments to estimate the power for all k. Table 1 contains representative results. Each table entry for the estimated power is the proportion r of experiments for which H_0 is rejected. The standard error of each estimate is $\sqrt{r(1 - r)/2000}$. Each column of the table is based on the same 2000 independent experiments. The results of Table 1 are expected since, as explained above, the BM, BM + area, and BM + maximum variance estimators are more sensitive to between-batch changes in mean than are the area and maximum estimators. An easy calculation gives $\mu^2/\sigma^2 = (1 - \phi)/(1 + \phi)$, and we see that the $k = 0$, 1, and 2 entries in Table 1 closely match the analytical results in Figures 2(a) and (b).

We also performed some Monte Carlo work for a similar, but less trivial, stochastic process.

EXAMPLE 3: Consider the stationary first-order exponential autoregressive process $Z_p = \phi Z_{p-1} + \epsilon_p$, where the ϵ_p's are independent random variables such that ϵ_p is exp(1) with probability $1 - \phi$, and $\epsilon_p = 0$ with probability ϕ, where $0 \le \phi < 1$ [14]. Under a setup similar to that of Example 2, we again

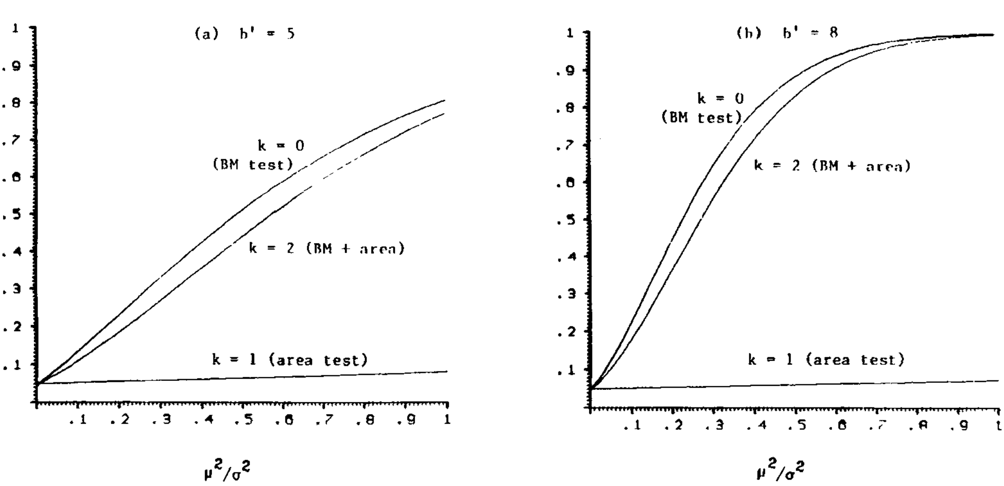

Figure 2. The power (9) of the tests described in Example 1.

Table 2. Estimated power of tests for shifted $M/M/1$.

k	$b' = 2$	$b' = 5$	$b' = 8$	$b' = 10$
0	0.045	0.222	0.384	0.396
1	0.088	0.109	0.122	0.126
2	0.070	0.214	0.357	0.391
3	0.151	0.210	0.246	0.268
4	0.117	0.275	0.411	0.464

used Monte Carlo simulation to compare the power for the cases $k = 0, 1, 2, 3, 4$. The results were almost exactly the same as those given in Example 2.

The next example gives area and maximum estimator results that are less disappointing.

EXAMPLE 4: Suppose we take a stationary $M/M/1$ queueing system with customer waiting times $Z_1, Z_2, \ldots, Z_n$, with $n = 8000$, arrival rate 0.5, and service rate 1.0. The stochastic process of interest is $X_p = Z_p + 1 - a_p$, $p = 1, 2, \ldots, 8000$, where $a_p = [1 - (p/4000)]^2$ for $p = 1, 2, \ldots, 4000$, and $a_p = 0$ for $p > 4000$. We conducted 2000 Monte Carlo runs, each consisting of $b = 20$ batches of $m = 400$ observations. The results are given in Table 2; they are not surprising in light of the fact that the transient-mean function (and thus much of the variance) disappears halfway through each run.

We have seen that the area and maximum estimators seem to be less sensitive to between-batch effects than the BM tests. We can get around this problem by implementing Schruben's original suggestion of dividing the simulation into two equal-size batches.

EXAMPLE 5: Consider the shifted $M/M/1$ process given in Example 4. We ran 2000 Monte Carlo simulations for the case $m = 4000$, $b = 2$, and $b' = 1$. The estimated power for the area estimator $R_{1,1}$ was 0.518, and the estimated power for the maximum estimator $R_{3,1}$ was 0.951, both of which are greater than any of the entries in Table 2. Of course, the BM procedures cannot even be used when $b' = 1$!

6. CONCLUSIONS

We have introduced a family of tests for detecting transient means in a simulated process. The process is divided into two nonoverlapping portions, from which two variance estimates are calculated. If these estimates are deemed to be significantly different, then we say that a transient mean is present, and would likely bias point estimators of the steady-state mean. The variance estimators we consider in this article are from batch means and standardized time series, but the use of other estimators is possible.

