SPSTS: A sequential procedure for estimating the steady-state mean using standardized time series

CHRISTOS ALEXOPOULOS¹
DAVID GOLDSMAN²
PENG TANG³
H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology,
Atlanta, Georgia 30332, USA

JAMES R. WILSON⁴
Edward P. Fitts Department of Industrial and Systems Engineering,
North Carolina State University,
Campus Box 7906,
Raleigh, NC 27695-7906, USA

¹E-mail: christos@isye.gatech.edu. Telephone: (404) 894-2361. Fax: (404) 894-2301.
²Fellow, Institute of Industrial Engineers. E-mail: sman@gatech.edu. Telephone: (404) 894-2365. Fax: (404) 894-2301.
³E-mail: peggypeng11@gmail.com. Telephone: (678) 608-7922. Fax: (404) 894-2301.
⁴Fellow, Institute of Industrial Engineers, and corresponding author. E-mail: jwilson@ncsu.edu. Telephone: (919) 515-6415. Fax: (919) 515-5281.
SPSTS: A Sequential Procedure for Estimating the Steady-State Mean Using Standardized Time Series

This article motivates, develops, and evaluates SPSTS, an automated sequential procedure for computing point and confidence-interval (CI) estimators for the steady-state mean of a simulation-generated process such that the final CI satisfies a user-specified relative precision requirement and approximately achieves a user-specified coverage probability. SPSTS uses standardized time series (STS) area estimators of the steady-state variance parameter (i.e., the sum of covariances at all lags). To estimate the end of the warm-up period, the von Neumann randomness test is applied sequentially to the signed areas corresponding to the relevant STS area variance estimators based on progressively larger batch sizes; and the final accepted batch size is taken as the length of the warm-up period. Then to determine a (larger) batch size sufficient to ensure adequate convergence of the STS area variance estimators to their (approximate) limiting scaled chi-squared distributions, the Shapiro-Wilk normality test is sequentially applied to the signed areas with progressively larger batch sizes. SPSTS’s performance is compared experimentally with that of recent batch-means methods using selected test problems of varying degrees of difficulty. SPSTS performed comparatively well in terms of its average required sample size as well as the coverage and average half-length of its delivered CIs.

Keywords: steady-state simulation; simulation output analysis; method of batch means; method of standardized time series; area variance estimator; nonoverlapping variance estimator; overlapping variance estimator

1. Introduction

Perhaps the most basic problem in simulation output analysis is to obtain valid point and confidence-interval (CI) estimators for the mean $\mu$ of a stationary simulation-generated process $X = \{X_k : k = 1, 2, \ldots \}$. If the simulation is in steady-state (warmed-up) operation, then to estimate $\mu$ we typically use the sample mean $\bar{X}_n = n^{-1} \sum_{k=1}^{n} X_k$ based on a simulation-generated time series of length $n$. Of course, a proper statistical analysis ought to also include a measure of the sample mean’s precision as an estimator of $\mu$; and so there is interest in estimating $\text{Var}[\bar{X}_n]$. This is often difficult because in practice one has to deal with the following deleterious effects: (i) initialization bias in the point estimator $\bar{X}_n$; (ii) marked departures from normality (e.g., pronounced skewness) in the marginal distribution of the $\{X_k\}$; and (iii) a high degree of correlation between successive observations. For example, if $X_k$ denotes the sojourn time of the $k$th workpiece passing through a heavily congested production system, then the marginal distribution of $X_k$ is typically skewed, and successive system sojourn times are highly correlated. Moreover, if the simulation is initialized in the empty-and-idle state or some other atypical initial condition, then a prolonged warm-up period may be required before the sojourn times exhibit steady-state behavior. These anomalies cause substantial difficulties not only in computing an accurate point estimator of $\text{Var}[\bar{X}_n]$ but also in computing a valid CI estimator of $\mu$.

Much research has been devoted to the problem of estimating $\text{Var}[\bar{X}_n]$ and the associated variance...
parameter $\sigma^2 \equiv \lim_{n \to \infty} n \text{Var}[\bar{X}_n]$. General methods for estimating $\text{Var}[\bar{X}_n]$ or $\sigma^2$ include the following: nonoverlapping batch means (NBM) (Fishman, 2001; Steiger and Wilson, 2001); overlapping batch means (OBM) (Meketon and Schmeiser, 1984); spaced batch means (Lada et al., 2008); spectral analysis (Heidelberger and Welch, 1981; Lada and Wilson, 2006; Lada et al., 2007); standardized time series (STS) (Schruben, 1983; Alexopoulos et al. 2007); and many other techniques described in Chapter 9 of Law (2015). As explained in Section 2, these techniques typically group a fixed number of observations into nonoverlapping or overlapping batches in an attempt to obtain low-bias, low-variance estimators for $\sigma^2$ and valid CI estimators for $\mu$, i.e., CIs that achieve a user-specified coverage probability. In particular, most estimators of $\sigma^2$ based on STS have better asymptotic properties (e.g., zero first-order bias and smaller mean squared error) than their NBM-based competitors, but often exhibit larger small-sample bias in the presence of positive autocorrelation due to substantial higher-order terms; see Aktaran-Kalaycı et al. (2007) and Theorem 1 below.

While fixed-sample-size estimators provide a good starting point for proper analysis of the key performance characteristics of the target system, they suffer from certain drawbacks that reduce their effectiveness. In advance of performing a simulation experiment using a fixed-sample-size procedure, it is generally impossible to guarantee that the final CI delivered by the procedure will achieve the desired coverage probability and will have a satisfactory relative precision (i.e., the ratio of its half-length to the magnitude of its midpoint). In particular, a fixed sampling budget $n$ may be insufficient to handle adequately the anomalous effects (i)–(iii) mentioned above. To the best of our knowledge, all methods for effectively eliminating these effects require a sufficiently large simulation budget that depends critically on the behavior of the target system; and in our experience, it is generally impossible to assign such a simulation budget in advance of performing any simulation experimentation. Hence, there is the need for intuitive, easy-to-use, yet rigorous sequential procedures that can systematically address these problems.

Since the mid-1990s, the literature has been enhanced by a variety of sequential NBM-based methods for computing valid CIs for $\mu$ such as the LABATCH.2 suite of algorithms of Fishman and Yarberry (1997), the ASAP3 procedure of Steiger et al. (2005), and the recent Skart procedure of Tafazzoli and Wilson (2011); these procedures are reviewed briefly in Section 4. The development and implementation of a sequential procedure involves trade-offs between the following competing objectives: (a) establishment of the asymptotic validity of the resulting CI for $\mu$ as the upper bound on its absolute or relative CI precision tends to zero; (b) validity of the CI in the case of no precision requirement; and (c) robust performance when tested on a variety of challenging simulation output processes. For instance, the methods in Fishman and Yarberry (1997) yield strongly consistent variance estimators (hence asymptotically valid CIs for $\mu$), but their implementation in a sequential fashion requires extensive user intervention. In fact, the experimentation in Steiger et al. (2005) illustrated that the LABATCH.2 methods perform poorly in the case of no precision requirement without such intervention. On the other hand, the ASAP3 and Skart methods meet objectives (b)–(c) and are considered as state-of-the-art amongst sequential NBM-based methods. But in order to meet objectives (b)–(c), both ASAP3 and Skart employ distributional approximations and adjustments related to the residual autocorrelation and skewness of the batch means; these approximations make the establishment of objective (a) an onerous, if not impossible, task.
Despite the attention devoted to STS-based estimators over the past three decades, the literature lacks automated sequential procedures involving such estimators. Nakayama (1994) proposes a two-stage procedure based on the STS area estimators reviewed in Section 2.3 of the current paper but provides no definitive method for selecting an initial batch size—a critical omission that severely limits the general applicability of the procedure. Moreover, the cursory performance evaluation of this procedure is confined to relatively “easy” versions of the $M/M/1$ queue-waiting-time process that are free of warm-up effects or pronounced autocorrelations as elaborated in Section 5.2.

In this article, we formulate and evaluate SPSTS, the first automated Sequential Procedure for computing valid CIs for the steady-state mean of a simulation-generated process using variance estimators based on STS. In particular, SPSTS employs batched area estimators due to Foley and Goldsman (1999) and has the following advantages:

- **It is relatively much simpler to implement** compared with competing sequential methods based on NBM, such as ASAP3 and Skart. Most importantly, many of the operating parameters necessary to apply the procedure have been pretuned based on their good performance across a wide variety of test problems with varying degrees of difficulty; thus, the number of decisions required for the user to run the procedure is minimized.

- **It overcomes the aforementioned small-sample bias of STS-based area estimators** by using a substantially larger initial batch size than its NBM-based competitors as well as aggressive statistical hypothesis testing in the early stages. These features of SPSTS induce larger sample sizes in the case of no precision requirement, but the resulting CIs for $\mu$ are tighter and have coverage properties that are comparable with those of ASAP3 and Skart. Substantial experimentation with test processes having challenging transient, distributional, or autocorrelation structures reveals that the estimated coverage of the CIs makes a graceful transition to the nominal coverage as the upper bound on the relative half-length of the CI for $\mu$ decreases from infinity (i.e., the case of no precision requirement) to values below 0.10; in those practically useful cases SPSTS frequently outperforms its NBM-based competitors.

- Although estimation of the variance parameter $\sigma^2$ is generally much more difficult than estimation of the steady-state mean $\mu$ (see, for example, Alexopoulos et al., 2007a and Lada et al., 2007), SPSTS has the potential to deliver readily both a point estimator and a valid CI estimator of $\sigma^2$. This property distinguishes SPSTS from its state-of-the-art NBM-based competitors.

The rest of this article is organized as follows. In Section 2 we present the fundamental assumptions and the resulting properties on which SPSTS is based. In Section 3 we explain the structure and operation of SPSTS. Section 4 provides an overview of ASAP3 and Skart, the NBM-based simulation analysis procedures whose performance is to be compared experimentally with that of SPSTS. Section 5 illustrates the performance of SPSTS in test processes that are often used as benchmark problems for evaluating output-analysis procedures. Section 6 summarizes our findings and conclusions. Alexopoulos et al. (2013) is a preliminary, abridged version of this article.
2. Background

The following discussion is necessary to establish basic notation, background, and motivation for the estimators and sequential procedures formulated in this article. Among the assumptions enumerated below, the first property is satisfied by a variety of stationary processes, including strongly mixing processes, associated stationary processes, and regenerative processes (Durrett, 2005; Glynn and Iglehart, 1990).

2.1. Assumptions

ASSUMPTION A.1 [Functional Central Limit Theorem (FCLT)]. Suppose that the process $X = \{X_1, X_2, \ldots \}$ is stationary. Define the covariance function $R_j \equiv \text{Cov}[X_1, X_{1+j}]$ for $j = 0, \pm 1, \pm 2, \ldots$, and assume $\sum_{j=-\infty}^{\infty} j^2 |R_j| < \infty$ so that we have $\sigma^2 = \lim_{n \to \infty} n \text{Var}[\bar{X}_n] = \sum_{j=-\infty}^{\infty} R_j < \infty$. Further, assume $\sigma^2 > 0$, and that the sequence of random functions $Y_n(t) \equiv \frac{\lfloor nt \rfloor (\bar{X}_{\lfloor nt \rfloor} - \mu)}{\sigma \sqrt{n}}$ for $t \in [0, 1]$ and $n = 1, 2, \ldots$ satisfies $Y_n(\cdot) \xrightarrow{n \to \infty} W(\cdot)$, where $\lfloor \cdot \rfloor$ is the greatest integer function; $W(\cdot)$ is a standard Brownian motion process on $[0, 1]$; and $\xrightarrow{n \to \infty}$ denotes weak convergence (as $n \to \infty$) in the Skorohod space $D[0, 1]$ of real-valued functions on $[0, 1]$ that are right-continuous with left-hand limits.

