Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

PERFORMANCE EVALUATION OF ASAP3 FOR STEADY-STATE OUTPUT ANALYSIS

Natalie M. Steiger

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

Emily K. Lada

SAS Institute Inc.
100 SAS Campus Drive, R5413
Cary, NC 27513-8617, U.S.A.

James R. Wilson

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

ABSTRACT

An experimental performance evaluation of ASAP3 is presented, including several queueing systems with characteristics typically encountered in practical applications of steady-state simulation analysis procedures. Based on the method of nonoverlapping batch means, ASAP3 is a sequential procedure designed to produce a confidence-interval estimator of a steady-state mean response that satisfies user-specified precision and coverage-probability requirements. ASAP3 is compared with its predecessor ASAP and the batch means procedure of Law and Carson (LC) in the following test problems: (a) queue waiting times in the $M/M/1$/LIFO, $M/H_2/1$, and $M/M/1$ queues with 80% server utilization; and (b) response (sojourn) times in a central server model of a computer system. Regarding conformance to the given precision and coverage-probability requirements, ASAP3 compared favorably with the ASAP and LC procedures. Regarding the average sample sizes needed to satisfy progressively more stringent precision requirements, ASAP3's efficiency was reasonable for the given test problems.

1 INTRODUCTION

In discrete-event simulation, we are often interested in estimating the steady-state mean μ_X of a stochastic output process $\{X_j : j = 1, 2, \ldots\}$ generated by a single, prolonged simulation run. Assuming the target process is stationary and given a time series of length n that is part of a single realization of this process, we see that a natural point estimator of μ_X is the sample mean, given by $\overline{X}(n) = n^{-1} \sum_{j=1}^{n} X_j$. We also require some indication of the precision of this point estimator; and typically we construct a confidence interval (CI) for μ_X with a user-specified probability $1 - \alpha$ of covering the point μ_X, where $0 < \alpha < 1$. The CI for μ_X should satisfy two criteria: (i) it is approximately valid—that is, its coverage probability is sufficiently close to the nominal level $1 - \alpha$; and (ii) it has sufficient precision—that is, it is narrow enough to be meaningful in the context of the application at hand.

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

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

and the grand mean of the individual batch means,

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

is used as a point estimator for μ_X (note that $\overline{Y}(m, k) = \overline{X}(n)$). We construct a CI estimator for μ_X that is centered on a point estimator like (2), where in practice we may exclude some initial batches to eliminate the effects of initialization bias.

If the batch size m is sufficiently large so that the batch means $\{Y_j(m) : j = 1, \ldots, k\}$ are approximately independent and identically distributed (i.i.d.) normal random variables with mean μ_X, then we can apply classical results concerning Student's t-distribution (see, for example, Alexopoulos and Goldsman 2004) to compute a confidence interval for μ_X from the batch means. If the original simulation-generated process $\{X_j : j = 1, \ldots, n\}$ is stationary and weakly dependent as specified, for example, in Theorem 1 of Steiger and Wilson (2001), then it follows that as $m \to \infty$ with k fixed so that $n \to \infty$, an asymptotically valid $100(1-\alpha)\%$ CI for μ_X is

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

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

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

is the sample variance of the k batch means for batches of size m.

Conventional NBM procedures such as ABATCH and LBATCH (Fishman and Yarberry 1997, Fishman 1998) and the procedure of Law and Carson (1979) are based on (3); and they are designed to determine the batch size, m, and the number of batches, k, that are required to satisfy approximately the assumption of i.i.d. normal batch means. If this assumption is satisfied exactly, then we will obtain a CI whose actual coverage probability is exactly equal to the nominal level $1-\alpha$.

By contrast, the more recent NBM procedures ASAP (Steiger 1999; Steiger and Wilson 1999, 2000, 2002a, 2002b) and ASAP2 (Steiger et al. 2002) are designed to determine a batch size and an initial warm-up period sufficient to ensure that batch means computed beyond the warm-up period are approximately multivariate normal with identically distributed marginals (that is, they approximate a stationary Gaussian process) but are not necessarily independent. If the resulting batch means are correlated, then the classical NBM t-ratio underlying (3) does not possess Student's t-distribution with $k-1$ degrees of freedom so that an appropriate modification of (3) is required to yield an approximately valid CI for μ_X.