The two most important performance criteria for statistical tests are validity and power. The theories of batch means and standardized time series are asymptotically valid under mild assumptions. Of course, asymptotic validity requires large sample sizes, but these are typically available in simulated time series. We derived analytic power results for the BM, area, and BM + area tests. Analytic power calculations for the maximum and BM + maximum tests are not tractable, so we resorted to empirical examples.

The limited empirical results presented in this article are in agreement with the more extensive studies in [2, 19, and 22]. In particular, the tests were valid in all of our examples. As the examples in Section 5 clearly show, there is not necessarily a choice of k, b, and b' that yields the most powerful test in all situations. However, the BM, BM + area, and BM + maximum tests fared particularly well when b and b' were large. Unfortunately, when we divided simulation output into many batches, the performance of the non-BM tests was disappointing. But as demonstrated by Example 5, a conservative yet powerful approach may be to use the area or maximum estimators after dividing the simulation into a smaller number of batches. In fact, in their empirical study, Cash et al. [2] recommended the compromise choice of $b \doteq 8$ (if the batch size m is "large enough" to meet the asymptotic requirements) and $b'/b \doteq 0.75$. Further, they state that the maximum and BM + max tests fare comparatively well in terms of power (although those tests sometimes yield higher size than desired).

APPENDIX

In this appendix, we justify Eq. (5): If $\overline{X}_{i,m} \sim \text{nor}(\mu(1 - \overline{a}_{i,m}), \sigma^2/m)$, $i = 1, 2, \ldots, b$, and if the $\overline{X}_{i,m}$'s are independent, then

$$m \sum_{i=1}^{b} (\overline{X}_{i,m} - \overline{\overline{X}}_n)^2 \sim \sigma^2\chi^2\left(b - 1, \frac{m\mu^2}{\sigma^2} \sum_{i=1}^{b} (\overline{a}_{i,m} - \overline{\overline{a}}_b)^2\right).$$

PROOF: (cf. Goldsman [8, Appendix A.3]): Consider independent $U_i \sim \text{nor}(\nu_i, \tau^2)$, $i = 1, 2, \ldots, b$. Define the $1 \times b$ vector $\mathbf{Z} \equiv \mathbf{UH}'$, where $\mathbf{H}$ is a nonrandom orthonormal $b \times b$ matrix whose last row is $(1/\sqrt{b}, \ldots, 1/\sqrt{b})$. (Such a matrix is easy to construct.) Then $Z_b = \sqrt{b}\,\overline{U}_b$, where $\overline{U}_b \equiv \Sigma_{i=1}^{b} U_i/b$. Since $\mathbf{H}$ is orthonormal,

$$\mathbf{ZZ}' = \mathbf{UH}'(\mathbf{UH}')' = \mathbf{UH}'\mathbf{HU}' = \mathbf{UU}'.$$

So $\Sigma_{i=1}^{b} Z_i^2 = \Sigma_{i=1}^{b} U_i^2$, and we obtain

$$S^2 \equiv \sum_{i=1}^{b} (U_i - \overline{U}_b)^2 = \sum_{i=1}^{b} U_i^2 - b\overline{U}_b^2 = \sum_{i=1}^{b} Z_i^2 - Z_b^2 = \sum_{i=1}^{b-1} Z_i^2.$$

Since $\mathbf{H}$ is nonrandom, $E[\mathbf{Z}] = E[\mathbf{UH}'] = \boldsymbol{\nu}\mathbf{H}'$. If we denote the $b \times b$ identity matrix by $\mathbf{I}_b$, then the covariance matrix of $\mathbf{Z}$ is

$$\begin{aligned} E[(\mathbf{Z} - \boldsymbol{\nu}\mathbf{H}')'(\mathbf{Z} - \boldsymbol{\nu}\mathbf{H}')] &= E[(\mathbf{UH}' - \boldsymbol{\nu}\mathbf{H}')'(\mathbf{UH}' - \boldsymbol{\nu}\mathbf{H}')] \\ &= E[\mathbf{H}(\mathbf{U} - \boldsymbol{\nu})'(\mathbf{U} - \boldsymbol{\nu})\mathbf{H}'] \\ &= \mathbf{HI}_b\tau^2\mathbf{H}' = \mathbf{I}_b\tau^2, \end{aligned}$$

because the U_i's are independent. Since each Z_i is a linear combination of normal random variables, we have $\mathbf{Z} \sim \text{nor}_b(\boldsymbol{\nu}\mathbf{H}', \mathbf{I}_b\tau^2)$. Thus, the Z_i's are independent normals.

