We formulate and evaluate an improved batch means procedure for steady-state simulation output analysis. This procedure delivers a confidence interval for an expected response that is centered on the sample mean of a portion of a simulation-generated time series and satisfies a user-specified absolute or relative precision requirement. We assume that the output process is 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 correlation-adjusted confidence interval. The latter adjustment 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. A simulation study demonstrates the performance improvements achieved by ASAP versus popular batch means procedures, especially with respect to confidence-interval coverage probability.

(Simulation; Statistical Analysis; Method of Batch Means; Steady-State Output Analysis)
1. Introduction

In discrete-event stochastic simulation, we are often interested in estimating the steady-state mean $\mu_X$ of a target output process $\{X_i : i \geq 1\}$ generated by a single prolonged simulation run. Given a time series of length $n$ generated by a simulation in steady-state operation, we see that a natural estimator of $\mu_X$ is the sample mean

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

When we require some indication of this estimator’s precision, typically we construct a confidence interval (CI) for $\mu_X$ at a user-specified coverage probability $1 - \alpha$, where $0 < \alpha < 1$. The CI for $\mu_X$ should satisfy two criteria: (a) it is narrow enough to be informative, and (b) its actual probability of covering the point $\mu_X$ is close to the nominal level $1 - \alpha$.

In the simulation analysis method of nonoverlapping batch means (NOBM), we divide the sequence of simulation-generated outputs $\{X_i : i = 1, \ldots, n\}$ into $k$ adjacent nonoverlapping batches each of size $m$. For simplicity, we assume that $n = km$; thus when $k$ is fixed and $m \to \infty$, we have $n \to \infty$. The sample mean for the $j$th 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 as a point estimator of $\mu_X$, we compute the grand mean of the individual batch means

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

(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.

We assume the target output process $\{X_i\}$ is stationary so that the joint distribution of the $X_i$’s is insensitive to time shifts. We 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 $\phi$-mixing; see Billingsley 1968) so that the lag-$q$ covariance

$$\gamma(q) \equiv \text{E}[(X_{i+q} - \mu_X)(X_i - \mu_X)] \quad \text{for} \quad q = 0, \pm 1, \pm 2, \ldots$$
satisfies \( \gamma(q) \to 0 \) as \(|q| \to \infty\). These weakly dependent processes typically obey a central limit theorem of the form

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

where

\[
\sigma^2 \equiv \lim_{n \to \infty} n \text{Var}\left[ \frac{X(n)}{n} \right] = \sum_{q=-\infty}^{\infty} \gamma(q)
\]

is the steady-state variance constant (as distinguished from the process variance \( \sigma^2_X \equiv \text{Var}[X_i] = \gamma(0) \)), and the symbol \( \xrightarrow{D_{n \to \infty}} \) denotes convergence in distribution. A sufficient condition for \( \sigma^2 \) to exist is that \( \sum_{q=-\infty}^{\infty} \gamma(q) < \infty \) (Anderson 1971). General conditions under which (2) holds are given, for example, in Theorem 20.1 of Billingsley (1968).

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 \( \mu_X \), then we can apply standard properties of the classical Student \( t \)-ratio to compute a confidence interval for \( \mu_X \) from the batch means—see, for example, of Steiger and Wilson (1999, 2001). The sample variance of the \( k \) batch means for batches of size \( m \) is

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

and the NOBM counterpart to the classical Student \( t \)-ratio is

\[
t = \frac{\overline{Y} - \mu_X}{\sqrt{S^2_{m,k}/k}} = \frac{\sqrt{k}\left[ \overline{Y}(m,k) - \mu_X \right]}{\sqrt{\text{Var}[Y(m)]}}.
\]

Under the same weak-dependence conditions that are sufficient to ensure (2) as \( m \to \infty \) with \( k \) fixed so that \( n \to \infty \), the NOBM \( t \)-ratio (3) converges in distribution to the Student \( t \)-distribution with \( k-1 \) degrees of freedom; and then an asymptotically valid 100(1 - \( \alpha \))% confidence interval for \( \mu_X \) is

\[
\overline{Y}(m,k) \pm t_{1-\alpha/2,k-1} \frac{S_{m,k}}{\sqrt{k}}.
\]
In Steiger and Wilson (2001), we present theoretical, numerical, and experimental results concerning the convergence properties of not only the vector of batch means but also the numerator and denominator of the NOBM t-ratio on the right-hand side of (3) when the number of batches $k$ is fixed and the batch size $m$ increases. From that work, we concluded that for the purpose of constructing a valid confidence interval for $\mu_X$ based on a variant of the NOBM method, the batch means often achieve approximate joint multivariate normality at a batch size that is substantially smaller than the batch size required to ensure approximate independence of the batch means; and a significant improvement over the classical NOBM confidence interval (4) can be achieved by taking advantage of this phenomenon. This observation motivated our development of ASAP—the Automated Simulation Analysis Procedure (ASAP) for steady-state simulation output analysis. ASAP addresses both the simulation start-up problem and the problem of correlated batch means, delivering a confidence interval for an expected response that is centered on the sample mean of a portion of a simulation-generated time series and that satisfies a user-specified precision requirement. Since ASAP is fully automated, the simulation practitioner need not repeatedly collect simulation data, manually analyze accumulated output, and then decide whether to collect more data.

The rest of this paper is organized as follows. In Section 2 we provide an overview of the operation of ASAP. In Section 3 we detail the statistical methods used by ASAP to test batch means for independence and multivariate normality. Section 4 explains how ASAP builds time series models for dependent normal batch means, and Section 5 shows how such a time series model is used to estimate an appropriate adjustment for confidence intervals derived from dependent normal batch means. In Section 6 we describe how ASAP controls the simulation run length in order to satisfy a user-specified precision requirement on the final confidence interval. Section 7 summarizes selected results from an extensive experimental evaluation of the performance of ASAP versus other widely used batch means procedures. Section 8 summarizes our main findings and discusses recommendations for future work.
2. **Overview of the Automated Simulation Analysis Procedure (ASAP)**

ASAP requires the following user-supplied inputs:

1. a simulation-generated realization \( \{X_j : j = 1, 2, \ldots, n\} \) of the target output process;

2. a confidence coefficient \( \alpha \) 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 \( \bar{Y} \).

ASAP delivers the following outputs:

1. a nominal \( 100(1 - \alpha)\% \) confidence interval for \( \mu_X \) having the form \( \bar{Y} \pm H \), where \( H \leq H^* \) or \( H \leq r^*|\bar{Y}| \), provided no additional simulation-generated observations are required; or

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

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 ignored, and the remaining 94 batch means are tested for independence. If the batch means fail the test for independence, then they are tested for joint multivariate normality. If the batch means fail the normality test, then the batch size is increased by a factor of \( \sqrt{2} \) and the process is repeated until the batch means finally pass one of the two tests.
Once the batch means pass either the independence test or the multivariate normality test, ASAP constructs a CI for $\mu_X$—either the usual NOBM CI (4) (if the batch means first pass the independence test) or a correlation-adjusted CI (if the batch means first pass the multivariate normality test), as described in Section 5. In subsequent iterations of ASAP that are performed to satisfy the user-specified precision requirement, the overall set of batch means is not retested for independence or multivariate normality—instead these iterations merely involve additional sampling, computing the additional batch means, and updating the CI whose form was previously determined until acceptable precision is finally achieved. Figure 1 displays an overall flow chart of ASAP. A formal algorithmic statement of ASAP is given in Figure 2. A standalone Windows-based version of ASAP and a user's manual are available in Steiger and Wilson (2000b). The following four sections explain in detail the development and operation of ASAP.