Both ASAP and ASAP2 are designed to adjust (3) so as to account for any correlations among the batch means that those procedures finally deliver; and the required correlation adjustment is based on an inverse Cornish-Fisher expansion for the classical NBM t-ratio. There is substantial experimental evidence that when ASAP or ASAP2 is applied with a user-specified absolute- or relative-precision requirement for the final delivered confidence interval, either procedure outperforms conventional NBM procedures such as ABATCH and LBATCH in a large class of steady-state simulation models (Steiger and Wilson 2002a, Steiger et al. 2002). However, when either ASAP or ASAP2 is applied without a precision requirement, the delivered confidence intervals may exhibit excessive variability in some applications—that is, the variance and coefficient of variation of the CI half-lengths may be unacceptably large (Steiger and Wilson 2002a; Steiger et al. 2002; Lada, Wilson, and Steiger 2003).

In this article we examine the performance of ASAP3, a refinement of ASAP and ASAP2 that retains the advantages of its predecessors but is specifically designed to prevent excessive CI variability even in the absence of a precision requirement. While Steiger et al. (2005a) summarize the performance of ASAP3 when it is applied to problems constituting a kind of "torture test" designed to elicit worst-case behavior, in this article we report the performance of ASAP3 in selected queueing-system simulations whose characteristics are more nearly typical of a broad class of steady-state simulation applications. In particular the following test processes are used: (a) queue waiting times in the $M/M/1$/LIFO, $M/H_2/1$, and $M/M/1$ queues with 80% server utilization; and (b) response (sojourn) times in a central server model of a computer system. Except for the $M/M/1$ queue waiting times, Steiger (1999) finds that both ASAP and the Law and Carson (LC) procedure exhibit problematic behavior in all these test processes; hence in this article we limit our experimental performance evaluation to a comparison of ASAP, ASAP3, and the LC procedure.

This paper is organized as follows. In §2 we provide a brief overview of ASAP3. In §3 we summarize some of the results of our experimental performance evaluation, and in §4 we summarize the results of an empirical efficiency analysis of ASAP3. Finally in §5 we present our main conclusions. Full details of the ASAP3 algorithm are available in Steiger et al. (2005a); see also Steiger et al. (2004). For complete details on the performance evaluation summarized in this article, see Steiger, Lada, and Wilson (2005b).

2 OVERVIEW OF ASAP3

Figure 1 displays a high-level flow chart of ASAP3. The procedure operates as follows. The series of simulation outputs is divided initially into $k = 256$ batches, each of a user-specified size m (where the default initial batch size $m = 16$); and the corresponding batch means are computed as in (1). The first four batches are ignored to reduce the potential effects of initialization bias, and the remaining $k' = k - 4 = 252$ batch means are organized into adjacent nonoverlapping groups, where each group consists of four consecutive batch means. We select every other group of four consecutive batch means to form a sample of 32 four-dimensional vectors that we will test for stationary multivariate normality. If this test is failed, then the batch size m is increased by the factor $\sqrt{2}$; additional data are obtained; and the process of computing 256 batch means with the new batch size and testing for multivariate normality proceeds as outlined above using all accumulated data. ASAP3 iteratively performs this sequence of steps, systematically decreasing the significance level δ for the multivariate normality test on successive iterations until

