Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

ASAP2: AN IMPROVED BATCH MEANS PROCEDURE FOR SIMULATION OUTPUT ANALYSIS

Natalie M. Steiger

Maine Business School
University of Maine
Orono, ME 04469-5723, U.S.A.

Emily K. Lada
James R. Wilson

Graduate Program in Operations Research
Department of Industrial Engineering
North Carolina State University
Raleigh, NC 27695-7906, U.S.A.

Christos Alexopoulos
David Goldsman

School of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332, U.S.A.

Faker Zouaoui

Sabre, Inc.
Research Group
Southlake, TX 76092, U.S.A.

ABSTRACT

We introduce ASAP2, an improved variant of the batch-means algorithm ASAP for steady-state simulation output analysis. ASAP2 operates as follows: the batch size is progressively increased until the batch means pass the Shapiro-Wilk test for multivariate normality; and then ASAP2 delivers a correlation-adjusted confidence interval. The latter adjustment is based on an inverted Cornish-Fisher expansion for the classical batch means t-ratio, where the terms of the expansion are estimated via a first-order autoregressive time series model of the batch means. ASAP2 is a sequential procedure designed to deliver a confidence interval that satisfies a prespecified absolute or relative precision requirement. When used in this way, ASAP2 compares favorably to ASAP and the well-known procedures ABATCH and LBATCH with respect to close conformance to the precision requirement as well as coverage probability and mean and variance of the half-length of the final confidence interval.

1 INTRODUCTION

In discrete-event simulation, we are often interested in estimating the steady-state mean μ_X of a stochastic output process $\{X_j : j \geq 1\}$ generated by a single, though long, simulation run. Assuming the target process is stationary and given a time series of length n from this process, we see that a natural estimator of μ_X is the sample mean, given by $\overline{X}(n) = n^{-1}\sum_{j=1}^{n} X_j$. We also require some indication of this estimator's precision; and typically a confidence interval (CI) for μ_X is constructed at a certain confidence level $1-\alpha$, where $0 < \alpha < 1$. The CI for μ_X should satisfy two criteria: (a) its actual coverage probability is close to the nominal level $1-\alpha$, and (b) it is narrow enough to be informative.

In the simulation analysis method of nonoverlapping batch means (NOBM), the sequence of simulation-generated outputs $\{X_j : j = 1, \ldots, n\}$ is divided into k adjacent nonoverlapping batches, each of size m. For simplicity, we assume that n is a multiple of m so that $n = km$. The sample mean for the jth batch is

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

and the grand mean of the individual batch means,

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

is used as an estimator for μ_X (note that $\overline{Y}(m,k) = \overline{X}(n)$). We construct a CI centered on an estimator like (1), where in practice we may exclude some initial batches to eliminate the effects of initialization bias.

If the batch size m is sufficiently large so that the batch means $\{Y_j(m) : 1 \leq j \leq k\}$ are approximately independent and identically distributed (i.i.d.) normal random variables with mean μ_X, then we can apply a classical result from statistics (see, for example, Steiger and Wilson 1999, 2000,

2001) to compute a confidence interval for μ_X from the batch means. The sample variance of the k batch means for batches of size m is

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

If the original simulation-generated process $\{X_j : j = 1,\ldots,n\}$ is stationary and weakly dependent as specified, for example, in Theorem 1 of Steiger and Wilson (2001), then it follows that as $m \to \infty$ with k fixed so that $n \to \infty$, an asymptotically valid $100(1-\alpha)\%$ confidence interval for μ_X is

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

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

Sequential NOBM procedures address the problem of determining the batch size, m, and the number of batches, k, that are required to satisfy approximately the assumptions of independence and normality of the batch means. If these assumptions are exactly satisfied, then we will obtain CIs whose actual coverage probability is exactly equal to the nominal coverage probability. In this paper we introduce ASAP2, an improved variant of the ASAP algorithm (Steiger and Wilson 1999, 2000, 2002) for analysis of steady-state simulation output; and we compare the performance of ASAP2 versus the original ASAP algorithm as well as the widely used NOBM procedures ABATCH and LBATCH (Fishman 1996; Fishman and Yarberry 1997; Fishman 1998).

The rest of this paper is organized as follows. A brief overview of ASAP2 is given in §2, and a detailed explanation of the steps of ASAP2 is given in §3. Some of the results of our performance evaluation of ASAP2 are presented in §4. Finally in §5 we summarize the main findings of this work.

2 OVERVIEW OF ASAP2

ASAP2 requires the following user-supplied inputs:

1. a simulation-generated output process $\{X_j : j = 1,2,\ldots,n\}$ from which the steady-state expected response μ_X is to be estimated;
2. a confidence coefficient α specifying that the desired confidence-interval coverage probability is $1-\alpha$; and
3. an absolute or relative precision requirement specifying the final confidence-interval half-length in terms of (a) a maximum absolute half-length H^*, or (b) a maximum relative fraction r^* of the magnitude of the final grand mean $\overline{Y}$.

ASAP2 delivers the following outputs:

1. a nominal $100(1-\alpha)\%$ confidence interval for μ_X having the form

$$\overline{Y} \pm H \quad \text{where} \quad H \le H^* \quad \text{or} \quad H \le r^*|\overline{Y}|, \qquad (3)$$

provided no additional simulation-generated observations are required; or
2. a new total sample size n to be supplied to the algorithm.

