*Proceedings of the 2007 Winter Simulation Conference*
*S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, eds.*

# SBatch: A SPACED BATCH MEANS PROCEDURE FOR SIMULATION ANALYSIS

Emily K. Lada

SAS Institute Inc.
100 SAS Campus Drive, R5413
Cary, NC 27513-8617, U.S.A.

James R. Wilson

Edward P. Fitts Department of
Industrial and Systems Engineering
North Carolina State University
Raleigh, NC 27695-7906, U.S.A.

## ABSTRACT

We discuss SBatch, a simplified procedure for steady-state simulation analysis that is based on spaced batch means, incorporating many advantages of its predecessors ASAP3 and WASSP while avoiding many of their disadvantages. SBatch is a sequential procedure designed to produce a confidence-interval estimator for the steady-state mean response that satisfies user-specified precision and coverage-probability requirements. First SBatch determines a batch size and an interbatch spacer size such that beyond the initial spacer, the spaced batch means approximately form a stationary first-order autoregressive process whose lag-one correlation does not significantly exceed 0.8. Next SBatch delivers a correlation-adjusted confidence interval based on the sample variance and lag-one correlation of the spaced batch means as well as the grand mean of all the individual observations beyond the initial spacer. In an experimental evaluation, SBatch compared favorably with ASAP3 and WASSP.

## 1 INTRODUCTION

In a nonterminating simulation, we are often interested in long-run (steady-state) average performance measures. Let $\{X_i : i = 1, 2, \ldots\}$ denote a stochastic process representing the sequence of outputs generated by a single run of a nonterminating probabilistic simulation. If the simulation is in steady-state operation, then the random variables $\{X_i\}$ will have the same steady-state cumulative distribution function (c.d.f.) $F_X(x) = \Pr\{X_i \le x\}$ for $i = 1, 2, \ldots,$ and for all real $x$.

Usually in a nonterminating simulation, we are interested in constructing point and confidence-interval (CI) estimators for some parameter of the steady-state c.d.f. $F_X(\cdot)$. In this work, we are primarily interested in estimating the steady-state mean, $\mu_X = \mathrm{E}[X] = \int_{-\infty}^{\infty} x\, dF_X(x)$; and we limit the discussion to output processes for which $\mathrm{E}[X_i^2] < \infty$ so that the process mean $\mu_X$ and process variance $\sigma_X^2 = \mathrm{Var}[X_i] = \mathrm{E}[(X_i - \mu_X)^2]$ are well defined. We let $n$ denote the length of the time series $\{X_i\}$ of outputs generated by a single, long run of the simulation.

In this article we discuss SBatch, a simplified procedure for steady-state simulation analysis that is based on the method of spaced batch means (Fox, Goldsman, and Swain 1991), incorporating many advantages of its predecessors ASAP3 (Steiger et al. 2004, 2005) and WASSP (Lada, Wilson, and Steiger 2003; Lada and Wilson 2006) while avoiding many of their disadvantages. Based on our experimentation with a diversity of test processes, we have concluded that SBatch has the nearly the same sampling efficiency as ASAP3 and the same ability to eliminate initialization bias as WASSP while being much simpler to implement and understand than either ASAP3 or WASSP.

The rest of this article is organized as follows. In §2 we review the simulation analysis method of spaced batch means, and in §3 we provide an overview of SBatch. Section 4 contains the detailed operational steps of SBatch. In §5 we summarize some of the results of our experimental performance evaluation; and in §6 we present our main conclusions. Lada and Wilson (2007) provide full details on the design and operation of SBatch and its experimental evaluation. The slides of the accompanying presentation are available via <www.ise.ncsu.edu/jwilson/files/sbatch07pr.pdf>.

## 2 THE METHOD OF SPACED BATCH MEANS

In the conventional method of spaced batch means (Fox, Goldsman, and Swain 1991), the output sequence $\{X_i : i = 1, 2, \ldots, n\}$ is divided into $k$ batches, each of size $m$ with a spacer of size $s$ preceding each batch, where both $s$ and $m$ are sufficiently large to ensure that the resulting spaced batch means are independent and identically distributed (i.i.d.) normal random variables. For the $j$th spaced batch, the sample mean is

$$\overline{X}_j(m,s) = \frac{1}{m} \sum_{i=m(j-1)+1+sj}^{j(m+s)} X_i \quad \text{for } j = 1, 2, \ldots, k; \quad (1)$$

and the average of all observations beyond the first spacer,

$$\overline{X} = \frac{1}{n-s}\sum_{i=s+1}^{n} X_i, \tag{2}$$

is the point estimator for $\mu_X$ based on the assumption that the first spacer $\{X_i : i = 1, 2, \ldots, s\}$ contains the entire warm-up period (if there is one). From (1) and (2), we can construct a CI estimator for $\mu_X$ using the (classical) approach detailed in Fox, Goldsman, and Swain (1991).

In this article, we develop a variant of the method of spaced batch means to handle the situation that the batch size $m$ and the spacer size $s$ are sufficiently large so that the resulting spaced batch means approximately constitute a stationary Gaussian process with mean $\mu_X$ but are not necessarily independent. In our experience such a situation is much easier to achieve in practice than other conditions that are required to apply other batch-means procedures. We compute the sample variance of the $k$ spaced batch means for batches of size $m$,

$$\widehat{\sigma}^2_{\overline{X}(m,s)} = \frac{1}{k}\sum_{j=1}^{k}\Big[\overline{X}_j(m,s) - \overline{\overline{X}}(m,k,s)\Big]^2, \tag{3}$$

where

$$\overline{\overline{X}}(m,k,s) = \frac{1}{k}\sum_{j=1}^{k}\overline{X}_j(m,s) \tag{4}$$

is the grand mean of the spaced batch means. For user-specified $\beta \in (0,1)$, we then compute an approximate $100(1-\beta)\%$ correlation-adjusted CI for $\mu_X$ as follows,

$$\overline{X} \pm t_{1-\beta/2,k-1}\sqrt{\frac{A\widehat{\sigma}^2_{\overline{X}(m,s)}}{k}}, \tag{5}$$