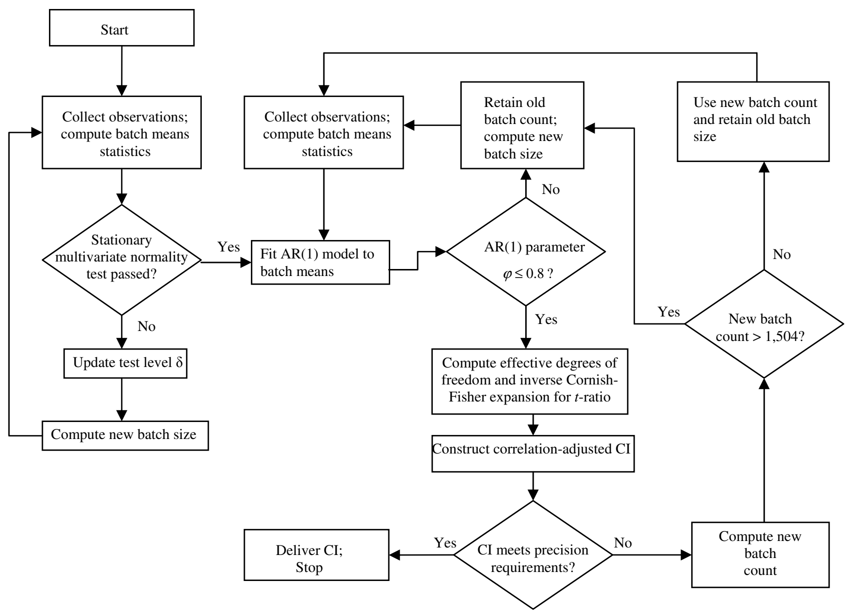

Figure 1: High-level Flow Chart of ASAP3

that test is finally passed. (See §3.1 of Steiger et al. 2005a for further explanation of this issue.)

Upon accepting the hypothesis of stationary multivariate normality of the batch means, we fit a first-order autoregressive (that is, AR(1)) time series model to the 252 batch means that remain after skipping the first group of four batch means. Adapting the notation of Box, Jenkins, and Reinsel (1994) to the nomenclature used here, we let $\{\widetilde{Y}_{j-4} \equiv Y_j(m) - \mu_X : j = 5, \ldots, k\}$ denote the corresponding deviations of the truncated batch means from the unknown steady-state mean μ_X. The ℓth observation of such an AR(1) process can be expressed as

$$\widetilde{Y}_\ell = \varphi \widetilde{Y}_{\ell-1} + a_\ell \quad \text{for} \quad \ell = 1, 2, \ldots, \tag{4}$$

where the autoregressive parameter $\varphi \in (-1, 1)$ and $\{a_\ell\}$ are i.i.d. normal residuals with mean 0 and variance σ_a^2.

After fitting the AR(1) model (4) to the truncated batch means $\{Y_j(m) : j = 5, \ldots, k\}$, we apply a normalizing arc sine transformation to the autoregressive parameter estimator $\widehat{\varphi}$ so as to test the null hypothesis that the correlation between adjacent batch means (that is, φ) is at most 0.8 versus the alternative hypothesis that $\varphi > 0.8$. We have found that the condition $\varphi > 0.8$ is associated with excessive variability in the CIs delivered by ASAP and ASAP2; for further explanation of this phenomenon, see §3.2 of Steiger et al. 2005a. If the null hypothesis is rejected, then the batch count is retained; the batch size m is increased by a factor projected to reduce the lag-one correlation between batch means to the threshold 0.8; additional data are obtained; and the process of computing batch means, fitting an AR(1) model, and testing the autoregressive parameter estimator proceeds as outlined above. ASAP3 iteratively performs the sequence of steps described in this paragraph until we finally obtain a batch size m for which we accept the null hypothesis of nonexcessive correlation between adjacent batch means.

Next ASAP3 constructs a CI for μ_X that has been adjusted to account for the remaining (nonexcessive) correlations between the k' batch means for batches of the current size m. The correlation adjustment uses an inverse Cornish-Fisher expansion (Stuart and Ord 1994) for the classical NBM t-ratio

$$t = \left[\,\overline{Y}(m,k') - \mu_X\,\right] \Big/ \sqrt{S^2_{m,k'}/k'}; \tag{5}$$

and the terms of this expansion are computed from the parameter estimates $\widehat{\varphi}$ and $\widehat{\sigma}_a^2$ that are obtained by fitting the AR(1) model (4) to the current set of k' truncated batch means. Based on this approach, a correlation-adjusted $100(1-\alpha)\%$ CI for μ_X is

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

where: $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution; $\widehat{\kappa}_2$ and $\widehat{\kappa}_4$ respectively denote estimators of the second and fourth cumulants of the t-ratio (5); $\widehat{\mathrm{Var}}[Y(m)]$ denotes an estimator of the variance of the batch means; and the statistics $\widehat{\kappa}_2$, $\widehat{\kappa}_4$, and $\widehat{\mathrm{Var}}[Y(m)]$ are computed from $\widehat{\varphi}$ and $\widehat{\sigma}_a^2$ as detailed in Steiger et al. (2005a).