**Figure 1 Overall Flow Chart of ASAP**

3. Testing Batch Means for Independence and Joint Normality

ASAP begins on iteration 1 with a user-specified initial batch size $m_1$ (by default $m_1 = 16$), requiring data for $k_1 = 96$ initial batches. The results of extensive experimentation show that
Figure 2 ASAP Algorithm

[0] Fix integers $m_1 \leftarrow$ user specified minimum batch size [default: $m_1 \leftarrow 16$], $i \leftarrow 1$, $k_1 \leftarrow 96$, $k_i' \leftarrow 94$, $k_i^+ \leftarrow 0$, $n_0 \leftarrow 1$, and $n_1 \leftarrow 96m_1$.

Fix confidence level $1 - \alpha$.
Set $\text{IndTestPassed} \leftarrow \text{‘no’}$ and $\text{MVTestPassed} \leftarrow \text{‘no’}$.
If relative precision required, then set $\text{RelPrec} \leftarrow \text{‘yes’}$ and $r^* \leftarrow$ fraction of magnitude of final sample mean that gives upper bound $H^*$ on acceptable CI half-length.
If absolute precision required, then set $\text{RelPrec} \leftarrow \text{‘no’}$ and $H^* \leftarrow$ absolute upper bound on acceptable CI half-length.

[1] Collect observations $X_{n_0}, \ldots, X_{n_i}$.
Compute: batch means $Y_1(m_i), \ldots, Y_{k_i}(m_i)$ and the grand mean

$$\bar{Y}(m_i, k_i') \leftarrow \frac{1}{k_i'm_i} \sum_{\ell=2m_i+1}^{n_i} X_\ell = \frac{1}{k_i'} \sum_{j=3}^{k_i} Y_j(m_i).$$

if $\text{MVTestPassed} = \text{‘yes’}$, then goto [6];
if $\text{IndTestPassed} = \text{‘yes’}$, then goto [3];
Test $H_{\text{ind}}$: \{$Y_3(m_i), \ldots, Y_{96}(m_i)$\} are i.i.d. using (5).
if $H_{\text{ind}}$ is rejected, then goto [5];

[2] $\text{IndTestPassed} \leftarrow \text{‘yes’}$;

[3] Calculate

$$S^2_{m_i, k_i'} \leftarrow \frac{1}{k_i'-1} \sum_{j=3}^{k_i} [Y_j(m_i) - \bar{Y}(m_i, k_i')]^2 \text{ and } H \leftarrow z_{1-\alpha/2} \frac{S_{m_i, k_i'}}{\sqrt{k_i'}}.$$ 

Construct CI

$$\bar{Y}(m_i, k_i') \pm H.$$

if $\text{RelPrec} = \text{‘yes’}$ then $H^* \leftarrow r^* |\bar{Y}(m_i, k_i')|$.
if $(H \leq H^*)$ or $(r^* = 0 \text{ and } H^* = 0)$, then deliver $\bar{Y} = \bar{Y}(m_i, k_i')$ and $H$ and stop.

[4] Calculate additional number of batches needed,

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

Set new number of batches $k_{i+1} \leftarrow k_i + k_i^+$, new batch size $m_{i+1} \leftarrow m_i$,
and new sample size $n_{i+1}' \leftarrow k_{i+1}m_{i+1}$,
if \( k_{i+1} \leq 1502 \) then
\[
k_{i+1}' \leftarrow k_i' + k_i^+,
\]
\[
n_0 \leftarrow n_i + 1,
k_{i+1} \leftarrow k_{i+1}m_{i+1},\text{ and } i \leftarrow i + 1;
goto [1];
\]
else
while \( k_{i+1} > 1502 \)
begin
\[
m_{i+1} \leftarrow \lfloor \sqrt{2m_{i+1}} \rfloor \text{ and } k_{i+1} \leftarrow \lceil n_{i+1}' / m_{i+1} \rceil;
\]
end
\[
n_0 \leftarrow n_i + 1,
k_{i+1}' \leftarrow k_i' + k_i^+,
k_{i+1} \leftarrow k_{i+1}m_{i+1},\text{ and } i \leftarrow i + 1;
goto [1];
\]
end

[5] Test \( H_{mvn} \):
\[
\{ [Y_{3+(j-1)6}(m_i), Y_{4+(j-1)6}(m_i), Y_{5+(j-1)6}(m_i), Y_{6+(j-1)6}(m_i)] : 1 \leq j \leq 16 \}
\]
are quadrivariate normal random vectors using (6);
if \( H_{mvn} \) is rejected, then
\[
m_i+1 \leftarrow \lfloor \sqrt{2m_i} \rfloor,
n_0 \leftarrow n_i + 1,
k_{i+1} \leftarrow 96m_{i+1},\text{ and } i \leftarrow i + 1;
goto [1];
\]
else
\[
\text{MVTestPassed} \leftarrow \text{‘yes’};
goto [6];
\]

[6] Fit an ARMA process to the batch means \( \{Y_3(m_i), \ldots, Y_k(m_i)\} \);
Compute variance estimators \( \hat{\text{Var}}[Y(m_i)] \) and \( \hat{\text{Var}}[Y(m_i, k_i')] \) from the ARMA fit;
Calculate estimates \( \hat{\kappa}_2 \) and \( \hat{\kappa}_4 \) of the second and fourth cumulants of the \( t \)-statistic and
\[
H \leftarrow \left[ z_{1-\alpha/2} \left( 1 + \frac{\hat{\kappa}_2 - 1}{2} - \frac{\hat{\kappa}_4}{8} \right) + \frac{\hat{\kappa}_4}{24} z_{1-\alpha/2}^3 \right] \sqrt{\frac{\hat{\text{Var}}[Y(m_i)]}{k_i'}};
\]
and construct the CI
\[
\bar{Y}(m_i, k_i') \pm H.
\]
if RelPrec='yes' then \( H^* \leftarrow r^* |\bar{Y}(m_i, k_i')| \).
if \( (H \leq H^*) \) or \( (r^* = 0 \text{ and } H^* = 0) \), then
\[
deliver \bar{Y} = \bar{Y}(m_i, k_i') \text{ and } H \text{ and stop;}
\]
else
goto [4].
ASAP performs well with this initial batch size, even for processes that are highly dependent and whose marginal distributions exhibit marked departures from normality. There were several reasons, which are detailed in the following discussion, for choosing an initial batch count of 96. While a total of $n_1 = k_1 m_1 = 1536$ observations may be more than are actually needed in a few applications, such a sample size is usually easy and inexpensive to generate.