where: $t_{1-\beta/2,k-1}$ is the $1-\beta/2$ quantile of Student's $t$-distribution with $k-1$ degrees of freedom; and the correlation adjustment $A$ is applied to the sample variance (3) to compensate for any residual correlation between the spaced batch means. The correlation adjustment $A$ is computed as

$$A = \frac{1+\widehat{\varphi}_{\overline{X}(m,s)}}{1-\widehat{\varphi}_{\overline{X}(m,s)}}, \tag{6}$$

where the standard estimator of the lag-one correlation of the spaced batch means is

$$\widehat{\varphi}_{\overline{X}(m,s)} = \widehat{\mathrm{Corr}}\big[\overline{X}_j(m,s), \overline{X}_{j+1}(m,s)\big] \tag{7}$$

$$= \frac{1}{k}\sum_{j=1}^{k-1}\frac{\big[\overline{X}_j(m,s) - \overline{\overline{X}}(m,k,s)\big]\big[\overline{X}_{j+1}(m,s) - \overline{\overline{X}}(m,k,s)\big]}{\widehat{\sigma}_{\overline{X}(m,s)}}.$$

## 3 OVERVIEW OF SBatch

Figure 1 depicts a high-level flowchart of SBatch. The following user-supplied inputs are required:

1. a simulation-generated output process from which the steady-state mean response is to be estimated;
2. the desired confidence-interval coverage probability $1-\beta$, where $0 < \beta < 1$; and
3. an absolute or relative precision requirement specifying the final confidence-interval half-length in terms of a maximum acceptable half-length $h^*$ (for an absolute precision requirement) or a maximum acceptable fraction $r^*$ of the magnitude of the confidence-interval midpoint (for a relative precision requirement).

SBatch returns the following outputs:

1. a nominal $100(1-\beta)\%$ confidence interval for the steady-state mean that satisfies the specified precision requirement, provided no additional data are required; or
2. a new, larger sample size to be supplied to the algorithm.

SBatch begins by dividing the initial simulation-generated output process of length $n = 16384$ observations into $k = 1024$ adjacent batches of size $m = 16$, with a spacer of initial size $s = 0$ observations preceding each batch. The randomness test of von Neumann (1941) is then applied to the initial set of batch means calculated for each batch. The primary purpose of the randomness test is to determine an appropriate data-truncation point beyond which all computed batch means are approximately independent of the simulation model's initial conditions. A second purpose of the randomness test is to construct a set of spaced batch means such that the interbatch spacer preceding each batch is sufficiently large to ensure all computed batch means are approximately independent and identically distributed, so that subsequently the spaced batch means can be tested for normality.

Each time the randomness test is failed, an additional batch is added to each spacer (up to a limit of 14 batches), and the randomness test is reperformed on the new reduced set of spaced batch means. If the randomness test is failed with a spacer consisting of 14 batches so that only 68 spaced batch means are used in the test, then the spacer size $s$ is reset to zero and both the batch size $m$ and the total sample size $n$ are increased by the factor $\sqrt{2}$; the required additional observations are obtained (by restarting the simulation if necessary); the augmented sample is rebatched into $k = 1024$ nonspaced batches of the new batch size $m$; and a new set of $k$ batch means is computed and tested for randomness with a spacer of size $s = 0$.

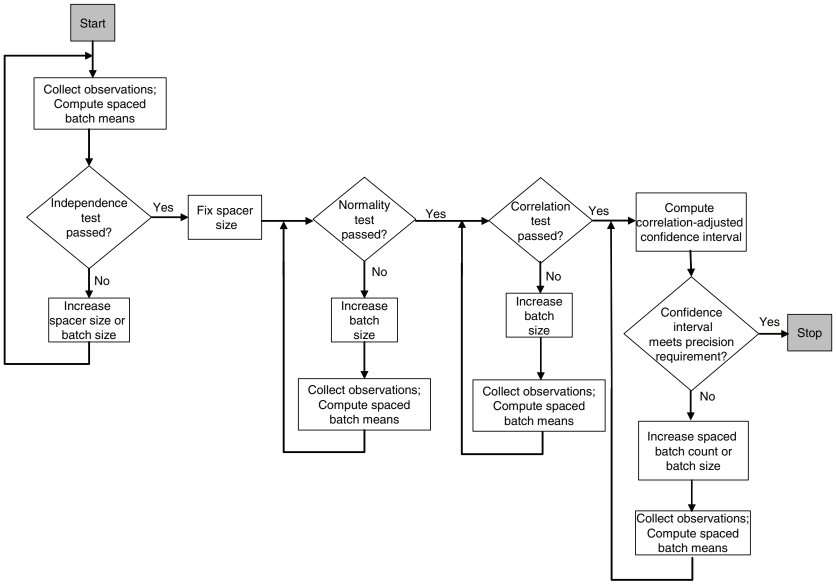
Start
Collect observations; Compute spaced batch means
Independence test passed?
Yes
No
Increase spacer size or batch size
Fix spacer size
Normality test passed?
Yes
No
Increase batch size
Collect observations; Compute spaced batch means
Correlation test passed?
Yes
No
Increase batch size
Collect observations; Compute spaced batch means
Compute correlation-adjusted confidence interval
Confidence interval meets precision requirement?
Yes
Stop
No
Increase spaced batch count or batch size
Collect observations; Compute spaced batch means

Figure 1: High-level flowchart of SBatch.

Once the randomness test is passed, the size of the spacer separating each batch is fixed, the observations $\{X_i : i = 1, \ldots, s\}$ composing the first spacer are discarded to eliminate any warm-up effects; and the resulting set of spaced batch means is tested for normality using the method of Shapiro and Wilk (1965). Each time the normality test is failed, the batch size $m$ is increased by the factor $\sqrt{2}$ up until the sixth consecutive failure. For each subsequent failure of the normality test (starting with iteration $q = 7$), the batch size $m$ is increased by the factor $\left\lfloor 2^{1/(q-4)} \right\rfloor$ for $q = 7, 8, \ldots$. A new set of spaced batch means is then computed using the final spacer size $s$ determined in the randomness test, and the normality test is repeated for the new set of spaced batch means.