If additional observations of the target process must be generated by the user's simulation model before a confidence interval with the required precision can be delivered, then ASAP2 must be called again with the additional data; and this cycle of simulation followed by analysis may be repeated several times before ASAP2 finally delivers a confidence interval.

On each iteration of ASAP2, the algorithm operates as follows. The simulation outputs are divided initially into a fixed number of batches (namely, $k = 256$ batches); and batch means are computed. The first four batches are discarded, and from the remaining $k' = k - 4 = 252$ batch means, we select every other group of four adjacent batch means to form a sample of 32 four-dimensional vectors of batch means that will be tested for joint multivariate normality. If the normality test is failed, then the batch size m is increased by a factor of $\sqrt{2}$ and the process is repeated until the normality test is passed.

Upon acceptance of the hypothesis of joint multivariate normality of the batch means, a CI is constructed—specifically, the correlation-adjusted CI (5) below based on k' batch means for batches of size m. The correlation correction uses an inverted Cornish-Fisher expansion (Stuart and Ord 1994) for the classical NOBM Student t-ratio

$$t = \frac{\overline{Y}(m,k') - \mu_X}{S_{m,k'}/\sqrt{k'}}; \qquad (4)$$

and the terms of this expansion are estimated by fitting an order-one autoregressive time-series model (Box, Jenkins and Reinsel 1994) to the set of k' retained batch means. Based on this approach, a correlation-adjusted $100(1-\alpha)\%$ confidence interval for μ_X is

$$\overline{Y}(m,k') \pm \left[\left(1 + \frac{\hat{\kappa}_2 - 1}{2} - \frac{\hat{\kappa}_4}{8}\right) z_{1-\alpha/2} + \frac{\hat{\kappa}_4}{24} z^3_{1-\alpha/2}\right] \times \sqrt{\frac{\widehat{\mathrm{Var}}[Y(m)]}{k'}}, \qquad (5)$$

where: $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution; $\hat{\kappa}_2$ and $\hat{\kappa}_4$ respectively denote estimators of the second and fourth cumulants of the t-

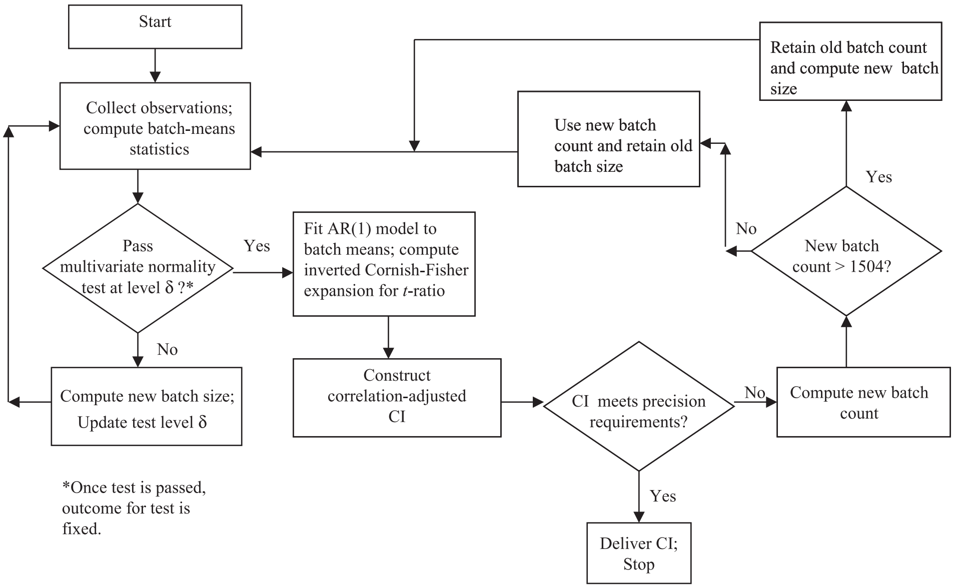

Figure 1: Flow Chart of ASAP2

ratio (4); $\widehat{\text{Var}}[Y(m)]$ denotes an estimator of the variance of the batch means; and all these statistics are based on fitting an order-one autoregressive time-series model to the (correlated) batch means process $\{Y_j(m) : j = 5, \ldots, k\}$ as detailed in §§3.2–3.3 below.

Subsequent iterations of ASAP2 that are performed to satisfy the user-specified precision requirement (if there is one) do not repeat testing for multivariate normality of the overall set of batch means. These subsequent iterations require additional sampling, computing the additional batch means, and reconstructing the CI, again discarding the first four batches of the overall data set (consisting of all original observations plus any additional observations required by ASAP2). Successive iterations of ASAP2 continue until the precision requirement is met.

A flow chart of ASAP2 is depicted in Figure 1. In the next section we provide complete details on the main steps in the operation of ASAP2.

3 DETAILED STEPS OF ASAP2

3.1 Testing Batch Means for Normality

ASAP2 begins on iteration 1 with a user-specified initial batch size m_1 (by default $m_1 = 16$), requiring data for $k_1 = 256$ initial batches. The results of extensive experimentation show that ASAP2 performs well with this initial batch size and batch count, even for processes that are highly dependent or whose marginal distributions exhibit marked departures from normality. While a total of $n_1 = k_1 m_1 = 4{,}096$ observations may exceed the user's precision requirement or computing budget in some applications, such an initial sample size is usually easy and inexpensive to generate.

