Proceedings of the 2000 Winter Simulation Conference
J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds.

EXPERIMENTAL PERFORMANCE EVALUATION OF BATCH MEANS PROCEDURES FOR SIMULATION OUTPUT ANALYSIS

Natalie M. Steiger

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

James R. Wilson

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

ABSTRACT

We summarize the results of an extensive experimental performance evaluation of selected batch means procedures for building a confidence interval for a steady-state expected simulation response. We compare the performance of the well-known ABATCH and LBATCH procedures versus ASAP, a recently proposed variant of the method of nonoverlapping batch means (NOBM) that 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 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. Applying ABATCH, ASAP, and LBATCH to the analysis of a suite of twenty test problems involving discrete-time Markov chains, time-series processes, and queueing systems, we found ASAP to deliver confidence intervals that not only satisfy a user-specified absolute or relative precision requirement but also frequently outperform the corresponding confidence intervals delivered by ABATCH and LBATCH with respect to coverage probability.

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 $\overline{X}(n) = n^{-1}\sum_{i=1}^{n} X_i$. 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 $0 < \alpha < 1$. The CI for μ_X should satisfy two criteria: (a) it is narrow enough to be informative, and (b) its actual coverage probability is close to the nominal level $1-\alpha$.

In the simulation analysis method of nonoverlapping batch means (NOBM), 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 $n = km$. The sample mean for the jth batch is

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

and the grand mean of the individual batch means,

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

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

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

$$S^2_{m,k} = \frac{1}{k-1}\sum_{j=1}^{k}\left[Y_j(m) - \overline{Y}(m,k)\right]^2. \tag{3}$$

As $m \to \infty$ with k fixed so that $n \to \infty$, an asymptotically valid $100(1-\alpha)\%$ confidence interval for μ_X is

$$\overline{Y}(m,k) \pm t_{1-\alpha/2,k-1}\frac{S_{m,k}}{\sqrt{k}}. \tag{4}$$

NOBM procedures address the problem of determining the batch size, m, and the number of batches, k, that are required to satisfy approximately the assumptions of independence and normality of the batch means. If these assumptions are exactly satisfied, then we will obtain CIs whose actual coverage probability is exactly equal to the nominal coverage probability. In this paper we present results of an experimental performance evaluation of ASAP, a new NOBM procedure for analysis of steady-state simulation output, versus the well-known NOBM procedures ABATCH and LBATCH (Fishman 1996; Fishman and Yarberry 1997; Fishman 1998). A brief overview of ASAP is given in the next section; a more complete description may be found in Steiger and Wilson (2000b).

2 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
$$\overline{Y} \pm H \quad \text{where} \quad H \le H^* \quad \text{or} \quad H \le r^*|\overline{Y}|, \tag{5}$$
provided no additional simulation-generated observations are required;
2. a new total sample size n to be supplied to the algorithm.

If additional observations of the target process must be generated by the user's simulation model before a confidence interval with the required precision can be delivered, then 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.

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 (4) (in the case of acceptance of independence) or a correlation-adjusted CI (6) (in the case of acceptance of multivariate normality). The correlation correction uses an inverted Cornish-Fisher expansion (Hall 1983; Kendall, Stuart and Ord 1987; Chien 1989) of the classical NOBM Student t-ratio $[\overline{Y}(m^*, k^*) - \mu_X]/[S_{m^*,k^*}/\sqrt{k^*}]$; and the terms of this expansion are estimated by fitting an autoregressive–moving average time-series model (Box, Jenkins and Reinsel 1994) to the final set of k^* batch means for batches of size m^*. Based on this approach, a correlation-adjusted $100(1-\alpha)\%$ confidence interval for μ_X is

$$\overline{Y}(m^*, k^*) \pm \left[z_{1-\alpha/2}\left(1 + \frac{\hat{\kappa}_2 - 1}{2} - \frac{\hat{\kappa}_4}{8}\right) + \frac{\hat{\kappa}_4}{24} z_{1-\alpha/2}^3 \right] \times \sqrt{\frac{\widehat{\mathrm{Var}}[Y(m^*)]}{k^*}}, \tag{6}$$

where $\hat{\kappa}_2$ and $\hat{\kappa}_4$ respectively denote estimators of the second and fourth cumulants of the usual NOBM Student t-ratio and $\widehat{\mathrm{Var}}[Y(m^*)]$ denotes an estimator of the variance of the batch means—and all these statistics are based on fitting a time-series model to the (correlated) batch means process.

