Proceedings of the 1999 Winter Simulation Conference
P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds.

IMPROVED BATCHING FOR CONFIDENCE INTERVAL CONSTRUCTION IN STEADY-STATE SIMULATION

Natalie M. Steiger

Department of Information Systems
and Operations Management
University of North Carolina at Greensboro
Greensboro, NC 27402-6165, U.S.A.

James R. Wilson

Department of Industrial Engineering
North Carolina State University
2401 Stinson Drive
Raleigh, NC 27695-7906, U.S.A.

ABSTRACT

We describe an improved batch-means procedure for building a confidence interval on a steady-state expected simulation response that is centered on the sample mean of a portion of the corresponding simulation-generated time series and satisfies a user-specified absolute or relative precision requirement. The theory supporting the new algorithm merely requires the output process to be weakly dependent (phi-mixing) so that for a sufficiently large batch size, the batch means are approximately multivariate normal but not necessarily uncorrelated. A variant of the method of nonoverlapping batch means (NOBM), the Automated Simulation Analysis Procedure (ASAP) operates as follows: the batch size is progressively increased until either (a) the batch means pass the von Neumann test for independence, and then ASAP delivers a classical NOBM confidence interval; or (b) the batch means pass the Shapiro-Wilk test for multivariate normality, and then ASAP delivers a corrected confidence interval. The latter correction is based on an inverted Cornish-Fisher expansion for the classical NOBM t-ratio, where the terms of the expansion are estimated via an autoregressive–moving average time series model of the batch means. An experimental performance evaluation demonstrates the advantages of ASAP versus other widely used batch-means procedures.

1 INTRODUCTION

In discrete-event simulation, we are often interested in estimating the steady-state mean μ_X of a stochastic output process $\{X_i : i \geq 1\}$ 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

$$\overline{X}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

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 $0 < \alpha < 1$. Normally, we would like the CI for μ_X to satisfy two criteria: (a) the CI is narrow enough to be informative, and (b) the actual coverage probability of the CI is close to the nominal coverage probability $1-\alpha$.

The usual method of CI construction from classical statistics, which assumes independent and identically distributed (i.i.d.) observations, is not directly applicable since observations of a simulation-generated output process are typically neither independent nor identically distributed. Several methods have been proposed for constructing CIs based on dependent observations, including the method of nonoverlapping batch means (NOBM).

In the NOBM method, the sequence of simulation-generated outputs $\{X_i : i = 1, \ldots, n\}$ is divided into k adjacent nonoverlapping batches, each of size m. For simplicity, we assume that n is a multiple of m so that $n = km$; thus when k is fixed and $m \to \infty$, we have $n \to \infty$. The sample mean, $Y_j(m)$, for the jth batch is calculated by

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

Then the grand mean $\overline{Y}(n,k)$ of the individual batch means, given by

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

is used as an estimator for μ_X (note that $\overline{Y}(n,k) = \overline{X}(n)$). Naturally, we seek to construct a CI centered on the estimator (2).

We will assume the selected output process $\{X_i\}$ is *stationary* (or stationary in the strict sense), that is, the joint distribution of the X_i's is insensitive to time shifts. We will also assume the process is *weakly dependent*, that is, X_i's widely separated from each other in the sequence are almost independent (in the sense of ϕ-*mixing*, see Billingsley (1968)) so that the lag-q covariance $\gamma(q) \to 0$ as q increases. These weakly dependent processes typically obey a Central Limit Theorem (CLT) for dependent processes of the form

$$\sqrt{n}\left[\overline{X}(n) - \mu_X\right] \underset{n\to\infty}{\xrightarrow{D}} N\left(0, \sigma^2\right),$$

$$\begin{aligned} \text{where } \sigma^2 &\equiv \lim_{n\to\infty} n\text{Var}\left[\overline{X}\,(n)\right] \\ &= \sum_{i=-\infty}^{\infty} \gamma(i) = \gamma(0) + 2\sum_{i=1}^{\infty} \gamma(i) \end{aligned}$$

is the steady-state variance constant (SSVC) (as distinguished from the process variance σ_X^2). A sufficient condition for the SSVC to exist is that $\sum_{i=-\infty}^{\infty} |\gamma(i)| < \infty$ (Anderson 1971). Note that $\gamma(0) = \text{Var}[X_i]$.

2 METHOD OF NONOVERLAPPING BATCH MEANS (NOBM)

Although some output analysis methods attempt to estimate the steady-state variance constant σ^2 for the construction of the CI, NOBM in its classical setting, i.e., when the number of batches is fixed, does not. NOBM seeks to make each batch a "repetition" of the experiment on the process. In order to achieve this, we assume that the batch size is sufficiently large so that the batch means $\left\{Y_j(m) : 1 \le j \le k\right\}$ are i.i.d. normal, $\left\{Y_j(m) : 1 \le j \le k\right\} \overset{\text{i.i.d.}}{\sim} N\left[\mu_X, \sigma^2(m)/m\right]$, where the symbol $\sim$ is read "is distributed as," $\sigma^2(m) = \gamma(0) + 2\sum_{q=1}^{m-1}\left(1 - \frac{q}{m}\right)\gamma(q)$, and $\text{Var}\left[Y_j(m)\right] = \sigma^2(m)/m$. It follows that $\lim_{m\to\infty} \sigma^2(m) = \sigma^2$ and $\text{Var}\left[Y_j(m)\right] \approx \sigma^2/m$, provided that m is sufficiently large.