On each iteration of ASAP2 that requires a normality test, the batch means are organized into groups of four adjacent quantities so that every other group can be tested for four-dimensional normality. Such an approach is tantamount to assuming that when the batch size is sufficiently large so that the batch means pass the test for multivariate normality, only dependence between batch means out to lag three is practically significant.

To address the start-up problem, we exclude the first group of four batches from the computation of overall statistics for the batch means; and in each normality test, we take a *spacer* (Fox, Goldsman, and Swain 1991) consisting of a group of four ignored batch means between each group that is to be tested for normality. Our computational experience with ASAP2 in a wide variety of applications has suggested that if the batch size is large enough for the spacer-separated four-dimensional vectors of adjacent batch means to pass the normality test, then the corresponding spacer also provides a reasonable start-up period (statistics clearing time) for eliminating the effects of initialization bias. Let $k_1' = k_1 - 4 = 252$ denote the initial number of batch means retained for confidence-interval construction from which we calculate the sample mean and variance

$$\overline{Y}(m_1, k_1') = \frac{1}{k_1'} \sum_{j=5}^{k_1} Y_j(m_1)$$

and
$$S^2_{m_1,k_1'} = \frac{1}{k_1' - 1} \sum_{j=5}^{k_1} \left[Y_j(m_1) - \overline{Y}(m_1, k_1') \right]^2,$$

respectively. (To simplify the subsequent notation, throughout the rest of this paper we define aggregate batch statistics like $\overline{Y}(m_1, k_1')$ and $S^2_{m_1,k_1'}$ to exclude the first four batches from the entire data set accumulated so far.)

The k_1' retained batch means $\{Y_j(m_1) : j = 5, \ldots, k_1\}$ are tested for multivariate normality by constructing 32 four-dimensional vectors $\{\mathbf{y}_\ell : \ell = 1, \ldots, 32\}$ as depicted in the following layout:

$$\left.\begin{array}{l}
\underbrace{Y_5(m_1), Y_6(m_1), Y_7(m_1), Y_8(m_1)}_{\text{1st } (4\times1) \text{ vector } \mathbf{y}_1},\\
\qquad\underbrace{Y_9(m_1), Y_{10}(m_1), Y_{11}(m_1), Y_{12}(m_1)}_{\text{ignored spacer}},\\
\underbrace{Y_{13}(m_1), Y_{14}(m_1), Y_{15}(m_1), Y_{16}(m_1)}_{\text{2nd } (4\times1) \text{ vector } \mathbf{y}_2},\\
\qquad\underbrace{Y_{17}(m_1), Y_{18}(m_1), Y_{19}(m_1), Y_{20}(m_1)}_{\text{ignored spacer}},\\
\qquad\cdots\\
\underbrace{Y_{253}(m_1), Y_{254}(m_1), Y_{255}(m_1), Y_{256}(m_1)}_{\text{32nd } (4\times1) \text{ vector } \mathbf{y}_{32}}.
\end{array}\right\} \qquad (6)$$

We apply the Shapiro-Wilk test for multivariate normality (Malkovich and Afifi 1973) to the resulting data set $\{\mathbf{y}_\ell : \ell = 1, \ldots, 32\}$. Although normality of each four-dimensional random vector $\mathbf{y}_\ell$ is not sufficient to ensure joint normality of all $k_1' = 252$ batch means (Stuart and Ord 1994, Exercise 15.20), our computational experience strongly suggests that this approach to testing for joint normality of the batch means has sufficient power to be effective in practical applications of ASAP and ASAP2 (Steiger 1999; Steiger and Wilson 1999, 2000, 2002).

Given a random sample $\{\mathbf{y}_\ell : \ell = 1, \ldots, g\}$ of q-dimensional response vectors, we perform the Shapiro-Wilk test for multivariate normality as follows. First we compute the sample statistics

$$\overline{\mathbf{y}} = g^{-1}\sum_{\ell=1}^{g}\mathbf{y}_\ell \quad\text{and}\quad \mathbf{A} = \sum_{\ell=1}^{g}(\mathbf{y}_\ell - \overline{\mathbf{y}})(\mathbf{y}_\ell - \overline{\mathbf{y}})^{\mathrm{T}}.$$

Throughout the rest of this section, we assume that the random matrix $\mathbf{A}$ is nonsingular with probability one. This property can be ensured, for example, by a mild technical requirement detailed by Tew and Wilson (1992, p. 91), provided the replication count $g > q$; and since we take $g = 32$ and $q = 4$ in ASAP2, with probability one we can identify the observation $\mathbf{y}^\dagger \in \{\mathbf{y}_\ell : \ell = 1, 2, \ldots, g\}$ for which

$$(\mathbf{y}^\dagger - \overline{\mathbf{y}})^{\mathrm{T}}\mathbf{A}^{-1}(\mathbf{y}^\dagger - \overline{\mathbf{y}}) = \max_{\ell=1,\ldots,g}\Big\{(\mathbf{y}_\ell - \overline{\mathbf{y}})^{\mathrm{T}}\mathbf{A}^{-1}(\mathbf{y}_\ell - \overline{\mathbf{y}})\Big\}.$$