Subsequent iterations of ASAP that are performed to satisfy the user-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.

A flow chart of ASAP is depicted in Figure 1. A formal algorithmic statement of ASAP is given in Steiger and Wilson (2000b). A standalone Windows-based version of ASAP and a user's manual are available in Steiger and Wilson (2000c).

3 PERFORMANCE EVALUATION FOR SELECTED NOBM PROCEDURES

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, Fish-

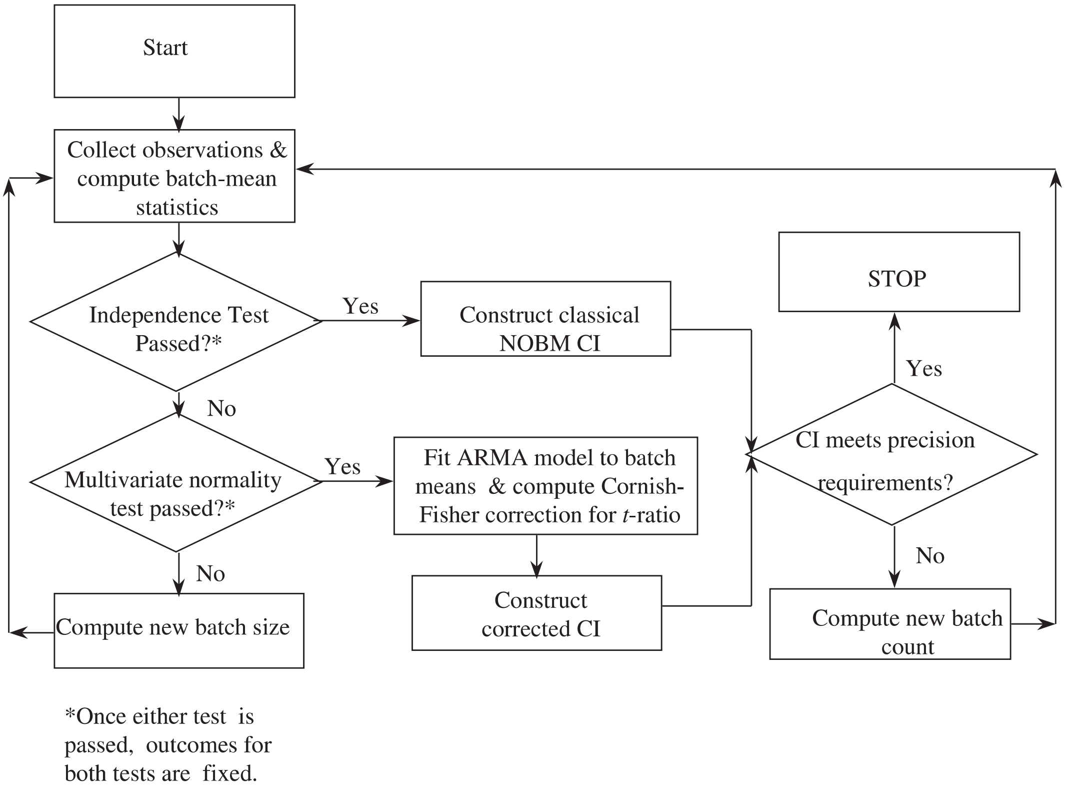
Start
Collect observations & compute batch-mean statistics
Independence Test Passed?*
Yes
Construct classical NOBM CI
No
Multivariate normality test passed?*
Yes
Fit ARMA model to batch means & compute Cornish-Fisher correction for t-ratio
No
Compute new batch size
Construct corrected CI
STOP
Yes
CI 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

man 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 structure 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 nine 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 eleven remaining test problems are not presented here because they contribute little additional insight into the relative performance of the algorithms. See Steiger (1999) 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 or the test for multivariate normality without considering a precision requirement;

