Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

SKART: A SKEWNESS- AND AUTOREGRESSION-ADJUSTED BATCH-MEANS PROCEDURE FOR SIMULATION ANALYSIS

Ali Tafazzoli
James R. Wilson

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

Emily K. Lada

Operations Research Department
SAS Institute Inc.
100 SAS Campus Drive
R5413
Cary, NC 27513, U.S.A.

Natalie M. Steiger

Maine Business School
University of Maine
5723 D. P. Corbett Bldg.
Orono, ME 04469, U.S.A.

ABSTRACT

We discuss Skart, an automated batch-means procedure for constructing a skewness- and autoregression-adjusted confidence interval for the steady-state mean of a simulation output process. Skart is a sequential procedure designed to deliver a confidence interval that satisfies user-specified requirements concerning not only coverage probability but also the absolute or relative precision provided by the half-length. Skart exploits separate adjustments to the half-length of the classical batch-means confidence interval so as to account for the effects on the distribution of the underlying Student's t-statistic that arise from nonnormality and autocorrelation of the batch means. Skart also delivers a point estimator for the steady-state mean that is approximately free of initialization bias. In an experimental performance evaluation involving a wide range of test processes, Skart compared favorably with other simulation analysis methods—namely, its predecessors ASAP3, WASSP, and SBatch as well as ABATCH, LBATCH, the Heidelberger-Welch procedure, and the Law-Carson procedure.

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 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 marginal 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 article, we are primarily interested in estimating the steady-state mean, $\mu_X = \mathrm{E}[X] = \int_{-\infty}^{\infty} x \, \mathrm{d}F_X(x)$; and we limit the discussion to output processes for which $\mathrm{E}[X_i^2] < \infty$ so that the marginal mean μ_X and 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 Skart, a procedure for steady-state simulation analysis that is based on the method of batch means and that incorporates many advantages of its predecessors—i.e., ASAP3 (Steiger et al. 2005); WASSP (Lada and Wilson 2006); and SBatch (Lada, Steiger, and Wilson 2008)—while avoiding many of their disadvantages. Based on our experimentation with a broad diversity of test processes, we reached the following conclusions: (i) Skart generally delivered closer conformance to the nominal CI coverage probability than its predecessors; (ii) Skart's sampling efficiency was about the same as that of ASAP3 and superior to that of WASSP and SBatch; (iii) Skart eliminated initialization bias about as well as its predecessors did; and (iv) Skart and Sbatch were simpler to implement and understand than ASAP3 and WASSP.

There is substantial experimental evidence that ASAP3 outperforms ABATCH and LBATCH (Steiger and Wilson 2002) and the Law-Carson procedure (Lada, Steiger, and Wilson 2006). Similarly Lada et al. (2007) provide good experimental evidence that WASSP outperforms the Heidelberger-Welch procedure. Thus we concluded that in most important respects, Skart compared favorably with many of the simulation analysis methods currently in use.

The rest of this article is organized as follows. In §2 we review the method of batch means, and in §3 we provide an overview of Skart. Section 4 details the operational steps of Skart. In §5 we summarize selected results from our experimental performance evaluation, and in §6 we present our main conclusions. A complete discussion of Skart, including the full performance evaluation, is available online via <www.ise.ncsu.edu/jwilson/files/skartsum.pdf>. The slides for the presentation accompanying this paper are available via <www.ise.ncsu.edu/jwilson/files/skartwsc08.pdf>.

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

2 THE METHOD OF BATCH MEANS

In the method of nonoverlapping batch means (NBM), the outputs $\{X_i : i = 1, \ldots n\}$ are divided into k adjacent nonoverlapping batches, each of size m, where we assume $n = km$ and both k and m are sufficiently large to ensure that the resulting batch means are at least approximately independent and identically distributed (i.i.d.) normal random variables. The sample mean for the jth batch is

$$Y_j(m) = \frac{1}{m} \sum_{i=m(j-1)+1}^{mj} X_i \quad \text{for} \quad j = 1, \ldots, k; \tag{1}$$

and the grand mean of the individual batch means,

$$\overline{Y} = \overline{Y}(m,k) = \frac{1}{k} \sum_{j=1}^{k} Y_j(m), \tag{2}$$

is used as a point estimator for μ_X. The objective is to construct a CI estimator for μ_X that is centered on a point estimator as in Equation (2), where in practice some initial observations (or batches) may be deleted (truncated) to eliminate the effects of initialization bias.

If after appropriate data truncation the output process $\{X_j : j = 1, \ldots, n\}$ is stationary and weakly dependent, then as $m \to \infty$ with k fixed so that $n = km \to \infty$, an asymptotically valid $100(1-\alpha)\%$ CI for μ_X is

$$\overline{Y}(m,k) \pm t_{1-\alpha/2,k-1} \frac{S_{m,k}}{\sqrt{k}}, \tag{3}$$

where $t_{1-\alpha/2,k-1}$ denotes the $1-\alpha/2$ quantile of Student's t-distribution with $k-1$ degrees of freedom, and

$$S^2_{m,k} = \frac{1}{k-1} \sum_{j=1}^{k} \left[Y_j(m) - \overline{Y}(m,k)\right]^2 \tag{4}$$

is the sample variance of the k batch means with batch size m. The main difficulty with conventional implementations of NBM—such as the procedure of Law and Carson (1979) and the ABATCH and LBATCH procedures of Fishman and Yarberry (1997)—is the lack of a reliable method for determining a sufficiently large batch size m so that the batch means $\{Y_j(m)\}$ are approximately uncorrelated and normal. For an elaboration of this issue, see Steiger et al. (2005) and Lada, Steiger, and Wilson (2006).