We can now apply a classical result from statistics to compute a confidence interval for μ_X from the batch means $\{Y_j(m) : 1 \le j \le k\}$. If $\left\{Z_j : 1 \le j \le k\right\} \overset{\text{i.i.d.}}{\sim} N\left(\mu_Z, \sigma_Z^2\right)$ so that the $\{Z_i\}$ constitute a random sample of size k from a normal distribution with mean μ_Z and variance σ_Z^2, then the sample mean $\overline{Z}(k)$ and the sample variance S_k^2 of the $\{Z_j\}$ are independent with

$$\overline{Z}(k) \sim N\left(\mu_Z, \frac{\sigma_Z^2}{k}\right), \tag{3}$$

$$\frac{(k-1)\,S_k^2}{\sigma_Z^2} \sim \chi_{k-1}^2, \tag{4}$$

and

$$\frac{\overline{Z}(k) - \mu}{\sqrt{S_k^2/k}} \sim t_{k-1}, \tag{5}$$

where t_{k-1} denotes the Student t-distribution with $k-1$ degrees of freedom and χ_{k-1}^2 denotes the chi-square distribution with $k-1$ degrees of freedom. We can then construct an exact $100(1-\alpha)$% CI for μ_Z of the form $\overline{Z}(k) \pm t_{1-\alpha/2,k-1} S_k/\sqrt{k}$. The sample variance of the k batch means of batches of size m is

$$S_{n,k}^2 = \frac{1}{k-1}\sum_{j=1}^{k}\left[Y_j(m) - \overline{X}(n)\right]^2. \tag{6}$$

Therefore the NOBM t-ratio equivalent to the ratio in (5) is

$$t = \frac{\overline{Y} - \mu}{\sqrt{S_{n,k}^2/k}} = \frac{\dfrac{\overline{Y} - \mu}{\sqrt{\text{Var}\left[\overline{X}(n)\right]}}\sqrt{\dfrac{k\text{Var}\left[\overline{X}(n)\right]}{\text{Var}\left[\overline{X}(m)\right]}}}{\sqrt{\dfrac{S_{n,k}^2}{\text{Var}\left[\overline{X}(m)\right]}}}, \tag{7}$$

where $\overline{X}(m) = Y(m)$ (the mean of a batch of size m), $\overline{X}(n) = \overline{Y}(n, k)$ ($= \overline{Y}$, the grand mean of $n = km$ observations organized into k nonoverlapping batches each of size m), and $S_{n,k}^2$ (the sample variance of the k batch means) are respectively defined by (1), (2), and (6). Replacing $\overline{Z}(k)$ by $\overline{X}(n)$ and S_k^2 by the sample variance of the batch means $S_{n,k}^2$ in (3)–(5), we have that (3)–(5) are approximately satisfied as the batch size m becomes sufficiently large while the batch count k is fixed. Then as $m \to \infty$ with k fixed so that $n \to \infty$, an asymptotically valid $100(1-\alpha)$% confidence interval for μ_X is

$$\overline{X}(n) \pm t_{1-\alpha/2,k-1}\frac{S_{n,k}}{\sqrt{k}}. \tag{8}$$

This CI is approximately valid when the batch count k is fixed and the batch size m becomes large because the batch means $Y_1(m), \ldots, Y_k(m)$ become almost independent (since the process is weakly dependent) and almost normally distributed (from an appropriate CLT for dependent processes). Thus the asymptotic validity of NOBM depends on both the assumption of approximate independence of the batch means and the assumption of the batch means being approximately normally distributed.

NOBM procedures address the problem of determining the batch size, m, and the number of batches, k, that are required to satisfy the assumptions of independence and normality. Theoretically, if these assumptions are satisfied, then we will get CIs whose actual coverage is close to the nominal coverage. In this paper we present a new procedure called Automated Simulation Analysis Procedure (ASAP) for implementing the NOBM procedure.

3 BASIS FOR THE AUTOMATED SIMULATION ANALYSIS PROCEDURE (ASAP)

Prior to developing a new procedure, we carried out theoretical and empirical analyses of the convergence properties of batch means in selected stochastic processes (Steiger 1999). The cases studied were chosen so that a variety of correlation structures and marginal distributions of the X_i's were represented. We concluded from the results of these analyses that if the vector of batch means has a multivariate normal distribution, then the first two moments of the square of the denominator of the classical batch means t-statistic (7) are close to the first two moments of a $\chi^2_{k-1}/(k-1)$ random variable. Additionally, although the numerator of the t-statistic (7) may not display the correct variance, i.e., the variance may not be equal to one, the multivariate normality of the batch means results in a numerator that is normally distributed with expected value zero.

We can also show that the numerator and squared denominator in (7) have zero correlation when the batch means are multivariate normal. Therefore, if we have a batch size large enough so that the batch means have a joint distribution that is approximately multivariate normal, then we may reasonably assume that the denominator of the t-statistic (7) possesses the required distribution and that the numerator and the denominator of the t-statistic are independent; and if these assumptions hold, then we can make a correction to the classical batch means confidence interval (8) to compensate for the failure of the numerator of the NOBM t-statistic to possess a variance of one.

The proposed correction to (8) is based on an inverted Cornish-Fisher expansion (Hall 1983) for the t-statistic in which the terms of the expansion are estimated by fitting an autoregressive–moving average (ARMA) time series model (Box and Jenkins 1976) to the series of final batch means. This approach should result in improved CI coverage at smaller batch sizes, even when the batch means do not appear to be independent. These considerations motivated the development of the new batch-means procedure that is described in the next section.

4 OVERVIEW OF ASAP

ASAP 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 $\overline{Y}$.

ASAP delivers the following outputs:

1. a nominal $100(1-\alpha)\%$ confidence interval for μ_X having the form
$$\overline{Y} \pm H \quad \text{where} \quad H \le H^* \quad \text{or} \quad H \le r^*|\overline{Y}|, \qquad (9)$$
provided no additional simulation-generated observations are required;
2. a new total sample size n to be supplied to the algorithm; or
3. the estimated final sample size N^*, final batch size m^*, and final batch count k^* required to deliver a valid confidence interval of the form (9) that satisfies the user-specified precision requirement.