If additional observations of the target process must be generated by the user's simulation model before a CI can be delivered that has the form (6) and the required precision, then ASAP3 estimates a new, larger sample size based on the ratio of the current iteration's CI half-length to the desired CI half-length as detailed in §3.4 of Steiger et al. (2005a). Then ASAP3 must be called again with the additional data; and this cycle of simulation followed by analysis may be repeated several times before ASAP3 finally delivers a CI with the required precision.

Subsequent iterations of ASAP3 that are performed to satisfy the user-specified precision requirement do not repeat the test of the overall set of batch means for stationary multivariate normality; but on every iteration of ASAP3, we fit an AR(1) process to the latest set of batch means, test the hypothesis that $\varphi \leq 0.8$, and if necessary increase the batch size by the updated factor that is currently projected to reduce the lag-one correlation between batch means to the threshold 0.8. Thus each additional iteration of ASAP3 that is performed solely to satisfy the precision requirement will involve the following operations: (i) obtaining additional simulation-generated data; (ii) recomputing the batch means with a new batch size or computing additional batch means of the same size; (iii) retesting the hypothesis that $\varphi \leq 0.8$ with progressively larger batch sizes until that hypothesis is accepted; and (iv) reconstructing the CI for μ_X and testing that CI for conformance to the user's precision requirement, if necessary computing the total sample size required for the next iteration of ASAP3. Successive iterations of ASAP3 involving operations (i)–(iv) above are performed until the precision requirement is met.

ASAP3 requires the following user-supplied inputs:

- a simulation-generated output process $\{X_j : j = 1, \ldots, n\}$ from which the steady-state expected response μ_X is to be estimated;
- the desired CI coverage probability $1 - \alpha$, where $0 < \alpha < 1$; and
- an absolute or relative precision requirement specifying the final confidence-interval half-length in terms of (i) a maximum acceptable half-length H^* (for an absolute precision requirement); or (ii) a maximum acceptable fraction r^* of the magnitude of the CI midpoint (for a relative precision requirement).

ASAP3 delivers the following outputs:

- a nominal $100(1 - \alpha)\%$ CI for μ_X that satisfies the specified absolute or relative precision requirement, provided no additional simulation-generated observations are required; or
- a larger total sample size n to be supplied to ASAP3 when it is executed again.

A formal algorithmic statement of ASAP3 along with a detail description of the algorithm is given in Steiger et al. (2004, 2005a). A stand-alone Windows-based version of ASAP3 and a user's manual are available online via Steiger et al. (2003).

3 EXPERIMENTAL PERFORMANCE EVALUATION

In Steiger et al. (2004, 2005a) we compared the performance of ASAP3 with that of ABATCH (Fishman and Yarberry 1997, Fishman 1998) and ASAP2 (Steiger et al. 2002) in a suite of test problems that were deliberately selected to provide extreme examples of correlated simulation output processes, some of which also exhibit substantial initialization bias or marked nonnormality—namely,

1. the $M/M/1$ queue waiting time process with a steady-state server utilization of 0.9 and an empty-and-idle initial condition;
2. the AR(1) process with autoregressive parameter value of 0.995, steady-state mean of 100, and initial condition of zero; and
3. the Autoregressive-to-Pareto (ARTOP) process obtained from a standardized, stationary version of process 2 above by inversion so that the resulting process has marginal distributions with finite mean and variance but with infinite skewness and kurtosis.

In this article we examine the performance of ASAP3 on test problems that are more nearly typical of practical applications, and we compare the performance of ASAP3 with that of ASAP and the LC procedure.