3 OVERVIEW OF SKART

In this article, we develop a variant of the method of batch means that effectively takes into account not only any deterministic trends or stochastic dependencies in the sequence of batch means but also their marginal skewness so as to determine sufficiently large values of the data-truncation point (statistics clearing time) and the batch size m such that the truncated batch means approximately constitute a stationary first-order autoregressive process with mean μ_X. In our experience this condition was much easier to achieve in practice than the conditions required to apply other batch-means procedures. Beyond the data-truncation point, we compute the sample variance of k' adjacent batch means for batch size m using Equation (4) with $k = k'$. (In the rest of this article, we always let k' denote the current number of truncated, adjacent batch means.) We deliver an asymptotically valid $100(1-\alpha)\%$ skewness- and autoregression-adjusted CI for μ_X having the form

$$\left[\overline{Y}(m,k') - G(L)\sqrt{\frac{AS^2_{m,k'}}{k'}},\ \overline{Y}(m,k') - G(R)\sqrt{\frac{AS^2_{m,k'}}{k'}}\right], \tag{5}$$

where

$$G(r) \equiv \frac{\sqrt[3]{1 + 6\beta(r-\beta)} - 1}{2\beta}, \text{ with } \beta = \frac{\widehat{\mathcal{B}}_{m,k''}}{6\sqrt{k''}} \tag{6}$$

and

$$\widehat{\mathcal{B}}_{m,k''} = \left\{ \begin{array}{l} \text{approximately unbiased estimator of marginal} \\ \text{skewness of each } Y_j(m) \text{ based on } k'' \text{ approx-} \\ \text{imately i.i.d. } \textit{spaced} \text{ batch means each with} \\ \text{batch size } m \text{ (see (20) below)}, \end{array} \right\}$$

and finally

$$L = t_{1-\alpha/2,k''-1} \quad \text{and} \quad R = t_{\alpha/2,k''-1}. \tag{7}$$

Thus we see that $G(L)$ and $G(R)$ are skewness-adjusted quantiles of Student's t-distribution for the left and right half-lengths of the proposed CI (this point is elaborated in §4.4); and the autoregression (correlation) adjustment A is applied to the sample variance $S^2_{m,k'}$ to compensate for any residual correlation between the batch means. The correlation adjustment A is computed as

$$A = \left[1 + \widehat{\varphi}_{Y(m)}\right] \Big/ \left[1 - \widehat{\varphi}_{Y(m)}\right], \tag{8}$$

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

$$\widehat{\varphi}_{Y(m)} = \widehat{\text{Corr}}\left[Y_j(m), Y_{j+1}(m)\right] = \tag{9}$$

$$\frac{1}{k'-1} \sum_{j=1}^{k'-1} \left[Y_j(m) - \overline{Y}(m,k)\right]\left[Y_{j+1}(m) - \overline{Y}(m,k)\right] \Big/ S^2_{m,k'}.$$

Figure 1 depicts a high-level flow chart of Skart. To invoke this procedure, the following user-supplied inputs are required:

1. A simulation-generated sequence $\{X_i : i = 1, \ldots, n\}$ of target responses whose steady-state mean μ_X is to be estimated;

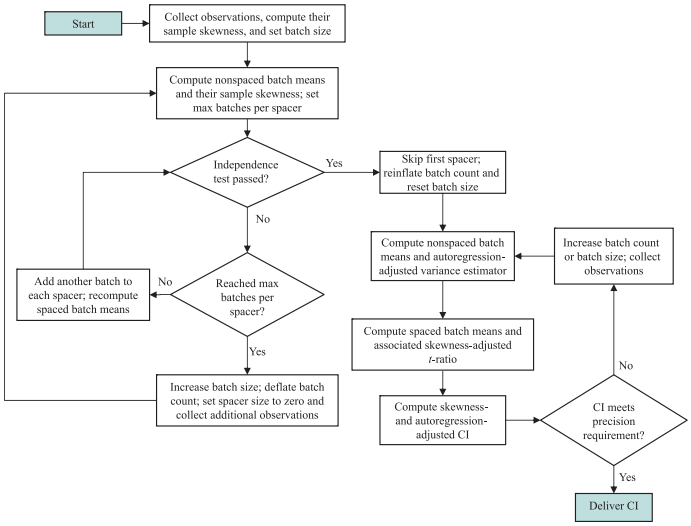

Figure 1: High-level flow chart of Skart.

2. The desired CI coverage probability $1-\alpha$, where $0 < \alpha < 1$; and
3. An upper bound H^* on the final CI half-length, where H^* is expressed (a) in absolute terms as the maximum acceptable half-length; or (b) in relative terms as the maximum acceptable fraction r^* of the magnitude of the CI midpoint.

Skart returns the following outputs:

1. A nominal $100(1-\alpha)\%$ CI for μ_X that satisfies the specified precision requirement; or
2. A new, larger sample size to be used by Skart.

If additional observations must be generated by the user's simulation model before a CI with the required precision can be delivered, then Skart must be called again with the additional data; and this cycle may be repeated several times before Skart finally delivers a CI.