Once the normality test is passed, we test the condition that 0.8 is an upper limit for the lag-one correlation of the resulting set of approximately normal, spaced batch means. Each time the correlation test is failed, the batch size $m$ is increased by 10%, the required additional observations are obtained (by restarting the simulation if necessary), a new set of spaced batch means is computed, and the correlation test is repeated for the new set of spaced batch means.

After the correlation test is passed, a correlation-adjusted confidence interval of the form (5) is computed, where the midpoint of the confidence interval is $\overline{X}$, the grand average of all the observations beyond the first spacer; and the confidence-interval half-length is $t_{1-\beta/2,k-1}\sqrt{A\widehat{\sigma}^2_{\overline{X}(m,s)}\big/k}$, which differs from the usual batch-means formulation by incorporating the correlation adjustment $A$ as well as the sample variance of the batch means. The correlation adjustment $A$ compensates for any residual correlation that may exist between the spaced batch means so that the final result is an approximately valid confidence interval for the steady-state mean.

The confidence interval (5) is then tested to determine if it satisfies a user-specified absolute or relative precision requirement. If the precision requirement is satisfied, then SBatch delivers the latest confidence interval and terminates. Otherwise, SBatch estimates the number of spaced batch means required to satisfy the precision requirement. If the estimated spaced batch count exceeds 1024, then the batch count is set to 1024 and a new batch size is estimated so that the total delivered sample size is the smallest possible value not less than the total required sample size. If necessary, additional observations are collected and a new set of spaced batch means is computed. The confidence interval (5) is recomputed and the precision requirement is retested.

## 4 DETAILED OPERATIONAL STEPS OF SBatch

### 4.1 The Test for Randomness

SBatch begins by dividing the initial sample $\{X_i : i = 1, \ldots, n\}$ of length $n = 16384$ observations into $k = 1024$ adjacent (nonspaced) batches of size $m = 16$. Batch means are computed according to (1) using initial spacers of size $s = 0$ for each of the $k$ batches. We then apply the randomness test of von Neumann (1941) to the resulting batch means $\left\{\overline{X}_1(m,s), \ldots, \overline{X}_k(m,s)\right\}$ by computing the ratio of the mean square successive difference of the batch means to the sample variance of the batch means. In SBatch we apply the von Neumann test for randomness iteratively to determine the size of an interbatch spacer that is sufficiently large to yield approximate independence of the corresponding spaced batch means. At the level of significance $\alpha_{\text{ran}} = 0.2$, we test the null hypothesis of independent, identically distributed spaced batch means,

$$\left\{ \overline{X}_j(m,s) : j = 1, \ldots, k \right\} \quad \text{are i.i.d.,} \tag{8}$$

by computing the test statistic,

$$C_k = 1 - \frac{\sum_{j=1}^{k-1} \left[ \overline{X}_j(m,s) - \overline{X}_{j+1}(m,s) \right]^2}{2\sum_{i=1}^{k} \left[ \overline{X}_i(m,s) - \overline{\overline{X}}(m,k,s) \right]^2}, \tag{9}$$

which is a relocated and rescaled version of the ratio of the mean square successive difference of the batch means to the sample variance of the batch means. Since SBatch's test for randomness always involves at least 68 batch means, we use a normal approximation to the null distribution of the test statistic (9); see Fishman and Yarberry (1997, p. 303). Let $z_\omega$ denote the $\omega$ quantile of the standard normal distribution for $0 < \omega < 1$. If

$$|C_k| \le z_{1-\alpha_{\text{ran}}/2} \sqrt{(k-2)\big/(k^2-1)}, \tag{10}$$

then the hypothesis (8) is accepted; otherwise (8) is rejected so that SBatch must increase the spacer size before retesting (8). As documented in Section 4.2 of Lada and Wilson (2006), we found that setting $\alpha_{\text{ran}} = 0.2$ works well in practice and provides an effective balance between errors of type I and II in testing the hypothesis (8).

If the $k = 1024$ batch means $\{\overline{X}_1(m,s), \overline{X}_2(m,s), \ldots, \overline{X}_k(m,s)\}$ with spacer size $s = 0$ pass the randomness test (8)–(10) at the level of significance $\alpha_{\text{ran}}$, then we set the number of batch means to be used in the normality test according to $k' \leftarrow 1024$; and we proceed to perform the normality test as detailed in Section 4.2 below. On the other hand, if the $k = 1024$ batch means with spacer size $s = 0$ fail the test for randomness, then we insert spacers each consisting of one ignored batch between the $k' \leftarrow 512$ remaining batches whose corresponding spaced batch means are to be retested for randomness; and thus the updated spacer size is $s \leftarrow m$ observations. That is, every other batch, beginning with the second batch, is retained for assignment as one of the spaced batch means; and the alternate batches are ignored.

We reapply the randomness test to the $k' = 512$ remaining spaced batch means $\left\{\overline{X}_1(m,s),\ \overline{X}_2(m,s),\ \ldots,\ \overline{X}_{512}(m,s)\right\}$ having batch size $m$ and spacers of size $s = m$. If the randomness retest is passed, then we proceed to perform the normality test in Section 4.3 with the current values of $s$ and $k'$; otherwise we add another ignored batch to each spacer so that the spacer size and number of remaining batches are updated according to

$$s \leftarrow s + m \quad \text{and} \quad k' \leftarrow \left\lfloor n\big/(m+s) \right\rfloor. \tag{11}$$

After executing (11), we have $k' = 341$ remaining spaced batch means $\left\{ \overline{X}_1(m,s), \overline{X}_2(m,s), \ldots, \overline{X}_{341}(m,s) \right\}$ with batch size $m$ and spacers of size $s = 2m$; and again the remaining spaced batch means are retested for randomness. This process is continued until one of the following conditions occurs: (*a*) the randomness test is passed; or (*b*) the randomness test is failed and in the update step (11), the batch count $k'$ drops below the lower limit of 68 batches. If condition (*a*) occurs, then we proceed to the normality test in Section 4.2 with the current values of $s$ and $k'$. On the other hand if condition (*b*) occurs, then the batch size $m$, the batch count $k$, the overall sample size $n$, and the spacer size $s$ are updated according to

$$m \leftarrow \left\lfloor \sqrt{2}m \right\rfloor, \quad k \leftarrow 1024, \quad n \leftarrow km, \quad \text{and} \quad s \leftarrow 0,$$

respectively; the required additional observations are obtained (by restarting the simulation if necessary) to complete the overall sample $\{X_i : i = 1, \ldots, n\}$; and then $k$ adjacent (nonspaced) batch means are computed from the overall sample according to (1).