We compute $Z_\ell \equiv (\mathbf{y}^\dagger - \overline{\mathbf{y}})^{\mathrm{T}}\mathbf{A}^{-1}(\mathbf{y}_\ell - \overline{\mathbf{y}})$ for $\ell = 1, 2, \ldots, g$; and we sort these auxiliary quantities in ascending order to obtain the corresponding order statistics $Z_{(1)} < Z_{(2)} < \cdots < Z_{(g)}$. Let $\{\beta_\ell : \ell = 1, 2, \ldots, g\}$ denote the associated coefficients of the univariate Shapiro-Wilk statistic for a random sample of size g (see Royston 1982a, 1982b). The null hypothesis of multinormal responses $\{\mathbf{y}_\ell\}$ is rejected at the level of significance δ $(0 < \delta < 1)$ if the multivariate Shapiro-Wilk statistic,

$$W^* = \frac{\left[\sum_{\ell=1}^{g}\beta_\ell Z_{(\ell)}\right]^2}{(\mathbf{y}^\dagger - \overline{\mathbf{y}})^{\mathrm{T}}\mathbf{A}^{-1}(\mathbf{y}^\dagger - \overline{\mathbf{y}})}, \qquad (7)$$

satisfies $W^* < w^*_\delta(q, g)$, the $100(1-\delta)\%$ quantile of the null distribution of W^*. (The null distribution of W^* is the c.d.f. $F_{W^*}(\cdot)$ of (7) when this statistic is based on a random sample of size g taken from a q-dimensional nonsingular normal distribution.)

On the ith iteration of ASAP2 for $i = 1, 2, \ldots,$ we let k_i and m_i respectively denote the batch count and the batch size. An additional iteration of ASAP2 will be required if the multivariate Shapiro-Wilk test yields a significant result (that is, the 32 four-dimensional vectors of batch means (6) fail the multivariate normality test) at the level of significance δ_i, where

$$\delta_i = \delta_1 \exp\Big[-\omega(i-1)^2\Big] \quad\text{for}\quad i = 1, 2, \ldots, \qquad (8)$$

with $\delta_1 = 0.10$ and $\omega = 0.18421$. If the test statistic W_i^* computed from (7) on iteration i corresponds to a P-value $F_{W^*}(W_i^*) < \delta_i$, then on iteration $i+1$ the batch size and batch count are respectively taken to be

$$m_{i+1} = \left\lfloor \sqrt{2}m_i \right\rfloor \text{ and } k_{i+1} = k_i \qquad (9)$$

so that the total required sample size is $n_{i+1} = m_{i+1}k_{i+1}$; and thus the user must provide the additional simulation responses $\big\{X_j : j = n_i + 1, n_i + 2, \ldots, n_{i+1}\big\}$ before executing iteration $i+1$ of ASAP2.

The scheme (6)–(9) is specifically designed so that ASAP2 avoids the excessive variability in the final sample size and confidence-interval half-length that we have sometimes observed with ASAP. Display (8) implies that for $i = 1, 2, \ldots, 6$, the significance level δ_i for the multivariate normality test has the following values: 0.10, 0.083, 0.048, 0.019, 0.0052, and 0.001; and on each iteration i beyond the sixth, δ_i declines by at least an order of magnitude.

3.2 Building an AR(1) Model for Dependent Normal Batch Means

If the batch means pass the test for joint multivariate normality, then we seek to adjust the classical batch means CI

(2) by taking into account the deviation of the distribution of the classical NOBM t-ratio (4) from the desired Student's t-distribution with $k'-1$ degrees of freedom. Our adjustment is based on an inverted Cornish-Fisher expansion for (4) that involves the first four cumulants of (4). In the next section, we develop expressions for the first four cumulants of (4) in terms of $\text{Var}[Y(m)]$ and $\text{Var}\big[\,\overline{Y}(m,k')\,\big]$. To compute sample estimators of $\text{Var}[Y(m)]$ and $\text{Var}\big[\,\overline{Y}(m,k')\,\big]$, we fit an order-one autoregressive (that is, AR(1)) time series model (Box, Jenkins, and Reinsel 1994) to the sequence of batch means $\{Y_j(m) : j = 5, \ldots, k\}$. (In this section we suppress the index i of the current iteration of ASAP2 to simplify the notation; no confusion can result from this simplification since the iteration index remains the same throughout the discussion.) For the batch means variance estimator $\widehat{\text{Var}}[Y(m)]$, we take the usual maximum likelihood estimator of the variance of the fitted AR(1) process (see Chapter 7 of Box, Jenkins, and Reinsel 1994); and for the grand mean variance estimator $\widehat{\text{Var}}\big[\,\overline{Y}(m,k')\,\big]$, we derive a similar statistic based on the estimated covariances between all relevant batch means expressed in terms of the maximum likelihood estimators of the parameters of the fitted AR(1) process.