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 ASAP must be called again with the additional data; and this cycle of simulation followed by analysis may be repeated several times before ASAP finally delivers a confidence interval.

A flow chart of ASAP is depicted in Figure 1. On each iteration of ASAP, the algorithm operates as follows. The simulation outputs are divided into a fixed number of batches (namely, 96 batches); and batch means are computed. The first two batches are discarded, and the remaining 94 batch means are tested for independence. If the test for independence fails, then the batch means are tested for joint multivariate normality. If the normality test fails, then the batch size is increased by a factor of $\sqrt{2}$ and the process is repeated until one of the tests is passed.

Upon acceptance of either the hypothesis of independence or the hypothesis of joint multivariate normality of the batch means, a CI is constructed—either the usual NOBM CI (8) (in the case of acceptance of independence) or a corrected CI (in the case of acceptance of multivariate normality). The correction uses an inverted Cornish-Fisher expansion (Hall 1983 and Kendall, Stuart and Ord 1987) of the NOBM t-statistic whose terms are estimated by fitting an ARMA model to the batch means process. Subsequent iterations of ASAP that are performed to satisfy the user-

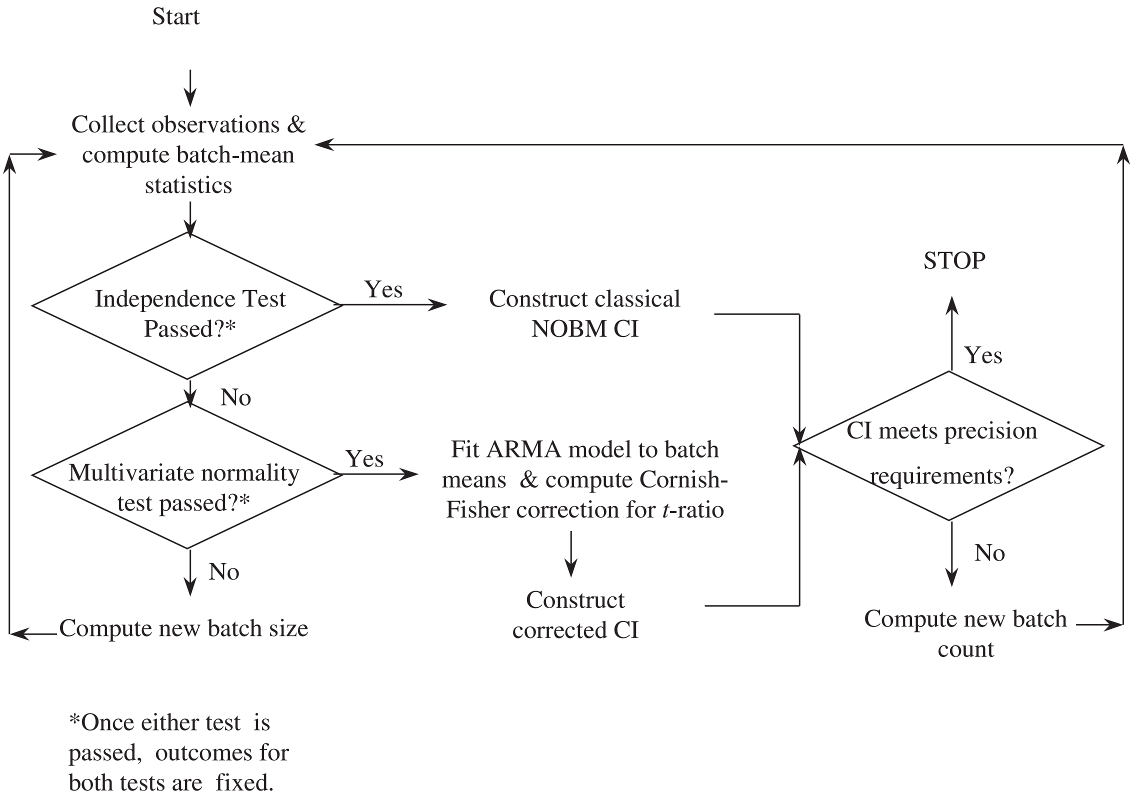
Start
Collect observations & compute batch-mean statistics
Independence Test Passed?*
Yes
Construct classical NOBM CI
No
Multivariate normality test passed?*
Yes
Fit ARMA model to batch means & compute Cornish-Fisher correction for t-ratio
No
Compute new batch size
Construct corrected CI
STOP
Yes
CI meets precision requirements?
No
Compute new batch count
*Once either test is passed, outcomes for both tests are fixed.

Figure 1: Flow Chart of ASAP

specified precision requirement (if there is one) do not repeat testing for independence or 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 two batches of the overall data set (consisting of all original observations plus any additional observations required by ASAP). Successive iterations of ASAP continue until the precision requirement is met.

Subsections 4.1–4.6 below provide some details on the main steps in the operation of ASAP. Steiger (1999) gives a complete description of ASAP.

4.1 Sample Size for First Iteration of ASAP

ASAP begins with an initial batch size $m_1 = 16$ and requires data for $k_1 = 96$ initial batches to be collected. The results of our tests and experiments with the algorithm show that ASAP performs well with this initial batch size, even for processes that are highly dependent and exhibit marked departures from normality. There were several reasons, which are presented in the following sections, for choosing an initial batch count of 96. While a total of $k_1 m_1 = 1536$ observations may be more than is actually needed in a few cases, such a sample size is usually easy and inexpensive to generate.