(b) ±15% precision—that is, we continued the simulation of each test problem until ASAP delivered a confidence interval (5) that satisfied the relative precision requirement with $r^* = 0.15$; and

(c) ±7.5% precision—that is, we continued the simulation of each test problem until ASAP delivered a confidence interval (5) that satisfied the relative precision requirement 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 §3.1.3, 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 0.03. 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.

3.1 Results for Selected Test Problems

3.1.1 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,

$$\mathbf{P} = \begin{array}{c} \\ 0 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix} \end{array} \quad \text{and} \quad \mathbf{h} = \begin{array}{c} \begin{array}{cc} 0 & \ 1 \end{array} \\ \begin{pmatrix} 5 & 10 \end{pmatrix} \end{array}. \tag{7}$$

The results for this problem are summarized in Table 1.

Table 1: Performance of Batch-Means Procedures for the 2-State DTMC Defined by (7) 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 adjusted CIs generated by ASAP: 0 and 100, respectively.

ASAP showed somewhat better confidence-interval coverage than did ABATCH and LBATCH in the case of the two-state Markov chain (7) with high positive correlation, especially in the cases of no precision requirement and a precision requirement of ±15%. For this model, ASAP delivered correlation-adjusted CIs based on a nonsignificant result from the test for multivariate normality (that is, the batch means passed the Shapiro-Wilk test 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.

3.1.2 Autoregressive Process

We also applied ABATCH, LBATCH, and ASAP to an autoregressive process of order one—that is, an AR(1) process given by

$$X_i = \mu_X + \varphi(X_{i-1} - \mu_X) + \epsilon_i \qquad \text{for} \quad i = 1, 2, \ldots, \tag{8}$$

where $|\varphi| < 1$ and the ϵ_i's are i.i.d. normal with mean zero and variance σ_ϵ^2 so that the X_i's have mean μ_X and variance $\sigma_X^2 = \sigma_\epsilon^2/(1-\varphi^2)$. To ensure that (8) defines a stationary process, we took $X_0 \sim N(\mu_X, \sigma_X^2)$. This process also has a geometrically declining positive autocorrelation function; but among the highly correlated processes tested, this process exhibits the most rapid convergence to the desired property of normally distributed batch means since in fact the batch means are exactly multivariate normal for every batch size (Kang and Schmeiser 1987). We present the AR(1) process as a test problem for which all three batch means algorithms performed well at small batch sizes. For our simulations, we chose $\varphi = 0.9$, $Z_0 \sim N(2.0, 5.263)$ and $\epsilon_i \sim N(0, 1)$; and this implies the steady-state mean $\mu_X = 2.0$. Table 2 summarizes the experimental results for this test problem.

3.1.3 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 autocorrelation function of the waiting time process decays very slowly with increasing lags; and (c) in steady-state operation the marginal distribution of waiting times has an exponential tail and is therefore markedly nonnormal. Because of these characteristics, we can expect slow convergence to the classical requirement that the batch means are independent and identically normally distributed. This test problem most dramatically displays one of the advantages of the ASAP algorithm—namely, that ASAP does not rely

Table 2: Performance of Batch-Means Procedures for the AR(1) Process (8) with $\varphi = 0.9$ and $\mu_X = 2.0$ Based on 100 Independent Replications of Nominal 90% Confidence Intervals

Precision	Procedure		
Requirement	LBATCH	ABATCH	ASAP†
NO PRECISION			
average sample size			1624
coverage	84%	86%	93%
avg. rel. precision	0.180	0.198	0.228
avg. CI half length	0.351	0.387	0.446
var. CI half length	0.002	0.005	0.006
±15% PRECISION			
average sample size			5862
coverage	88%	88%	93%
avg. rel. precision	0.104	0.107	0.123
avg. CI half length	0.208	0.212	0.244
var. CI half length	0.001	0.001	0.001
±7.5% PRECISION			
average sample size			24860
coverage	87%	88%	91%
avg. rel. precision	0.053	0.054	0.059
avg. CI half length	0.106	0.108	0.118
var. CI half length	0.0003	0.0003	0.0003

†No. of classical and adjusted CIs generated by ASAP: 2 and 98, respectively.