ASSUMPTION A.2. The weight function $f(\cdot)$ on $[0, 1]$ has a continuous second derivative and is normalized so that $\int_0^1 f^\prime(s)f(s)\lfloor \min(s,t) - st \rfloor ds dt = 1$. We define $F(t) \equiv \int_0^t f(s) ds$ and $\bar{F}(t) \equiv \int_0^t F(s) ds$ for $t \in [0, 1]$; and we let $F \equiv F(1)$ and $\bar{F} \equiv \bar{F}(1)$.

Assumption A.1 implies that if $\bar{X}_n$ and $\hat{\sigma}^2$ are asymptotically independent as $n \to \infty$ and if $\hat{\sigma}^2 \xrightarrow{n \to \infty} \sigma^2 \chi^2_v / v$, where $\chi^2_v$ denotes a chi-squared distribution with $v$ degrees of freedom, then

$$\bar{X}_n \pm t_{v,1-\alpha/2} \sqrt{\hat{\sigma}^2 / n} \quad (1)$$

is an asymptotically valid 100(1 - $\alpha$)% CI for $\mu$, where $t_{v,\beta}$ is the $\beta$ quantile of Student’s $t$ distribution with $v$ degrees of freedom depending on the type of the estimator. The absolute precision of the CI (1) is its half-length, while the relative precision of (1) is the ratio of its half-length to $|\bar{X}_n|$.

We define the unknown constants $\gamma_k = 2 \sum_{j=1}^{\infty} j^k R_j$, $k = 1, 2, \ldots$. Song and Schmeiser (1995) use the terms “balance point” and “center of gravity” to refer to the ratio $\gamma_1/(R_0 + \gamma_0) = \gamma_1/\sigma^2$ as a geometric characterization of the associated autocorrelation function. We also use the following notation: (i) $p(n) = O(q(n))$ means that there are positive constants $c$ and $n_0$ such that $0 \leq p(n) \leq cq(n)$ for all $n \geq n_0$; and (ii) $p(n) = o(q(n))$ means that $\lim_{n \to \infty} p(n)/q(n) = 0$.

2.2. Nonoverlapping Batch Means

The concept of batching has a long history in the simulation output analysis literature. Suppose we form $b$ nonoverlapping batches, each consisting of $m$ observations (so that $n = bm$). Specifically,
batch $i$ consists of $\{X_{(i-1)m+j} : j = 1, \ldots, m\}$. Finally, for $i = 1, \ldots, b$ and $k = 1, \ldots, m$, let \( \bar{X}_{i,k} \equiv k^{-1} \sum_{\ell=1}^{k} X_{(i-1)m+\ell} \).

If the process $X$ satisfies Assumption A.1, then as $m \to \infty$ the nonoverlapping batch means $\bar{X}_{i,m}$ for $i = 1, \ldots, b$ become uncorrelated (Law and Carson, 1979) and normally distributed. The NBM estimator for $\sigma^2$ is defined by

$$\mathcal{N}(b, m) \equiv \frac{m}{b-1} \sum_{i=1}^{b} (\bar{X}_{i,m} - \bar{X}_n)^2, \tag{2}$$

that is, $m$ times the sample variance of the batch means. If $b$ is fixed, then $\mathcal{N}(b, m) \xrightarrow{m \to \infty} \sigma^2 \chi^2_{b-1}/(b-1)$, and the usual NBM CI estimator for the mean,

$$\bar{X}_n \pm t_{b-1,1-a/2} \sqrt{\frac{\mathcal{N}(b, m)}{n}},$$

is asymptotically valid as $m \to \infty$ (see, for example, Steiger and Wilson, 2001). Under some additional mixing assumptions, one can also show that

$$\mathbb{E}[\mathcal{N}(b, m)] = \sigma^2 - \frac{(b+1)\gamma_1}{bm} + o\left(\frac{1}{m}\right), \tag{3}$$

and

$$\text{Var}[\mathcal{N}(b, m)] = \frac{2\sigma^4 (b+1)}{(b-1)^2} + O\left(\frac{1}{bm^{1/4}}\right) + o\left(\frac{1}{n}\right);$$

see Alexopoulos and Goldsman (2004). If both $m$ and $b$ tend to $\infty$, then the last two equations imply weak consistency for $\mathcal{N}(b, m)$ because $\text{MSE}[\mathcal{N}(b, m)] \to 0$.

### 2.3. Batched STS Area Estimators

The STS for the sample $\{X_1, \ldots, X_n\}$ is defined by

$$T_n(t) = \frac{|nt| (\bar{X}_n - \bar{X}_{|nt|})}{\sigma \sqrt{n}} \quad \text{for } t \in [0, 1] \text{ and } n = 1, 2, \ldots. \tag{4}$$

Under Assumption A.1, it can be shown that

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

where $\mathcal{B}(\cdot)$ is a standard Brownian bridge process that is independent of $\mathcal{W}(1)$, so that $T_n(\cdot)$ and $\bar{X}_n$ are asymptotically independent (Schruben, 1983).

**Remark 1.** Schruben (1983) coined the term “standardized time series” to refer to the continuous-time stochastic process (4) derived from the time series $\{X_k : k = 1, \ldots, n\}$. Some authors also use this term to refer to the NBM and OBM methods of steady-state simulation analysis; see, for example, Glynn and Iglehart (1990). However, in most of recent simulation literature a distinction is made between simulation analysis methods based purely on batch means and those methods based on the standardized time series.
As in Section 2.2, we form \( b \) nonoverlapping batches, each consisting of \( m \) observations. The STS from batch \( i \) is

\[
T_{i,m}(t) = \frac{[mt](\bar{X}_{i,m} - \bar{X}_{i,1}mt)}{\sigma \sqrt{m}} \quad \text{for } t \in [0, 1], m = 1, 2, \ldots, \text{ and } i = 1, \ldots, b.
\]

Under Assumption A.1, we have

\[
(\sqrt{m}(\bar{X}_{1,m} - \mu), \ldots, \sqrt{m}(\bar{X}_{b,m} - \mu); \sigma T_{1,m}, \ldots, \sigma T_{b,m}) \xrightarrow{m \to \infty} (\sigma \xi_1, \ldots, \sigma \xi_b; \sigma B_0, \ldots, \sigma B_{b-1}).
\]

(5)

where the \( \xi_i \) are independent and identically distributed (i.i.d.) standard normal random variables, and \( B_s(\cdot) \) denotes a standard Brownian bridge defined in terms of \( \mathcal{W}(\cdot) \) on \([s, s+1]\), for \( s \in [0, b-1] \). That is,

\[
B_s(t) \equiv t[\mathcal{W}(s+1) - \mathcal{W}(s)] - [\mathcal{W}(s+t) - \mathcal{W}(s)] \quad \text{for } t \in [0, 1] \text{ and } s \in [0, b-1].
\]

It follows immediately that the Brownian bridges \( B_0(\cdot), B_1(\cdot), \ldots, B_{b-1}(\cdot) \) are independent.

We define the *signed areas* associated with the STSs computed from each batch using the weight function \( f(\cdot) \) as follows:

\[
Z_i(f; m) = \frac{1}{m} \sum_{\ell=1}^{m} f(\ell/m) \sigma T_{i,m}(\ell/m)
\]

\[
= \frac{1}{m^{3/2}} \sum_{\ell=1}^{m} f(\ell/m) \ell(\bar{X}_{i,m} - \bar{X}_{i,\ell}) \quad \text{for } i = 1, \ldots, b.
\]

(6)

Equation (5) ensures that

\[
\left[ Z_1(f; m), Z_2(f; m), \ldots, Z_b(f; m) \right] \xrightarrow{m \to \infty} \sigma [\xi_1, \xi_2, \ldots, \xi_b]
\]

(7)

so that the \( \{Z_i(f; m) : i = 1, \ldots, b\} \) become i.i.d. \( \sigma N(0, 1) \) as \( m \to \infty \) (Alexopoulos et al. 2007, Goldsman et al. 1990).

The *batched area* estimator for \( \sigma^2 \) is the average of the squares of the signed areas:

\[
\mathcal{A}(f; b, m) = \frac{1}{b} \sum_{i=1}^{b} Z_i^2(f; m) \xrightarrow{m \to \infty} \sigma^2 \chi_b^2 / b.
\]

(8)

Under Assumptions A.1 and A.2 and some mixing conditions, for fixed \( b \),

\[
\mathbb{E}[\mathcal{A}(f; b, m)] = \sigma^2 - \left[ \frac{(F - \bar{F})^2 + \bar{F}^2}{2m} \right]_{\gamma_1} + o\left(\frac{1}{m}\right);
\]

(9)
and if the random variables \( \{A^2(f; b, m) : m = 1, 2, \ldots \} \) are uniformly integrable (Durrett, 2005), then we have \( \lim_{m \to \infty} \text{Var}[A(f; b, m)] = 2a^4/b \) (Foley and Goldsman, 1999; Goldsman et al. 1990). If one chooses weights having \( F = F^* = 0 \), then the resulting area estimator is first-order unbiased since its bias is \( o(1/m) \). Damerdji and Goldsman (1995) gave conditions such that the estimator \( A(f; b, m) \) is strongly consistent as both \( b, m \to \infty \) in an appropriate fashion.

Popular choices of weight functions are

\[
\begin{align*}
    f_0(t) &= \sqrt{12} \quad \text{(Schruben, 1983)}, \\
    f_2(t) &= \sqrt{840(3t^2 - 3t + 1/2)} \quad \text{(Goldsman et al. 1990), and} \\
    f_{\cos, j}(t) &= \sqrt{8\pi j \cos(2\pi jt)}, \; j = 1, 2, \ldots \quad \text{(Foley and Goldsman, 1999)}
\end{align*}
\]  

(9)

Aktaran-Kalaycı et al. (2007) give detailed expressions for the expectations of \( A(f_0; b, m) \), \( A(f_2; b, m) \), and \( A(f_{\cos, 1}; b, m) \). In particular, Schruben’s estimator has substantial first-order bias given by \(-3\gamma_1/m\) (about 3 times that of the NBM estimator), while the remaining estimators are first-order unbiased and have comparable second-order bias terms. Foley and Goldsman (1999) show that the area estimators arising from the weight functions \( \{f_{\cos, j}(:, j = 1, 2, \ldots \} \) are also asymptotically independent. Thus we can average the first \( k \) of these area estimators to obtain a first-order unbiased estimator for \( \sigma^2 \) with more degrees of freedom:

\[
\bar{A}_k(f_{\cos}; b, m) = \frac{1}{k} \sum_{j=1}^{k} A(f_{\cos, j}; b, m) \xrightarrow{m \to \infty} \sigma^2 \chi_{bk}^2 / (bk) \quad \text{for} \; k = 1, 2, \ldots \ .
\]

Unfortunately, the applicability of the last asymptotic result under small-to-moderate batch sizes has limitations. The following theorem gives detailed expressions for the bias of the batched area estimators \( A(f_{\cos, j}; b, m) \) for \( j = 1, 2, \ldots \). This result extends Equation (13) of Aktaran-Kalaycı et al. (2007), which gives the bias of \( A(f_{\cos, 1}; b, m) \). The proof can be found in the Appendix.