The first two batches of data are excluded from the computations of batch means statistics in an effort to address the start-up problem. Let $k_1' = k_1 - 2 = 94$ denote the number of batch means retained for confidence-interval construction from which we calculate the sample mean and variance

$$\overline{Y}(m_1, k_1') = \frac{1}{k_1'} \sum_{j=3}^{k_1} Y_j(m_1) \quad \text{and} \quad S^2_{m_1,k_1'} = \frac{1}{k_1'-1} \sum_{j=3}^{k_1} [Y_j(m_1) - \overline{Y}(m_1, k_1')]^2,$$

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 two batches from the entire data set accumulated so far.) The $k_1'$ retained batch means $\{Y_j(m_1) : j = 3, \ldots, k_1\}$ are tested for independence using von Neumann’s ratio of the sample mean-square successive difference to the sample variance (von Neumann 1941, Fishman 1978),

$$C_{k_1'} = 1 - \frac{\sum_{j=3}^{k_1-1} [Y_j(m_1) - Y_{j+1}(m_1)]^2}{2(k_1'-1)S^2_{m_1,k_1'}}.$$

If the batch means $\{Y_j(m_1) : j = 3, \ldots, k_1\}$ are i.i.d. normal and $k_1' \geq 25$, then the critical values of the standardized test statistic

$$Q_{k_1'} = C_{k_1'} / \sqrt{(k_1' - 2)/(k_1')^2 - 1}$$

are extremely close to the critical values of the $N(0, 1)$ distribution (Anderson 1971, Fishman 1978); and substantial positive (respectively, negative) lag-one correlation between the batch means results in significant positive (respectively, negative) values of the standardized test statistic (5) (Fishman 1978). In our experience the test statistic (5) based on a sample of size $k_1' = 94$ has sufficient power to detect a wide variety of departures from randomness; and this is one of the main reasons for the relatively large number of batches used in ASAP.
Since the studies detailed in Steiger and Wilson (2000a, 2001) show that the correlation function of the batch means process is not always a monotone decreasing function of the lag separating the batch means, we chose to use a two-sided test for the independence of the batch means with size $\alpha_{\text{ind}} = 0.20$. Thus we reject the hypothesis that the batch means $\{Y_j(m_1) : j = 3, \ldots, k_1\}$ are independent at the level of significance $\alpha_{\text{ind}}$ if $|Q_{k'_1}| > \Phi^{-1}(1 - \alpha_{\text{ind}}/2)$, where $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal distribution function. This is comparable to using a one-sided test of size $\alpha_{\text{ind}} = 0.10$ in the LBATCH and ABATCH procedures (Fishman and Yarberry 1997), where the hypothesis of independence is rejected if $Q_{k'_1} > \Phi^{-1}(1 - \alpha_{\text{ind}})$. If the $k'_1 = 94$ batch means $\{Y_j(m_1) : j = 3, \ldots, k_1\}$ pass the independence test, then the classical batch means confidence interval (4) is constructed with midpoint and half-length given respectively by $\bar{Y} = \bar{Y}(m_1, k'_1)$ and $H = t_{1-\alpha/2, k'_1-1} S_{m_1,k'_1}/\sqrt{k'_1}$. In this situation no adjustment is made to the confidence interval because presumably none is needed when the batch means are independent (Fishman 1978, Fishman and Yarberry 1997).

If the batch means fail the test for independence, then ASAP constructs 16 vectors each consisting of four adjacent batch means. Two batch means between each set of four are ignored in an effort to obtain approximately independent four-dimensional vectors of batch means as depicted in the following layout:

$$
\begin{align*}
&Y_3(m_1), Y_4(m_1), Y_5(m_1), Y_6(m_1), Y_7(m_1), Y_8(m_1), Y_9(m_1), Y_{10}(m_1), Y_{11}(m_1), Y_{12}(m_1), \\
&\text{1st (4x1) vector } y_1 \\
&\text{ignored} \\
&Y_{13}(m_1), Y_{14}(m_1), \ldots, Y_{93}(m_1), Y_{94}(m_1), Y_{95}(m_1), Y_{96}(m_1). \\
&\text{2nd (4x1) vector } y_2 \\
&\text{ignored} \\
&\text{16th (4x1) vector } y_{16}
\end{align*}
$$

We apply the Shapiro-Wilk test for multivariate normality (Malkovich and Afifi 1973) to the resulting 16 vectors, each consisting of four adjacent batch means. This is another reason for the choice of 96 initial batches. Although joint normality of these selected sets of $r = 4$ adjacent batch means is not sufficient to ensure joint normality of all $k'_1 = 94$ batch means (Kendall, Stuart and Ord 1987, p. 504), the results of an extensive experimental evaluation of ASAP’s performance strongly suggest that testing for joint normality in $r \times 1$ vectors
each composed of \( r = 4 \) adjacent batch means and separated by two batches yields good performance in many applications of ASAP.

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

\[
\bar{y} = g^{-1} \sum_{i=1}^{g} y_i \quad \text{and} \quad A = \sum_{i=1}^{g} (y_i - \bar{y})(y_i - \bar{y})^T.
\]

Throughout the rest of this discussion, we assume that \( A \) 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 \( y^\dagger \in \{y_i : i = 1, 2, \ldots, g\} \) for which

\[
(y^\dagger - \bar{y})^T A^{-1} (y^\dagger - \bar{y}) = \max_{i=1,\ldots,g} \left\{ (y_i - \bar{y})^T A^{-1} (y_i - \bar{y}) \right\}.
\]

We compute \( Z_i \equiv (y^\dagger - \bar{y})^T A^{-1} (y_i - \bar{y}) \) 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 \( \{\beta_i : i = 1, 2, \ldots, g\} \) 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

\[
W^* = \frac{\left[ \sum_{i=1}^{g} \beta_i Z_{(i)} \right]^2}{(y^\dagger - \bar{y})^T A^{-1} (y^\dagger - \bar{y})} \tag{6}
\]

(Malkovich and Afifi 1973). The null hypothesis of multinormal responses \( \{y_i\} \) is rejected at the level of significance \( \alpha_{mvn} \) \((0 < \alpha_{mvn} < 1)\) if \( W^* < w^*_{\alpha_{mvn}}(r, g) \), where \( w^*_{\alpha_{mvn}}(r, g) \) denotes the quantile of order \( \alpha_{mvn} \) 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).

During preliminary experimentation with ASAP on the same six processes that were used in our investigation of the convergence properties of batch means (Steiger 1999, Steiger and Wilson 2001), we found that ASAP delivered the best overall results in terms of confidence-interval coverage probability when we set \( \alpha_{ind} = 0.20 \) and \( \alpha_{mvn} = 0.10 \). Subsequently we
used these values in the comprehensive experimental performance evaluation of ASAP that is described in Section 7. All our experimentation with ASAP indicates that the procedure works well in a broad class of stochastic simulations when we take $\alpha_{\text{ind}} = 0.20$ and $\alpha_{\text{mvn}} = 0.10$. Gaining a more complete understanding of the joint effect of these two parameters on the performance of ASAP is the subject of ongoing research.

On the $i$th iteration of ASAP 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 ASAP will be required if the following conditions occur simultaneously on iteration $i$:

- a) the independence test yields a significant result (that is, the batch means fail the independence test) at the level of significance $\alpha_{\text{ind}}$ when this test is applied to the last $k'_i = k_i - 2$ batch means for batches of size $m_i$; and then

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

For iteration $i + 1$ the batch size and batch count are respectively taken to be

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

so that 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.


If the batch means fail the test for independence but pass the test for joint multivariate normality, then we seek to adjust the classical batch means CI (4) by taking into account
the deviation of the distribution of the classical NOBM ratio (3) from the desired Student’s $t$-distribution with $k-1$ degrees of freedom. Our adjustment is based on an inverted Cornish-Fisher expansion for (3) that involves the first four cumulants of (3). In the next section, we develop expressions for the first four moments and cumulants of (3) in terms of $\text{Var}[Y(m)]$ and $\text{Var}[\overline{Y}(m, k)]$. To compute sample estimators of $\text{Var}[Y(m)]$ and $\text{Var}[\overline{Y}(m, k)]$, we fit an autoregressive–moving average (ARMA) time series model (Box, Jenkins, and Reinsel 1994) to the sequence of batch means $\{Y_j(m) : j = 3, \ldots, k\}$. (In this section we suppress the index $i$ of the current iteration of ASAP to simplify the notation; no confusion can result from this simplification since the iteration remains the same throughout the discussion.) For the batch means variance estimator $\hat{\text{Var}}[Y(m)]$, we take the usual maximum likelihood estimator of the variance of the fitted ARMA process; and for the grand mean variance estimator $\hat{\text{Var}}[\overline{Y}(m, k)]$, we derive a similar statistic based on the estimated covariances between all relevant batch means expressed in term of the maximum likelihood estimators of the parameters of the fitted ARMA process.