Now, $S^2/\tau^2 = \sum_{i=1}^{b-1} (Z_i/\tau)^2 \sim \chi^2(b - 1, \delta)$, where δ is computed as follows: If $Q \sim \chi^2(b - 1, \delta)$, then $E[Q] = b - 1 + \delta$ [16, p. 315]. Letting $\bar{\nu}_b \equiv \sum_{i=1}^{b} \nu_i/b$, we get

$$
\begin{aligned}
E[S^2] &= \sum_{i=1}^{b} E[U_i^2] - bE[\overline{U}_b^2] \\
&= \sum_{i=1}^{b} \{\text{var}(U_i) + E^2[U_i]\} - b\{\text{var}(\overline{U}_b) + E^2[\overline{U}_b]\} \\
&= b\tau^2 + \sum_{i=1}^{b} \nu_i^2 - \tau^2 - b\bar{\nu}_b^2 \\
&= (b - 1)\tau^2 + \sum_{i=1}^{b} (\nu_i - \bar{\nu}_b)^2.
\end{aligned}
$$

ACKNOWLEDGMENTS

We are grateful to Professor Lionel Weiss for suggestions concerning (5). We also thank the Associate Editor and a referee for valuable comments. The work of David Goldsman and James J. Swain was partially supported by National Science Foundation Grant No. DDM-9012020.

REFERENCES

[1] Billingsley, P., *Convergence of Probability Measures*, Wiley, New York, 1968.
[2] Cash, C.R., Nelson, B.L., Long, J.M., Dippold, D.G., and Pollard, W.P., "Evaluation of Tests for Initial-Condition Bias," *Proceedings of the 1992 Winter Simulation Conference*, 1992, pp. 577–585.
[3] Chance, F., "A Historical Review of the Initial Transient Problem in Discrete Event Simulation Literature," Technical Report, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1993.
[4] Crane, M.A., and Iglehart, D.L., "Simulating Stable Stochastic Systems: III. Regenerative Processes and Discrete-Event Simulations," *Operations Research* **23**, 33–45 (1975).
[5] Fishman, G.S., *Concepts and Methods in Discrete Event Digital Simulation*, Wiley, New York, 1973.
[6] Fishman, G.S., *Principles of Discrete Event Simulation*, Wiley, New York, 1978.
[7] Glynn, P.W., and Iglehart, D.L., "Simulation Output Analysis Using Standardized Time Series," *Mathematics of Operations Research*, **15**, 1–16 (1990).
[8] Goldsman, D., "On Using Standardized Time Series to Analyze Stochastic Processes," Ph.D. thesis, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, New York, 1984.
[9] Goldsman, D., Kang, K., and Seila, A.F., "Cramér-von Mises Variance Estimators for Simulations," Technical Report, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 1993.

[10] Goldsman, D., Meketon, M.S., and Schruben, L.W., "Properties of Standardized Time Series Weighted Area Variance Estimators," *Management Science*, **36**, 602–612 (1990).

[11] Goldsman, D., and Schruben, L.W., "New Confidence Interval Estimators Using Standardized Time Series," *Management Science*, **36**, 393–397 (1990).

[12] Heidelberger, P., and Welch, P.D., "A Spectral Method for Confidence Interval Generation and Run-Length Control in Simulations," *Communications of the ACM*, **24**, 233–245 (1981).

[13] Johnson, N.L., and Kotz, S., *Distributions in Statistics—Continuous Univariate Distributions—2*, Wiley, New York, 1970.

[14] Lewis, P.A.W., "Simple Models for Positive-Valued and Discrete-Valued Time Series with ARMA Correlation Structure," in P.K. Krishnaiah (Ed.), *Multivariate Analysis—V*, North Holland, New York, 1980, pp. 151–166.

[15] Meketon, M.S., and Schmeiser, B.W., "Overlapping Batch Means: Something for Nothing?" *Proceedings of the 1984 Winter Simulation Conference*, 1984, pp. 227–230.

[16] Rohatgi, V.K., *An Introduction to Probability Theory and Mathematical Statistics*, Wiley, New York, 1976.

[17] Schriber, T.J., and Andrews, R.W., "ARMA-Based Confidence Intervals for Simulation Output Analysis," *American Journal of Mathematical and Management Sciences*, **4**, 345–373 (1984).

[18] Schruben, L.W. "Control of Initialization Bias in Multivariate Simulation Response," *Communications of the ACM*, **24**, 246–252 (1981).

[19] Schruben, L.W., "Detecting Initialization Bias in Simulation Output," *Operations Research*, **30**, 569–590 (1982).

[20] Schruben, L.W., "Confidence Interval Estimation Using Standardized Time Series," *Operations Research*, **31**, 1090–1108 (1983).

[21] Schruben, L.W., and Goldsman, D., "Initialization Effects in Computer Simulation Experiments," Technical Report No. 594, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY, 1985.

[22] Schruben, L.W., Singh, H., and Tierney, L., "Optimal Tests for Initialization Bias in Simulation Output," *Operations Research*, **31**, 1167–1178 (1983).

[23] Wilson, J.R., and Pritsker, A.A.B., "A Survey of Research on the Simulation Startup Problem," *Simulation*, **31**, 55–58 (1978).

[24] Wilson, J.R., and Pritsker, A.A.B., "Evaluation of Startup Policies in Simulation Experiments," *Simulation*, **31**, 79–89 (1978).

Manuscript received January 9, 1990
Revised manuscript received June 1993
Accepted July 14, 1993