**Theorem 1.** Suppose that \( X = \{X_1, X_2, \ldots \} \) is covariance stationary with covariance function \( \{R_i\} \) satisfying \( |R_i| = O(\delta^i) \) for \( i = 1, 2, \ldots \), where \( \delta \in (0, 1) \). Then for \( j \) fixed and \( \ll m \), the expected value of the batched area estimator \( A(f_{\cos, j}; b, m) \) is given by

\[
E[A(f_{\cos, j}; b, m)] = \sigma^2 + \frac{\pi^2 j^2(\sigma^2 - 6\gamma_2)}{3m^2} + \frac{2\pi^2 j^2(2\gamma_3 + \gamma_1)}{3m^3} + \frac{\pi^4 j^4(\sigma^2 + 10\gamma_4 - 10\gamma_2)}{15m^4} + O(m^{-5}).
\]

**Remark 2.** Under the hypothesis of Theorem 1 that \( |R_i| = O(\delta^i) \) for \( i = 1, 2, \ldots \), where \( \delta \in (0, 1) \), Equation (3) can be sharpened to \( E[N(b, m)] = \sigma^2 - (b + 1)\gamma_1/(bm) + O(\delta^m) \); see Corollary 3 of Aktaran-Kalaycı et al. (2007).

Comparisons between the bias expression for \( A(f_{\cos, j}; b, m) \) in Theorem 1 and the refined bias expression for \( N(b, m) \) in Remark 2 reveal several interesting findings for processes with positive autocorrelation (as in Example 1 below), where one often has \( \sigma^2 \ll \gamma_1 \ll \gamma_2 \ll \cdots \): (a) the negative first-order bias of \( N(b, m) \) can be dominated by the second-order bias of \( A(f_{\cos, j}; b, m) \) for small values of the batch size \( m \), but this “anomaly” vanishes as \( m \) becomes sufficiently large (see various plots in Aktaran-Kalaycı et al.,
2007); and (b) the second-order term of the bias of \( A(f_{\cos,j};b,m) \) becomes more negative as \( j \) increases. Notice that the powers of \( j \) occur in all terms of order \( \geq 2 \).

Algorithm SPSTS uses these findings in three different ways. First, it starts with a substantially larger batch size than its NBM-based competitors. Second, it employs the signed areas in Equation (6) corresponding to the first two weights \( f_{\cos,1}(\cdot) \) and \( f_{\cos,2}(\cdot) \) in all statistical tests for removing the bias due to the initial conditions and assessing the independence and normality of the signed areas. The incorporation of the estimators based on the weight \( f_{\cos,2}(\cdot) \) was supported by substantial experimentation associated with the simulation study in Section 5, which revealed that the signed areas \( f_{Z_i}(f_{\cos,2};b,m) \) frequently require larger batch sizes to pass the hypothesis tests than do \( f_{Z_i}(f_{\cos,1};b,m) \). (We could have adopted the estimator \( A(f_2; b, m) \) in place of \( A(f_{\cos,2}; b, m) \), but we preferred to exploit the second-order bias of the latter estimator as a device for ensuring adequate batch sizes in the early stages of testing the associated signed areas; moreover we preferred \( A(f_{\cos,2}; b, m) \) because it is asymptotically independent of \( A(f_{\cos,1}; b, m) \).) Third, in its final stages SPSTS uses the maximum of \( A(f_{\cos,j}; b, m) \), \( j = 1, 2 \), and the OBM estimator in Section 2.4 as a “safeguard” against the effects of the potential negative bias of the variance estimators.

Example 1. Consider an \( M/M/1 \) queueing system with arrival rate \( \omega \) and traffic intensity \( \rho < 1 \). Let \( X_k \) be the time spent in queue by entity (customer) \( k \) prior to receiving service, let \( S_k \) be the service time for entity \( k \), and let \( A_k \) be the time between arrivals \( k - 1 \) and \( k \). If the delay of the first entity is generated from the cumulative distribution function (c.d.f.)

\[
\Pr(X_1 \leq x) = \begin{cases} 
0, & \text{if } x < 0, \\
1 - \rho, & \text{if } x = 0, \\
1 - \rho \exp[-\omega(1-\rho)x/\rho], & \text{if } x > 0,
\end{cases}
\]

while the remaining delays are generated using Lindley’s recursion

\[
X_k = \max\{X_{k-1} + S_{k-1} - A_k, 0\}, \quad k = 2, 3, \ldots,
\]

then the process \( \{X_1, X_2, \ldots\} \) is stationary with mean \( \mu = \rho^2/\left[\omega(1-\rho)\right] \) and variance parameter \( \sigma^2 = \rho^3(2 + 5\rho - 4\rho^2 + \rho^3)/[\omega^2(1-\rho)^4] \); see, for example, Steiger and Wilson (2001). We consider the case when \( \omega = 0.8 \) and \( \rho = 0.8 \); hence \( \mu = 4 \) and \( \sigma^2 = 1976 \).

Figure 1 plots the exact expected values of \( A(f_{\cos,j}; 1, m) \) for \( j = 1, \ldots, 8 \) and \( m = 2^p, \quad p = 7, \ldots, 16 \), based on Equation (16) from the Appendix and the following expression for the covariance function of \( M/M/1 \) queue waiting times:

\[
R_i = \frac{1 - \rho^2}{2\pi \omega^2} \int_0^{r} \frac{t^{i+3/2}(r-t)^{1/2}}{(1-t)^3} \, dt \quad \text{for } i = 1, 2, \ldots, \tag{10}
\]

where \( r = 4\rho/[(1+\rho)^2] \); see Equation (18) of Aktaran-Kalayci et al. (2007) and the references cited therein. The effect of \( j \) with respect to the magnitude of the negative bias of \( A(f_{\cos,j}; 1, m) \) is apparent.
Fig. 1. Expected values of area variance estimators based on 8 cosine weight functions for an $M/M/1$ queue-waiting-time process with $\rho = 0.8, \sigma^2 = 1976, b = 1$, and sample sizes $2^\ell, \ell = 7, \ldots, 16$.

2.4. Overlapping Batch Means

In various NBM procedures, von Neumann’s (1941) randomness test is used to determine a batch size that is sufficiently large to ensure that the resulting batch means are approximately i.i.d. normal with mean $\mu$. Unfortunately in large-scale applications, we often find that the resulting batch means exhibit nonnegligible correlation and skewness (Steiger et al. 2005; Tafazzoli and Wilson, 2011). Such correlation usually causes systematic underestimation of $\text{Var} [\bar{X}_n]$ (that is, negative bias); and this effect coupled with nonnegligible skewness of the batch means can cause the final CI coverage to fall substantially below its nominal level.

ASAP3 and Skart sought to estimate this correlation and compensate for it by modeling the batch means as a stationary AR(1) process for a sufficiently large batch size; moreover Skart also aimed at estimating the skewness of the batch means and then to compensate for it by applying a suitable adjustment to the final CI. This approach proved to be effective in a broad diversity of applications.

In the development of SPSTS, we observed similar problems with the von Neumann randomness test and the Shapiro–Wilk normality test when those tests were applied to the signed areas so as to determine a batch size $m$ that is sufficiently large to ensure that the $\{Z_i(f;m) : i = 1, \ldots, b\}$ are approximately i.i.d. normal with variance $\sigma^2$. In particular, the analysis surrounding Theorem 1 in the preceding section demonstrated that the STS area estimators $A(f;b,m)$ of $\sigma^2$ often exhibit negative second-order bias; and this effect sometimes causes the final CI coverage to fall below its nominal level. To address this issue, we also consider variance estimators based on overlapping batch means, which for very small sample sizes can exhibit lower (though still significant) bias than do other estimators; but generally speaking, as sample sizes increase a bit, the various first-order-unbiased estimators have better bias properties than OBM estimators.

Given a “shift” $d \in \{1, \ldots, m - 1\}$, we define the overlapping batches $\{X_{(i-1)d+1}, \ldots, X_{(i-1)d+m}\}$ for $i = 1, \ldots, q$, where $q \equiv \lceil (n-m)/d \rceil + 1$. Within this context we define $b \equiv n/m$, with the understanding
that it does not denote the number of batches. The respective overlapping batch means are

\[ \bar{X}_{i,m;d} = \frac{1}{m} \sum_{j=1}^{m} X_{(i-1)d+j} \quad \text{for} \quad i = 1, \ldots, q, \]

and the OBM estimator for \( \sigma^2 \) is

\[ \bar{O}(b,m;d) = \frac{m}{q-1} \sum_{i=1}^{q} (\bar{X}_{i,m;d} - \bar{X})^2. \]  \hspace{0.1mm} (11)

The assignment \( d = 1 \) yields the estimator of Meketon and Schmeiser (1984). Under the assumptions for the NBM estimator (2), it can be shown that the OBM variance estimator has about the same mean as NBM, but \( \text{Var}[\bar{O}(b,m;1)] \approx \frac{3}{4} \text{Var}[\bar{N}(b,m)] \). Welch (1987) noted that the NBM and OBM estimators are related to classical spectral estimators and showed that partial overlapping not only saves computational effort, but retains most benefits of extreme overlapping. For example, \( \text{Var}[\bar{O}(b,m;m/4)] \approx 0.69 \text{Var}[\bar{N}(b,m)] \) and \( \text{Var}[\bar{O}(b,m;m/2)] \approx 0.75 \text{Var}[\bar{N}(b,m)] \).