If the batch means pass the test for multivariate normality detailed in Section 3, then an ARMA process of at most order 2 is fitted to the set of $k' = 94$ batch means. Obtaining a good result from the ARMA fitting process generally requires over 50 observations (Box, Jenkins, and Reinsel 1994, p. 17), which is the final reason for choosing an initial number of batches close to 100. The following ARMA models may be fitted to the batch means $\{Y_j(m) : j = 3, \ldots, k\}$: AR(1), AR(2), MA(1), MA(2), and ARMA(1,1). Adapting the notation in Box, Jenkins, and Reinsel (1994) to the notation used here, we let $\{\tilde{Y}_j(m) \equiv Y_j(m) - \mu_X : j = 3, \ldots, k\}$ denote the corresponding deviations from the steady-state mean $\mu_X$. The $\ell$th observation of an AR(1) process can be expressed as

$$\tilde{Y}_\ell = \varphi_1 \tilde{Y}_{\ell-1} + a_\ell \quad \text{for} \quad \ell = 1, 2, \ldots, \quad (7)$$

where $\varphi_1$ is the autoregressive parameter and $a_\ell$ is an independent normal “shock” with mean zero and variance $\sigma_a^2$. Similarly, for the second-order autoregressive process, AR(2), we have $\tilde{Y}_\ell = \varphi_1 \tilde{Y}_{\ell-1} + \varphi_2 \tilde{Y}_{\ell-2} + a_\ell$ for $\ell = 1, 2, \ldots$. Fits of first- and second-order moving
average processes are also performed, where $\bar{Y}_\ell = a_\ell - \theta_1 a_{\ell-1}$ for $\ell = 1, 2, \ldots$, represents the MA(1) process; and $\bar{Y}_\ell = a_\ell - \theta_1 a_{\ell-1} - \theta_2 a_{\ell-2}$ for $\ell = 1, 2, \ldots$, is the MA(2) process. We also fit an ARMA(1,1) time series model, given by $\bar{Y}_\ell = \varphi_1 \bar{Y}_{e-1} + a_\ell - \theta_1 a_{\ell-1}$ for $\ell = 1, 2, \ldots$.

We believe this collection of ARMA time series models provides a sufficiently broad diversity of correlation structures to account for the principal features of the dependencies between batch means that affect the performance of ASAP—and the experimental results presented in Section 7 provide substantial evidence to support this belief.

IMSL routines (IMSL Problem Solving Software Systems 1987) are used to estimate the autoregressive–moving average parameters, the residual variance, $\sigma_a^2$, and the process variance, $\operatorname{Var}[\bar{Y}_\ell]$, for the five ARMA models. Then the “best” fit of the five models is chosen. Preference is given to the AR(1) model (7). If the MA(1) model has a residual variance less than or equal to 75% of the residual variance from the AR(1) model, then it is chosen as the “best-fitting” first-order model. A second-order model (that is, AR(2), MA(2), or ARMA(1,1)) is chosen only if its residual variance is less than or equal to 50% of the residual variance of the “best-fitting” first-order model. Although we found this purely ad hoc approach for fitting an ARMA model to the batch means yielded more parsimonious models with much less computational overhead than a model selection procedure based on Akaike’s Information Criterion (Akaike 1974), we found the performance of ASAP (in terms of coverage probability for the delivered confidence intervals) was nearly the same with both approaches.

The estimators of $\operatorname{Var}[Y(m)]$ and the parameters from the ARMA fit are then used to estimate $\widehat{\operatorname{Var}}[\bar{Y}(m, k')]$:

$$\widehat{\operatorname{Var}}[\bar{Y}(m, k')] = \frac{1}{k'} \sum_{q=-k'+1}^{k'-1} \left(1 - \frac{|q|}{k'}\right) \hat{\gamma}_m(q),$$

where $\hat{\gamma}_m(q)$ denotes the estimated lag-$q$ covariance of the batch means $\{Y_j(m) : j = 3, \ldots, k\}$ based on the fitted time series model. For an AR(1) process (7), the covariance at lag $q$ is given by $\operatorname{Cov}(\bar{Y}_\ell, \bar{Y}_{\ell+q}) = \varphi_1 |q| \sigma_a^2 / (1 - \varphi_1^2)$, for $q = 0, \pm 1, \pm 2, \ldots$. Thus if (7) is an adequate
model of the batch means process for batches of size \( m \) and if \( \hat{\varphi}_1 \) and \( \hat{\sigma}_a \) denote the usual maximum likelihood estimates of \( \varphi_1 \) and \( \sigma_a \) respectively (Box, Jenkins, and Reinsel 1994, Chapter 7), then the estimated covariances in (8) are

\[
\hat{\gamma}_m(q) = \frac{\hat{\varphi}_1|q|}{1 - \hat{\varphi}_1^2} \hat{\sigma}_a^2 \quad \text{for} \quad q = 0, \pm 1, \pm 2, \ldots ;
\] (9)

and a similar approach is taken when other ARMA models are fitted to the batch means process for a given batch size \( m \). In practice, we have found that ASAP almost always uses the result (9) for the fitted AR(1) model to compute the overall variance estimator (8). See Steiger (1999) for complete details on the time series estimation techniques used in ASAP.

5. Confidence Interval Adjustment for Dependent Batch Means

In this section, we formulate an adjustment to the usual CI (4) that accounts for dependency between the batch means. The adjustment is based on the first four noncentral moments of the \( t \)-ratio (3). To simplify the discussion, we let

\[
N \equiv \frac{\sqrt{k \left( \bar{Y}(m, k) - \mu_X \right)}}{\sqrt{\text{Var}[\bar{Y}(m)]}} \quad \text{and} \quad D \equiv \sqrt{\frac{S_{m,k}^2}{\text{Var}[\bar{Y}(m)]}}
\] (10)

respectively denote the numerator and denominator of the \( t \)-ratio on the right-hand side of equation (3). To compute the moments of (3), we make the following key assumptions.

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

\( A_2 \): As defined in (10), the numerator \( N \) and denominator \( D \) of the \( t \)-ratio (3) are independent.

\( A_3 \): The squared denominator \( D^2 \) of the \( t \)-ratio (3) is distributed as \( \chi^2_{k-1}/(k - 1) \).