4.2 Testing Batch Means for Independence

ASAP uses the von Neumann ratio of the sample mean-square successive difference to the sample variance (von Neumann 1941, Fishman 1978) to test for independence of the batch means. For a sample of k observations, $Z_1, Z_2, \ldots, Z_k$, this ratio is

$$C_k = 1 - \frac{\sum_{j=1}^{k-1} \left(Z_j - Z_{j+1}\right)^2}{2\sum_{j=1}^{k} \left(Z_j - \overline{Z}\right)^2}. \qquad (10)$$

The null hypothesis of this test is that the Z_j's are i.i.d. If the Z_j's are normally distributed, then under H_0, $C_k \overset{\cdot}{\sim} N\big(0, (k-2)/(k^2-1)\big)$, for k as small as 8. If the Z_j's are nonnormal, then under H_0, C_k has mean zero. Furthermore, as the sample size (in the case of batch means, k is the number of batches) increases, the variance of $V_k \equiv C_k/\sqrt{(k-2)/(k^2-1)}$ approaches one and the skewness and excess kurtosis converge to zero (Fishman 1978). Therefore, if k is large, then the large sample properties suggest that we can approximate the distribution of V_k with the $N(0,1)$ distribution, provided that the Z_j's are i.i.d. The critical values for V_k with $k \geq 25$ are extremely close to the critical values of the $N(0,1)$ (Anderson 1971). We note here that the more observations used for the test, the more powerful the test is, i.e. the more capable the test is of detecting type II errors. The relative power of the von Neumann test with large k is one reason for starting with a batch count of 96.

Our studies of the batch means process reveal that correlation between batch means is not always a monotone decreasing function of the batch size. Therefore, we chose to use a two-sided test for the independence of the batch means with size $\alpha_{\text{ind}} = 0.20$. The first two batches of data are excluded from computations of batch-means statistics in an effort to overcome the initial bias problem. Let $k_1^* = k_1 - 2 = 94$ denote the number of batch means retained for confidence-interval construction. The k_1^* retained batch means are tested for independence using von Neumann's ratio (10). If the $k_1^* = 94$ batch means pass the independence test, then the classical batch means confidence interval (8) is constructed with midpoint $\overline{Y}$and half-length

$$H = t_{1-\alpha/2, k_1^*-1} \frac{S_{n,k_1^*}}{\sqrt{k_1^*}}.$$

No correction is made to the confidence interval because presumably none is needed if the batch means are independent (Fishman 1978, Fishman and Yarberry 1997).

4.3 Testing Batch Means for Joint Normality

If the test for independence fails, then ASAP tests the batch means for joint normality in the following manner. First, $g = 16$ vectors each consisting of $r = 4$ adjacent batch means are constructed. Two batch means between each set of four are ignored in an effort to obtain approximately independent 4-dimensional vectors of batch means, i.e.,

$$\begin{gathered}
\underbrace{Y_3(m), Y_4(m), Y_5(m), Y_6(m)}_{\text{1st } (4\times1) \text{ vector } \mathbf{y}_1}, \underbrace{Y_7(m), Y_8(m)}_{\text{ignored}}, \\
\underbrace{Y_9(m), Y_{10}(m), Y_{11}(m), Y_{12}(m)}_{\text{2nd } (4\times1) \text{ vector } \mathbf{y}_2}, \underbrace{Y_{13}(m), Y_{14}(m)}_{\text{ignored}}, \\
\ldots, \underbrace{Y_{93}(m), Y_{94}(m), Y_{95}(m), Y_{96}(m)}_{\text{16th } (4\times1) \text{ vector } \mathbf{y}_{16}}. \qquad (11)
\end{gathered}$$

We apply the Shapiro-Wilk test for multivariate normality (Malkovich and Afifi 1973, Tew and Wilson 1992) to the resulting sample $g = 16$ vectors, each consisting of $r = 4$ adjacent batch means. Although joint normality of these selected sets of 4 adjacent batch means is not sufficient to ensure joint normality of all 96 batch means (see exercise 15.20 on p. 504 of Kendall, Stuart and Ord 1987), the results reported of an extensive experimental evaluation of ASAP's performance strongly suggest that testing for joint quadrivariate normality in adjacent batch means yields good performance in many situations.

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

$$\bar{\mathbf{y}} = g^{-1} \sum_{i=1}^{g} \mathbf{y}_i \quad \text{and} \quad \mathbf{A} = \sum_{i=1}^{g} (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^{\mathrm{T}}.$$

Throughout the rest of this discussion, we assume that $\mathbf{A}$ is nonsingular with probability one. This property can be ensured, for example, by a mild technical requirement detailed by Tew and Wilson (1992), provided the replication count $g > r$; and since we take $r = 4$ and $g = 16$ in ASAP, with probability one we can identify the observation $\mathbf{y}^\dagger \in \{\mathbf{y}_i : i = 1, 2, \ldots, g\}$ for which

$$(\mathbf{y}^\dagger - \bar{\mathbf{y}})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{y}^\dagger - \bar{\mathbf{y}}) = \max_{i=1,\ldots,g} \left\{ (\mathbf{y}_i - \bar{\mathbf{y}})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}) \right\}.$$

We compute $Z_i \equiv (\mathbf{y}^\dagger - \bar{\mathbf{y}})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}})$ for $i = 1, 2, \ldots, g$,and we sort these quantities in ascending order to obtain the corresponding order statistics $Z_{(1)} < Z_{(2)} < \cdots < Z_{(g)}$. Let $\{a_i : i = 1, 2, \ldots, g\}$ denote the coefficients of the univariate Shapiro-Wilk statistic for a random sample of size g (see Royston 1982a, 1982b). The multivariate Shapiro-Wilk statistic is then given by

$$W^* = \frac{\left[\sum_{i=1}^{g} a_i Z_{(i)}\right]^2}{(\mathbf{y}^\dagger - \bar{\mathbf{y}})^{\mathrm{T}} \mathbf{A}^{-1} (\mathbf{y}^\dagger - \bar{\mathbf{y}})}$$

(Malkovich and Afifi 1973). The null hypothesis of multinormal responses $\{\mathbf{y}_i\}$ is rejected at the α level of significance $(0 < \alpha < 1)$ if $W^* < w_\alpha^*(r,g)$, where $w_\alpha^*(r,g)$ denotes the quantile of order α for the null distribution of W^* (that is, the distribution of W^* when this statistic is based on a random sample of size g taken from an r-dimensional nonsingular normal distribution).