Skart also has a nonsequential mode of operation in which the user merely supplies a data set of fixed size and requests a CI with a specific coverage probability based on the entire data set. In this situation Skart either delivers the requested CI or advises the user of the number of additional observations that are required to build the requested CI. A public-domain version of Skart is available online via the Web page <www.ise.ncsu.edu/jwilson/page3>.

4 DETAILED OPERATIONAL STEPS OF SKART

4.1 Initialization Step

To invoke Skart, the user must provide a data set that is at least large enough to satisfy Skart's minimal requirement for a simulation-generated time series $\{X_j : j = 1, \ldots, n\}$ of length $n \leftarrow 1{,}280$. This requirement is arguably not too far from the minimal sample size required for meaningful analysis of a time series with any of the following properties: (i) a nontrivial deterministic trend (initialization bias); (ii) a nontrivial stochastic dependency structure (autocorrelation function); or (iii) nonnormal distributional characteristics.

To determine the initial value of the batch size m, Skart computes the sample skewness of the last $\ell \leftarrow 1{,}024$ observations in the initial sample,

$$\widehat{\mathcal{B}} \leftarrow \frac{\ell}{(\ell-1)(\ell-2)} \sum_{j=n-\ell+1}^{n} \left(X_j - \bar{X}\right)^3 / S^3,$$

where

$$\bar{X} \leftarrow \frac{1}{\ell} \sum_{j=n-\ell+1}^{n} X_j, \; S^2 \leftarrow \frac{1}{\ell-1} \sum_{j=n-\ell+1}^{n} \left(X_j - \bar{X}\right)^2,$$

respectively denote the sample mean and variance of the last ℓ observations in the initial sample.

If $\left|\widehat{\mathcal{B}}\right| \leq 4$, then Skart takes the initial batch size $m \leftarrow 1$; otherwise, Skart takes $m \leftarrow 16$. Then Skart initializes the current batch count $k \leftarrow 1{,}280$ and computes the current sample size $n \leftarrow km$, requesting additional observations from the user if the initially supplied data set does not contain at least n observations. This method for assigning the initial batch size and then increasing the initial sample size from $n = 1{,}280$ to $n = 16 \times 1{,}280 = 20{,}480$ if necessary is designed to ensure that in subsequent steps of Skart, the skewness of the batch means has sufficiently small magnitude so that the proposed skewness adjustment to the

classical batch-means Student's t-statistic will be effective; see §§4.2–4.5.

Next Skart divides the data set $\{X_i : i = 1, \ldots, n\}$ into k batches of size m, with the initial spacer consisting of $d \leftarrow 0$ ignored batches preceding each "spaced" batch from which we compute a batch mean; and the corresponding "spaced" batch means $\{Y_1(m), \ldots, Y_k(m)\}$ are computed using (1). (The index d will be used in §4.2, when Skart determines the number of spacers per batch that are required to ensure the spaced batch means are approximately i.i.d.) Then Skart determines the maximum number of batches per spacer, d^*, to be used in the test for randomness of the spaced batch means in §4.2 based on the sample skewness of the current set of batch means. By default we take $d^* \leftarrow 10$. To get a more accurate estimator $\widehat{\mathcal{B}}_m$ of the marginal skewness of the batch means with current batch size m and to reduce the effect of any initial transient on the sample skewness computation, Skart skips the first 20% of the current set of batch means and computes the skewness only using the last 80% of the batch means as follows:

$$\ell \leftarrow \lfloor 0.8k \rfloor, \quad \bar{Y}(m, \ell) \leftarrow \frac{1}{\ell} \sum_{j=k-\ell+1}^{k} Y_j(m), \tag{10}$$

$$S^2_{m,\ell} \leftarrow \frac{1}{\ell - 1} \sum_{j=k-\ell+1}^{k} \left[Y_j(m) - \bar{Y}(m, \ell) \right]^2, \tag{11}$$

$$\widehat{\mathcal{B}}_m \leftarrow \frac{\ell}{(\ell - 1)(\ell - 2)} \sum_{j=k-\ell+1}^{k} \frac{\left[Y_j(m) - \bar{Y}(m, \ell) \right]^3}{S^3_{m,\ell}}. \tag{12}$$

To handle effectively an output process whose marginal skewness has excessive magnitude, Skart exploits the broadly applicable property that as $m \to \infty$, the $\{Y_j(m)\}$ are asymptotically i.i.d. normal; and since the normal distribution is determined by its moments, the skewness of the batch means tends to zero as $m \to \infty$. Hence, if the sample skewness $\widehat{\mathcal{B}}_m$ of the $\{Y_j(m)\}$ in (12) satisfies

$$\left| \widehat{\mathcal{B}}_m \right| > 0.5, \tag{13}$$

then Skart sets $d^* \leftarrow 3$ as the maximum number of batches allowed per spacer to be used in the test for randomness of the spaced batch means. By doing this, Skart forces the randomness test to increase the batch size more frequently for skewed processes as explained in the next section.

To complete initialization, we take $\alpha_{\text{ran}} \leftarrow 0.2$ as the randomness test size and $b \leftarrow 0$ as the number of times the batch count has been deflated in the randomness test.