If the batch means pass the test for multivariate normality detailed in §3.1, then an AR(1) process is fitted to the set of $k' = 252$ batch means. This is based on all our previous computational experience with the original ASAP algorithm (Steiger 1999; Steiger and Wilson 1999, 2000, 2002). Generally, however, identification and estimation of autoregressive–time series models should be based on at least 50 and preferably 100 or more observations (Box, Jenkins, and Reinsel 1994, p. 17); and this is one of the reasons that ASAP2 requires an initial batch count of 256. Adapting the notation in Box, Jenkins, and Reinsel (1994) to the notation used here, we let $\{\tilde{Y}_{j-4} \equiv Y_j(m) - \mu_X : j = 5, \ldots, k\}$ denote the corresponding deviations from the steady-state mean μ_X. The ℓth observation of an AR(1) process can be expressed as

$$\tilde{Y}_\ell = \varphi_1 \tilde{Y}_{\ell-1} + a_\ell \quad \text{for} \quad \ell = 1, 2, \ldots, \tag{10}$$

where φ_1 is the autoregressive parameter and a_ℓ is an independent normal "shock" with mean zero and variance σ_a^2.

The estimators of $\text{Var}[Y(m)] = \text{Var}[\tilde{Y}_\ell]$ and the other parameters of the AR(1) model (10) are then used to estimate $\text{Var}\big[\,\overline{Y}(m,k')\,\big]$:

$$\widehat{\text{Var}}\big[\,\overline{Y}(m,k')\,\big] = \frac{1}{k'} \sum_{q=-k'+1}^{k'-1} \left(1 - \frac{|q|}{k'}\right) \widehat{\gamma}_m(q), \tag{11}$$

where $\widehat{\gamma}_m(q)$ denotes the estimated lag-q covariance of the batch means $\{Y_j(m) : j = 5, \ldots, k\}$ based on the fitted time series model. For an AR(1) process (10), the covariance at lag q is given by $\text{Cov}\big[\tilde{Y}_\ell, \tilde{Y}_{\ell+q}\big] = \varphi_1^{|q|}\sigma_a^2 \big/ \big(1 - \varphi_1^2\big)$, for $q = 0, \pm1, \pm2, \ldots$. Thus if (10) is an adequate model of the batch means process for batches of size m and if $\widehat{\varphi}_1$ and $\widehat{\sigma}_a$ denote the usual maximum likelihood estimates of φ_1 and σ_a respectively (Box, Jenkins, and Reinsel 1994, Chapter 7), then the estimated covariances in (11) are

$$\widehat{\gamma}_m(q) = \frac{\widehat{\varphi}_1^{|q|}}{1 - \widehat{\varphi}_1^2}\widehat{\sigma}_a^2 \quad \text{for} \quad q = 0, \pm1, \pm2, \ldots. \tag{12}$$

See Steiger (1999) for complete details on the time series estimation techniques used in ASAP2.

3.3 Confidence Interval for Dependent Batch Means

In this section, we formulate an adjustment to the usual CI (2) that accounts for dependency between the batch means. The adjustment is based on the first four cumulants of the usual t-ratio (4) on which the classical confidence interval (2) is built. To simplify the discussion, we let

$$N \equiv \frac{\sqrt{k'}\big[\,\overline{Y}(m,k') - \mu_X\,\big]}{\sqrt{\text{Var}[\,Y(m)\,]}}, \quad D \equiv \sqrt{\frac{S_{m,k'}^2}{\text{Var}[\,Y(m)]}} \tag{13}$$

respectively denote the numerator and denominator of the t-ratio (4) based on k' batch means for batches of size m. To compute the moments of (4), we make the following key assumptions.

- A_1: The batch means have a joint multivariate normal distribution.
- A_2: As defined in (13), the numerator N and denominator D of the t-ratio (4) are independent.
- A_3: The squared denominator D^2 of the t-ratio (4) is distributed as $\chi^2_{k'-1}/(k'-1)$.

Note that if A_1 holds and the batch means are independent, then A_2 and A_3 follow immediately. The basis for A_1 is ASAP2's test for multivariate normality; moreover, some theoretical and experimental evidence for the reasonableness of A_2 and A_3 can be found, respectively, in equation (19) and in Figures 9–10 of Steiger and Wilson (2001).

Exploiting assumptions A_1–A_3, we first derive expressions for the first four cumulants κ_1, κ_2, κ_3, and κ_4 of the NOBM t-ratio (4). From A_1–A_3 it follows that

$$\kappa_p = 0 \text{ for } p = 1, 3 \text{ and } k' \geq 5, \tag{14}$$

$$\kappa_2 = \frac{k'(k'-1)\text{Var}\big[\,\overline{Y}(m,k')\,\big]}{(k'-3)\text{Var}[Y(m)]} \text{ for } k' \geq 4, \tag{15}$$

$$\kappa_4 = \frac{2(k')^2(k'-1)^2\text{Var}^2\big[\,\overline{Y}(m,k')\,\big]}{(k'-3)^2(k'-5)\text{Var}^2[Y(m)]} \text{ for } k' \geq 6. \tag{16}$$

See Steiger (1999) or Steiger and Wilson (2002) for a detailed justification of (14)–(16). In terms of these cumulants, we obtain the following adjusted $100(1-\alpha)\%$ confidence interval for μ_X:

$$\overline{Y}(m,k') \pm h'(z_{1-\alpha/2})\frac{S_{m,k'}}{\sqrt{k'}},$$

where
$$h'(z_{1-\alpha/2}) = \left(\kappa_1 - \frac{\kappa_3}{6}\right) + \left(1 + \frac{\kappa_2 - 1}{2} - \frac{\kappa_4}{8} + \frac{5\kappa_3^2}{36}\right) z_{1-\alpha/2} + \frac{\kappa_3}{6} z_{1-\alpha/2}^2 + \left(\frac{\kappa_4}{24} - \frac{\kappa_3^2}{18}\right) z_{1-\alpha/2}^3. \quad (17)$$

The result (17) is obtained from the inverted Cornish-Fisher expansion (6.56) of Stuart and Ord (1994) based on a standard normal density.