Theoretical and experimental evidence from Steiger and Wilson (2001) supports the reasonableness of these assumptions.
Exploiting assumptions A1–A3, we first derive expressions for appropriate moments of $N$ and $D$ separately. First we observe that assumption A1 yields

$$E[N] = E[N^3] = 0 \tag{11}$$

since $N$ is normal with mean zero when its component batch means are jointly normal. An immediate consequence of the definition of $N$ is that

$$E[N^2] = \frac{k}{\text{Var}[Y(m)\]} E\left\{ \left[ \bar{Y}(m,k) - \mu_X \right]^2 \right\} = \frac{k \text{Var}[\bar{Y}(m,k)]}{\text{Var}[Y(m)]}. \tag{12}$$

Assumption A1 ensures that the fourth moment of $N$ is

$$E[N^4] = \frac{3k^2 \text{Var}^2[\bar{Y}(m,k)\]}{\text{Var}^2[Y(m)]} \tag{13}$$

since $[\bar{Y}(m,k) - \mu_X]/\text{Var}^{1/2}[\bar{Y}(m,k)] \sim N(0,1)$ when the batch means are multivariate normal, and the fourth moment of a $N(0,1)$ random variable is 3. Under assumption A3, straightforward calculation reveals that

$$E[D^{-p}] = \frac{\Gamma\left(\frac{k-p-1}{2}\right)(\frac{k-1}{2})^{p/2}}{\Gamma\left(\frac{k-1}{2}\right)} \text{ for } p = 1, 2, \ldots \text{ and } k \geq p + 2. \tag{14}$$

Using (11)–(14) and assumption A2, we compute the first four noncentral moments of (3). From (11), (14), and assumption A2, we have

$$\mu_p \equiv E[t^p] = E[N^p] E[D^{-p}] = 0 \text{ for } p = 1, 3 \text{ and } k \geq 5. \tag{15}$$

From (12), (14), and assumption A2, we obtain

$$\mu_2 \equiv E[t^2] = E[N^2] E[D^{-2}] = \frac{k \text{Var}[\bar{Y}(m,k)\]}{\text{Var}[Y(m)\]} \frac{(k-1)}{(k-3)} \text{ for } k \geq 4. \tag{16}$$

Similarly, from (13), (14), and assumption A2 we see that

$$\mu_4 \equiv E[t^4] = E[N^4] E[D^{-4}] = \frac{3k^2 \text{Var}^2[\bar{Y}(m,k)\]}{\text{Var}^2[Y(m)]} \frac{(k-1)^2}{(k-3)(k-5)} \text{ for } k \geq 6. \tag{17}$$

The adjusted confidence intervals used in ASAP are based on an inverted Cornish-Fisher expansion for the NOBM $t$-ratio (3) that involves the cumulants $\kappa_1$, $\kappa_2$, $\kappa_3$ and $\kappa_4$ of (3).
From (15), (16), and (17) together with Section 3.14 of Kendall, Stuart, and Ord (1987), we have

\[ \kappa_p = \mu_p = 0 \text{ for } p = 1, 3 \text{ and } k \geq 5, \quad (18) \]

\[ \kappa_2 = \mu_2 = \frac{k \text{Var}[\bar{Y}(m, k)](k - 1)}{\text{Var}[Y(m)](k - 3)} \text{ for } k \geq 4, \quad (19) \]

and

\[ \kappa_4 = \mu_4 - \mu_2^2 = \frac{2k^2(k - 1)^2 \text{Var}^2[\bar{Y}(m, k)]}{(k - 3)^2(k - 5)\text{Var}^2[Y(m)]} \text{ for } k \geq 6. \quad (20) \]

In terms of the cumulants (18)–(20), we obtain the following adjusted 100(1 – α)% confidence intervals for \( \mu_X \) based on \( k \) batch means for batches of size \( m \)

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

where

\[ h'(z_{1 - \alpha/2}) = z_{1 - \alpha/2} + (\kappa_1 - \kappa_3/6) + [(\kappa_2 - 1)/2 - \kappa_4/8] z_{1 - \alpha/2} \]

\[ + (\kappa_3/6) z_{1 - \alpha/2}^2 + (\kappa_4/24) z_{1 - \alpha/2}^3 \]

(Hall 1983, Chien 1989).

Adapting these results to each iteration of ASAP, we see that under the assumptions A1–A3 the first four cumulants of the ASAP \( t \)-ratio \( \{\bar{Y}(m, k') - \mu_X \}/[S_{m,k'}/\sqrt{k'}] \) are given by (18)–(20) when \( k \) is replaced by \( k' \). In the expressions (19) and (20) for \( \kappa_2 \) and \( \kappa_4 \), we replace the quantities \( \text{Var}[Y(m)] \) and \( \text{Var}[\bar{Y}(m, k)] \) respectively by the variance estimators \( \hat{\text{Var}}[Y(m)] = \hat{\gamma}_m(0) \) and \( \hat{\text{Var}}[\bar{Y}(m, k')] \) that are obtained from standard results similar to (9) and from relation (8) by fitting an ARMA process to the batch means \( \{Y_j(m) : j = 3, \ldots, k\} \); and this procedure yields the following approximate 100(1 – α)% confidence interval for \( \mu_X \):

\[ \bar{Y}(m, k') \pm \left[ z_{1 - \alpha/2} \left( 1 + \frac{\hat{\kappa}_2 - 1}{2} - \frac{\hat{\kappa}_4}{8} \right) + \frac{\hat{\kappa}_4}{24} z_{1 - \alpha/2}^3 \right] \sqrt{\frac{\text{Var}[Y(m)]}{k'}}. \quad (21) \]
6. Fulfilling the Precision Requirement

The final step in ASAP is to determine if the constructed confidence interval meets the user’s precision requirement. 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 multivariate normality test (that is, the batch means pass the test for multivariate normality). If the relevant requirement

$$H \leq H^* \quad \text{or} \quad H \leq r^*|\overline{Y}|$$

(22)

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 (22) is not satisfied on iteration $i$ of ASAP, then the procedure estimates the number of additional batches $k_i^+$ required to satisfy (22) using batch size $m_i$

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

(23)

thus on iteration $i + 1$ of ASAP the batch count and batch size are tentatively taken to be $k_{i+1} \leftarrow k_i + k_i^+$ and $m_{i+1} \leftarrow m_i$ so that the total required sample size on iteration $i + 1$ is tentatively taken to be $n'_{i+1} \leftarrow k_{i+1}m_{i+1}$.

For ease in coding ASAP, we specified an upper limit on the number of batches that the algorithm could require. Preliminary experiments with ASAP revealed that, except in the most difficult cases, 1500 batches were sufficient to meet relative precision requirements as small as ±5%. If the total number of batches $k_{i+1}$ exceeds 1502, then a new tentative batch size, $m_{i+1} \leftarrow \lfloor \sqrt{2}m_{i+1} \rfloor$, is calculated and a new tentative number of batches $k_{i+1} \leftarrow \lfloor n'_{i+1}/m_{i+1} \rfloor$ is computed. The tentative batch size $m_{i+1}$ is repeatedly increased by a factor of $\sqrt{2}$ and the resulting $k_{i+1}$ is recomputed until $k_{i+1} \leq 1502$. Thus if the user-specified precision requirement (22) is not satisfied on iteration $i$ of ASAP, then iteration $i + 1$ will be required in which the number of batches $k_{i+1} \leq 1502$ and the total sample size is finally taken to be $n_{i+1} \leftarrow m_{i+1}k_{i+1}$; 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 of the entire simulation-generated data set are again omitted from the calculation of the sample mean. The batch means of $k'_{i+1} = (k_{i+1} - 2)$ 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 $k'_{i+1}$ batches of size $m_{i+1}$; moreover, in this situation new estimates of $\text{Var}[Y(m_{i+1})]$, $\text{Var}[Y(m_{i+1}, k'_{i+1})]$, $\kappa_2$ and $\kappa_4$ are computed, and the updated CI (21) is constructed. If the classical NOBM confidence interval (4) was used on the previous iteration, then (4) is updated using the current batch means for $k'_{i+1}$ batches of size $m_{i+1}$. If the precision requirement (22) is satisfied on iteration $i + 1$ of ASAP, then the algorithm terminates, returning $\bar{Y} = \bar{Y}(m_{i+1}, k'_{i+1})$ and $H$. If the required precision is not achieved on iteration $i + 1$ of ASAP, then the rest of iteration $i + 1$ of ASAP proceeds along the same lines as described above.