Table 3: 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 adjusted CIs generated by ASAP: 4 and 96, respectively.

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 (6).

As can be seen from Table 3, 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 ±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 4 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 similar to the one used for ASAP. After we performed the simulation with an initial run length of 1536 observations, 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 of 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 of the observations accumulated so far. This 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 4 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 3 and 4 highlight the performance advantages achieved by ASAP without requiring analysis or manual intervention by the user.

From Table 4 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 in Table 4 are much smaller than those required by ASAP to achieve the same precision. For example, the

Table 4: 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 Requirement	Procedure	
	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

Table 5: Steady-State Expected Waiting Time in Selected Queueing Systems

System	Utilization τ	Expected Waiting Time
$M/M/1$	0.9	9.00
$M/M/1$ LIFO	0.8	3.20
$M/M/1$ SIRO	0.8	3.20
$M/M/1/M/1$	0.8	6.40

Table 6: Performance of Batch-Means Procedures for the M/M/1 LIFO Queue Waiting Time Process with $\tau = 0.8$ Based on 100 Independent Replications of Nominal 90% Confidence Intervals

Precision Requirement	Procedure		
	LBATCH	ABATCH	ASAP†
NO PRECISION			
avg. sample size			5025
coverage	72%	75%	72%
avg. rel. precision	0.209	0.223	0.210
avg. CI half length	0.645	0.693	0.652
var. CI half length	0.070	0.113	0.074
±15% PRECISION			
avg. sample size			14317
coverage	80%	81%	77%
avg. rel. precision	0.135	0.143	0.119
avg. CI half length	0.426	0.451	0.372
var. CI half length	0.009	0.013	0.004
±7.5% PRECISION			
average sample size			57539
coverage	86%	89%	82%
avg. rel. precision	0.073	0.075	0.062
avg. CI half length	NA	0.239	0.196
var. CI half length	NA	0.002	0.0006

†No. of classical and adjusted CIs generated by ASAP: 92 and 8, respectively.

average sample size used by ABATCH for the waiting time process in the $M/M/1$ queue with utilization $\tau = 0.9$ and a precision requirement of ±7.5% is approximately 397,387. This is considerably less than the average sample size of 815,755 required by ASAP. For a precision requirement of ±7.5% and 90% confidence-interval coverage probability, Whitt's (1989) approximation for estimating the 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 latter result suggests that ASAP yields adequate sample sizes when a precision requirement is specified.

Our experimental performance evaluation also included the eight queueing systems used by Law and Carson (1979). In this paper we discuss the results obtained for three of these queueing systems: (a) the $M/M/1$ LIFO queue with server utilization $\tau = 0.8$; (b) the $M/M/1$ queue with service in random order (SIRO) and $\tau = 0.8$; and (c) the tandem $M/M/1/M/1$ queue with $\tau = 0.8$. The simulations of all queueing systems were started empty and idle. The steady-state expected waiting times for these systems are given in Table 5.

From Table 6 we see that ASAP, ABATCH, and LBATCH performed similarly for the $M/M/1$ LIFO queue with $\tau = 0.8$, showing some evidence of undercoverage for the no-precision requirement and for the precision requirement of ±15%. All the algorithms showed adequate coverage for the precision requirement of ±7.5%. In 92 of 100 replications of the algorithm, ASAP delivered classical (unadjusted) CIs of the form (4).