At this point we reperform the entire randomness-testing procedure, starting with the current set of $k = 1024$ adjacent (nonspaced) batch means of the current batch size $m$ with spacer size $s = 0$. If the randomness test is passed, then we set $k' \leftarrow k$ and proceed to the normality test in Section 4.2; otherwise we repeat the steps outlined in the two immediately preceding paragraphs. Once the randomness test is passed, we finalize the spacer size $s$ so that we have a set of $k'$ approximately i.i.d. spaced batch means, where $68 \le k' \le 1024$.

This approach to handling the simulation start-up problem is very similar to the approach used in WASSP. Through extensive experimentation with the WASSP algorithm as well as with SBatch, we found this approach to be effective in determining an appropriate spacer size $s$ so that:

1. The observations $\{X_1, X_2, \ldots, X_s\}$ constituting the first spacer can be regarded as containing the warm-up period because the spaced batch means beyond the first spacer do not exhibit significant departures from randomness—that is, they do not exhibit a deterministic trend or any type of stochastic dependence on the simulation's initial conditions.
2. The spaced batch means computed beyond the first spacer are approximately i.i.d. and thus can be meaningfully tested for normality using standard goodness-of-fit tests.

There are a few key differences between the methods used in WASSP and SBatch for eliminating initialization bias, however. WASSP requires an initial sample of size $n = 4096$ and allows a maximum of 9 batches per spacer, resulting in a final spaced batch count in the range $25 \le k' \le$

256. By contrast, SBatch requires an initial sample of size $n = 16384$ and allows up to 14 batches per spacer, resulting in a final spaced batch count in the range $68 \le k' \le 1024$. By increasing the initial sample size (and consequently the minimum number of spaced batch means $k'$), we have increased the sensitivity of both the randomness test and the subsequent test for normality that are applied to the resulting set of spaced batch means.

Increasing the sensitivity of the normality test is critical to the effectiveness of SBatch since the delivered CI (5) is derived from classical results based on Student's $t$-distribution that require the basic observations to be a stationary Gaussian process. The WASSP algorithm, on the other hand, is less sensitive to departures from normality of its batch means because (*a*) in general WASSP requires substantially larger final sample sizes, thus ensuring a central-limit effect in its final point estimator of $\mu_X$; and (*b*) WASSP merely exploits approximate normality of the batch means to ensure that the log-periodogram of the batch means is well behaved and therefore yields a better estimator of the batch-means log-spectrum at zero frequency.

### 4.2 Testing Spaced Batch Means for Normality

To test the spaced batch means for normality, we use the Shapiro-Wilk test (Shapiro and Wilk 1965) because it is a well-established, powerful, and widely used test for departures from normality in a data set consisting of i.i.d. observations (Royston 1982, 1993, 1995). Thus in SBatch we apply the Shapiro-Wilk test to the $k'$ spaced batch means with the final spacer size $s$ determined in the preceding test for randomness. To assess the normality of the sample $\left\{\overline{X}_1(m,s), \ldots, \overline{X}_{k'}(m,s)\right\}$, we start by sorting the observations in ascending order to obtain the order statistics $\overline{X}_{(1)}(m,s) \le \overline{X}_{(2)}(m,s) \le \cdots \le \overline{X}_{(k')}(m,s)$. The Shapiro-Wilk test statistic is then computed as follows,

$$W = \frac{\left\{\sum_{\ell=1}^{\lfloor k'/2\rfloor} \zeta_{k'-\ell+1}\left[\overline{X}_{(k'-\ell+1)}(m,s) - \overline{X}_{(\ell)}(m,s)\right]\right\}^2}{\sum_{\ell=1}^{k'}\left[\overline{X}_{\ell}(m,s) - \overline{\overline{X}}(m,k,s)\right]^2}, \qquad (12)$$

where the coefficients $\{\zeta_{k'-\ell+1} : \ell = 1, \ldots, \lfloor k'/2\rfloor\}$ are evaluated using the algorithm of Royston (1982). The test statistic $W$ is then compared to the $\alpha_{\text{nor}}$ quantile $w_{\alpha_{\text{nor}}}$ of the distribution of $W$ under the null hypothesis of i.i.d. normal batch means,

$$\left\{\overline{X}_j(m,s) : j = 1, \ldots, k'\right\} \overset{\text{i.i.d.}}{\sim} N\left[\mu_X, \sigma^2_{\overline{X}(m,s)}\right]. \qquad (13)$$

If $W \le w_{\alpha_{\text{nor}}}$, then at the level of significance $\alpha_{\text{nor}}$ we reject the hypothesis (13).

For the first iteration of the normality test, the iteration counter is set to $q \leftarrow 1$ and the level of significance for the Shapiro-Wilk test is $\alpha_{\text{nor}}(1) \leftarrow 0.05$. In general if on the $q$th iteration of the normality test (12)–(13) the hypothesis (13) is accepted at the level of significance

$$\alpha_{\text{nor}}(q) \leftarrow \alpha_{\text{nor}}(1)\exp\left[-0.184206(q-1)^2\right], \qquad (14)$$

then we proceed to the test for correlation as outlined in Section 4.3; otherwise, we repeat the following steps until the batch means finally pass the normality test (12)–(13):

1. The iteration counter $q$, the batch size $m$, and overall sample size are increased according to

$$\left.\begin{array}{l} q \leftarrow q+1, \\ m \leftarrow \left\lfloor 2^{1/(\max\{q-4,2\})} m \right\rfloor, \quad \text{and} \\ n \leftarrow k'(s+m), \end{array}\right\}$$

   respectively; and the required additional observations are obtained (by restarting the simulation if necessary) to complete the overall sample $\{X_i : i = 1, \ldots, n\}$.
2. The overall data set $\{X_1, \ldots, X_n\}$ is reorganized into $k'$ spaced batches of size $m$ so that successive batches are separated from each other by spacers each consisting of $s$ observations.
3. The spaced batch means $\left\{\overline{X}_j(m,s) : j = 1, \ldots, k'\right\}$ are recomputed according to (1).
4. The level of significance $\alpha_{\text{nor}}(q)$ for the current iteration $q$ of the Shapiro-Wilk test is reset according to (14).
5. The $k'$ spaced batch means $\left\{\overline{X}_j(m,s) : j = 1, \ldots, k'\right\}$ are retested for normality at the level of significance $\alpha_{\text{nor}}(q)$.