For the multivariate normality test in ASAP, we used the size $\alpha_{\text{mvn}} = 0.10$ based on results of experimenting with the parameter α_{mvn}. In practice, ASAP appears to perform well even where there is mild departure from multivariate normality of the batch means.

4.4 Additional Iterations of ASAP

On the ith iteration of ASAP, we let k_i and m_i respectively denote the batch count and the batch size; and we take $k_1 = 96$, $m_1 = 16$ on the first iteration of the procedure.

An additional iteration of ASAP will be required if the following conditions occur on iteration i:

a) the independence test (see Section 4.2) yields a significant result (that is, independence is rejected) at the level of significance α_{ind} when this test is applied to the k_i batch means for batches of size m_i; and then

b) the multivariate Shapiro-Wilk test (see Section 4.3) yields a significant result (that is, multivariate normality is rejected) at the level of significance α_{mvn} when this test is applied to the corresponding sample of size 16 consisting of four-dimensional random vectors formed from adjacent batch means.

Now if conditions a) and b) both occur on iteration i of ASAP, then iteration $i+1$ will be required in which the batch size and batch count are respectively taken to be

$$m_{i+1} = \lfloor \sqrt{2} m_i \rfloor \text{ and } k_{i+1} = k_i$$

so the total required sample size is $n_{i+1} = m_{i+1}k_{i+1}$; and thus the user must generate the additional simulation responses $\{X_j : j = n_i + 1, n_i + 2, \ldots, n_{i+1}\}$ before executing iteration $i+1$ of ASAP. We chose to increase the batch size by the factor $\sqrt{2}$ at each iteration so that the total sample size would double on every other iteration.

4.5 Inverted Cornish-Fisher Correction for Dependent Normal Batch Means

If the batch means have failed the test for independence but have passed the test for joint multivariate normality, then we would like to make a correction to the classical batch means CI to adjust for the correlation between the batch means. A standard way of adjusting CIs for nonnormality is to use some version of an inverted Edgeworth expansion for the corresponding point estimator of the parameter of interest (Hall 1983, Kendall, Stuart and Ord 1987). ASAP uses an inverted Cornish-Fisher expansion of the NOBM t-statistic (7) that is in terms of the statistic's first four cumulants. Expressions for these cumulants involve $\text{Var}[\overline{X}(m)]$ and $\text{Var}[\overline{X}(n)]$. Therefore, in order to compute sample estimators of these cumulants, we must have sample estimators of $\text{Var}[\overline{X}(m)]$ and $\text{Var}[\overline{X}(n)]$.

If the hypothesis of multivariate normality in Section 4.3 is accepted, then to obtain sample estimators of $\text{Var}[\overline{X}(m)]$ and $\text{Var}[\overline{X}(n)]$ ASAP first fits an ARMA process of at most order 2 to the set of $k^* = 94$ batch means. Obtaining a good result from the ARMA fitting process generally requires over 50 observations (Box and Jenkins 1976, p. 18), which is the final reason for choosing an initial number of batches close to 100. Fits of five possible ARMA models are attempted: AR(1), AR(2), MA(1), MA(2), and ARMA(1,1). IMSL routines (IMSL Problem Solving Software Systems 1987) are used to estimate the autoregressive–moving average parameters, the residual variance, σ_a^2, and the process variance, $\sigma_Y^2 = \text{Var}[Y_\ell]$, for the five ARMA models. Then the “best" fit of the five is chosen. Preference is given to the AR(1) model. An alternate model is used only if it has a significantly smaller residual variance than the AR(1) model.

The estimators of $\text{Var}[\overline{X}(m)]$ and the parameters from the ARMA fit are then used to estimate $\text{Var}[\overline{X}(n)]$:

$$\widehat{\text{Var}}\left[\overline{X}(n)\right] = \frac{1}{k^*} \sum_{q=-k^*+1}^{k^*-1} \left(1 - \frac{|q|}{k^*}\right) \widehat{\gamma}_m(q), \tag{12}$$

where $\widehat{\gamma}_m(q)$ denotes the estimated lag-q covariance of the batch means Y_j, $j = 3, \ldots, k$ based on the fitted time series model.

The derivation of the terms in the inverted Cornish-Fisher expansion is based on the following three assumptions:

A_1: The batch means have a joint multivariate normal distribution.

A_2: The numerator and denominator of the t-ratio (7) are independent.

A_3: The square of the denominator of the t-ratio (7) is distributed as $\chi^2_{k-1}/(k-1)$.

Assumption A_1 is based on using a batch size large enough to yield a nonsignificant result for the multivariate Shapiro-Wilk test as described in Section 4.3. Assumption A_2 is supported by the result that if the batch means have a multivariate normal distribution, then the numerator and squared denominator of the t-statistic (7) are uncorrelated. Finally, results of our studies of the convergence properties of the numerator and squared denominator of the NOBM t-ratio (7) suggest that the square of the denominator has approximately achieved the distribution of a $\chi^2_{k-1}/(k-1)$ variate if the batch size m is large enough to result in approximate multivariate normality of the batch means.