From Table 7 we see that for the $M/M/1$ SIRO queue with server utilization $\tau = 0.8$, ASAP displayed better coverages than LBATCH and ABATCH in the cases with no precision requirement and with a precision requirement of ±15%. With a precision requirement of ±7.5%, ASAP, LBATCH, and ABATCH all delivered coverage close to the nominal level.

From Table 8 we see that for the $M/M/1/M/1$ queue with server utilization $\tau = 0.8$, ASAP performed better than ABATCH and LBATCH for the no-precision requirement and for the precision requirement of ±15%. With a precision requirement of ±7.5%, ASAP, LBATCH, and ABATCH all delivered coverage close to the nominal level.

Table 7: Performance of Batch-Means Procedures for the M/M/1 SIRO Queue Waiting Time Process with $\tau = 0.8$ Based on 100 Independent Replications of Nominal 90% Confidence Intervals

Precision Requirement	Procedure LBATCH	ABATCH	ASAP†
NO PRECISION			
avg. sample size			6481
coverage	66%	76%	88%
avg. rel. precision	0.168	0.199	0.264
avg. CI half length	0.534	0.652	0.901
var. CI half length	0.053	0.196	0.800
±15% PRECISION			
average sample size			32544
coverage	81%	82%	92%
avg. rel. precision	0.093	0.102	0.115
avg. CI half length	0.299	0.328	0.370
var. CI half length	0.0064	0.0092	0.0062
±7.5% PRECISION			
average sample size			116925
coverage	86%	88%	86%
avg. rel. precision	0.054	0.054	0.056
avg. CI half length	0.171	0.180	0.179
var. CI half length	0.0015	0.0017	0.0009

†No. of classical and adjusted CIs generated by ASAP: 29 and 71, respectively.

Table 8: 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

Precision Requirement	Procedure LBATCH	ABATCH	ASAP†
NO PRECISION			
avg. sample size			3152
coverage	65%	75%	85%
avg. rel. precision	0.162	0.222	0.454
avg. CI half length	1.053	1.476	3.250
var. CI half length	0.119	0.585	14.06
±15% PRECISION			
avg. sample size			46610
coverage	80%	80%	93%
avg. rel. precision	0.070	0.074	0.103
avg. CI half length	0.438	0.465	0.649
var. CI half length	0.016	0.018	0.030
±7.5% PRECISION			
avg sample size			117339
coverage	85%	87%	90%
avg. rel. precision	0.042	0.044	0.050
avg. CI half length	0.266	0.281	0.318
var. CI half length	0.005	0.005	0.008

†No. of classical and adjusted CIs generated by ASAP: 6 and 93, respectively.

3.1.4 Computer Models

Law and Carson (1979) also tested their sequential output analysis procedure on queueing network models of computer systems. The first is a time-shared model with a single central processing unit (CPU) and J terminals (jobs). Each terminal “thinks" for a period of time that is exponentially distributed with rate μ_1 and then sends a job to the CPU with a service time that is exponentially distributed with rate μ_2. The jobs join a queue at the CPU, which allocates a maximum “time slice" of s^* time units to each job in FIFO order. If the remaining service time s of a job is less than s^*, then the job spends s time units plus a fixed overhead of h time units at the CPU and then returns to the terminal. If $s > s^*$, then the CPU spends $s^* + h$ time units processing the job; and then the job returns to the end of the queue. This process is continued until the job is finished, and then it returns to the terminal. The process of interest is the response time of the jobs $\{R_i : i \geq 1\}$—that is, R_i is the elapsed time between the instant that job i joins the CPU queue at the end of a “thinking" period and the instant that job i completes its last time slice of service on the CPU. We chose the same parameters that Law and Carson used, i.e., $J = 35$, $\mu_1 = 1/25$, $\mu_2 = 5/4$, $s^* = 0.1$, and $h = 0.015$. (See Law and Carson 1979 for a more complete description of the model.)