We found through extensive experimentation with WASSP that the scheme in Equation (14) for decreasing the significance level $\alpha_{\text{nor}}(q)$ by at least an order of magnitude on each iteration $q$ of the normality test for which $q \ge 7$ works well for a wide variety of simulation-generated output processes (see Section 4.3 of Lada and Wilson 2006). However, experimentation with SBatch indicates that simply decreasing the significance level on each iteration of the normality test is not sufficient to control excessive growth of the batch sizes required to achieve approximate normality of the spaced batch means. As discussed in the previous section, SBatch starts with a larger batch count to increase the sensitivity of the normality test by enforcing a minimum of 68 spaced batch means for the normality test as opposed to 25 spaced batch means for WASSP. However, increasing the sensitivity of the normality test also increases the variability of the final sample size in applications involving highly nonnormal data. Consequently, instead of using the WASSP approach of increasing the batch size by the factor $\sqrt{2}$ each time the normality test is failed, SBatch increases the batch size by the factor $\sqrt{2}$ for the first six consecutive failures of the normality test; and for each failure after that, the batch size is increased according to

$$m \leftarrow \lfloor 2^{1/(q-4)} m \rfloor \quad \text{for} \quad q = 7, 8, \ldots .$$

This modification to the batch-size inflation factor balances the need for avoiding gross departures from normality of the batch means and avoiding excessive growth in the batch sizes necessary to ensure approximate normality of the batch means.

### 4.3 The Test for Correlation

When estimating CIs based on spaced batch means, we have found that we must avoid situations in which $\varphi_{\overline{X}(m,s)}$, the lag-one correlation of the spaced batch means, is so close to one as to induce excessive variability in the correlation-adjusted CI (5). Furthermore, in our prior research on ASAP and ASAP2, we found that if the nonspaced batch means $\{ \overline{X}_j(m) \}$ obtained by taking $s = 0$ in (1) have lag-one correlation $\varphi_{\overline{X}(m)}$ that substantially exceeds 0.8, then in the no precision case, CIs delivered by ASAP or ASAP2 can be extremely unstable, with excessive values for the mean, variance, or coefficient of variation of the CI half-lengths. An explanation for this phenomenon is detailed in Appendix A of Steiger et al. (2005). To avoid this phenomenon, the ASAP3 algorithm was designed to include a method that uses the arc sine transformation of $\widehat{\varphi}_{\overline{X}(m)}$, the standard estimator of $\varphi_{\overline{X}(m)}$, to check the condition that a 97.5% upper confidence limit for $\sin^{-1}[\varphi_{\overline{X}(m)}]$ does not exceed the threshold $\sin^{-1}(0.8)$.

Based on all the foregoing considerations, the SBatch algorithm applies a similar correlation test to the approximately normal, spaced batch means. In particular, we have found that SBatch delivers reasonably stable, well-behaved CIs if the batch size $m$ is taken sufficiently large to ensure that

$$\varphi_{\overline{X}(m,s)} \equiv \mathrm{Corr}\big[ \overline{X}_j(m,s), \overline{X}_{j+1}(m,s) \big] \le 0.8. \qquad (15)$$

To test the hypothesis (15) at the level of significance $\alpha_{\mathrm{corr}} = 0.025$, we seek a one-sided upper CI for $\varphi_{\overline{X}(m,s)}$ that is based on $\widehat{\varphi}_{\overline{X}(m,s)}$ as defined in (7) and that, with probability no less than $1 - \alpha_{\mathrm{corr}} = 0.975$, falls at or below the limit 0.8 when (15) holds. Since $k' \ge 68$ on every iteration of SBatch, we use the approximation

$$\sin^{-1}\big[\widehat{\varphi}_{\overline{X}(m,s)}\big] \stackrel{\cdot}{\sim} N\Big\{ \sin^{-1}\big[\varphi_{\overline{X}(m,s)}\big], 1/k' \Big\}$$

(Steiger et al. 2005, p. 52) to test the hypothesis (15) at the level of significance $\alpha_{\mathrm{corr}} = 0.025$ by checking for the condition that the $100(1 - \alpha_{\mathrm{corr}})\%$ one-sided upper confidence limit for $\sin^{-1}\big[\varphi_{\overline{X}(m,s)}\big]$ does not exceed the threshold $\sin^{-1}(0.8) = 0.927$. If on a particular iteration of the correlation test within SBatch we find

$$\widehat{\varphi}_{\overline{X}(m,s)} \le \sin\left[ 0.927 - \frac{z_{1-\alpha_{\mathrm{corr}}}}{\sqrt{k'}} \right], \qquad (16)$$

where $z_{1-\alpha_{\mathrm{corr}}} = z_{0.975} = 1.96$, then we conclude that the current batch size $m$ satisfies (15); and we proceed to the construction of a correlation-adjusted CI for $\mu_X$ in the next step of SBatch.

If the condition (16) is not satisfied, then we increase the batch size $m$ and the overall sample size $n$ according to

$$m \leftarrow \lfloor 1.1m \rfloor \quad \text{and} \quad n \leftarrow k'(s + m).$$

Additional simulation-generated observations are collected if necessary and a new set of spaced batch means is computed according to (1) using the new batch size $m$ and the final spacer size $s$ determined in the randomness test. Then the hypothesis (15) is retested.