When the batch size \( m \) is sufficiently large, we can use Theorem 2.1 of Box (1954) to approximate the distribution of \( \bar{O}(b,m;d) \) by a weighted sum of i.i.d. \( \chi^2 \) random variables. Then the method of Satterthwaite (1941) yields

\[ \bar{O}(b,m;d) \sim \text{E}[\bar{O}(b,m;d)] \chi^2_{\nu}/\nu, \quad \text{where} \quad \nu = \left\lceil \frac{2E\text{E}^2[\bar{O}(b,m;d)]}{\text{Var}[\bar{O}(b,m;d)]} \right\rceil. \]  \hspace{0.1mm} (12)

\( \lceil z \rceil \) denotes rounding \( z \) to the nearest integer, and \( \nu \) is the “effective” degrees of freedom. In particular, for large \( m \) we have

\[ \nu \approx \begin{cases} \lceil 1.5b \rceil & \text{for} \quad d = 1 \\ \lceil 1.45b \rceil & \text{for} \quad d = m/4 \\ \lceil 1.33b \rceil & \text{for} \quad d = m/2. \end{cases} \]

Using Equation (12), we obtain the following approximate 100\((1 - \alpha)\)% CI for \( \mu \):

\[ \bar{X}_n \pm t_{\nu,1-\alpha/2} \sqrt{\frac{\bar{O}(b,m;d)}{n}}. \]

3. SPSTS: A Sequential Procedure Based on Standardized Time Series

In this section, we develop and evaluate SPSTS, the first sequential procedure for estimating the steady-state mean \( \mu \) based on standardized time series formed by nonoverlapping batches. The methodology is based primarily on the signed areas \( \{Z_i(f_{\cos,j};m) : i = 1, \ldots, b\} \) for \( j = 1, 2 \) as defined in Equation (6), and on the asymptotic result given in Equation (7). SPSTS has two advantages over recent methods for estimating \( \mu \) based on batch means and the associated Student’s \( t \)-statistic: it is relatively much simpler to implement (see Section 4 for an elaboration of this point); and it can also be used to construct valid CIs for the variance parameter \( \sigma^2 \) (Alexopoulos et al. 2007a). (In a distribution-free tabular CUSUM chart...
for statistical process control of autocorrelated processes, Lee et al. (2009) also used a version of the signed areas \( \{Z_i(f_2; m) : i = 1, \ldots, b\} \) for batch-size determination.) A high-level flowchart of SPSTS is depicted in Figure 2.

![Flowchart of SPSTS](image)

**Fig. 2.** High-level flow chart of SPSTS.

SPSTS starts with \( b = 32 \) nonoverlapping batches of size \( m_0 = 256 \) so that the initial simulation-generated time series \( \{X_k\} \) must have length at least \( n = bm = 8192 \). Step [1] of SPSTS uses a preliminary Shapiro–Wilk test to assess the normality of the sums \( Y_i = Z_i(f_{\cos,1}; m) + Z_i(f_{\cos,2}; m) \) for \( i = 1, \ldots, b \) at the significance level \( \beta = 0.20 \). If this test fails to reject normality, then the algorithm proceeds with the current batch size; otherwise, it increases the initial batch size to 2048. This step aims at avoiding data waste when each \( X_k \) is already approximately normal—for example, (a) \( X_k \) is the sum of some more basic individual observations such as the queue waiting times experienced by customer \( k \) in passing through a lengthy series of service centers composing the system of interest; or (b) \( X_k \) is the time-weighted average of a response observed in continuous time such as the queue length at a particular service center accumulated over the \( k \)th time interval \((k - 1)\Delta, k\Delta\) of fixed length \( \Delta \).

Designed to overcome the small-sample bias issues of STS area estimators, the approach of SPSTS to assigning the initial batch size is based on the following: (i) the theoretical results detailed in Aktaran-Kalaycı et al. (2007) as well as the preceding discussion of our Theorem 1 and Remark 2; and (ii) the extensive performance evaluation reported in Alexopoulos et al. (2007a). With respect to its method for
initial batch-size assignment, SPSTS stands in sharp contrast to NBM-based procedures such as ASAP3 and Skart, which start with many batches of small size and take advantage of the property that as the batch size $m$ increases, nonoverlapping batch means generally converge to approximate normality and therefore usually pass the normality test faster than do the signed estimators $\{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b\}$. As outlined in Section 1 and discussed in more detail in Sections 4.1 and 4.2, ASAP3 and Skart proceed by fitting an AR(1) model to the batch means process and deliver CIs for $\mu$ based on adjustments for residual correlation and skewness of the batch means. These steps lead to more-complex algorithmic structures. The potentially large initial batch size used by SPSTS can lead to significantly larger average final sample sizes than ASAP3 and Skart in the case of no precision requirement; but because of this larger initial batch size, SPSTS avoids the need to apply complex distributional approximations and CI adjustments.

To determine a subsequent batch size $m$ that is sufficiently large so that the $\{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b\}$ for $j = 1, 2$ are approximately i.i.d. $\sigma N(0,1)$ (that is, an adequate degree of convergence to the asymptotic result (7) has been achieved), SPSTS uses sequential versions of the von Neumann randomness test and the Shapiro–Wilk normality test. If the $\{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b\}$ fail iteration $\ell$ of the von Neumann randomness test ($\ell = 1, 2, \ldots$) at the significance level $\beta \psi(\ell)$, then the batch size $m$ is increased by a factor of about $\sqrt{2}$, the iteration counter $\ell$ is incremented, and the normality test is performed again. We use $\beta = 0.20$ and $\psi(\ell) \equiv \exp[-\eta(\ell - 1)^{\theta}]$; and the current values of the parameters $\eta = 0.184206$ and $\theta = 2$ as specified in Lada and Wilson (2006). These assignments allow the significance level to decrease rapidly from the “aggressive” value of 0.20 to the relatively small value of 0.002 after 6 iterations. Specifically, steps [2]–[3] of SPSTS apply the normality test to each of the signed area sequences $\{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b\}$ for $j = 1$ and then for $j = 2$. Let $m_j$ denote the smallest batch size at which the $\{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b\}$ pass this test for $j = 1$ and $j = 2$. We proceed with the conservative batch size $m = \max\{m_1, m_2\}$.

Step [4] of SPSTS addresses the simulation start-up problem by truncating (deleting) the first batch of observations. The idea behind this truncation is that once the signed areas pass the normality test in steps [2]–[3] of SPSTS, one can assume that any effects of initialization bias are effectively confined to the first batch so that statistics computed from subsequent batches are free of initialization bias.

Steps [5]–[6] of SPSTS apply the Shapiro–Wilk normality test. If the $\{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b\}$ fail iteration $\ell$ of the normality test (we reset $\ell = 1$ when the new test begins) at the significance level $\delta \psi(\ell)$, with $\delta = 0.20$, then the batch size $m$ is increased by a factor of about $\sqrt{2}$, the iteration counter $\ell$ is incremented, and the normality test is performed again. As with step [2] of SPSTS, for $j = 1$ and then for $j = 2$, we let $m_j'$ denote the smallest batch size at which the $\{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b\}$ pass the Shapiro–Wilk test. Then we move to the next step of SPSTS with the batch size $m = \max\{m_1', m_2'\}$.

The variance estimate in step [7] of SPSTS is the conservative maximum of the area estimates based on the cosine weights $f_{\cos,j}$ ($j = 1, 2$) and the OBM estimate in Equation (11) with $d = m/4$. The motivation for the use of the OBM estimator was given in Section 2.4; and its contribution to the estimated coverage of the final CI delivered by step [8] of SPSTS will be assessed experimentally at the end of Section 5.5. We point out that the use of the maximum estimator in this step does not affect the asymptotic validity of the final CI as the batch size $m \to \infty$. This validity follows from Equations (8) and (12). The accuracy of the
approximation (12) was assessed experimentally by Alexopoulos et al. (2007a).

Given the user-specified relative precision \( r^* \geq 0 \), SPSTS sequentially increases the total sample size \( n \), the batch size \( m \), and the batch count \( b \) in a way similar to ASAP3 and Skart. Specifically, it sets an upper limit on the batch count \( (b^* = 64) \) and estimates the number of batches \( b' \) required to satisfy the precision requirement under the current batch size. The batch count is set at \( b = \min\{b', b^*\} \). If \( b \) equals \( b' \), then the batch size remains intact; otherwise, the batch size is increased by a factor equal to the median value of \( \{1.05, b'/b, 2\} \). The formulation of the latter batch-size inflation factor is based on our extensive computational experience with ASAP3 and Skart; and it is designed to avoid the following problems: (i) numerous iterations of SPSTS in which the batch size is increased by a relatively small amount so that no substantive change in the CI’s relative precision is achieved; and (ii) iterations in which the usual projected increase in sample size could be excessive, especially in the early iterations of SPSTS. This updating scheme in step [7] of SPSTS ensures that the half-length of the final CI for \( \mu \) does not exceed \( r^*|\bar{X}_n| \) in terms of the final point estimator \( \bar{X}_n \) of \( \mu \). When \( r^* = \infty \) (i.e., in the case of no precision requirement), SPSTS terminates as soon as the signed areas based on the cosine weights \( f_{\cos,1}(\cdot) \) and \( f_{\cos,2}(\cdot) \) pass both tests for randomness and normality. Finite values of \( r^* \) should be reasonably small, say \( r^* \leq 0.15 \) (Law, 2015, pp. 506–507). A formal algorithmic statement of SPSTS is given below.

**Algorithm SPSTS**

---

[0] **Initialization:** Set \( \alpha, \beta = 0.20, \delta = 0.20, b^* = 64 \), and relative error upper bound \( r^* \). Define the function \( \psi(\ell) \equiv \exp\left[-\eta(\ell - 1)^6\right], \ell = 1, 2, \ldots, \) where \( \eta = 0.184206 \) and \( \theta = 2 \).

[1] Generate \( b = 32 \) batches of size \( m_0 = 256 \). Compute the sums \( Y_i = Z_i(f_{\cos,1}; m) + Z_i(f_{\cos,2}; m) \) of the signed areas corresponding to the weights \( f_{\cos,1}(\cdot) \) and \( f_{\cos,2}(\cdot) \) for \( i = 1, \ldots, b \). Apply the two-sided Shapiro–Wilk test for normality to the \( \{Y_i\} \) at the level of significance \( \beta = 0.20 \).

If the test fails to reject normality of the \( \{Y_i\} \), then set \( m \leftarrow m_0 \); otherwise, set \( m \leftarrow 2048 \) and generate \( b(m - m_0) \) additional observations.

Set \( j \leftarrow 1 \) and \( \ell \leftarrow 1 \).

[2] **Until** von Neumann’s test fails to reject independence using the current weight function \( f_{\cos,j}(\cdot) \):

- compute the signed areas \( \{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b\} \) for the weight function \( f_{\cos,j}(\cdot) \);
- assess the independence of \( \{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b\} \) using von Neumann’s two-sided randomness test at significance level \( \beta \psi(\ell) \);
- set \( \ell \leftarrow \ell + 1 \); generate \( b(\lceil m\sqrt{2}\rceil - m) \) additional observations; set \( m \leftarrow \lceil m\sqrt{2}\rceil \);

[3] Set \( m_j \leftarrow m \); reset \( m \leftarrow 2048 \) and \( \ell \leftarrow 1 \); update \( j \leftarrow j + 1 \). If \( j \leq 2 \), then repeat step [2]; otherwise, take \( m \leftarrow \max\{m_1, m_2\} \) and \( n \leftarrow bm \).
[4] Remove the first batch \( \{X_1, \ldots, X_m\} \); reindex the remaining sample and the signed areas as \( \{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b - 1; j = 1, 2\} \); collect a new batch of size \( m \); compute \( \{Zb(f_{\cos,j}; m) : j = 1, 2\} \) from the new batch; set \( j \leftarrow 1 \) and \( \ell \leftarrow 1 \).