Table 9 summarizes the results for the time-shared computer system. For this test problem, ASAP exhibits almost ideal behavior. The sample size for the no-precision case is small, but coverage is acceptable. With increasingly stringent precision requirements, the CIs become smaller and less variable and the coverage improves. For this model, ASAP delivered correlation-adjusted CIs in all 100 replications of the procedure.

Table 9: Performance of Batch-Means Procedures for the Time-Shared Model Based on 100 Independent Replications of Nominal 90% Confidence Intervals

Precision Requirement	Procedure LBATCH	ABATCH	ASAP†
NO PRECISION			
avg. sample size			1765
coverage	73%	79%	92%
avg. rel. precision	0.113	0.139	0.183
avg. CI half length	0.900	1.102	1.468
var. CI half length	0.016	0.050	0.139
±15% PRECISION			
avg. sample size			4496
coverage	83%	84%	94%
avg. rel. precision	0.088	0.095	0.121
avg. CI half length	0.722	0.772	0.985
var. CI half length	0.016	0.021	0.021
±7.5% PRECISION			
avg. sample size			18747
coverage	86%	87%	98%
avg. rel. precision	0.046	0.048	0.061
avg. CI half length	0.380	0.392	0.500
var. CI half length	0.003	0.003	0.003

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

The second computer model used by Law and Carson (1979) consists of a central server (CPU) and $M - 1$ peripheral units labeled 2 through M. The system has a fixed number jobs, N, 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 a peripheral unit. The probability that the job is routed to unit i from the CPU is p_i, $i = 2, \ldots, M$. 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 its arrival at the CPU queue and its departure from the system. Law and Carson chose to simulate this model for four cases. Table 10 displays the system parameters for cases 2 and 3 of the central-server model used by Law and Carson (1979). In both these cases we see that: μ_1, the service rate at the CPU is 1.0; p_1, the probability that the job leaves the system after service at the CPU, is zero; and the number of peripheral units is two. In model 2 the steady-state utilization of the CPU and peripheral units 1 and 2 are 0.8, 0.8, and 0.8, respectively. In model 3 these steady-state utilizations are 0.44, 0.88, and 0.88, respectively.

Table 10: Parameters for the Selected Central Server Models with $M = 3$, $\mu_1 = 1.0$, and $p_1 = 0$

						Expected	Initial
Model	N	μ_2	μ_3	p_2	p_3	Response Time	State
2	8	0.50	0.50	0.5	0.5	10.000	(1,1,6)
3	8	0.45	0.05	0.9	0.1	18.279	(5,1,2)

From Table 11 we see that ASAP achieved good coverage for central server model 2, satisfying the precision requirements of ±15% and ±7.5% without increasing the sample size beyond that used in the no-precision case. Table 12 reveals that in central server model 3, 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.

4 CONCLUSIONS

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 being achieved at smaller batch sizes than approximate in-

Table 11: Performance of Batch-Means Procedures for the Central Server Model 2 Based on 100 Independent Replications of Nominal 90% Confidence Intervals

Precision	Procedure		
Requirement	LBATCH	ABATCH	ASAP†
NO PRECISION			
avg. sample size			1580
coverage	87%	90%	86%
avg. rel. precision	0.039	0.040	0.039
avg. CI half length	0.388	0.400	0.387
var. CI half length	0.002	0.005	0.003
±15% PRECISION			
avg. sample size			1580
coverage	87%	90%	86%
avg. rel. precision	0.039	0.040	0.039
avg. CI half length	0.388	0.400	0.387
var. CI half length	0.002	0.005	0.003
±7.5% PRECISION			
average sample size			1580
coverage	87%	90%	86%
avg. rel. precision	0.039	0.040	0.039
avg. CI half length	0.388	0.400	0.387
var. CI half length	0.002	0.005	0.003

†No. of classical and adjusted CIs generated by ASAP: 67 and 33, respectively.