Iterations that are performed for the purpose of fulfilling the precision requirement do not involve retesting the batch means for independence or multivariate normality. Once a sample size has been determined based on either the construction of a classical NOBM CI or an adjusted CI, then further iterations employ the same type of CI. We felt that the precision requirement should not affect the mode of CI construction; rather the dependency structure and distributional characteristics of the observations should be the basis for the type of CI ultimately constructed. In cases where we are required to take large sample sizes and therefore have large final batch sizes, the corresponding batch means may indeed appear to be approximately independent. However, as can be theoretically anticipated, we found that in such cases the adjustment to the final CI was generally very small.

7. Experimental Performance Evaluation

To evaluate the performance of ASAP 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 ASAP together with the ABATCH and LBATCH algorithms (Fishman 1996, 1998; Fishman and Yarberry 1997) to a suite of twenty test problems. This suite includes some standard problems used for testing simulation output analysis procedures, some problems which more closely resemble real-world applications, and some problems possessing characteristics which we believe will stress any output analysis procedure—namely, a pronounced, slowly decaying correlation function or markedly nonnormal marginal distributions (or both). Included in our twenty test problems are the fourteen stochastic models that Law and Carson (1979) used to test their batch means algorithm. In this section we summarize the results of our experimentation on four of the test problems. The steady-state mean response is available analytically for each of these test problems; thus we were able to evaluate the performance of ABATCH, ASAP, and LBATCH in terms of actual versus nominal coverage probabilities for the confidence intervals delivered by each of these procedures. Experimental results for the sixteen remaining test problems are not presented here because they contribute little additional insight into the relative performance of the algorithms. See Steiger (1999) or Steiger and Wilson (2000a) for complete details on the experimental performance evaluation for all twenty test problems.

For each test problem to be simulated, we performed 100 independent replications of each batch means procedure to construct nominal 90% confidence intervals that satisfy three different precision requirements:

(a) no precision requirement—that is, we continued the simulation of each test problem until ASAP delivered a confidence interval based on 94 batches of the size at which the batch means passed either the statistical test for independence (5) or the test for multivariate normality (6) without considering a precision requirement;

(b) ±15% precision—that is, we continued the simulation of each test problem until ASAP delivered a CI that satisfied (22) with $r^* = 0.15$; and

(c) ±7.5% precision—that is, we continued the simulation of each test problem until ASAP
delivered a CI that satisfied (22) with \( r^* = 0.075 \).

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 ASAP. Based on all our computational experience with ASAP, we believe that the results given below are typical of the performance of ASAP that can be expected in many practical applications. For a number of reasons elaborated in §7.1.2, it is not clear that a similar statement can be made about ABATCH and LBATCH; nevertheless, the results given below do provide an arguably fair basis for comparing the performance of ABATCH, LBATCH, and ASAP. Since each confidence interval with a nominal coverage probability of 90% was replicated 100 times, the standard error of each coverage estimator is approximately 3%. 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 ASAP versus ABATCH and LBATCH on many of the test problems.

7.1. Results for Selected Test Problems

7.1.1. Discrete-Time Markov Chain

The first test problem consists of a cost function defined on a simple two-state discrete-time Markov chain (DTMC) whose one-step transition probability matrix and cost function are, respectively,

\[
P = \begin{pmatrix}
0 & 1 \\
0.99 & 0.01
\end{pmatrix}
\quad \text{and} \quad
h = \begin{pmatrix}
0 & 1 \\
5 & 10
\end{pmatrix}.
\]

(24)

The results for this test problem are summarized in Table 1.

ASAP showed somewhat better confidence-interval coverage than did ABATCH and LBATCH in the case of the two-state Markov chain (24), especially in the cases of no precision requirement and a precision requirement of ±15%. For this test problem, ASAP delivered correlation-adjusted CIs based on a nonsignificant result from the test for multi-
Table 1  Performance of Batch Means Procedures for the Two-State DTMC Defined by (24) Based on 100 Independent Replications of Nominal 90% Confidence Intervals

<table>
<thead>
<tr>
<th>Precision Requirement</th>
<th>Procedure</th>
<th>LBATCH</th>
<th>ABATCH</th>
<th>ASAP&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO PRECISION</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. sample size</td>
<td></td>
<td>3036</td>
<td></td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td></td>
<td>70%</td>
<td>85%</td>
<td>96%</td>
</tr>
<tr>
<td>avg. CI half-length</td>
<td></td>
<td>0.515</td>
<td>0.642</td>
<td>1.20</td>
</tr>
<tr>
<td>var. CI half-length</td>
<td></td>
<td>0.009</td>
<td>0.012</td>
<td>0.172</td>
</tr>
<tr>
<td>±15% PRECISION</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. sample size</td>
<td></td>
<td>5171</td>
<td></td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td></td>
<td>72%</td>
<td>81%</td>
<td>96%</td>
</tr>
<tr>
<td>avg. CI half-length</td>
<td></td>
<td>0.045</td>
<td>0.053</td>
<td>0.906</td>
</tr>
<tr>
<td>var. CI half-length</td>
<td></td>
<td>0.011</td>
<td>0.010</td>
<td>0.023</td>
</tr>
<tr>
<td>±7.5% PRECISION</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. sample size</td>
<td></td>
<td>22711</td>
<td></td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td></td>
<td>81%</td>
<td>86%</td>
<td>99%</td>
</tr>
<tr>
<td>avg. CI half-length</td>
<td></td>
<td>0.253</td>
<td>0.284</td>
<td>0.438</td>
</tr>
<tr>
<td>var. CI half-length</td>
<td></td>
<td>0.003</td>
<td>0.003</td>
<td>0.006</td>
</tr>
</tbody>
</table>

<sup>a</sup>No. of classical and corrected CIs generated by ASAP: 0 and 100, respectively.

variate normality (that is, the batch means passed the Shapiro-Wilk test (6) for multivariate normality) on all 100 replications of ASAP. The CIs from ASAP are wider than those from ABATCH and LBATCH, which is necessary for the improved coverage. However, the coefficient of variation of the CI half-lengths delivered by ASAP are smaller than those delivered by LBATCH and ABATCH.

### 7.1.2. Queueing Systems

We applied ABATCH, LBATCH, and ASAP to the waiting time process in the $M/M/1$ queue with server utilization $\tau = 0.9$ and an empty-and-idle initial condition. This is a particularly difficult test problem for several reasons: (a) the initialization bias is large
and decays relatively slowly (Wilson and Pritsker 1978); (b) in steady-state operation the correlation 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. The experimental results for the $M/M/1$ queue waiting times are summarized in Table 2. This test problem most dramatically displays one of the advantages of the ASAP algorithm—namely, that ASAP does not rely solely on the von Neumann (1941) test for independence. In fact, in 96 out of 100 replications of the procedure, ASAP delivered correlation-adjusted CIs of the form (21).

As can be seen from Table 2, ASAP substantially outperforms ABATCH and LBATCH for the case of no precision requirement. As we demand more precision, we are of course forced to perform more sampling. For the precision requirement of $\pm 7.5\%$, the three algorithms gave similar results. This suggests that ABATCH and LBATCH will give satisfactory results 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 ASAP is that it usually determines a sample size sufficient to yield acceptable results, even when no precision requirement is specified.