[5] Until the Shapiro–Wilk test fails to reject normality using the current weight function \( f_{\cos,j}() \):

- assess the multivariate normality of \( \{Z_i(f_{\cos,j}; m) : i = 1, \ldots, b\} \) using the Shapiro–Wilk one-sided test at significance level \( \delta \psi(\ell) \);
- set \( \ell \leftarrow \ell + 1 \); generate \( b \left( \left\lceil m \sqrt{2} \right\rceil - m \) additional observations; set \( m \leftarrow \left\lceil m \sqrt{2} \right\rceil \);

[6] Set \( m'_j \leftarrow m \); reset \( m \leftarrow \max\{m_1, m_2\} \) and \( \ell \leftarrow 1 \); update \( j \leftarrow j + 1 \). If \( j \leq 2 \), then repeat step [5]; otherwise, take \( m \leftarrow \max\{m'_1, m'_2\} \), and set \( n \leftarrow bm \).

[7] Until the relative half-length \( h(b, m, \alpha)/|\bar{X}_n| \leq r^* \):

[a] Compute the CI midpoint \( \bar{X}_n \) and half-length \( h(b, m, \alpha) \):

- update the truncated grand mean to reflect the new total sample size \( \bar{X}_n \leftarrow n^{-1} \sum_{k=1}^{n} X_k \);
- compute the updated batched area estimates with the current values of \( b \) and \( m \):

\[
\mathcal{A}(f_{\cos,j}; b, m) = \frac{1}{b} \sum_{i=1}^{b} Z_i^2(f_{\cos,j}; m) \quad \text{for } j = 1, 2;
\]

- compute the updated OBM variance estimate \( O(b, m; m/4) \) using Equation (11) with shift \( d = m/4 \) and the current values of \( n \) and \( m \);
- set

\[
\hat{V}(b, m) \leftarrow \max \{ O(b, m; m/4), \mathcal{A}(f_{\cos,j}; b, m) : j = 1, 2 \}
\]

and compute the half-length

\[
h(b, m, \alpha) \equiv t_{\nu, 1-\alpha/2} \sqrt{\hat{V}(b, m) / n},
\]

where

\[
\nu \approx \begin{cases} 
\lceil b \rceil & \text{if } \hat{V}(b, m) = \mathcal{A}(f_{\cos,j}; b, m) \\
\lceil 1.45b \rceil & \text{if } \hat{V}(b, m) = O(b, m; m/4);
\end{cases}
\]

[b] If \( h(b, m, \alpha)/|\bar{X}_n| > r^* \), then

- estimate the number of batches of the current size required to satisfy the precision requirement,

\[
b' = \left\lceil b \left( \frac{h(b, m, \alpha)}{r^* \bar{X}_n} \right)^2 \right\rceil.
\]
where \([\cdot]\) denotes the “ceiling” function;

- update the batch count \(b\), the batch size \(m\), and the total sample size \(n\) as follows:

  \[
  b \leftarrow \min \{ b', b^* \},
  \]

  \[
  m \leftarrow \begin{cases} 
  m & \text{if } b = b', \\
  \lceil m \times \text{mid} \{ 1.05, (b'/b), 2 \} \rceil & \text{if } b < b'.
  \end{cases}
  \]

  \[
  n \leftarrow bm;
  \]

- generate the required additional observations.

**End If**

[8] Deliver the 100(1 - \(\alpha\))% CI \(\bar{X}_n \pm h(b, m, \alpha)\).

**Remark 3.** The factor of \(\sqrt{2}\) applied to the increase of the batch size in steps [2] and [5] of SPSTS allows computational savings because accumulators collected in iteration \(\ell\) are used to compute the \(\{Z_i(f_{\cos,j};m)\}\) in iteration \(\ell + 2\). Further, for each \(i\) and \(m\), the quantities \(Z_i(f_{\cos,1};m)\) and \(Z_i(f_{\cos,2};m)\) are computed in parallel. On the other hand, the OBM estimator in Equation (11) can be updated on the fly in constant time.

As we mentioned in Section 2, the algorithms in the LABATCH.2 suite of Fishman and Yarberry (1997) are the only known algorithms with overall linear time complexity and sublinear \(O(\log_2 n)\) space complexity. In our setting, for fixed \(n\), the OBM and area estimators \(A(f_{\cos,j}; b, m)\) can be computed in \(O(n)\) time and \(O(m)\) space (Meketon and Schmeiser, 1984; Foley and Goldsman, 1999). While the latter is an upper bound per iteration, the overall time complexity for Algorithm SPSTS remains an open problem that we plan to address elsewhere.

**Remark 4.** In addition to the final 100(1 - \(\alpha\))% CI for \(\mu\) delivered in Step [8] of SPSTS, we can also compute an approximate 100(1 - \(\alpha\))% CI for \(\sigma^2\) from the results of the last iteration of SPSTS as follows:

\[
\frac{\nu \mathcal{O}(b, m; m/4)}{\chi_{1-\alpha/2,\nu}^2} \leq \sigma^2 \leq \frac{\nu \mathcal{O}(b, m; m/4)}{\chi_{\alpha/2,\nu}^2},
\]

where \(\nu = \lceil 1.45b \rceil\) and \(\chi_{\beta,\nu}^2\) denotes the \(\beta\)-quantile of the \(\chi^2\) distribution. As detailed in Section 4.2 of Alexopoulos et al. (2007), Equation (13) follows from Equation (12); detailed asymptotic properties of OBM variance estimators are described in Theorem 2 and Equation (42) of Aktaran-Kalaycı et al. (2009). An alternative CI based on the area estimate \(A(f_{\cos,j}; b, m)\) for \(j = 1, 2\) can be obtained based on Equation (8). In follow-up work, we will extend SPSTS to deliver simultaneous approximate 100(1 - \(\alpha\))% CIs for \(\mu\) and \(\sigma^2\) that satisfy user-specified precision requirements for each parameter separately.
4. Simulation Analysis Methods to Be Compared with SPSTS

4.1. Overview of the ASAP3 Procedure

ASAP3 (Steiger et al. 2005) is arguably the first fully sequential batch means algorithm designed to obtain point and CI estimators for the steady-state mean with user-specified requirements concerning coverage probability and absolute or relative precision. The procedure starts with \( b = 256 \) batches, each of size \( m = 16 \). The first four batches are ignored to reduce the potential effects due to initialization bias, and the remaining 252 batch means are organized into adjacent nonoverlapping groups of four consecutive batch means. The set of 32 four-dimensional vectors consisting of every other group is tested for stationary multivariate normality using the Shapiro–Wilk test (Malkovich and Afifi, 1973). If this test fails, then the batch size is increased by a factor of \( \sqrt{2} \) and the above sequence of steps is repeated with progressively decreasing significance level for the Shapiro–Wilk test (this decrease aims at controlling excessive variability in the final sample size in applications with no precision requirement).

Once the hypothesis of stationary multivariate normality is accepted, ASAP3 fits a first-order autoregressive (AR(1)) model to the sequence of 252 batch means and applies a normalizing arc sine transformation to the autoregressive parameter estimator \( \hat{\phi} \) in order to test the hypothesis that the correlation \( \rho \) between the batch means is at most 0.8. If the one-sided hypothesis is rejected, then the batch size is increased by an appropriate multiplier as detailed in Appendix B of Steiger et al. (2005).

Next, ASAP3 constructs a CI for the mean that has been adjusted to account for correlations between the batch means; this adjustment uses an inverse Cornish–Fisher expansion (Stuart and Ord, 1994) for the classical Student’s \( t \)-ratio based on the batch means. If additional observations of the output process must be obtained before a CI that meets the user-specified precision can be delivered, then ASAP3 estimates a new, larger sample size using an approach similar to step [7] of algorithm SPSTS in Section 3.

In an extensive experimental performance evaluation, Steiger et al. (2005) found that ASAP3 compares favorably with other batch-means procedures (specifically, LABATCH.2 as well as ASAP3’s predecessors ASAP and ASAP2) with respect to the following: (a) conformance with the user-specified CI precision and coverage-probability requirements; and (b) the mean and variance of the delivered CI half-length.

\( b = 256 \) and batch size \( m = 16 \). The first four batches are skipped to reduce the initial bias and for the remaining 252 batch means, every four consecutive ones are grouped into a four-dimensional vector. It selects every other group of these vectors to pass the Shapiro–Wilk test for four-dimensional normality. If this test is failed, additional data are acquired to ensure that the batch size \( m \) is increased by the factor \( \sqrt{2} \) and the batch count is back to the initial 256. ASAP3 iteratively performs the sequence of steps until it finally passes the Shapiro–Wilk test. During the procedure, the significant level \( \delta \) is controlled to decrease systematically on successive iterations so as to avoid excessive variability in the final sample size in an application with no precision requirement.

4.2. Overview of the Skart Procedure

Tafazzoli and Wilson (2011) formulated Skart as an automated sequential batch means procedure for constructing a skewness- and autoregression-adjusted CI for the steady-state mean of a simulation output pro-
cess. A Visual Basic implementation of Skart is available online via www.ise.ncsu.edu/jwilson. Similar to ASAP3, Skart also consists of three main steps.

Skart starts with a sample of size 1280 and computes an estimate $\hat{\gamma}$ of the marginal skewness based on the last 1024 observations. If $|\hat{\gamma}| > 4$, then it proceeds with $b = 1280$ batches of size $m = 16$; otherwise, it sets the initial batch size equal to unity. In the first loop of iterations, von Neumann’s randomness test is sequentially applied to spaced batch means with increasing sizes for each batch and its preceding spacer until the spaced batch means finally pass the test; then the initial spacer is removed to eliminate any warm-up effects.

In the second step, from the truncated time series of the original (unbatched) observations, nonspaced batch means are computed using the batch size from the first step; then from estimates of the variance, skewness, and lag-one correlation of the nonspaced batch means, the classical batch means CI for $\mu$ is adjusted as follows: (a) the autocorrelation adjustment is based on an AR(1) model for the nonspaced batch means; and (b) the skewness adjustment is based on a Cornish–Fisher expansion for the classical batch means Student’s $t$-statistic.

In the third step, if the precision requirement is satisfied by the CI from the second step, then the CI is delivered, and Skart terminates; otherwise, the required simulation run length is estimated, the batch count and batch size are suitably increased within certain limits on their growth in one iteration of Skart, and the first two steps are repeated. As in SPSTS, the method for estimating the new batch count and batch size required to satisfy the precision requirement is based on the analogous method developed for ASAP3.

Tafazzoli et al. (2011) conducted an extensive experimental performance evaluation of Skart on a wide range of test processes with characteristics that are typical of large-scale practical applications as well as test processes that were deliberately designed to provide an extreme “stress test” of Skart. The experimental results show that Skart compares favorably with other steady-state simulation analysis procedures—namely, its predecessors WASSP (Lada and Wilson, 2006; Lada et al., 2007), ASAP3, and SBatch (Lada et al., 2008), as well as sequential versions of LABATCH.2, the procedure of Law and Carson (1979), and the spectral procedure of Heidelberger and Welch (1981).

Tafazzoli et al. (2011a) also developed N-Skart, a nonsequential version of Skart that works with a fixed-size data set. In an experimental performance evaluation involving the same test processes as for Skart, the authors find that N-Skart outperforms LABATCH.2.