Table 12: Performance of Batch-Means Procedures for the Central Model 3 Based on 100 Independent Replications of Nominal 90% Confidence Intervals

Precision	Procedure		
Requirement	LBATCH	ABATCH	ASAP†
NO PRECISION			
avg. sample size			2277
coverage	75%	79%	78%
avg. rel. precision	0.073	0.076	0.074
avg. CI half length	1.33	1.40	1.35
var. CI half length	0.107	0.163	0.135
±15% PRECISION			
avg. sample size			2277
coverage	75%	79%	78%
avg. rel. precision	0.073	0.076	0.074
avg. CI half length	1.33	1.40	1.35
var. CI half length	0.107	0.163	0.135
±7.5% PRECISION			
avg. sample size			3389
coverage	75%	76%	79%
avg. rel. precision	0.060	0.062	0.058
avg. CI half length	1.08	1.11	1.05
var. CI half length	0.037	0.047	0.028

†No. of classical and adjusted CIs generated by ASAP: 76 and 24, respectively.

dependence (Steiger and Wilson 2000a) *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 experimental evaluation reveals the main advantage of ASAP—it performs with reasonable reliability in highly dependent simulation output processes. In these cases, ASAP determines sample sizes that are sufficient for achieving adequate CI coverage but that are not excessively large. Taken as a whole, the results of the experimental performance evaluation reported in this paper strongly suggest that significant improvements in the performance of batch means procedures can be achieved using the approach of ASAP for constructing correlation-adjusted confidence intervals in situations for which it is difficult to identify a batch size sufficiently large to ensure approximate independence of the batch means. We are continuing to explore and refine this approach to the analysis of steady-state simulation outputs.

BIBLIOGRAPHY

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time series analysis: Forecasting and control*. 3d ed. 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. 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.

Kang, K., and B. Schmeiser. 1987. Properties of batch means from stationary time series. *Operations Research Letters* 6 (1): 19–24.

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

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

Steiger, N. M. 1999. Improved batching for confidence interval construction in steady state simulation. Doctoral dissertation, Department of Industrial Engineering, North Carolina State University, Raleigh, North Carolina. Available on-line via `<http://www.lib.ncsu.edu/etd/public/etd-19231992992670/etd.pdf>` [accessed June 18, 2000].

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

Steiger, N. M., and J. R. Wilson. 2000a. Convergence properties of the batch means method for simulation output analysis. *INFORMS Journal on Computing*, in review. Available on-line via `<ftp://ftp.ncsu.edu/pub/eos/pub/jwilson/ibssa1v6.pdf>` [accessed June 18, 2000].

Steiger, N. M., and J. R. Wilson. 2000b. An improved batch means procedure for simulation output analysis. *Operations Research*, in review. Available on-line via `<ftp://ftp.ncsu.edu/pub/eos/pub/jwilson/asaporv12.pdf>` [accessed June 18, 2000].

Steiger, N. M., and J. R. Wilson 2000c. ASAP software and user's manual [online]. Department of Industrial Engineering, North Carolina State University, Raleigh, North Carolina. Available on-line via `<ftp://ftp.ncsu.edu/pub/eos/pub/jwilson/installasap.exe>` [accessed June 18, 2000].

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.

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

AUTHOR BIOGRAPHIES

NATALIE M. STEIGER is an Assistant Professor of Production and Operations Management in the University of Maine Business School. In 2000 she received the IIE Pritsker Doctoral Dissertation Award (second place) for work that is partially documented in this paper. She is a member of IIE and INFORMS.

JAMES R. WILSON is Professor and Head of the Department of Industrial Engineering at North Carolina State University. From 1988 to 1992, he served as a departmental editor of *Management Science*; and since 1997, he has served as an area editor of *ACM Transactions on Modeling and Computer Simulation*. He has also held various offices in INFORMS–College on Simulation; and he currently serves as the corepresentative of that organization to the Board of Directors of the Winter Simulation Conference. He is a member of ASA, ACM, IIE, and INFORMS. His e-mail address is `<jwilson@eos.ncsu.edu>`, and his web page is `<www.ie.ncsu.edu/jwilson>`.