4.2 The Test for Randomness

In this step, Skart applies the randomness test of von Neumann (1941) to the current set of $k' \leftarrow k = 1{,}280$ batch means by computing the ratio of the mean square successive difference of the batch means to the sample variance of the batch means. Skart applies this test iteratively to determine the size of an interbatch spacer that is sufficiently large to yield approximate independence of the corresponding spaced batch means and consequently to determine a proper batch count, batch size, and data truncation point beyond which all the computed batch means are approximately independent of the simulation's initial conditions as well as being i.i.d.—that is, the spaced batch means constitute a random sample from a common distribution.

At the significance level α_{ran}, we test the null hypothesis that the current spaced batch means are i.i.d.,

$$\left\{ Y_{j(d+1)}(m) : j = 1, \ldots, k' \right\} \quad \text{are i.i.d.,} \tag{14}$$

by computing their grand mean,

$$\bar{Y}(m, k', d) = \frac{1}{k'} \sum_{j=1}^{k'} Y_{j(d+1)}(m),$$

and the corresponding randomness test statistic,

$$C_{k'} \leftarrow 1 - \frac{\sum_{j=1}^{k'-1} \left[Y_{j(d+1)}(m) - Y_{(j+1)(d+1)}(m) \right]^2}{2 \sum_{j=1}^{k'} \left[Y_{j(d+1)}(m) - \bar{Y}(m, k', d) \right]^2}, \tag{15}$$

which is a relocated and rescaled version of the ratio of the mean square successive difference of the spaced batch means to the sample variance of the spaced batch means. Since Skart's test for randomness usually involves a large number of spaced batch means, we use a normal approximation to the null distribution of the test statistic (15); see Fishman and Yarberry (1997, p. 303). Let z_ω denote the ω quantile of the standard normal distribution for $0 < \omega < 1$. If

$$\left| C_{k'} \right| \leq z_{1-\alpha_{\text{ran}}/2} \sqrt{(k' - 2) / \left[(k')^2 - 1 \right]}, \tag{16}$$

then the hypothesis (14) is accepted; otherwise (14) is rejected so that Skart must increase the spacer size before retesting (14). 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 (14).

If the randomness test is passed at significance level α_{ran} with $d = 0$ and $k = 1{,}280$ so that the current set of batch means is $\{Y_1(m), Y_2(m), \ldots, Y_{1280}(m)\}$, then we fix the batch count $k' \leftarrow 1{,}280$ and the batch size m and proceed to the variance-adjustment step as detailed in §4.3. Otherwise, we insert spacers each consisting of one ignored batch between the $k' \leftarrow 1280/2 = 640$ remaining batches and increment the number of batches per spacer, $d \leftarrow d + 1$. Thus every other batch, beginning with the second batch, is retained as one of the spaced batch means; and the alternate batches are ignored. Now, we

retest the corresponding set of $k' = 640$ spaced batch means $\{Y_2(m), Y_4(m), \ldots, Y_{1280}(m)\}$ for randomness by reperforming (14)–(16) with batch size m and $d = 1$ batch per spacer.

If the latest retest of (14) using (16) is passed, then we move to the variance-adjustment step of §4.3 with the current values of k', m, and d; otherwise, we add another ignored batch to each spacer so that the total number of batches per spacer and the number of spaced batches are respectively updated according to

$$d \leftarrow d + 1, \quad \text{and} \quad k' \leftarrow \lfloor n / \{(d+1)m\} \rfloor. \tag{17}$$

If the update step (17) has been executed for the first time, then we now have $k' = 426$ spaced batch means $\{Y_3(m), Y_6(m), \ldots, Y_{1278}(m)\}$ with batch size m and $d = 2$ batches per spacer so that the spacer size is $dm = 2m$. Each time (17) is executed, the sequence of steps in the three immediately preceding paragraphs (i.e., (14)–(16)) is reperformed until one of the following conditions occurs:

(i) The randomness test (16) is failed and in the update step (17), we get $d > d^*$ so that the total number of batches per spacer exceeds its upper limit.
(ii) The randomness test is passed.

When condition (i) occurs, the batch size m is increased (inflated), the batch count k is decreased (deflated), and Skart's other status variables are updated as follows:

$$\left.\begin{array}{c} m \leftarrow \lceil \sqrt{2}m \rceil, \quad k \leftarrow \lceil 0.9k \rceil, \quad n \leftarrow km, \\ d \leftarrow 0, \quad b \leftarrow b + 1, \quad \text{and} \quad d^* \leftarrow 10. \end{array}\right\} \tag{18}$$

The required additional observations are obtained (from the original data set, or by restarting the simulation if necessary) to complete the overall sample $\{X_i : i = 1, \ldots, n\}$; and then k nonspaced batch means are computed from the overall sample according to (1). The updated sample skewness of the new nonspaced batch means is computed using (10)–(12). If (13) is satisfied, then Skart takes $d^* \leftarrow 3$ as the maximum number of batches per spacer; otherwise $d^* \leftarrow 10$.

If condition (i) above occurs so that the update step of the previous paragraph (including (18)) is executed, then Skart reperforms the entire randomness-testing procedure (14)–(17), starting with the current set of k nonspaced batch means of the current batch size m so that the current spacers each contain $d = 0$ batches. Skart repeats the steps outlined in the five preceding paragraphs (starting with (14)) until condition (ii) above finally occurs.