5. Performance Evaluation of SPSTS

This section contains an empirical study designed to test the performance of SPSTS. Specifically, we studied five output processes presenting a variety of statistical challenges, including a pronounced initial transient and varying degrees of nonnormality and correlation between responses: an AR(1) process with very high autocorrelation; the queue-waiting-time processes in $M/M/1$, $M/H_2/1$, and $M/M/1/LIFO$ systems; and an AR(1)-to-Pareto (ARTOP) process. The last four processes have highly skewed marginal distributions—this should force failures of the preliminary Shapiro–Wilk test in step [1]. On the other hand, the normality of the AR(1) process should allow SPSTS to proceed with the smaller batch size of 256. The results for ASAP3 and Skart are taken from Table 1 of Tafazzoli et al. (2011). The results for SPSTS and Skart are
based on 1000 independent replications of each procedure (with resulting coverage standard errors of less than 0.001), whereas the results for ASAP3 are based on 400 replications.

5.1. AR(1) Process

Consider the first-order autoregressive process defined via the relation

$$X_k = \mu + \rho(X_{k-1} - \mu) + \varepsilon_k \quad \text{for } k = 1, 2, \ldots,$$

(14)

where $\mu = 100$, $\rho = 0.995$, the errors $\{\varepsilon_k\}$ are i.i.d. standard normal random variables, and the initial state is $X_0 = 0$. The steady-state distribution of this process is normal with mean $\mu$ and standard deviation $\sigma_X = 1/\sqrt{1-\rho^2} = 10.0125$; hence the initial state is located about 10 standard deviations below the steady-state mean. It turns out that $\sigma^2 = \sigma_X^2(1+\rho)/(1-\rho) = 39,999.8$.

The duration of the initial transient and the magnitude of the autocorrelation of the process (14) were designed to assess the ability of SPSTS to “eliminate” initialization bias as well as to assess the effectiveness of steps [2]–[3] of SPSTS. The experimental results in Table 1 contain estimates under more-stringent precision requirements than the other processes under study. The preliminary normality test in step [1] of SPSTS frequently failed to reject the null hypothesis, thereby setting the initial batch size to 256 for steps [2]–[3] of SPSTS. However, to pass the randomness test in steps [2]–[3] of SPSTS, the high degree of correlation in this process typically forced the batch size to values of 2048 or beyond. In all cases, the CIs generated by SPSTS exhibited significant overcoverage. In the case of no precision requirement, SPSTS was outperformed by both Skart and ASAP3 with regard to sampling efficiency. As the relative precision became significantly tighter, SPSTS caught up with its competitors with regard to all performance measures.

We can also examine Table 1 for purposes of conducting a cursory experimental evaluation of the asymptotic behavior of SPSTS as the upper bound $r^*$ on the relative CI precision approaches zero. In particular, the last three cases in the table show that as $r^*$ decreases from 0.9375% to 0.5% to 0.25%, the average sample size grows in an anticipated fashion while the CI coverage probability decreases to the point where it slightly exceeds the nominal value.

5.2. M/M/1 Queue-Waiting-Time Process

Consider the waiting-time process $\{X_k\}$ for the $M/M/1$ queueing system described in Example 1 with arrival rate $\lambda = 0.9$ and traffic intensity $\rho = 0.9$, so that $\mu = 9$ and $\sigma^2 = 35,901$. Assume that the system starts empty ($X_1 = 0$).

Table 2 reports results on the following performance characteristics for SPSTS, Skart, and ASAP3 for CIs with no precision requirement and with relative precisions of 7.5% and 3.75%: empirical coverage, average sample size until procedure termination, average CI half-length, and sample standard deviation of CI half-length. For the specific $M/M/1$ process under study, the CI coverage probabilities for SPSTS usually met or (slightly) exceeded the user-specified nominal levels. As expected, in the case of no precision requirement this coverage was achieved at the cost of substantially larger sample sizes than were required by Skart and ASAP3. However, SPSTS is on equal footing with Skart and ASAP3 with respect to the average
Table 1. Experimental results for the SPSTS algorithm for an AR(1) process

<table>
<thead>
<tr>
<th>Precision Rqt.</th>
<th>Performance Measure</th>
<th>Nominal 90% CIs</th>
<th>Nominal 95% CIs</th>
<th>Nominal 95% CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SPSTS Skart ASAP3</td>
<td>SPSTS Skart ASAP3</td>
<td>SPSTS Skart ASAP3</td>
</tr>
<tr>
<td>none</td>
<td>CI coverage (%)</td>
<td>95.3 92.9 95.5</td>
<td>98.1 96.6 98.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>85.797 21.537 41.076</td>
<td>85.797 21.537 41.076</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>1.260 2.544 2.33</td>
<td>1.513 3.092 2.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>St.Dev. CI half-length</td>
<td>0.414 0.394 0.394</td>
<td>0.496 0.476 0.52</td>
<td></td>
</tr>
<tr>
<td>±3.75%</td>
<td>CI coverage (%)</td>
<td>95.3 93.5 87.0</td>
<td>98.1 97.1 98.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>1.258 2.672 3.210</td>
<td>1.507 3.172 2.820</td>
<td></td>
</tr>
<tr>
<td></td>
<td>St.Dev. CI half-length</td>
<td>0.414 0.394 0.394</td>
<td>0.496 0.476 0.52</td>
<td></td>
</tr>
<tr>
<td>±1.875%</td>
<td>CI coverage (%)</td>
<td>95.0 94.2 93.5</td>
<td>97.8 96.2 99.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>86.729 46.747 57.449</td>
<td>88.556 68.636 101.526</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>1.204 1.746 1.650</td>
<td>1.400 1.729 1.770</td>
<td></td>
</tr>
<tr>
<td></td>
<td>St.Dev. CI half-length</td>
<td>0.261 0.124 0.206</td>
<td>0.232 0.144 0.110</td>
<td></td>
</tr>
<tr>
<td>±0.9375%</td>
<td>CI coverage (%)</td>
<td>93.5 93.8 94.0</td>
<td>96.4 97.1 97.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>149,541 170,792 229,730</td>
<td>243,900 231,873 254,920</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.869 0.902 0.830</td>
<td>0.827 0.903 0.896</td>
<td></td>
</tr>
<tr>
<td></td>
<td>St.Dev. CI half-length</td>
<td>0.083 0.037 0.102</td>
<td>0.094 0.028 0.046</td>
<td></td>
</tr>
<tr>
<td>±0.5%</td>
<td>CI coverage (%)</td>
<td>92.2 94.7 95.2</td>
<td>96.5 97.5 98.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>527,222</td>
<td>753,941</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.472</td>
<td>0.473</td>
<td></td>
</tr>
<tr>
<td></td>
<td>St.Dev. CI half-length</td>
<td>0.025</td>
<td>0.021</td>
<td></td>
</tr>
<tr>
<td>±0.25%</td>
<td>CI coverage (%)</td>
<td>91.1 91.2 91.5</td>
<td>97.0 97.5 98.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>2,144,624</td>
<td>3,079,783</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.236</td>
<td>0.236</td>
<td></td>
</tr>
<tr>
<td></td>
<td>St.Dev. CI half-length</td>
<td>0.011</td>
<td>0.011</td>
<td></td>
</tr>
</tbody>
</table>

CI half-length, i.e., the ratio of the average half-lengths under Skart and SPSTS is close to the square root of the ratio of the average sample sizes required by SPSTS and Skart, respectively. Under the rather stringent precision requirements of 7.5% and 3.75%, the sample sizes required by SPSTS were about the same as those required for Skart, but somewhat larger than the corresponding sample sizes for ASAP3. Notice that the CIs delivered by SPSTS were more variable than their counterparts from ASAP3 and Skart.

Table 2. Experimental results for the SPSTS algorithm for an $M/M/1$ system with traffic intensity 0.9

<table>
<thead>
<tr>
<th>Precision Rqt.</th>
<th>Performance Measure</th>
<th>Nominal 90% CIs</th>
<th>Nominal 95% CIs</th>
<th>Nominal 95% CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SPSTS Skart ASAP3</td>
<td>SPSTS Skart ASAP3</td>
<td>SPSTS Skart ASAP3</td>
</tr>
<tr>
<td>none</td>
<td>CI coverage (%)</td>
<td>91.5 87.6 87.5</td>
<td>95.2 93.9 91.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>207,657 42,369 31,181</td>
<td>207,657 42,369 31,181</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.753 1.767 2.072</td>
<td>0.905 2.298 2.521</td>
<td></td>
</tr>
<tr>
<td></td>
<td>St.Dev. CI half-length</td>
<td>0.191 0.508 0.590</td>
<td>0.230 1.137 0.731</td>
<td></td>
</tr>
<tr>
<td>±7.5%</td>
<td>CI coverage (%)</td>
<td>91.2 91.1 89.5</td>
<td>94.5 95.9 94.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>287,477 302,305 287,568</td>
<td>414,087 431,677 382,958</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.600 0.635 0.627</td>
<td>0.603 0.637 0.632</td>
<td></td>
</tr>
<tr>
<td></td>
<td>St.Dev. CI half-length</td>
<td>0.072 0.037 0.048</td>
<td>0.064 0.035 0.045</td>
<td></td>
</tr>
<tr>
<td>±3.75%</td>
<td>CI coverage (%)</td>
<td>89.1 92.0 89.5</td>
<td>94.6 96.0 93.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>1,134,947 1,105,417 969,011</td>
<td>1,555,347 1,586,267 1,341,522</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.305 0.321 0.320</td>
<td>0.311 0.321 0.321</td>
<td></td>
</tr>
<tr>
<td></td>
<td>St.Dev. CI half-length</td>
<td>0.028 0.014 0.020</td>
<td>0.021 0.014 0.020</td>
<td></td>
</tr>
</tbody>
</table>
5.3. $M/H_2/1$ Queue-Waiting-Time Process

Consider the waiting-time process in an $M/H_2/1$ queueing system with an empty-and-idle initial condition, an arrival rate of 1.0, and a hyperexponential service-time distribution that is a mixture of two exponential distributions with mixing probabilities $p = (5 + \sqrt{15})/10 \approx 0.887$ and $1 - p$, and with respective means $1/(2p\tau)$ and $1/[2(1 - p)\tau]$, where $\tau = 1.25$. Hence the service times have a mean of 0.8 and a coefficient of variation (standard deviation over mean) of 2.0 (see Appendix 2 of Lada et al., 2006). In steady-state operation this system has a server utilization of 80% and a mean queue waiting time of 8.0, so that $\mu = 8.0$ and $\sigma^2 = 24,204.8$.

Table 3 reports results of the three methods for three levels of the precision requirement. For the specific $M/H_2/1$ process under study, the CI coverage probabilities for SPSTS always slightly exceeded the user-specified nominal levels. As anticipated, in the case of no precision requirement, the average sample sizes required by SPSTS were substantially larger than the two competitors. Under the relative precision requirements of 7.5% and 3.75%, the sample sizes and the CI coverage probabilities of SPSTS were about the same as those of Skart, but somewhat higher than the corresponding results of ASAP3. Again, the variability of the SPSTS CI half-length is larger than the other two competitors when the relative precision is relatively high ($r^*$ is small).