### 4.4 Computing the Correlation-Adjusted Confidence Interval

At this point, we have a time series of $k'$ spaced batch means that is approximately a stationary Gaussian process. This time series can be used to construct a CI for $\mu_X$ that has been adjusted to compensate for the remaining (nonexcessive) correlations between the $k'$ spaced batch means for batches of the current size $m$. The correlation adjustment $A$ has the form (6) (where in the computing formulas (3) and (7) we take $k = k'$) so that a correlation-adjusted $100(1 - \beta)\%$ CI for the steady-state mean is

$$\overline{X} \pm t_{1-\beta/2, k'-1} \sqrt{\frac{A \widehat{\sigma}^2_{\overline{X}(m,s)}}{k'}}. \qquad (17)$$

The justification for (17) is given in the Appendix.

### 4.5 Fulfilling the Precision Requirement

Let $H$ denote the half-length of the confidence interval (17). If (17) satisfies the precision requirement

$$H \le H^*, \qquad (18)$$

where $H^*$ is given by

$$H^* \leftarrow \begin{cases} \infty, & \text{for no user-spec. prec. level,} \\ r^* \big| \overline{X} \big|, & \text{for a user-spec. rel. prec. level } r^*, \\ h^*, & \text{for a user-spec. abs. prec. level } h^*, \end{cases} \qquad (19)$$

then SBatch terminates, delivering the CI (17)

If the precision requirement (18) is not satisfied, then we estimate $k^*$, the total number of spaced batches of the current batch size $m$ that are needed to satisfy (18),

$$k^* \leftarrow \left\lceil \left(H/H^*\right)^2 k' \right\rceil; \quad \text{and we take} \quad k' \leftarrow \min\left\{k^*, 1024\right\}.$$

Thus $k^*(s+m)$ is our latest estimate of the total sample size needed to satisfy the precision requirement; and an upper bound of 1024 is imposed on the actual spaced batch count $k'$ for the next iteration of SBatch.

The new batch size $m$ for the next iteration of SBatch is assigned according to

$$m \leftarrow \left\lceil \left(k^*/k'\right)(s+m) \right\rceil - s,$$

so that the total sample size $n$ is increased approximately by the factor $(H/H^*)^2$. On the next iteration of SBatch, the total sample size, including the warm-up period, is thus given by

$$n \leftarrow k'(s+m),$$

where $s$ was finalized in the randomness test. The additional observations are obtained by restarting the simulation or by retrieving extra data from storage; then the next iteration of SBatch is performed by computing a new CI for $\mu_X$ using (17).

## 5 EXPERIMENTAL RESULTS

To evaluate the performance of SBatch with respect to the coverage probability of its CIs, the mean and variance of the half-length of its CIs, and the associated sample sizes, we applied SBatch together with ASAP3 and WASSP to a large suite of test problems. The experimental design includes some problems typically used to "stress-test" simulation analysis procedures and some problems more closely resembling real-world applications. To demonstrate the robustness of SBatch, we limit our discussion here to an $M/M/1$ queue waiting time process for a system with an empty-and-idle initial condition, an interarrival rate of 0.9, and a service rate of 1.0. In this system the steady-state server utilization is 0.9 and the steady-state expected waiting time in the queue is $\mu_X = 9$.

The $M/M/1$ queue waiting time process is a particularly difficult test problem for several reasons: (*i*) the magnitude of the initialization bias is substantial and decays relatively slowly; (*ii*) in steady-state operation the autocorrelation function of the waiting time process decays very slowly with increasing lags; and (*iii*) in steady-state operation the marginal distribution of waiting times has an exponential tail and is therefore markedly nonnormal. Because of these characteristics, we can expect slow convergence to the classical requirement that the batch means are i.i.d. normal. This test problem clearly reveals one of the principal advantages that SBatch shares with ASAP3—namely, that SBatch does not require the final batch means to be independent or even approximately independent. The steady-state mean response is available analytically for this test problem; thus we were able to evaluate the performance of SBatch, ASAP3, and WASSP in terms of actual versus nominal coverage probabilities for the CIs delivered by each of these procedures. Experimental results for the other test problems may be found in Lada and Wilson (2007).

Our experiments included 1000 independent replications of SBatch and WASSP and 400 independent replications of ASAP3 to construct nominal 90% and 95% CIs that satisfied four different precision requirements. For the case of no precision requirement, we took $H^* = \infty$ in (18) so that SBatch delivers the CI (17) using the batch count and batch size required to pass the randomness, normality, and correlation tests. For the cases of the precision requirements ±15%, ±7.5%, and ±3.75%, we continued the simulation of each test problem until SBatch delivered a CI of the form (17) that satisfied the stopping criterion in (18) and (19) with $r^* = 0.15$, 0.075, and 0.0375, respectively.

For each CI that was replicated 400 (respectively, 1000) times, the standard error of the coverage estimator for CIs with nominal 90% coverage probability is approximately 1.5% (respectively, 0.95%); and for CIs with nominal 95% coverage probability, the standard error of the coverage estimator is approximately 1.1% (respectively, 0.69%). As explained below, these levels of precision in the estimation of coverage probabilities turned out to be sufficient to draw meaningful conclusions about the performance of SBatch compared with that of ASAP3 and WASSP.

Table 1 summarizes the experimental performance of the procedures SBatch, ASAP3, and WASSP when they were applied to the waiting times in the $M/M/1$ queue. As can be seen from this table, all three procedures performed reasonably well. The results in Table 1 suggest that as the precision level $r^*$ becomes progressively smaller, both SBatch and ASAP3 deliver CIs whose coverage probabilities converge to their nominal levels, while WASSP delivers CIs with some overcoverage; moreover, in this situation WASSP appears to require substantially larger sample sizes than are required by SBatch or ASAP3.