If the condition (ii) above occurs, then we proceed to the correlation-adjustment step detailed in §4.3 with the current value of d as the number of batches per spacer and the following values for the batch size and batch count:

$$k' \leftarrow \left\lceil k'(1/0.9)^b \right\rceil \quad \text{and} \quad m \leftarrow \max\left\{\lfloor n'/k' \rfloor, m\right\}, \tag{19}$$

where the batch count k' is reinflated by the factor $(1/0.9)^b$ to compensate for the total number of times the batch count was deflated in successive iterations of (18).

Skart's approach for determining a data-truncation point (statistics clearing time, warm-up period) and a proper batch size and batch count is similar to the approaches used in WASSP and Sbatch. The observations $\{X_1, X_2, \ldots, X_{dm}\}$ 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. Moreover, the spaced batch means computed beyond the first spacer are approximately i.i.d. and thus can be meaningfully used for computing the skewness adjustment to the Student's t-statistic underlying the classical method of batch-means.

However, there are a few key differences between the randomness-testing methods used in WASSP and Sbatch on the one hand and Skart on the other. In Skart, the value of d^*, the maximum number of batches per spacer, is dynamically adjusted based on the sample skewness of the most recently computed set of nonspaced batch means, whereas WASSP and Sbatch use (different) fixed values for d^*. Skart also decreases the initial count of the nonspaced batch means each time the randomness testing is restarted with (18), but in WASSP and SBatch the batch count is constant.

4.3 Autocorrelation Adjustment for the Variance Estimator

Constructing a valid CI for μ_X requires an approximately unbiased estimator for the variance of the grand mean $\bar{Y}(m, k')$, where we take $k = k'$ in (2) to indicate that the batch means have been suitably truncated to eliminate initialization bias. In the classical NBM method, $S^2_{m,k'}/k'$ is taken as the estimator of $\mathrm{Var}\left[\bar{Y}(m, k')\right]$; and thus the conventional $100(1-\alpha)\%$ CI (3) centered at $\bar{Y}(m, k')$ is taken to have half-length $t_{1-\alpha/2,k'-1} S_{m,k'}/\sqrt{k'}$ on the assumption that the batch means are approximately i.i.d. normal variates. In practice, however, we have found that the batch means are rarely uncorrelated even when they are approximately normal (Steiger et al. 2005); and in general $S^2_{m,k'}/k'$ can be a badly biased estimator of $\mathrm{Var}[\bar{Y}(m, k')]$—especially when the original simulation output process $\{X_i\}$ has a positive autocorrelation function that declines slowly with increasing lags. Like SBatch, Skart applies the autocorrelation adjustment A given by (8) to the variance estimator $S^2_{m,k'}/k'$ to compensate for any residual correlation between the truncated batch means.

To compute the autocorrelation adjustment, Skart first uses the batch count k' and the batch size m determined

in (19) to perform the following operations on the data set $\{X_i : i = 1, \ldots, n\}$ accumulated so far: (a) Skart skips the first $w = d \times m$ observations to eliminate the effects of initialization bias; and (b) Skart computes the corresponding nonspaced batch means $\{Y_j(m) : j = 1, \ldots, k'\}$ from the truncated data set. In general the resulting nonspaced batch means will have a nonnegligible autocorrelation structure; moreover, they sometimes exhibit significant departures from normality. If the $\{Y_j(m)\}$ constitute a stationary process (a property that the randomness test was designed to ensure), then usually the batch-means process can be adequately modeled by an autoregressive–moving average (ARMA) process, at least for the purpose of estimating the autocorrelation structure of the batch means (Box, Jenkins, and Reinsel 1994; Steiger and Wilson 2002). Skart exploits the property that the lag-one autoregressive parameter is the dominant factor in the ARMA model for the $\{Y_j(m)\}$ as $m \to \infty$; and hence Skart fits a first-order autoregressive (AR(1)) model to the $\{Y_j(m)\}$, estimating the associated autoregressive parameter by $\widehat{\varphi}_{Y(m)}$ as given in (9). Then $\widehat{\varphi}_{Y(m)}$ is used to compute the correlation adjustment A as given in (8) so that $AS^2_{m,k'}/k'$ is an approximately unbiased estimator of $\mathrm{Var}\big[\overline{Y}(m, k')\big]$.

4.4 Skewness Adjustment to Student's t-Statistic

When the truncated, nonspaced batch means $\{Y_j(m) : j = 1, \ldots, k'\}$ exhibit significant departures from normality, Skart applies an appropriate adjustment to the usual critical value of Student's t-distribution to yield a valid CI for μ_X. In §§4.1–4.2, Skart inflates the batch size for a highly skewed process to mitigate at least partially the effect of nonnormality of the batch means on the associated NBM Student's t-statistic. In some simulation applications, however, conditions such as high congestion can induce substantial skewness in the batch means (of flow times, for example) even for batch sizes that are sufficiently large to ensure the batch means are nearly uncorrelated. Therefore, the skewness adjustment that Skart applies in this step can be crucial in delivering a CI with good coverage. Moreover, we have found that the batch-size increases imposed in previous steps of Skart are necessary to ensure that the skewness of the batch means has sufficiently small magnitude so the skewness adjustment is effective.

In this step of Skart, we adapt the skewness adjustment developed by Willink (2005). The adjustment is based the modified t-statistic of Johnson (1978) in which key terms of a Cornish-Fisher expansion involve the marginal skewness of the basic data items going into the t-statistic (in this case, batch means). It must be recognized that Willink's skewness adjustment is based on the assumption that the basic data items going into the t-statistic are i.i.d.; but the experimental performance evaluation shows the effectiveness of this adjustment in the operation of Skart when the basic data items are correlated batch means.