<table>
<thead>
<tr>
<th>Precision Rqt.</th>
<th>Performance Measure</th>
<th>Nominal 90% CIs</th>
<th>Nominal 95% CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPSTS</td>
<td>Skart</td>
<td>ASAP3</td>
</tr>
<tr>
<td>none</td>
<td>CI coverage (%)</td>
<td>93.6</td>
<td>90.0</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>182,377</td>
<td>30,379</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.676</td>
<td>1.808</td>
</tr>
<tr>
<td></td>
<td>StDev. CI half-length</td>
<td>0.161</td>
<td>0.680</td>
</tr>
<tr>
<td>±7.5%</td>
<td>CI coverage (%)</td>
<td>92.7</td>
<td>91.3</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>259,137</td>
<td>255,363</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.534</td>
<td>0.566</td>
</tr>
<tr>
<td></td>
<td>StDev. CI half-length</td>
<td>0.065</td>
<td>0.033</td>
</tr>
<tr>
<td>±3.75%</td>
<td>CI coverage (%)</td>
<td>91.3</td>
<td>91.8</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>967,847</td>
<td>929,527</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.273</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>StDev. CI half-length</td>
<td>0.022</td>
<td>0.014</td>
</tr>
</tbody>
</table>

5.4. $M/M/1/LIFO$ Queue-Waiting-Time Process

We now assess the performance of SPSTS based on the queue-waiting-time process in an $M/M/1/LIFO$ queueing system, with mean interarrival time of 1.0, mean service time of 0.8, and LIFO service discipline. Again, $X_k$ is the time spent in queue by entity $k$ prior to service, and we assume that $X_1 = 0$. In steady-state operation, this system has a server utilization of $\tau = 0.8$ and a mean queue waiting time $\mu = 3.2$.

The reason for studying this process is that in steady state, the observations $X_k$ are highly nonnormal with an autocorrelation function that does not exhibit a geometric decay rate; and relatively large batch sizes are required to achieve approximate normality of the batch means (see the discussion in Tafazzoli et al. 2011,
Section 4.6).

Table 4 summarizes the performance of SPSTS, Skart, and ASAP3. (No results were available for ASAP3 under the 3.75% relative precision requirement.) For this LIFO system, the CI coverage probabilities for SPSTS always exceeded the user-specified nominal levels. As anticipated, in the case of no precision requirement, the average sample sizes required by SPSTS were substantially larger than for the other two methods, but the respective CI half-lengths were significantly smaller. Under the precision requirement of 7.5%, the sample sizes required by SPSTS were larger than those of Skart and ASAP3. Under the 3.75% relative precision requirement, the sample sizes of SPSTS were about 20% larger than Skart’s with nearly the same CI coverage probabilities and average CI half-lengths.

Table 4. Experimental results for the SPSTS algorithm for an $M/M/1/LIFO$ system with traffic intensity 0.8

<table>
<thead>
<tr>
<th>Precision Rqt.</th>
<th>Performance Measure</th>
<th>Nominal 90% CIs</th>
<th>Nominal 95% CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPSTS</td>
<td>Skart</td>
<td>ASAP3</td>
</tr>
<tr>
<td>None</td>
<td>CI coverage (%)</td>
<td>94.2</td>
<td>85.6</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>145,613</td>
<td>21,176</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.180</td>
<td>0.514</td>
</tr>
<tr>
<td></td>
<td>StdDev. CI half-length</td>
<td>0.039</td>
<td>0.296</td>
</tr>
<tr>
<td>±7.50%</td>
<td>CI coverage (%)</td>
<td>93.9</td>
<td>91.6</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>146,917</td>
<td>81,441</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.177</td>
<td>0.224</td>
</tr>
<tr>
<td></td>
<td>StdDev. CI half-length</td>
<td>0.035</td>
<td>0.017</td>
</tr>
<tr>
<td>±3.75%</td>
<td>CI coverage (%)</td>
<td>93.3</td>
<td>91.6</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>377,243</td>
<td>305,903</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.104</td>
<td>0.114</td>
</tr>
<tr>
<td></td>
<td>StdDev. CI half-length</td>
<td>0.014</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

5.5. AR(1)-to-Pareto (ARTOP) Process

The final test process was an AR(1)-to-Pareto process with location parameter $\lambda > 0$ and shape parameter $\zeta > 0$. This process is generated as follows: one starts with the stationary AR(1) “base process” $\{Z_\ell : \ell = 1, 2, \ldots\}$,

$$Z_\ell = \rho Z_{\ell-1} + \varepsilon_\ell,$$

where $Z_0 \sim N(0, 1)$ and $\{\varepsilon_\ell : \ell = 1, 2, \ldots\} \sim N(0, \sigma^2_\varepsilon)$ is a white-noise process with variance $\sigma^2_\varepsilon = 1 - \rho^2$. Then we feed the base process into the standard normal c.d.f. $\Phi(\cdot)$ to obtain a sequence of correlated uniform $U(0, 1)$ random variables $\{U_k = \Phi(Z_k) : k = 1, 2, \ldots\}$. Then the process $\{U_k : k = 1, 2, \ldots\}$ is treated as input to the inverse of the Pareto c.d.f.

$$F_X(x) \equiv \Pr\{X \leq x\} = \begin{cases} 1 - (\lambda/x)^\zeta & \text{for } x \geq \lambda \\ 0 & \text{for } x < \lambda \end{cases}$$
to generate the ARTOP process \( \{X_k : k = 1, 2, \ldots \} \) according to the relation

\[
X_k = F_X^{-1}(U_k) = F_X^{-1}(\Phi(Z_k)) = \frac{\lambda}{1 - \Phi(Z_k)}^{1/\xi} \quad k = 1, 2, \ldots.
\]

The mean and variance of the ARTOP process are respectively given by

\[
\mu_X = \mathbb{E}[X_k] = \frac{\xi \lambda (\xi - 1)^{-1}}{\zeta} \quad \text{for } \zeta > 1 \quad \text{and} \quad \sigma_X^2 = \text{Var}[X_k] = \frac{\lambda^2 (\xi - 1)^{-2}(\xi - 2)^{-1}}{\zeta} \quad \text{for } \zeta > 2.
\]

The parameters of the Pareto distribution were set according to \( \xi = 2.1 \) and \( \lambda = 1 \), and the lag-one correlation in the base AR(1) process was set to \( \rho = 0.995 \). This yielded an ARTOP process whose marginal distribution has mean, variance, skewness, and kurtosis, respectively, given by \( \mu_X = 1.9091, \sigma_X^2 = 17.3554, \text{E}\{[(X_k - \mu)/\sigma_X]^3\} = \infty, \) and \( \text{E}\{[(X_k - \mu)/\sigma_X]^4\} = \infty \). The variance parameter \( \sigma^2 = 1612.78 \).

The most difficult aspect of this system is that the marginals are highly nonnormal, and their distribution has a very heavy tail. To put SPSTS on equal footing with Skart, we initialized the AR(1) process in state \( Z_0 = 3.4 \). This yields an initial state \( X_0 = F_X^{-1}[\Phi(Z_0)] = 43.5689 \), which is 10 standard deviations above the mean \( \mu \). On the other hand, ASAP3 was initialized in steady state.

In this setting, the estimated CI coverage probabilities were remarkably close to those obtained by Skart and significantly better than those delivered by ASAP3. As expected, in the case of no precision requirement, SPSTS required significantly larger sample sizes. Notice that under the 7.5% and 3.75% relative precision requirements, SPSTS exhibited slightly better sampling efficiency than Skart with smaller average sample sizes and CIs whose half-lengths had similar average size and variability.

### Table 5. Experimental results for the SPSTS algorithm based on an ARTOP process

<table>
<thead>
<tr>
<th>Precision Rqt.</th>
<th>Performance Measure</th>
<th>Nominal 90% CIs</th>
<th>Nominal 95% CIs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPSTS</td>
<td>Skart</td>
<td>ASAP3</td>
</tr>
<tr>
<td>none</td>
<td>CI coverage (%)</td>
<td>90.3</td>
<td>88.3</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>308.994</td>
<td>37.923</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.112</td>
<td>0.640</td>
</tr>
<tr>
<td></td>
<td>StdDev. CI half-length</td>
<td>0.028</td>
<td>0.788</td>
</tr>
<tr>
<td>±7.50%</td>
<td>CI coverage (%)</td>
<td>89.3</td>
<td>88.3</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>325.164</td>
<td>333.666</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.110</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>StdDev. CI half-length</td>
<td>0.021</td>
<td>0.017</td>
</tr>
<tr>
<td>±3.75%</td>
<td>CI coverage (%)</td>
<td>89.0</td>
<td>91.1</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>1,037,704</td>
<td>1,098,130</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.062</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td>StdDev. CI half-length</td>
<td>0.008 &lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>
as in step [7] of the original SPSTS procedure;

- $\hat{V}(b, m) = A(f_{\cos,1}; b, m)$ in the first alternative version of step [7]; and

- $\hat{V}(b, m) = O(b, m; m/4)$ in the second alternative version of step [7].

These results show that the CIs based on a single variance estimator fail to achieve the nominal CI coverage. We point out that we experienced similar behaviors with the previous four systems. In all our experimentation with SPSTS using a single variance estimator, we observed a substantial percentage of delivered CIs that were “near misses”—that is, the true value of $\mu$ was just outside the final CI delivered by SPSTS.

Lada et al. (2013) observe a similar phenomenon in their development of ARD, an automated sequential version of the replication-deletion method for estimating $\mu$ (Law, 2015, Section 9.5.2). Lada et al. (2013) achieve substantially improved CI coverage with a version of ARD in which the relevant estimator of the variance of the grand average of the truncated sample means computed across all $q_0$ replications each with truncation point $w^*$ and run length $n^*$ is taken to be the maximum of the following: (i) the usual replication-deletion estimator of the variance of the grand average of the truncated sample means; and (ii) a batch-means estimator of $\sigma^2/[q_0(n^* - w^*)]$ computed from all the nonoverlapping batches beyond the truncation point $w^*$ within each of the $q_0$ replications.

In the case of SPSTS, we found that using Equation (15) to compute the variance estimator $\hat{V}(b, m)$ effectively converted most “near misses” into CIs that just barely covered $\mu$; and this improved performance of the CIs delivered by SPSTS was usually accompanied by a relatively small increase in the CI’s half-length. Table 6 illustrates this phenomenon.

6. Discussion and Conclusions

In this paper, we introduced SPSTS, the first automated fully sequential procedure for computing point and confidence interval estimators for the steady-state mean of a simulation output process based on STS area variance estimators. SPSTS compensates for the small-sample bias of the STS area estimators with a larger initial sample size and statistical hypothesis tests based on two estimators corresponding to orthogonal

<table>
<thead>
<tr>
<th>Precision Rqt.</th>
<th>Performance Measure</th>
<th>CI coverage (%)</th>
<th>$\hat{V}(b, m) = A(f_{\cos,1}; b, m)$</th>
<th>$\hat{V}(b, m) = O(b, m; m/4)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>CI coverage (%)</td>
<td>89.3</td>
<td>84.4</td>
<td>86.5</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>308,994</td>
<td>308,994</td>
<td>308,994</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.112</td>
<td>0.100</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>StDev. CI half-length</td>
<td>0.028</td>
<td>0.027</td>
<td>0.026</td>
</tr>
<tr>
<td>3.75%</td>
<td>CI coverage (%)</td>
<td>89.2</td>
<td>84.8</td>
<td>86.6</td>
</tr>
<tr>
<td></td>
<td>Avg. sample size</td>
<td>1,037,704</td>
<td>800,774</td>
<td>818,874</td>
</tr>
<tr>
<td></td>
<td>Avg. CI half-length</td>
<td>0.062</td>
<td>0.061</td>
<td>0.063</td>
</tr>
<tr>
<td></td>
<td>StDev. CI half-length</td>
<td>0.008</td>
<td>0.009</td>
<td>0.008</td>
</tr>
</tbody>
</table>
cosine weight functions. However, SPSTS is significantly simpler than its NBM-based state-of-the-art competitors and does not rely on distributional approximations and adjustments for correlation and skewness of the batch means.