Table 3 displays the additional results for the $M/M/1$ queue waiting times with $\tau = 0.9$ obtained through standalone application of LBATCH and ABATCH operating with a stopping rule based on a user-specified precision requirement for the final confidence interval. After we performed the simulation with an initial run length of 1536 observations (the same initial sample size required for ASAP), we applied the precision requirement to the final CI constructed by LBATCH or ABATCH. If the precision requirement was not satisfied, then we calculated an estimate similar to (23) for the number of additional observations needed to satisfy the precision requirement, we generated the additional observations, and we executed LBATCH or ABATCH again with all the accumulated observations. This
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

<table>
<thead>
<tr>
<th>Precision Requirement</th>
<th>Procedure</th>
<th>LBATCH</th>
<th>ABATCH</th>
<th>ASAP$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO PRECISION</td>
<td>avg. sample size</td>
<td>44%</td>
<td>60%</td>
<td>83%</td>
</tr>
<tr>
<td></td>
<td>coverage</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>avg. CI half-length</td>
<td>1.70</td>
<td>2.67</td>
<td>11.8</td>
</tr>
<tr>
<td></td>
<td>var. CI half-length</td>
<td>0.683</td>
<td>3.92</td>
<td>523.0</td>
</tr>
<tr>
<td>±15% PRECISION</td>
<td>avg. sample size</td>
<td></td>
<td></td>
<td>298950</td>
</tr>
<tr>
<td></td>
<td>coverage</td>
<td>79%</td>
<td>80%</td>
<td>88%</td>
</tr>
<tr>
<td></td>
<td>avg. CI half-length</td>
<td>0.543</td>
<td>0.613</td>
<td>0.783</td>
</tr>
<tr>
<td></td>
<td>var. CI half-length</td>
<td>0.027</td>
<td>0.039</td>
<td>0.082</td>
</tr>
<tr>
<td>±7.5% PRECISION</td>
<td>avg. sample size</td>
<td></td>
<td></td>
<td>815755</td>
</tr>
<tr>
<td></td>
<td>coverage</td>
<td>88%</td>
<td>90%</td>
<td>94%</td>
</tr>
<tr>
<td></td>
<td>avg. CI half-length</td>
<td>0.353</td>
<td>0.382</td>
<td>0.413</td>
</tr>
<tr>
<td></td>
<td>var. CI half-length</td>
<td>0.012</td>
<td>0.039</td>
<td>0.018</td>
</tr>
</tbody>
</table>

$^a$No. of classical and corrected CIs generated by ASAP: 4 and 96, respectively.

process was repeated until the final CI delivered by LBATCH or ABATCH satisfied the precision requirement. Although LBATCH and ABATCH were not necessarily designed to be used in this way, we believe that this stopping rule is a natural approach to planning steady-state simulations and that the results in Table 3 provide a more complete perspective on the relative performance of LBATCH and ABATCH versus ASAP. Since our applications of ABATCH and LBATCH were completely automated in order to perform 100 replications of each procedure, we did not manually analyze the convergence of the sample estimators delivered by LBATCH and ABATCH on each application of these procedures along the lines suggested in Fishman (1998). We believe that the results of Tables 2 and 3 highlight the
performance advantages achieved by ASAP without requiring analysis or manual intervention by the user.

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

<table>
<thead>
<tr>
<th>Precision Requirement</th>
<th>Procedure</th>
<th>LBATCH</th>
<th>ABATCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO PRECISION</td>
<td>avg. sample size</td>
<td>1536</td>
<td>1536</td>
</tr>
<tr>
<td></td>
<td>coverage</td>
<td>35%</td>
<td>54%</td>
</tr>
<tr>
<td></td>
<td>avg. CI half-length</td>
<td>1.648</td>
<td>2.882</td>
</tr>
<tr>
<td></td>
<td>var. CI half-length</td>
<td>0.552</td>
<td>4.250</td>
</tr>
<tr>
<td>±15% PRECISION</td>
<td>avg. sample size</td>
<td>34349</td>
<td>50910</td>
</tr>
<tr>
<td></td>
<td>coverage</td>
<td>65%</td>
<td>77%</td>
</tr>
<tr>
<td></td>
<td>avg. CI half-length</td>
<td>1.071</td>
<td>1.080</td>
</tr>
<tr>
<td></td>
<td>var. CI half-length</td>
<td>0.0513</td>
<td>0.0336</td>
</tr>
<tr>
<td>±7.5% PRECISION</td>
<td>avg. sample size</td>
<td>227987</td>
<td>397387</td>
</tr>
<tr>
<td></td>
<td>coverage</td>
<td>80%</td>
<td>81%</td>
</tr>
<tr>
<td></td>
<td>avg. CI half-length</td>
<td>0.551</td>
<td>0.553</td>
</tr>
<tr>
<td></td>
<td>var. CI half-length</td>
<td>0.005</td>
<td>0.007</td>
</tr>
</tbody>
</table>

From Table 3 we see that for $M/M/1$ queue waiting times with server utilization $\tau = 0.9$, 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 in Table 3 are much smaller than those required by ASAP to achieve the same precision. For example, an average sample size of 397,387 was used by ABATCH to satisfy a precision requirement of $\pm 7.5\%$. This is considerably less than the average sample size of 815,755 required by ASAP. For a precision requirement of $\pm 7.5\%$ and 90% confidence-interval coverage probability, Whitt’s
(1989) approximation for estimating the run length required by a queueing simulation yields an estimated sample size of 855,238 for this test problem. This latter result suggests that ASAP yields adequate sample sizes when a precision requirement is specified.

Table 4 displays results for the tandem $M/M/1/M/1$ queue with server utilization $\tau = 0.8$ (expected response $\mu_\lambda = 6.40$) and an empty-and-idle initial condition. This case is one of the eight queueing systems from Law and Carson (1979) which were included in our evaluation. From Table 4 we see that for the $M/M/1/M/1$ queue, ASAP performed better than ABATCH and LBATCH for the no-precision requirement and for the precision requirement of $\pm 15\%$. With a precision requirement of $\pm 7.5\%$, ASAP, LBATCH, and ABATCH all delivered coverage close to the nominal level—although again these results are based on sample sizes determined by running ASAP first.

7.1.3. Computer Models

One of the computer models used by Law and Carson (1979) consists of a central server (the CPU, or workcenter 1) and $M - 1$ peripheral units referred to as workcenters 2 through $M$. The system has a fixed number jobs, $J$, in it. When a job is finished at the CPU, it leaves the system with probability $p_1$ and is immediately replaced with another job at the CPU queue. If the job does not leave the system, then it is routed to peripheral unit $i$ with probability $p_i$ for $i = 2, \ldots, M$. After getting service at one of the peripheral units, the job leaves the system and is immediately replaced by a job joining the CPU queue. The process of interest is the response time of a job, i.e., the time between the job’s arrival at the CPU queue and its departure from the system. Law and Carson chose to simulate this model for four cases. We present results of applying ASAP, LBATCH, and ABATCH for Law and Carson’s (1979) third case in which there are $J = 8$ jobs and $M = 3$ workcenters with respective service rates $\mu_1 = 1.0$, $\mu_2 = 0.45$, and $\mu_3 = 0.05$. Initially there are 5 jobs at the CPU, 1 job at peripheral unit 2, and 2 jobs at peripheral unit 3. In this system the steady-state utilizations at the three workcenters are 0.44, 0.88, and 0.88, respectively; and
Table 4  Performance of Batch Means Procedures for the $M/M/1/M/1$ Queue Waiting Time Process with $\tau = 0.8$ Based on 100 Independent Replications of Nominal 90% Confidence Intervals