To obtain an approximately unbiased estimator of the marginal skewness of the current set of truncated, nonspaced batch means $\{Y_j(m) : j = 1, \ldots, k'\}$, Skart computes this skewness estimator from approximately i.i.d. spaced batch means constituting a subset of the current set of nonspaced batch means. From the randomness test in §4.2, we concluded that spacers consisting of w observations are sufficiently large to ensure approximate independence of the corresponding spaced batch means. Thus from the current set of k' nonspaced batch means, we can extract the spaced batch means $\big\{Y_1(m), Y_{d'+2}(m), \ldots, Y_{(k''-1)(d'+1)+1}(m)\big\}$, where $d' = \lceil w/m \rceil$ is the number of batches per spacer and $k'' = 1 + \lfloor (k'-1)/(d'+1) \rfloor$ is the total number of spaced batch means. For simplicity in the following discussion, we let $Y_j(m, d') \equiv Y_{(j-1)(d'+1)+1}(m)$ $(j = 1, \ldots, k'')$ denote the associated spaced batch means with the following approximately unbiased estimators of their required marginal moments: the grand mean

$$\overline{Y}(m, k'', d') \leftarrow \frac{1}{k''} \sum_{j=1}^{k''} Y_j(m, d');$$

the sample variance

$$S^2_{m,k'',d'} \leftarrow \frac{1}{k''-1} \sum_{j=1}^{k''} \big[Y_j(m, d') - \overline{Y}(m, k'', d')\big]^2;$$

and sample third central moment

$$\mathcal{T}_{m,k'',d'} \leftarrow \frac{k''}{(k''-1)(k''-2)} \sum_{j=1}^{k''} \big[Y_j(m, d') - \overline{Y}(m, k'', d')\big]^3.$$

Skart uses these statistics to calculate $G(L)$ and $G(R)$, the skewness-adjusted quantiles of Student's t-ratio for the left and right half-lengths of the proposed CI. The function $G(\cdot)$ is defined by taking

$$\widehat{\mathcal{B}}_{m,k''} \leftarrow \mathcal{T}_{m,k'',d'} \big/ S^3_{m,k'',d'} \tag{20}$$

in (6); and then L and R are defined by (7). Thus Skart provides the correlation- and skewness-adjusted CI in (5).

4.5 Fulfilling the Precision Requirement

The final step of Skart is to determine whether the constructed CI satisfies the user specified precision requirement. The half-length of the CI (5) is given by

$$H \leftarrow \max\left\{ |G(L)| \sqrt{\frac{AS^2_{m,k'}}{k'}},\ \ |G(R)| \sqrt{\frac{AS^2_{m,k'}}{k'}} \right\},$$

where L and R are defined in (7), the maximum of the two half lengths. If the CI (5) satisfies the precision requirement

$$H \le H^*, \tag{21}$$

where H^* is given by

$$H^* \leftarrow \begin{cases} \infty, & \text{for no user-spec. prec. level,} \\ r^* |\bar{X}|, & \text{for a user-spec. rel. prec. level } r^*, \\ h^*, & \text{for a user-spec. abs. prec. level } h^*, \end{cases} \quad (22)$$

then Skart terminates, delivering the confidence interval (5).

If the precision requirement $H \leq H^*$ is not satisfied, then Skart estimates the total number of nonspaced batches of the current batch size that are needed to satisfy the precision requirement, $k^* \leftarrow \lceil (H/H^*)^2 k' \rceil$; and thus $k^* m$ is our latest estimate of the total sample size beyond the truncation point that is needed to satisfy the precision requirement. The batch count k' is set for the next iteration of Skart as follows: $k' \leftarrow \min\{k^*, 1{,}024\}$, where 1,024 is the upper bound on the number of batch means used in Skart. Our experiments showed that in those situations requiring more than 1,024 batches to achieve the desired precision, we could generally obtain better performance (in terms of the final required sample size) by increasing the batch size rather than increasing the batch count. If the projected total number of batches $k^* > 1{,}024$, then on the next iteration of Skart we take $k' \leftarrow 1{,}024$ and we update the batch size according to

$$m \leftarrow \lceil m \times \text{mid}\{1.02, (k^*/k'), 2\} \rceil,$$

which is assigned to satisfy the precision requirement based on an approximation to the (complicated) way in which the half-length H of the CI (5) depends on the batch size. This approximation is explained in detail in §3.4 of Steiger et al. (2005). We constrain the batch-size inflation factor to lie between the limits 1.02 and 2 to avoid an excessive number of iterations of Skart or an excessive total sample size.

On the next iteration of Skart, the total sample size including the warm-up period is thus given by $n \leftarrow (d + k')m$, where d was finalized in the randomness test. The additional simulation-generated observations are obtained by restarting the simulation or by retrieving extra data from storage; and then the next iteration of Skart is performed by computing a new CI for μ_X using Equation (5) after recomputing the correlation- and skewness-adjustments, and $S^2_{m,k'}$ for the new set of truncated, nonspaced batch means.

5 EXPERIMENTAL RESULTS

To evaluate the performance of Skart 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 Skart together with ASAP3, WASSP, and SBatch 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 Skart, we limit our discussion here to two test problems—namely, queue waiting times in the $M/M/1$ and $M/M/1$/LIFO queues.