Exploiting our approximations for the first four cumulants of the t-ratio (4) based on (14)–(16), we compute the final confidence interval delivered by ASAP2 as follows. In the expressions (15) and (16) for κ_2 and κ_4, we replace the quantities $\text{Var}[Y(m)]$ and $\text{Var}\big[\,\overline{Y}(m,k')\,\big]$ by the corresponding variance estimators $\widehat{\text{Var}}[Y(m)] = \widehat{\gamma}_m(0)$ and $\widehat{\text{Var}}\big[\,\overline{Y}(m,k')\,\big]$ that are respectively obtained from relations (12) and (11) by fitting an AR(1) process to the batch means $\{Y_j(m) : j = 5,\ldots,k\}$; and this procedure yields the approximate $100(1-\alpha)\%$ confidence interval (5) for μ_X.

3.4 Fulfilling the Precision Requirement

The final step in ASAP2 is to determine if the constructed confidence interval satisfies the user's precision requirement. The confidence interval is based on a nonsignificant result from the multivariate normality test (that is, the batch means pass the test for multivariate normality). If the relevant precision requirement

$$H \le H^* \quad \text{or} \quad H \le r^*|\overline{Y}| \quad (18)$$

is satisfied, then ASAP2 terminates, returning a confidence interval with midpoint $\overline{Y}$ and half-length H. If the precision requirement (18) is not satisfied on iteration i of ASAP2, then the procedure estimates the number of additional batches k_i^+ required to satisfy (18) using batch size m_i,

$$k_i^+ = \left\lceil \left(\frac{H}{H^*}\right)^2 k_i' \right\rceil - k_i'. \quad (19)$$

To simplify the operation of ASAP2, we specified an upper limit on the number of batches that the algorithm may require. Preliminary experiments with ASAP2 revealed that no substantial improvements in the performance of the procedure could be achieved by setting the upper limit on the batch count much above 1,500. If the projected total number of batches $k_i + k_i^+$ exceeds 1,504, then a new batch size is calculated and the batch count remains fixed so that

$$\text{If } k_i + k_i^+ > 1{,}504, \quad \text{then} \quad m_{i+1} = \lfloor (H/H^*)\, m_i \rfloor \text{ and } k_{i+1} = k_i.$$

If, however, the projected total number of batches does not exceed 1,504, then a new batch count is calculated and the batch size remains fixed so that

$$\text{If } k_i + k_i^+ \le 1{,}504, \quad \text{then} \quad k_{i+1} = k_i + k_i^+ \text{ and } m_{i+1} = m_i.$$

Thus if the user-specified precision requirement (18) is not satisfied on iteration i of ASAP2, then iteration $i+1$ will be required in which the number of batches $k_{i+1} \le 1{,}504$ and the total sample size is finally taken to be

$$n_{i+1} = m_{i+1} k_{i+1};$$

and the user must provide the additional simulation responses $\{X_j : j = n_i{+}1, n_i{+}2, \ldots, n_{i+1}\}$ before executing iteration $i+1$ of ASAP2.

The user then performs iteration $i+1$ of ASAP2 with the values of m_{i+1}, k_{i+1}, and n_{i+1} for the batch size, batch count, and total sample size, respectively. The first four batches of the entire simulation-generated data set are again omitted from the calculation of the overall sample mean. The batch means for $k'_{i+1} = k_{i+1} - 4$ batches of size m_{i+1} are computed. Then an updated AR(1) fit is made using k'_{i+1} batches of size m_{i+1}; moreover, in this situation new estimates of $\text{Var}[Y(m_{i+1})]$, $\text{Var}\big[\,\overline{Y}(m_{i+1}, k'_{i+1})\,\big]$, κ_2, and κ_4 are computed; and the updated CI (5) is constructed. If the precision requirement (18) is satisfied on iteration $i+1$ of ASAP2, then the algorithm terminates, returning a confidence interval with midpoint $\overline{Y} = \overline{Y}(m_{i+1}, k'_{i+1})$ and the associated half-length H. If the required precision is not achieved on iteration $i+1$ of ASAP2, then the rest of iteration $i+1$ of ASAP2 proceeds along the same lines as described above.