The following are the cumulants κ_1, κ_2, κ_3 and κ_4 of the NOBM t-ratio (7) computed under assumptions A_1–A_3:

$$\kappa_1 = \kappa_3 = 0, \tag{13}$$

$$\kappa_2 = \frac{k\text{Var}[\overline{X}(n)](k-1)}{\text{Var}[\overline{X}(m)](k-3)}, \tag{14}$$

and

$$\kappa_4 = \frac{2k^2(k-1)^2\mathrm{Var}^2[\overline{X}(n)]}{(k-3)^2(k-5)\mathrm{Var}^2[\overline{X}(m)]}. \tag{15}$$

Based on an inverted Cornish-Fisher expansion for the classical NOBM t-ratio, an adjusted $100(1-\alpha)\%$ confidence interval for μ_X is

$$\left[\overline{X}(n) - h'(z_{1-\alpha/2})\frac{S_{n,k}}{\sqrt{k}},\ \ \overline{X}(n) - h'(-z_{1-\alpha/2})\frac{S_{n,k}}{\sqrt{k}}\right],$$

where $n = km$ and

$$\begin{aligned} h'(z_{1-\alpha/2}) = z_{1-\alpha/2} &+ (\kappa_1 - \kappa_3/6) \\ &+ [(\kappa_2-1)/2]\, z_{1-\alpha/2} + (\kappa_3/6)z^2_{1-\alpha/2}, \end{aligned}$$

for the first-order pivot, or

$$\begin{aligned} h'(z_{1-\alpha/2}) = z_{1-\alpha/2} &+ (\kappa_1 - \kappa_3/6) \\ &+ [(\kappa_2-1)/2 - \kappa_4/8]\, z_{1-\alpha/2} \\ &+ (\kappa_3/6)z^2_{1-\alpha/2} + (\kappa_4/24)z^3_{1-\alpha/2}, \end{aligned}$$

for the second-order pivot, and κ_i denotes the ith cumulant of the t-statistic, for $i = 1, 2, 3, 4$ (Hall 1983, Chien 1989).

Under the assumptions A$_1$–A$_3$, the first four cumulants of the t-ratio are given by (13)–(15). By substituting the variance estimator $\widehat{\mathrm{Var}}[\overline{X}(m)]$ from the ARMA fit for $\mathrm{Var}[\overline{X}(m)]$ and by substituting the variance estimator $\widehat{\mathrm{Var}}[\overline{X}(n)]$ of display (12) for $\mathrm{Var}[\overline{X}(n)]$ in the expressions (14) and (15) for κ_2 and κ_4, we obtain the following approximate $100(1-\alpha)\%$ confidence intervals for μ_X:

$$\overline{X}(n) \pm z_{1-\alpha/2}\left(1 + \frac{\hat{\kappa}_2 - 1}{2}\right)\sqrt{\frac{\widehat{\mathrm{Var}}[\overline{X}(m)]}{k}}, \tag{16}$$

or

$$\overline{X}(n) \pm \left[z_{1-\alpha/2}\left(1 + \frac{\hat{\kappa}_2 - 1}{2} - \frac{\hat{\kappa}_4}{8}\right) + \frac{\hat{\kappa}_4}{24}z^3_{1-\alpha/2}\right] \times\sqrt{\frac{\widehat{\mathrm{Var}}[\overline{X}(m)]}{k}} ,\tag{17}$$

depending on which pivot is used.

4.6 Fulfilling the Precision Requirement

The final step in ASAP is to determine if the confidence interval that was constructed meets the user's requirement for precision. The confidence interval may be the one based on a nonsignificant result from the independence test (that is, the batch means pass the test for independence) or the adjusted CI based on a nonsignificant result from the test for multivariate normality of the batch means (that is, the batch means pass the test for multivariate normality). If the relevant requirement

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

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

$$k_i^+ = \left\lceil \left(\frac{H}{H^*}\right)^2 k_i \right\rceil - k_i\ ;$$

thus on iteration $i+1$ of ASAP the batch count and batch size are $k_{i+1} \leftarrow k_i + k_i^+$ and $m_{i+1} \leftarrow m_i$ so that the total required sample size on iteration $i+1$ is $n_{i+1} \leftarrow m_{i+1}k_{i+1}$; and thus the user must generate the additional simulation responses $\{X_j : j = n_i{+}1, n_i{+}2, \ldots, n_{i+1}\}$ before executing iteration $i+1$ of ASAP.

The user then performs iteration $i+1$ of ASAP 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 two batches are again omitted from the calculation of the sample mean. The batch means of $(k_{i+1} - 2)$ batches of size m_{i+1} are computed. If an ARMA model was used in constructing an adjusted CI on the previous iteration, then an updated ARMA fit is made using $(k_{i+1} - 2)$ batches of size m_{i+1}; moreover, in this situation new estimates of $\mathrm{Var}[\overline{X}(m)]$, $\mathrm{Var}[\overline{X}(n)]$, κ_2 and κ_4 are computed, and the CI (16) or (17) is constructed. If the CI for the previous iteration was based on batch means that passed the independence test, then the classical NOBM confidence interval (8) is constructed with the batch means of $(k_{i+1} - 2)$ batches of size m_{i+1}. If the precision requirement (18) is satisfied on iteration $i+1$ of ASAP, then the algorithm terminates, returning $\overline{Y}$ and H. If the required precision is not achieved on iteration $i+1$, ASAP estimates a new number of batches and sample size to be used for the next iteration.