## 6 CONCLUSIONS

SBatch is a completely automated spaced batch means procedure for constructing an approximate confidence interval for the steady-state mean of a simulation output process. The SBatch procedure incorporates many advantages of its predecessors ASAP3 and WASSP, such as sampling efficiency and the ability to effectively eliminate initialization bias. While our extensive performance evaluation of SBatch indicates that it compares favorably with ASAP3 and WASSP, the primary advantage of SBatch is that it is much simpler to understand and implement than either ASAP3 or

Table 1: Performance of SBatch, WASSP, and ASAP3 in the $M/M/1$ queue waiting time process with 90% server utilization.

| Precision | | Nominal 90% CIs | | | Nominal 95% CIs | | |
|---|---|---|---|---|---|---|---|
| Requirement | Performance Measure | SBatch | WASSP | ASAP3 | SBatch | WASSP | ASAP3 |
| None | # replications | 1,000 | 1,000 | 400 | 1,000 | 1,000 | 400 |
| | CI coverage | 87.1% | 87.7% | 87.5% | 91.6% | 93.4% | 91.5% |
| | Avg. sample size | 54,371 | 18,090 | 31,181 | 54,371 | 17,971 | 31,181 |
| | Avg. CI half-length | 1.3864 | 3.0715 | 2.0719 | 1.6578 | 3.9987 | 2.5209 |
| | Var. CI half-length | 0.2603 | 2.0026 | 0.3478 | 0.3725 | 3.6999 | 0.5350 |
| ±15% | # replications | 1,000 | 1,000 | 400 | 1,000 | 1,000 | 400 |
| | CI coverage | 86.6% | 87.2% | 91% | 91.2% | 93% | 95.5% |
| | Avg. sample size | 66,719 | 92,049 | 103,742 | 88,447 | 143,920 | 140,052 |
| | Avg. CI half-length | 1.1556 | 1.1103 | 1.1820 | 1.2046 | 1.1342 | 1.2059 |
| | Var. CI half-length | 0.0396 | 0.0387 | 0.0259 | 0.0263 | 0.0314 | 0.0205 |
| ±7.5% | # replications | 1,000 | 1,000 | 400 | 1,000 | 1,000 | 400 |
| | CI coverage | 88.8% | 90.4% | 89.5% | 94% | 97% | 94% |
| | Avg. sample size | 278,642 | 388,000 | 287,568 | 403,844 | 598,020 | 382,958 |
| | Avg. CI half-length | 0.6141 | 0.5866 | 0.6273 | 0.6160 | 0.5950 | 0.6324 |
| | Var. CI half-length | 0.0055 | 0.0072 | 0.0023 | 0.0056 | 0.0056 | 0.0020 |
| ±3.75% | # replications | 1,000 | 1,000 | 400 | 1,000 | 1,000 | 400 |
| | CI coverage | 89.8% | 94.0% | 89.5% | 95.2% | 97.7% | 93.5% |
| | Avg. sample size | 1,151,178 | 1,518,400 | 969,011 | 1,618,147 | 2,361,300 | 1,341,522 |
| | Avg. CI half-length | 0.3081 | 0.3060 | 0.3200 | 0.3076 | 0.3060 | 0.3210 |
| | Var. CI half-length | 0.0014 | 0.0008 | 0.0004 | 0.0014 | 0.0007 | 0.0004 |

WASSP. Additional experimental results, follow-up papers and revised software, will be available on the website <www.ise.ncsu.edu/jwilson>.

# A APPENDIX: JUSTIFICATION FOR THE CONFIDENCE INTERVAL (17)

Steiger et al. (2005) find that for adjacent (nonspaced) batch means (i.e., $s = 0$), if the batch size $m$ is sufficiently large to satisfy (15), then the stochastic behavior of the truncated, nonspaced batch means

$$\overline{X}_j(m) \equiv \frac{1}{m}\sum_{i=1}^{m} X_{s+(j-1)m+i} \quad \text{for} \quad j = 1, \ldots, k$$

is closely approximated by that of a stationary AR(1) process,

$$\overline{X}_j(m) - \mu_X = \varphi_{\overline{X}(m)}\big[\,\overline{X}_{j-1}(m) - \mu_X\,\big] + \varepsilon_j$$

for $j = 1, 2, \ldots$, where the autoregressive parameter $\varphi_{\overline{X}(m)}$ satisfies $\big|\varphi_{\overline{X}(m)}\big| < 1$ and the residuals $\{\varepsilon_j : j = 1, 2, \ldots\}$ are i.i.d. normal with mean zero and variance $\sigma^2_{\overline{X}(m)}\big[1 - \varphi^2_{\overline{X}(m)}\big]$. In this situation the autocorrelation function (ACF) of the nonspaced batch means is given by

$$\rho_{\overline{X}(m)}(\ell) \equiv \mathrm{Corr}\big[\,\overline{X}_j(m), \overline{X}_{j+\ell}(m)\,\big] = \varphi^{|\ell|}_{\overline{X}(m)} \tag{20}$$

for lag $\ell = 0, \pm1, \pm2, \ldots$.

To compute an approximately unbiased estimator of $\mathrm{Var}\big[\,\overline{\overline{X}}(m,k,s)\,\big]$, we derive an approximation to the ACF of the spaced batch means based on the ACF (20) of the nonspaced batch means. If spacer size is an integral multiple of the batch size so that $s = rm$ for some positive integer $r$, then it follows from (20) that

$$\begin{aligned}\rho_{\overline{X}(m,s=rm)}(\ell) &\equiv \mathrm{Corr}\big[\,\overline{X}_1(m, s = rm), \overline{X}_{1+|\ell|}(m, s = rm)\big] \\ &= \mathrm{Corr}\big[\,\overline{X}_1(m), \overline{X}_{1+|\ell|(r+1)}(m)\,\big] \\ &= \varphi^{|\ell|(r+1)}_{\overline{X}(m)} \quad \text{for} \quad \ell = 0, \pm1, \pm2, \ldots. \end{aligned} \tag{21}$$

In view of (21), we make the key assumption that for any nonnegative integer value of $s$, we can take $r = s/m$ (a rational number) to obtain the approximate result

$$\rho_{\overline{X}(m,s)}(\ell) \approx \varphi^{|\ell|\left(\frac{s}{m}+1\right)}_{\overline{X}(m)} = \varphi^{|\ell|(m+s)/m}_{\overline{X}(m)} \tag{22}$$

for $\ell = 0, \pm1, \pm2, \ldots$. Letting $\varphi_{\overline{X}(m,s)} \equiv \rho_{\overline{X}(m,s)}(1) = \varphi^{(m+s)/m}_{\overline{X}(m)}$ and taking $\ell = 1$ in (22), we have $\varphi_{\overline{X}(m)} \approx \varphi^{m/(m+s)}_{\overline{X}(m,s)}$; and inserting this latter result back into (22), we have

$$\rho_{\overline{X}(m,s)}(\ell) \approx \left[\varphi^{m/(m+s)}_{\overline{X}(m,s)}\right]^{|\ell|(m+s)/m} = \varphi^{|\ell|}_{\overline{X}(m,s)} \tag{23}$$

for $\ell = 0, \pm1, \pm2, \ldots$.