The steady-state mean response is available analytically for these test problems; thus we were able to evaluate the performance of Skart, ASAP3, WASSP, and SBatch in terms of actual versus nominal coverage probabilities for the CIs delivered by each of these procedures.

Our experiments included 1,000 independent replications of Skart and SBatch and 400 independent replications of ASAP3 to construct nominal 90% and 95% CIs that satisfied different precision requirements. In case of WASSP, we had 1,000 independent replications available for the $M/M/1$ queue, but only 400 replications were available for the $M/M/1$/LIFO queue. For the case of no precision requirement, we took $H^* = \infty$ in (22) so Skart delivers the CI (5) using the batch count and batch size required to pass the randomness test. For the cases of the precision requirements $\pm 15\%$, $\pm 7.5\%$, and $\pm 3.75\%$, we continued the simulation of each test problem until Skart delivered a CI of the form (5) that satisfied the stopping criterion in (21) and (22) 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 Skart compared with that of ASAP3, WASSP, and SBatch.

5.1 *M/M/1* Queue Waiting Times

The first test process is the sequence of queue waiting times for an $M/M/1$ queue 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 queue waiting time is $\mu_X = 9$.

The $M/M/1$ queue waiting time process is a particularly difficult test problem for several reasons: (i) in steady-state operation the autocorrelation function of the waiting time process decays very slowly with increasing lags; and (ii) 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.

Table 1 summarizes the experimental performance of the procedures Skart, ASAP3, WASSP, and SBatch when they were applied to the waiting times in the $M/M/1$ queue. As can be seen from this table, all four procedures performed reasonably well. To put these figures in the proper perspective, note that the corresponding results for LBATCH,

Table 1: Performance of Skart, SBatch, WASSP, and ASAP3 in the $M/M/1$ queue waiting time process with 90% server utilization and empty-and-idle initial condition

Precision	Performance	Nominal 90% CIs				Nominal 95% CIs			
Requirement	Measure	Skart	SBatch	WASSP	ASAP3	Skart	SBatch	WASSP	ASAP3
None	# replications	1,000	1,000	1,000	400	1,000	1,000	1,000	400
	CI coverage	88.3%	87.1%	87.7%	87.5%	93%	91.6%	93.4%	91.5%
	Avg. sample size	42,833	54,371	18,090	31,181	42,833	54,371	17,971	31,181
	Avg. CI half-length	1.8008	1.3864	3.0715	2.0719	2.2515	1.6578	3.9987	2.5209
	Var. CI half-length	0.2810	0.2603	2.0026	0.3478	0.6426	0.3725	3.6999	0.5350
±15%	# replications	1,000	1,000	1,000	400	1,000	1,000	1,000	400
	CI coverage	89.3%	86.6%	87.2%	91%	94%	91.2%	93%	95.5%
	Avg. sample size	82,812	66,719	92,049	103,742	116,545	88,447	143,920	140,052
	Avg. CI half-length	1.1985	1.1556	1.1103	1.1820	1.2287	1.2046	1.1342	1.2059
	Var. CI half-length	0.0212	0.0396	0.0387	0.0259	0.0148	0.0263	0.0314	0.0205
±7.5%	# replications	1,000	1,000	1,000	400	1,000	1,000	1,000	400
	CI coverage	92%	88.8%	90.4%	89.5%	96.2%	94%	97%	94%
	Avg. sample size	296,158	278,642	388,000	287,568	438,392	403,844	598,020	382,958
	Avg. CI half-length	0.6369	0.6141	0.5866	0.6273	0.6349	0.6160	0.5950	0.6324
	Var. CI half-length	0.0013	0.0055	0.0072	0.0023	0.0012	0.0056	0.0056	0.0020
±3.75%	# replications	1,000	1,000	1,000	400	1,000	1,000	1,000	400
	CI coverage	92.1%	89.8%	94%	89.5%	96.7%	95.2%	97.7%	93.5%
	Avg. sample size	1,121,182	1,151,178	1,518,400	969,011	1,571,975	1,618,147	2,361,300	1,341,522
	Avg. CI half-length	0.3212	0.3081	0.3060	0.3200	0.3212	0.3076	0.3060	0.3210
	Var. CI half-length	0.0002	0.0014	0.0008	0.0004	0.0002	0.0014	0.0007	0.0004

ABATCH, the Law-Carson procedure, and the Heidelberger-Welch procedure are inferior to most of the results in Table 1. From Table 2 of Steiger and Wilson (2002) for example, ABATCH delivered the following coverage probabilities for nominal 90% CIs with the indicated relative precision levels: (i) no precision, 60%; (ii) ±15% precision, 72%; and (iii) ±7.5% precision, 82%. From Table 4 of Lada, Steiger, and Wilson (2006), the corresponding coverage probabilities for the Law-Carson procedure are 85%, 85%, and 87%, respectively. From Table 2 of Lada et al. (2007), the corresponding coverage probabilities for the Heidelberger-Welch procedure are 67.8%, 76%, and 77%.

The results in Table 1 suggest that as the precision level r^* becomes progressively smaller, Skart, ASAP3, and SBatch 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 Skart, ASAP3, or SBatch. Overall, Skart achieved reasonable conformance to the requested coverage probabilities at all precision levels.