5 PERFORMANCE EVALUATION FOR ASAP

We tested ASAP on many problems representing various types of stochastic processes. In this section we discuss the results of this experimentation on two of the processes tested. The steady-state mean is available analytically in these models. Therefore, we were able to evaluate the performance of ASAP in terms of actual coverage versus nominal coverage.

The first test case we present is a process defined by a real-valued function on a simple 2-state Discrete Time Markov Chain (DTMC) whose one-step probability transition matrix and cost vector associated with the states are respectively given by

$$\mathbf{P} = \begin{array}{cc} & \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix} \end{array} \quad \text{and} \quad \mathbf{h} = \begin{array}{c} \begin{array}{cc} 0 & \ 1 \end{array} \\ \begin{pmatrix} 5 & 10 \end{pmatrix} \end{array}. \qquad (19)$$

The second case is the waiting time process in the $M/M/1$ queue with utilization $\tau = 0.9$. Both of these processes display high dependency. Also the marginal distribution of the X_i's in the waiting time process of the $M/M/1$ queue has an exponential tail and is therefore markedly nonnormal. We believe the characteristics of high dependency and markedly nonnormal marginal distribution should stress any output analysis procedure.

We made 100 independent simulations of these two systems and attempted to construct nominal 90% confidence intervals for three cases:

(i) no precision requirement, i.e., we terminated the procedure when a CI was constructed based on 94 batches of the size at which the batch means passed either the statistical test for independence or the test for multivariate normality;

(ii) ±15% precision so that $r^* = 0.15$ in (18); and

(iii) ±7.5% precision so that $r^* = 0.075$ in (18).

This enabled us to estimate the actual coverage of CIs constructed via the ASAP algorithm. We also tested the performance of the LBATCH and ABATCH algorithms (Fishman 1996, Fishman and Yarberry, 1997) for comparison. Since LBATCH and ABATCH do not explicitly determine a sample size, we passed to the LBATCH and ABATCH algorithms the same data sets used by ASAP. Therefore, the CIs computed by the LBATCH and ABATCH algorithms are based on the same sample sizes and data sets used for ASAP. We do not mean to imply that these results are what one could expect routinely from either LBATCH or ABATCH, but only what one could expect by applying LBATCH and ABATCH to the data sets that resulted from first using the ASAP algorithm.

Tables 1 and 2 display in detail the results of our tests. Table 3 displays the additional results obtained through standalone application of LBATCH and ABATCH to waiting times in the $M/M/1$ queue with $\tau = 0.9$ when LBATCH and ABATCH operate with a stopping rule based on a user-specified precision requirement for the final confidence interval. We began the experiments for these systems with a sample size of 1536 (the same sample size required for the first iteration of ASAP). We then applied a stopping rule

Table 1: Performance of Batch-Means Procedures for the 2-State DTMC Defined by (19) Based on 100 Independent Replications of Nominal 90% Confidence Intervals

Precision	Procedure		
Requirement	LBATCH	ABATCH	ASAP†
NO PRECISION			
avg. sample size			3036
coverage	70%	85%	96%
avg. rel. precision	0.069	0.086	0.159
avg. CI half-length	0.515	0.642	1.20
var. CI half-length	0.009	0.012	0.172
±15% PRECISION			
avg. sample size			5171
coverage	72%	81%	96%
avg. rel. precision	0.060	0.070	0.120
avg. CI half-length	0.045	0.053	0.906
var. CI half-length	0.011	0.010	0.023
±7.5% PRECISION			
avg. sample size			22711
coverage	81%	86%	99%
avg. rel. precision	0.034	0.038	0.059
avg. CI half-length	0.253	0.284	0.438
var. CI half-length	0.003	0.003	0.006

†No. of classical and corrected CIs generated by ASAP: 0 and 100, respectively.

Table 2: Performance of Batch-Means Procedures for the M/M/1 Queue Waiting Time Process with $\tau = 0.9$ Based on 100 Independent Replications of Nominal 90% Confidence Intervals

Precision	Procedure		
Requirement	LBATCH	ABATCH	ASAP†
NO PRECISION			
avg. sample size			7719
coverage	44%	60%	83%
avg. rel. precision	0.202	0.301	1.088
avg. CI half-length	1.70	2.67	11.8
var. CI half-length	0.683	3.92	523.0
±15% PRECISION			
avg. sample size			298950
coverage	79%	80%	88%
avg. rel. precision	0.061	0.069	0.089
avg. CI half-length	0.543	0.613	0.783
var. CI half-length	0.027	0.039	0.082
±7.5% PRECISION			
avg. sample size			815755
coverage	88%	90%	94%
avg. rel. precision	0.039	0.043	0.046
avg. CI half-length	0.353	0.382	0.413
var. CI half-length	0.012	0.039	0.018

†No. of classical and corrected CIs generated by ASAP: 4 and 96, respectively.

Table 3: Performance of LBATCH and ABATCH under a Relative Precision Requirement for $M/M/1$ Queue with $\tau = 0.9$ Based on 100 Independent Replications of Nominal 90% Confidence Intervals