4 PERFORMANCE EVALUATION FOR SELECTED NOBM PROCEDURES

To evaluate the performance of ASAP2 with respect to the coverage probability of its confidence intervals, the mean and variance of the half-length of its confidence intervals, and its total sample size, we applied ASAP2 together with the ABATCH, LBATCH, and ASAP algorithms (Fishman 1996; Fishman and Yarberry 1997; Steiger and Wilson 1999, 2000, 2002) to the queue waiting time process for the $M/M/1$

queue with server utilization of 0.9 and an empty-and-idle initial condition. This is a particularly difficult test problem for several reasons: (a) the magnitude of the initialization bias is substantial and decays relatively slowly; (b) in steady-state operation the autocorrelation function of the waiting time process decays very slowly with increasing lags; and (c) 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 independent and identically normally distributed. This test problem most dramatically displays one of the advantages of the ASAP2 algorithm—namely, that ASAP2 does not rely on any test for independence of the batch means.

The steady-state mean response is available analytically for this test problem; thus we were able to evaluate the performance of ASAP, ASAP2, ABATCH, and LBATCH in terms of actual versus nominal coverage probabilities for the confidence intervals delivered by each of these procedures.

We performed 400 independent replications of each batch means procedure to construct nominal 90% confidence intervals that satisfy three different precision requirements:

(1) no precision requirement—that is, we continued the simulation of each test problem until ASAP2 delivered a confidence interval (5) based on 256 batches of the size at which the batch means passed the statistical test (7) for multivariate normality without considering a precision requirement;
(2) $\pm 15\%$ precision—that is, we continued the simulation of each test problem until ASAP2 delivered a confidence interval (3) that satisfied the relative precision requirement (18) with $r^* = 0.15$; and
(3) $\pm 7.5\%$ precision—that is, we continued the simulation of each test problem until ASAP2 delivered a confidence interval (3) that satisfied the relative precision requirement (18) with $r^* = 0.075$.

In addition to the experimentation using the ASAP2 algorithm, we performed 400 independent replications of the original ASAP algorithm under the same precision requirements described above. However, for case (a) above (that is, no precision requirement), we continued the simulation of each test problem until ASAP delivered a confidence interval based on 96 batches of the size at which the batch means passed either the statistical test for independence or for multivariate normality, as prescribed by the original ASAP algorithm (Steiger 1999; Steiger and Wilson 2000, 2002). Since ABATCH and LBATCH do not explicitly determine a sample size, we passed to the ABATCH and LBATCH algorithms the same data sets used by ASAP2. Based on all our computational experience with ASAP and ASAP2, we believe that the results given below are typical of the performance of ASAP and ASAP2 that can be expected in many practical applications. On the other hand, ABATCH and LBATCH are nonsequential procedures whose proper operation may require direct interaction with the user (Fishman 1998); and thus it is not clear that the following results exemplify the performance of ABATCH and LBATCH in practical applications. Nevertheless, we believe that the results given below provide some basis for comparing the performance of ABATCH, LBATCH, ASAP, and ASAP2. Since each confidence interval with a nominal coverage probability of 90% was replicated 400 times, the standard error of each coverage estimator is approximately 1.5%. As explained below, this level of precision in the estimation of coverage probabilities turns out to be sufficient to reveal significant differences in the performance of ASAP2 versus ASAP, ABATCH, and LBATCH in the test problem presented here.

As can be seen from Table 1, ASAP2 outperforms ABATCH and LBATCH with respect to confidence interval coverage for all three precision requirements. As we demand more precision, we are of course forced to perform more sampling. The results in Table 1 suggest that ABATCH and LBATCH will give satisfactory coverage if these procedures are supplied with an adequate amount of data; however, ABATCH and LBATCH provide no mechanism for determining the amount of data that should be used. A desirable feature of ASAP2 is that it usually determines a sample size sufficient to yield acceptable results. It should be recognized, however, that ASAP2 was designed for use with a user-specified precision requirement; and in the absence of a precision requirement, ASAP2-generated confidence intervals can be highly variable in their half-lengths. If a user-specified precision requirement is imposed after using ASAP2 to generate an initial or "pilot" confidence interval without a precision requirement, then in our computational experience, the resulting follow-up confidence interval will exhibit the same stability that we have observed in all our other applications of ASAP2 with a user-specified (nonvacuous) precision requirement.

By comparing the performance of ASAP2 versus the performance of the original ASAP algorithm given in the rightmost column in Table 1, we see that ASAP2 frequently outperforms ASAP in small-sample applications.

5 CONCLUSIONS

Building on the promising results we obtained with ASAP in a broad diversity of stochastic systems, we have introduced ASAP2 as an improved batch-means procedure for steady-state simulation output analysis. Our extensive experimental performance evaluation of ASAP2 indicates that it outperforms LBATCH and ABATCH in virtually all the stochastic systems to which we have applied all three procedures. Moreover, built into ASAP2 is an enhanced test

Table 1: Performance of Batch-Means Procedures for the $M/M/1$ Queue Waiting Time Process with Traffic Intensity $\tau = 0.9$ Based on 400 Independent Replications of Nominal 90% Confidence Intervals