Thus the assumption that the nonspaced batch means constitute an AR(1) process coupled with the approximation (22) for any positive integer values of $m$ and $s$ yields the same type of exponentially decaying autocorrelation function for the spaced batch means as for the nonspaced batch means. Note, moreover, that the lag-one autocorrelation $\varphi_{\overline{X}(m,s)}$ of the spaced batch means depends on both the batch size $m$ and the spacer size $s$.

It is easily proved that

$$\begin{aligned} B(m,k,s) &\equiv 1 + 2\sum_{\ell=1}^{k-1}\left(1-\ell/k\right)\varphi^{\ell}_{\overline{X}(m,s)} \\ &= \frac{1+\varphi_{\overline{X}(m,s)}}{1-\varphi_{\overline{X}(m,s)}} - \frac{2\varphi_{\overline{X}(m,s)}\left[1-\varphi^{k}_{\overline{X}(m,s)}\right]}{k\left[1-\varphi_{\overline{X}(m,s)}\right]^2}; \end{aligned}$$

and in the following analysis, we use the approximation

$$B(m,k,s) \approx \frac{1+\varphi_{\overline{X}(m,s)}}{1-\varphi_{\overline{X}(m,s)}} \quad \text{for large } k. \tag{24}$$

From (23), (24), and standard results for the AR(1) process, we have

$$\begin{aligned} \operatorname{Var}\left[\overline{\overline{X}}(m,k,s)\right] &= \left\{\frac{\operatorname{Var}\left[\overline{X}_1(m,s)\right]}{k}\right\} B(m,k,s) \qquad (25) \\ &\approx \left\{\frac{\operatorname{Var}\left[\overline{X}_1(m,s)\right]}{k}\right\}\left[\frac{1+\varphi_{\overline{X}(m,s)}}{1-\varphi_{\overline{X}(m,s)}}\right] \end{aligned}$$

for large $k$; and thus (25) is the basis for the sample estimator

$$\widehat{\operatorname{Var}}\left[\overline{\overline{X}}(m,k,s)\right] = \frac{A\widehat{\sigma}^2_{\overline{X}(m,s)}}{k}. \tag{26}$$

Thus the half-length $H$ of (17) is appropriate for a CI centered at $\overline{\overline{X}}(m,k,s)$; and since $\overline{X}$ has the same mean and smaller variance than that of $\overline{\overline{X}}(m,k,s)$, it follows that the CI (17) will have coverage no less than that of the confidence interval $\overline{\overline{X}}(m,k,s) \pm H$. ■

## REFERENCES

Fishman, G. S., and L. S. Yarberry. 1997. An implementation of the batch means method. *INFORMS Journal on Computing* 9 (3): 296–310.

Fox, B., D. Goldsman, and J. J. Swain. 1991. Spaced batch means, *Operations Research Letters* 10:255–263.

Lada, E. K., and J. R. Wilson. 2006. A wavelet-based spectral procedure for steady-state simulation analysis. *European Journal of Operational Research* 174:1769–1901.

Lada, E. K., and J. R. Wilson. 2007. SBatch: A spaced batch means procedure for steady-state simulation analysis. Technical report, Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina. Available online via <www.ise.ncsu.edu/jwilson/files/sbatchtr.pdf> [accessed June 20, 2007].

Lada, E. K., J. R. Wilson, and N. M. Steiger. 2003. A wavelet-based spectral method for steady-state simulation analysis. In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 422–430. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online via <www.informs-sim.org/wsc03papers/052.pdf> [accessed July 15, 2006].

Lada, E. K., J. R. Wilson, N. M. Steiger, and J. A. Joines. 2007. Performance of a wavelet-based spectral procedure for steady-state simulation analysis. *INFORMS Journal on Computing* 19 (2): 150–160.

Royston, J. P. 1982. Algorithm AS 181. The $W$ test for normality. *Applied Statistics* 31:176–180.

Royston, P. 1993. A toolkit for testing for non-normality in complete and censored samples. *The Statistician* 42 (1): 37–43.

Royston, P. 1995. Remark AS R94: A remark on algorithm AS 181: The $W$-test for normality. *Applied Statistics* 44 (4): 547–551.

Shapiro, S. S., and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52 (3–4): 591–611.

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, and D. Goldsman. 2004. Steady-state simulation analysis using ASAP3. In *Proceedings of the 2004 Winter Simulation Conference*, ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 672–680. Available online via <www.informs-sim.org/wsc04papers/081.pdf> [accessed June 20, 2007].

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, and D. Goldsman. 2005. ASAP3: A batch means procedure for steady-state simulation output analysis. *ACM Transactions on Modeling and Computer Simulation* 15 (1): 39–73.

von Neumann, J. 1941. Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics* 12 (4): 367–395.

## AUTHOR BIOGRAPHIES

**EMILY K. LADA** is an operations research development tester at the SAS Institute. She is a member of IIE and INFORMS. Her e-mail address is <Emily.Lada@sas.com>.

**JAMES R. WILSON** is professor and head of the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He is a member of AAUW, ACM, and ASA, and he is a Fellow of IIE and INFORMS. His e-mail address is <jwilson@eos.ncsu.edu>, and his web page is <www.ie.ncsu.edu/jwilson>.