Precision	Procedure	
Requirement	LBATCH	ABATCH
NO PRECISION		
avg. sample size	1536	1536
coverage	35%	54%
avg. rel. precision	0.204	0.338
avg. CI half-length	1.648	2.882
var. CI half-length	0.552	4.250
±15% PRECISION		
avg. sample size	34349	50910
coverage	65%	77%
avg. rel. precision	0.121	0.125
avg. CI half-length	.1.071	1.080
var. CI half-length	0.0513	0.0336
±7.5% PRECISION		
avg. sample size	227987	397387
coverage	80%	81%
avg. rel. precision	0.062	0.062
avg. CI half-length	0.551	0.553
var. CI half-length	0.005	0.007

similar to the one used for ASAP. After the initial run with 1536 observations was made, the final CI constructed by LBATCH or ABATCH was examined to see if the precision requirement was met. If not, then we calculated an estimate of additional observations that would be required, we generated the additional observations, and we executed LBATCH or ABATCH again with all of the observations accumulated so far. This process was repeated until the final CI delivered by LBATCH or ABATCH met the precision requirement. We realize that LBATCH and ABATCH were not really designed to be used in this way, but using such stopping rules is a natural approach to planning steady-state simulations; and we believe that the results in Table 3 provide a more complete perspective on the relative performance of LBATCH and ABATCH versus ASAP. No effort was made to analyze the convergence of the sample estimators from LBATCH and ABATCH, as is suggested in Fishman (1998). We only include these results to highlight the performance advantages achieved by ASAP without requiring analysis or manual intervention by the user.

ASAP showed somewhat better coverage than did LBATCH in the case of the 2-state DTMC (19) with high positive correlation, especially in the cases of no precision requirement and ±15% precision requirement. For this model, ASAP constructed adjusted CIs based on a nonsignificant result on the test for multivariate normality (i.e., the batch means passed the Shapiro-Wilk test for multivariate normality) in all 100 cases. The CIs from ASAP are wider than those from LBATCH and ABATCH, which is necessary for the improved coverage. However, the coefficient of variation of the CI half-lengths are smaller than those from LBATCH and ABATCH.

As stated previously the waiting time process in the $M/M/1$ queue with $\tau = 0.9$ is a difficult case in that the lag-1 correlation of the observations is close to one, the correlation function decays slowly, and the marginal distribution of X_i has an exponential tail. Because of these characteristics, we can expect slow convergence to both conditions for ensuring the validity of the batch means method, i.e., that the batch means are independent and identically normally distributed. This case most dramatically displays one of the advantages of the ASAP algorithm, i.e., it does not rely solely on the von Neumann test for independence. In fact, in 96 replications, ASAP constructed adjusted CIs based on a nonsignificant result from the test for multivariate normality.

As can be seen from Table 2, ASAP substantially outperforms LBATCH and ABATCH for the case of no precision requirement. As we demand more precision, we are of course forced to perform more sampling. At the precision requirement of ±7.5% the three algorithms give similar results. This implies that LBATCH and ABATCH will give satisfactory results if supplied with an adequate amount of data. However, LBATCH and ABATCH provide no mechanism for assessing the amount of data which should be used. This emphasizes a desirable feature of ASAP—that it determines a sample size that gives acceptable results, even when no precision is specified.

From Table 3 we see that in the $M/M/1$ queue with $\tau = 0.9$, if LBATCH and ABATCH are run until a certain precision requirement is met, coverage is severely degraded, especially when the precision requirement is so “loose" that it leads to relatively little additional sampling. Note that the sample sizes are much smaller than those required by ASAP to achieve the same precision. For example, the average sample size used by ABATCH for the waiting time process in the $M/M/1$ queue with utilization $\tau = 0.9$ and ±7.5% precision is approximately 397,387. This is considerably less than the average sample size of 815,755 used by ASAP. For ±7.5% precision and 90% confidence, Whitt’s (1989) approximation for estimating required run lengths of queueing simulations yields an estimated sample size of 855,238 for the waiting time process in the $M/M/1$ queue with $\tau = 0.9$. This strongly suggests that ASAP yields adequate sample sizes when a precision requirement is specified.

6 CONCLUSIONS

The advantages of ASAP may be summarized as follows:

- ASAP addresses the initial bias problem.
- ASAP is fully automated, since it
  - specifies initial sample size,

  - determines final sample size,
  - delivers CI of prespecified precision, and
  - requires no intervention or analysis by the user.
- Although it is heuristic, ASAP has some theoretical basis.
- ASAP does not rely solely on tests for independence.
- ASAP gives good results for highly dependent processes.
- ASAP delivers stable CIs with close to nominal coverage.

BIBLIOGRAPHY

Anderson, T. W. 1971. *The statistical analysis of time series*. New York: John Wiley & Sons, Inc.

Billingsley, P. 1968. *Convergence of probability measures*. New York: John Wiley & Sons, Inc.

Box, G. E. P., and G. M. Jenkins. 1976. *Time series analysis: Forecasting and control*. San Francisco: Holden-Day, Inc.

Chien, C. 1989. Small sample theory for steady state confidence intervals. Technical Report No. 37, Department of Operations Research, Stanford University, Stanford, California. U.S. Army Research Contract DAAL-**-K-0063

Fishman, G. S. 1978. Grouping observations in digital simulation. *Management Science* 24:510–521.

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.

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

Hall, P. 1983. Inverting an Edgeworth expansion. *Annals of Statistics* 11 (2): 560–576.

IMSL, Inc. 1987. *User's manual: STAT/LIBRARY.* Houston: IMSL, Inc.

Kendall, M., A. Stuart, and J. K. Ord. 1987. *Kendall's advanced theory of statistics*. New York: Oxford University Press.

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* 31:115–124.

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.

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.

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

Whitt, W. 1989. Planning queueing simulations. *Management Science* 35:1341–1366.

AUTHOR BIOGRAPHIES

NATALIE M. STEIGER is an Assistant Professor in the Department of Information Systems and Operations Management at the University of North Carolina at Greensboro. She is a member of IIE and INFORMS.

JAMES R. WILSON is Professor and Head the Department of Industrial Engineering at North Carolina State University. Currently he serves as a corepresentative of the INFORMS–College on Simulation to the WSC Board of Directors. He is a member of ASA, ACM, IIE, and INFORMS.