Precision Requirement	Procedure LBATCH	ABATCH	ASAP2	ASAP
NO PRECISION				
avg. sample size			22554	5873
coverage	65%	72%	88%	81%
avg. rel. precision	0.175	0.209	0.579	1.31
avg. CI half-length	1.660	1.980	6.440	18.2
var. CI half-length	1.500	1.716	167.0	3506
±15% PRECISION				
avg. sample size			93374	117856
coverage	75%	81%	90%	93%
avg. rel. precision	0.089	0.103	0.135	0.13
avg. CI half-length	0.788	0.912	1.184	1.19
var. CI half-length	0.026	0.028	0.025	0.022
±7.5% PRECISION				
avg. sample size			281022	321468
coverage	80%	85%	92%	93%
avg. rel. precision	0.057	0.063	0.070	0.069
avg. CI half-length	0.511	0.563	0.628	0.62
var. CI half-length	0.004	0.005	0.002	0.003

for normality of the batch means; and this feature should enable ASAP2 to avoid the anomalous behavior of ASAP that we have observed in some stochastic systems wherein significant departures from normality of the batch means are observed even for batch sizes sufficiently large to ensure negligible dependence between the batch means. Although such situations occur relatively infrequently in our computational experience, they should be properly handled by a comprehensive, general-purpose batch-means algorithm for steady-state simulation output analysis. Based on our preliminary experimentation, ASAP2 appears to have this property. We are continuing the refinement and experimental evaluation of ASAP2; and future developments concerning ASAP2, including technical reports, papers submitted to archival journals, software, and corrections will be available on the websites <www.isye.gatech.edu/˜christos> and <www.isye.gatech.edu/˜sman> as well as <www.ie.ncsu.edu/jwilson>.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under grant number DMI-9900164.

REFERENCES

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time series analysis: Forecasting and control*. 3d ed. Englewood Cliffs, New Jersey: Prentice Hall.

Fishman, G. S. 1996. *Monte Carlo: Concepts, algorithms, and applications*. New York: Springer-Verlag.

Fishman, G. S. 1998. LABATCH.2: Software for statistical analysis of simulation sample path data. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 131–139. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available via <www.informs-cs.org/wsc98papers/017.PDF> [accessed June 25, 2000].

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

Malkovich, J. F., and A. A. Afifi. 1973. On tests for multivariate normality. *Journal of the American Statistical Association* 68:176–179.

Royston, J. P. 1982a. An extension of Shapiro and Wilk's *W* test for normality to large samples. *Applied Statistics* 31 (2): 115–124.

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

Steiger, N. M. 1999. Improved batching for confidence interval construction in steady-state simulation. Doctoral dissertation, Department of Industrial Engineering, North Carolina State University, Raleigh, North Carolina. Available from <www.lib.ncsu.edu/etd> [full web address is <http://www.lib.ncsu.edu/etd/public/

etd-19231992992670/etd.pdf>; accessed June 18, 2000].

Steiger, N. M., and J. R. Wilson. 1999. Improved batching for confidence interval construction in steady-state simulation. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 442–451. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available from <www.informs-cs.org> [full web address is <www.informs-cs.org/wsc99papers/061.PDF>; accessed March 21, 2000].

Steiger, N. M., and J. R. Wilson. 2000. Experimental performance evaluation of batch-means procedures for simulation output analysis. In *Proceedings of the 2000 Winter Simulation Conference*, ed. J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, 627–636. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available from <www.informs-cs.org> [full web address is <www.informs-cs.org/wsc00papers/084.PDF>; accessed January 6, 2001].

Steiger, N. M., and J. R. Wilson. 2001. Convergence properties of the batch means method for simulation output analysis. *INFORMS Journal on Computing* 13 (4): 277–293.

Steiger, N. M., and J. R. Wilson. 2002. An improved batch means procedure for simulation output analysis. *Management Science* in review. Available via <ftp.ncsu.edu/pub/eos/pub/jwilson> [full web address is <ftp://ftp.ncsu.edu/pub/eos/pub/jwilson/asaporv12.pdf>; accessed January 16, 2002].

Stuart, A. and J. K. Ord. 1994. *Kendall's advanced theory of statistics, volume 1: Distribution theory*. 6th ed. London: Edward Arnold.

Tew, J. D., and J. R. Wilson. 1992. Validation of simulation analysis methods for the Schruben-Margolin correlation-induction strategy. *Operations Research* 40 (1): 87–103.

AUTHOR BIOGRAPHIES

NATALIE M. STEIGER is an Assistant Professor of Production and Operations Management in the University of Maine Business School. Her e-mail address is <nsteiger@maine.edu>.

EMILY K. LADA is a Ph.D. student in the Operations Research Graduate Program at North Carolina State University. Her e-mail address is <eklada@eos.ncsu.edu>, and her web page is <www4.ncsu.edu/~eklada/>.

JAMES R. WILSON is Professor and Head of the Department of Industrial Engineering at North Carolina State University. His e-mail address is <jwilson@eos.ncsu.edu>, and his web page is <www.ie.ncsu.edu/jwilson>.

CHRISTOS ALEXOPOULOS is an Associate Professor in the School of Industrial & Systems Engineering at Georgia Tech. His e-mail address is <christos@isye.gatech.edu>, and his web page is <www.isye.gatech.edu/~christos>.

DAVID GOLDSMAN is a Professor in the School of Industrial & Systems Engineering at Georgia Tech. His e-mail address is <sman@isye.gatech.edu>, and his web page is <www.isye.gatech.edu/~sman>.

FAKER ZOUAOUI is an operations research consultant in the Research Group at Sabre, Inc. His e-mail address is <faker.zouaoui@sabre.com>.