5.2 *M/M/1/LIFO* Queue Waiting Times

The second test process included is the sequence of queue waiting times for the $M/M/1/\text{LIFO}$ queueing system with mean interarrival time of 1.0, mean service time of 0.8, and an empty-and-idle initial condition. Thus in steady-state operation this system has a server utilization of 0.8 and a mean queue waiting time $\mu_X = 3.20$. This test problem was selected mainly because in steady-state operation, batch means computed from the waiting times are highly skewed even for batch sizes that are sufficiently large to ensure the batch means are nearly uncorrelated. Table 2 summarizes the experimental performance of the queue waiting time process in the $M/M/1/\text{LIFO}$ queueing system. These results show that Skart has much better sampling efficiency compared with WASSP and SBatch, especially at the less-stringent precision levels. From Table 2 of Lada, Steiger, and Wilson (2006), the Law-Carson procedure delivered the following coverage probabilities for nominal 90% CIs: (i) no precision, 64%; (ii) ±15% precision, 76%; and (iii) ±7.5% precision, 84%. All in all, the we judged the performance of Skart to be superior to its competitors in this test problem.

6 CONCLUSIONS

Skart is a completely automated batch-means procedure for constructing an approximate CI for the steady-state mean of a simulation output process. Skart incorporates some advantages of its predecessors ASAP3, WASSP, and SBatch (such as the sampling efficiency of ASAP3 and the ability of WASSP and SBatch to eliminate initialization bias effectively) while exploiting separate adjustments to the classical batch-means confidence interval based on the corresponding effects of nonnormality and correlation of the delivered batch means. Our extensive performance evaluation of Skart indicates that it compares favorably with its predecessors as well as LBATCH, ABATCH, the Law-Carson procedure, and the Heidelberger-Welch procedure.

Table 2: Performance of Skart, WASSP, ASAP3, and SBatch in the $M/M/1$/LIFO queue waiting time process with 80% server utilization

Precision Requirement	Performance Measure	Nominal 90% CIs				Nominal 95% CIs			
		Skart	SBatch	WASSP	ASAP3	Skart	SBatch	WASSP	ASAP3
None	# replications	1,000	1,000	400	400	1,000	1,000	400	400
	CI coverage	87.5%	91.4%	93%	87%	93.5%	95.9%	96%	92.5%
	Avg. sample size	24,076	117,416	125,517	53,958	24,076	117,416	124,202	53,958
	Avg. CI half-length	0.5053	0.1891	0.2650	0.1060	0.649	0.2255	0.3350	0.3120
	Var. CI half-length	0.1310	0.0057	0.0230	0.1060	0.2230	0.0080	0.0310	0.0080
±15%	# replications	1,000	1,000	400	400	1,000	1,000	400	400
	CI coverage	92.6%	91.3%	90.7%	86.8%	95.3%	94%	95.2%	92.8%
	Avg. sample size	30,094	118,209	124,512	54,017	38,192	119,903	126,682	54,265
	Avg. CI half-length	0.3920	0.1812	0.2490	0.2600	0.4174	0.2117	0.2960	0.3080
	Var. CI half-length	0.0044	0.0021	0.0110	0.0040	0.0028	0.0022	0.0110	0.0050
±7.5%	# replications	1,000	1,000	400	400	1,000	1,000	400	400
	CI coverage	91.7%	89.5%	90.2%	87.5%	95.3%	95.4%	96.2%	92.5%
	Avg. sample size	83,781	126,961	152,355	68,325	122,285	134,123	194,590	90,911
	Avg. CI half-length	0.2225	0.1734	0.1860	0.2190	0.2235	0.1996	0.1990	0.2260
	Var. CI half-length	0.0003	0.0012	0.0020	0.0005	0.0002	0.0011	0.0010	0.0003

REFERENCES

Box, G. E. P., Jenkins, G. M., and G. C. Reinsel. 1994. *Time Series Analysis: Forecasting and Control*. 3rd ed. Englewood Cliffs, NJ: Prentice Hall, Inc.

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

Johnson, N. J. 1978. Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association* 73(363): 536–544.

Lada, E. K., N. M. Steiger, and J. R. Wilson. 2006. Performance evaluation of recent procedures for steady-state simulation analysis. *IIE Transactions* 38:711–727.

Lada, E. K., N. M. Steiger, and J. R. Wilson. 2008. SBatch: A spaced batch means procedure for steady-state simulation analysis. *Journal of Simulation* forthcoming. Available online via <www.ise.ncsu.edu/jwilson/files/lada08jos.pdf> [accessed May 8, 2008].

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., 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.

Law, A. M., and J. S. Carson. 1979. A sequential procedure for determining the length of a steady-state simulation. *Operations Research* 27(5): 1011–1025.

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.

Steiger, N. M., and J. R. Wilson. 2002. An improved batch means procedure for simulation output analysis. *Management Science* 48 (12): 1569–1586.

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.

Willink R. 2005. A confidence interval and test for the mean of an asymmetric distribution. *Communications in Statistics—Theory and Methods* 34: 753–766.

AUTHOR BIOGRAPHIES

ALI TAFAZZOLI is a Ph.D. Candidate at Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. He is a member of IIE and INFORMS. His e-mail address is <atafazz@ncsu.edu>.

JAMES R. WILSON is professor 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@ncsu.edu>, and his web page is <www.ise.ncsu.edu/jwilson>.

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>.

NATALIE M. STEIGER is an associate professor of production and operations management in the University of Maine Business School. She is a member of IIE and INFORMS. Her e-mail address is <nsteiger@maine.edu>.