<table>
<thead>
<tr>
<th>Precision Requirement</th>
<th>Procedure</th>
<th>LBATCH</th>
<th>ABATCH</th>
<th>ASAP$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO PRECISION</td>
<td>avg. sample size</td>
<td></td>
<td></td>
<td>3152</td>
</tr>
<tr>
<td></td>
<td>coverage</td>
<td>65%</td>
<td>75%</td>
<td>85%</td>
</tr>
<tr>
<td></td>
<td>avg. CI half length</td>
<td>1.053</td>
<td>1.476</td>
<td>3.250</td>
</tr>
<tr>
<td></td>
<td>var. CI half length</td>
<td>0.119</td>
<td>0.585</td>
<td>14.06</td>
</tr>
<tr>
<td>±15% PRECISION</td>
<td>avg. sample size</td>
<td></td>
<td></td>
<td>46610</td>
</tr>
<tr>
<td></td>
<td>coverage</td>
<td>80%</td>
<td>80%</td>
<td>93%</td>
</tr>
<tr>
<td></td>
<td>avg. CI half length</td>
<td>0.438</td>
<td>0.465</td>
<td>0.649</td>
</tr>
<tr>
<td></td>
<td>var. CI half length</td>
<td>0.016</td>
<td>0.018</td>
<td>0.030</td>
</tr>
<tr>
<td>±7.5% PRECISION</td>
<td>avg sample size</td>
<td></td>
<td></td>
<td>117339</td>
</tr>
<tr>
<td></td>
<td>coverage</td>
<td>85%</td>
<td>87%</td>
<td>90%</td>
</tr>
<tr>
<td></td>
<td>avg. CI half length</td>
<td>0.266</td>
<td>0.281</td>
<td>0.318</td>
</tr>
<tr>
<td></td>
<td>var. CI half length</td>
<td>0.005</td>
<td>0.005</td>
<td>0.008</td>
</tr>
</tbody>
</table>

$^a$No. of classical and corrected CIs generated by ASAP: 6 and 93, respectively.

the steady-state expected response time for each job is 18.279.

Table 5 reveals that in this case, the coverage losses incurred with all three procedures are serious but not catastrophic. We also ran this model with a precision requirement of ±2% and observed 85% coverage for the nominal 90% CIs constructed by ASAP. In this system LBATCH and ABATCH perform similarly to ASAP. For results from the three algorithms on the other computer models used by Law and Carson (1979), see Steiger and Wilson (2000a) and Steiger (1999).
Table 5  Performance of Batch Means Procedures for the Central Server Model 3
Based on 100 Independent Replications of Nominal 90% Confidence Intervals

<table>
<thead>
<tr>
<th>Precision Requirement</th>
<th>Procedure</th>
<th>LBATCH</th>
<th>ABATCH</th>
<th>ASAP&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO PRECISION</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. sample size</td>
<td></td>
<td></td>
<td></td>
<td>2277</td>
</tr>
<tr>
<td>coverage</td>
<td></td>
<td>75%</td>
<td>79%</td>
<td>78%</td>
</tr>
<tr>
<td>avg. CI half length</td>
<td></td>
<td>1.33</td>
<td>1.40</td>
<td>1.35</td>
</tr>
<tr>
<td>var. CI half length</td>
<td></td>
<td>0.107</td>
<td>0.163</td>
<td>0.135</td>
</tr>
<tr>
<td>±15% PRECISION</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. sample size</td>
<td></td>
<td>75%</td>
<td>79%</td>
<td>78%</td>
</tr>
<tr>
<td>coverage</td>
<td></td>
<td>1.33</td>
<td>1.40</td>
<td>1.35</td>
</tr>
<tr>
<td>avg. CI half length</td>
<td></td>
<td>0.107</td>
<td>0.163</td>
<td>0.135</td>
</tr>
<tr>
<td>var. CI half length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>±7.5% PRECISION</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. sample size</td>
<td></td>
<td>75%</td>
<td>76%</td>
<td>79%</td>
</tr>
<tr>
<td>coverage</td>
<td></td>
<td>1.08</td>
<td>1.11</td>
<td>1.05</td>
</tr>
<tr>
<td>avg. CI half length</td>
<td></td>
<td>0.037</td>
<td>0.047</td>
<td>0.028</td>
</tr>
<tr>
<td>var. CI half length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>No. of classical and corrected CIs generated by ASAP: 76 and 24, respectively.

7.2. Computational Overhead of ASAP

The 20 test problems and 100 replications of each test problem provided us with the opportunity to observe the performance of ASAP in 2000 applications. The most computationally intensive portion of the ASAP algorithm is the batching process. A delay in response time was observed when the sample size on an iteration was large. In the worst case—that is, in the test problem requiring the largest sample sizes as displayed in Table 2—the average sample size for the no-precision scenario is 7719, which implies about 6 iterations involving 96 batches of up to an average of 80 observations were required before one of the statistical tests yielded a nonsignificant result.

In almost all instances, once the batch means passed either the independence test or
the multivariate normality test, ASAP required only one additional iteration to satisfy the precision requirement. Therefore, the number of batching operations, and hence the computation time, is not onerous even in the most difficult cases. Given a sample size, any batching scheme would require similar time to execute. For instance, with large sample sizes, LBATCH and ABATCH involve many iterations, each of which calls for rebatching the observations (Fishman 1998, Fishman and Yarberry 1997). The other operations involved in the ASAP algorithm—testing for independence and multivariate normality, fitting the ARMA model, and calculation of the correlation-adjusted CI—require negligible computer time.

8. Conclusions and Recommendations

Batching schemes to date have ignored the question of normality based on the assumption that if the batch size is large enough for the batch means to be approximately independent, then the batch size is large enough for the batch means to be approximately normally distributed. These schemes have focused on selecting a batch size large enough to achieve near independence of the batch means. The method of determining whether the batch means are independent varies from scheme to scheme. ABATCH and LBATCH, for instance, rely on the von Neumann test for independence. ASAP is the first method to recognize the frequently occurring phenomenon of approximate multivariate normality of the batch means being achieved at a smaller batch size than is required to achieve approximate independence of the batch means insofar as these properties affect the performance of NOBM analysis procedures; and ASAP exploits this phenomenon when it is detected so as to compensate for any remaining dependence between the batch means.

The results presented in this paper and in Steiger (1999) and Steiger and Wilson (2000a) suggest further research involving refinements to the structure and operation of ASAP. We believe that additional improvements in the performance of ASAP may be achieved by replacing assumption A_3 with a more realistic hypothesis from which more accurate estimates of the first four cumulants of the NOBM t-ratio (3) may be derived. Since ASAP showed
the poorest coverage in cases where the majority of the CIs were classical NOBM CIs (based on the acceptance of the hypothesis of independence), further investigation of the role of the independence test in the ASAP algorithm may produce an alternative that will yield better coverage in these cases.

Acknowledgments

The authors thank David Goldsman (Georgia Institute of Technology) and Stephen D. Roberts (North Carolina State University) for many enlightening discussions on this paper.

References


Wilson, J. R., and A. A. B. Pritsker.  1978. Evaluation of startup policies in simulation experiments.  SIMULATION 31(3) 79–89.