An empirical evaluation involving three queueing systems, an AR(1) process, and an ARTOP process, revealed that SPSTS performed reasonably well compared with ASAP3 and Skart in terms of CI estimated coverage, average required sample size, and average CI half-length. Despite the substantially larger average run length in the case of no precision requirement, SPSTS delivered CIs with very good estimated coverage and meaningful half-lengths (with average relative half-length under 10%). As the upper bound on the relative CI precision dropped to values below 5%, SPSTS often outperformed ASAP3 and Skart with regard to sampling efficiency, in particular for the challenging ARTOP process with infinite marginal skewness and kurtosis. Such performance is quite remarkable given the maturity of the ASAP3 and Skart methods (of course, the development of SPSTS has taken advantage of knowledge accumulated during experimentation with NBM-based procedures over the last 20+ years).

In the design of SPSTS, we have tried to bear in mind the following maxim of Albert Einstein: “If you are out to describe the truth, leave elegance to the tailor.” The objective of SPSTS is to deliver practically useful point and CI estimators of the steady-state mean for a reasonably broad class of stochastic simulations; and we believe there is good evidence that this objective has been achieved to a certain extent.

The following tasks will be undertaken in the near future: (i) development and incorporation of STS area estimators with negligible small-sample bias; (ii) development of sequential procedures based on STS area estimators computed from overlapping batches of observations (Alexopoulos et al., 2007a,?); (iii) development of sequential procedures for estimating the variance parameter $\sigma^2$; and (iv) enhanced experimental evaluation based on queueing networks such as the Central Server Model 3 of Law and Carson (1979).

Acknowledgments

This work was partially supported by the National Science Foundation grants CMMI-1233141/1232998. We thank Dr. Kemal Dingec for his computational assistance.

References


Appendix

Proof of Theorem 1  Parts of this proof run parallel to Theorem 3 and the corresponding examples in Aktaran-Kalaycı et al. (2007). We use the notation $f_j^*(t) = f_{\cos,j}(t)$.

By software such as Mathematica (Wolfram Research, Inc., 2014), we have

$$\sum_{k=1}^{m} \frac{k}{m} f_j^*(\frac{k}{m}) = \sqrt{2}\pi j \quad \text{and} \quad \sum_{k=\ell}^{m} f_j^*(\frac{k}{m}) = \sqrt{2}\pi j \left[ 1 - \sin\left(\frac{\pi j(2\ell-1)}{m}\right) \csc\left(\frac{\pi j}{m}\right) \right],$$

so that

$$h(\ell) = \sum_{k=1}^{m} \frac{k}{m} f_j^*(\frac{k}{m}) - \sum_{k=\ell}^{m} f_j^*(\frac{k}{m}) = \sqrt{2}\pi j \sin\left(\frac{\pi j(2\ell-1)}{m}\right) \csc\left(\frac{\pi j}{m}\right), \quad \ell = 1, 2, \ldots, m.$$

Then as in Foley and Goldsman (1999), the expected value of the estimator can be written as

$$\mathbb{E}[A(f_j^*,b,m)] = \frac{1}{m^3} \left[ R_0 \sum_{\ell=1}^{m} h^2(\ell) + 2 \sum_{i=1}^{m-1} \sum_{\ell=1}^{m-i} h(\ell)h(\ell+i) \right]$$

$$= \frac{\pi^2 j^2 \csc^2\left(\frac{\pi j}{m}\right)}{m^2} \left( R_0 + \frac{2}{m} \sum_{i=1}^{m-1} \left[ (m-i) \cos\left(\frac{2\pi ji}{m}\right) + \sin\left(\frac{2\pi ji}{m}\right) \csc\left(\frac{2\pi ji}{m}\right) \right] R_i \right), \quad (16)$$

with the last step following via Mathematica.
At this point, recall the following Taylor series expansions:

$$
\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + \sum_{n=3}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!},
$$  \hspace{1cm} (17)

$$
\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \sum_{n=3}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!},
$$  \hspace{1cm} (18)

$$
\csc(x) = \frac{1}{x} + \frac{x}{6} + \sum_{n=3}^{\infty} \frac{(-1)^{n+1} (2^{2n} - 2) B_{2n} x^{2n-1}}{(2n)!}, \quad 0 < |x| < \pi,
$$  \hspace{1cm} (19)

where the $B_{2n}$’s are Bernoulli numbers. By a well-known relationship between $B_{2n}$ and the $\zeta$-function (see en.wikipedia.org/wiki/Riemann_zeta_function), the last term in (19) can be bounded by

$$
\sum_{n=3}^{\infty} \frac{2 (2^{2n} - 2) \zeta(2n) x^{2n-1}}{(2\pi)^{2n}} \leq C_0 \frac{\zeta(x)}{x^2} \leq C_1 |x|^5, \hspace{1cm} (20)
$$

where $C_0$ and $C_1$ are large-enough positive constants. Thus, for $j$ fixed and $m \ll m$, (19) and (20) imply

$$
\frac{\pi^2 j^2 \csc^2(\pi/j)}{m^2} = \frac{\pi^2 j^2}{m^2} \left( \frac{m}{\pi j} + \frac{\pi j}{6m} + \frac{7\pi^3 j^3}{360m^3} + O(m^{-5}) \right)^2 = 1 + \frac{\pi^2 j^2}{3m^2} + \frac{\pi^4 j^4}{15m^4} + O(m^{-6}).
$$  \hspace{1cm} (21)

By (17)–(20), we have

$$
(m - i) \cos(\frac{\pi j}{m}) + \sin(\frac{\pi j}{m}) \csc(\frac{\pi j}{m})
= (m - i) \left[ 1 - \frac{2\pi^2 j^2 i^2}{m^2} + \frac{2\pi^4 j^4 i^4}{3m^4} + \sum_{n=3}^{\infty} \frac{(-1)^n (\frac{\pi j}{m})^{2n}}{(2n)!} \right]
+ \left[ \frac{2\pi j i}{m} - \frac{4\pi^3 j^3 i^3}{3m^3} + \frac{4\pi^5 j^5 i^5}{15m^5} + \sum_{n=3}^{\infty} \frac{(-1)^n (\frac{\pi j}{m})^{2n+1}}{(2n+1)!} \right] \left[ \frac{m}{2\pi j} + \frac{\pi j}{3m} + \frac{7\pi^3 j^3}{45m^3} + O(m^{-5}) \right]
= m - \frac{2\pi^2 j^2 i^2}{m^2} + \frac{2\pi^4 j^4 i^4}{3m^4} + O(m^{-4})
+ (m - i) \sum_{n=3}^{\infty} \frac{(-1)^n (\frac{\pi j}{m})^{2n}}{(2n)!} + \left[ \frac{m}{2\pi j} + O(m^{-1}) \right] \sum_{n=3}^{\infty} \frac{(-1)^n (\frac{\pi j}{m})^{2n+1}}{(2n+1)!}.
$$  \hspace{1cm} (22)

We will now bound an expression that is similar to the remainder terms in (22). Suppose $q \in \mathbb{N}$ and $c \in \mathbb{R}$. We assume that $|R_i| = O(\delta^i)$, for some $\delta \in (0, 1)$. Then by the usual Stirling’s approximation,
which is actually a lower bound for $n!$ for all $n \geq 1$, we have for large-enough positive constant $C_2$,

$$
\sum_{i=1}^{m-1} \sum_{n=k}^{\infty} i^q \left( \frac{ci}{m} \right)^n \frac{1}{n!} |R_i| \leq C_2 \sum_{i=1}^{m-1} i^q \delta \sum_{n=k}^{\infty} \left( \frac{ci}{m} \right)^n \frac{(n/2)^n}{\sqrt{2\pi n}} \leq \frac{C_2}{\sqrt{2\pi k}} \left( \sum_{n=k}^{\infty} \left( \frac{ci}{n} \right)^k \sum_{n=k}^{\infty} \left( \frac{ci}{n} \right)^n \right) \leq C_3 \sum_{i=1}^{m-1} i^q \delta \sum_{n=k}^{\infty} \left( \frac{ci}{i} \right)^n \leq \frac{C_3}{m^{k-1}} \sum_{i=1}^{m-1} i^q + k \delta \leq O(m^{-(k-1)}).
$$

As special cases, we have

$$
\sum_{i=1}^{m-1} \sum_{n=k}^{\infty} i^q \left( \frac{ci}{m} \right)^2 \frac{1}{(2n)!} |R_i| = O(m^{-(2k-1)}) \quad \text{and} \quad \sum_{i=1}^{m-1} \sum_{n=k}^{\infty} i^q \left( \frac{ci}{m} \right)^{2n+1} \frac{1}{(2n+1)!} |R_i| = O(m^{-2k}).
$$

Plugging (21) and (22) into (16), we obtain

$$
E[A(f_j^*; b, m)] = \left( 1 + \frac{\pi^2}{3m^2} j^2 + \frac{\pi^4}{15m^4} j^4 + O(m^{-6}) \right) \left\{ R_0 + \sum_{i=1}^{m-1} R_i \left( 1 - \frac{2\pi^2}{m^2} i^2 j^2 + \frac{2\pi^4 j^4}{3m^4} i^4 + O(m^{-5}) \right) \right.
$$

$$
+ \left( 1 - \frac{i}{m} \right) \sum_{n=3}^{\infty} \frac{(-1)^n \left( \frac{2\pi j i}{m} \right)^{2n}}{(2n)!} + \left( \frac{1}{2\pi j} + O(m^{-2}) \right) \sum_{n=3}^{\infty} \frac{(-1)^n \left( \frac{2\pi j i}{m} \right)^{2n+1}}{(2n+1)!} \right\}.
$$

Finally, by Corollary 1 of Aktaran-Kalaycı et al. (2007), which states that $2 \sum_{i=1}^{m-1} i^j R_i = \gamma_j + O(m^j \delta^m)$, and application of (23) with $k = 3$, we have

$$
E[A(f_j^*; b, m)] = \left( 1 + \frac{\pi^2}{3m^2} j^2 + \frac{\pi^4}{15m^4} j^4 + O(m^{-6}) \right)
$$

$$
\times \left[ \alpha^2 - \frac{2\pi^2}{m^2} j^2 \gamma^2 + \frac{2\pi^2 j^2 (2\gamma + \gamma^1)}{3m^3} + \frac{2\pi^4 j^4 \gamma^4}{3m^4} + O(m^{-5}) \right]. \quad \square
$$