Steiger (1999) identifies three queueing systems for which estimation of the steady-state mean sojourn or queueing times are particularly problematic for both ASAP and the LC procedure: the Central Server Model 3 of Law and Carson (1979); the $M/M/1$/LIFO queue with server utilization 0.8; and the $M/H_2/1$ queueing system with server utilization 0.8. Although they are analytically tractable, these systems exemplify the output-analysis problems arising in a broad class of steady-state simulation applications; thus they facilitate a direct comparison of the performance of ASAP3 with that of the LC procedure as well as a demonstration of the performance improvements achieved by ASAP3 over the original ASAP algorithm.

The following subsections summarize the results for the three simulation systems listed above, as well as the results for queue waiting times for the $M/M/1$ queueing system with server utilization 0.8. For each system, 400 replications of ASAP3 were performed for nominal 90% and 95% CIs. For the ASAP and LC procedures, the available experimental results for the four selected systems only include 100 replications of nominal 90% CIs. The coverage estimators for ASAP3's CIs have a standard error of approximately 1.5% for nominal 90% CIs and a standard error of approximately 1% for nominal 95% CIs. The coverage estimators for ASAP's and Law and Carson's nominal 90% CIs have a standard error of 3%.

3.1 Central Server Model 3 of Law and Carson (1979)

Central Server Model 3 of Law and Carson (1979) is one of four computer-system models used by Steiger (1999) to evaluate the performance of ASAP and the LC procedure. This model consists of a CPU (the central server) and two peripheral units so that there are $M = 3$ service centers in this system. The system has a fixed number jobs, $N = 8$, in it. When a job is finished at the CPU, it leaves the system with probability p_1 (in this case $p_1 = 0.0$) 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, $i = 2, \ldots, M$; and we take $p_1 = 0.9$ and $p_2 = 0.1$ in this example. After getting service at one of the peripheral units at rate μ_i (so that the corresponding service time is exponentially distributed with mean $1/\mu_i$), the job leaves the system and is immediately replaced by a job joining the CPU queue. Law and Carson (1979) used the service rates $\mu_1 = 1$, $\mu_2 = 0.45$, and $\mu_3 = 0.05$ in this system. 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; and the corresponding steady-state expected value is 18.279. The system's initial condition consisted of 5, 1, and 2 customers at service centers 1, 2, and 3, respectively.

Out of a suite of twenty test problems used to evaluate the performance of the original ASAP algorithm, Steiger (1999) reports that the Central Server Model 3 was the only test problem for which the LC procedure significantly outperformed ASAP. Table 1 shows that the undercoverage problems experienced with ASAP were largely eliminated by ASAP3. The coverages for the ASAP3 and LC procedures were comparable for this system; however ASAP3 required significantly larger sample sizes on average. Asymptotically, ASAP3 performed very well, delivering confidence intervals that were close to the nominal level.

3.2 $M/M/1$/LIFO Queue

In the following discussion, we let M and H_2 denote, respectively, the exponential and hyperexponential distributions with coefficients of variation 1 and 2. In this section we present results for the queue waiting time process in the $M/M/1$/LIFO queue with server utilization $\tau = 0.8$ and last-in-first-out (LIFO) queueing discipline. The mean interarrival time is 1, and no customers are present at time zero. The steady-state mean waiting time in the queue for this system is 3.20.

Table 1: Performance of Batch Means Procedures for the Queue Waiting Time Process in Central Server Model 3 of Law and Carson (1979) Based on Independent Replications of Nominal 90% and 95% CIs

Precision	Nominal 90% CIs			95% CIs
Requirement	ASAP	ASAP3	LC	ASAP3
NO PRECISION				
# replications	100	400	100	400
avg. sample size	2,277	18,447	1,152	18,447
coverage	78.0%	87.0%	87.0%	92.3%
avg. rel. precision	0.074	0.032	0.097	0.038
avg. CI half length	1.350	0.581		0.694
var. CI half length	0.135	0.033		0.047
±15% PRECISION				
# replications	100	400	100	400
avg. sample size	2,277	18,447	1,152	18,447
coverage	78.0%	87.0%	87.0%	92.3%
avg. rel. precision	0.074	0.032	0.097	0.038
avg. CI half length	1.350	0.581		0.694
var. CI half length	0.135	0.033		0.047
±7.5% PRECISION				
# replications	100	400	100	400
avg. sample size	3,389	18,447	3,740	18,460
coverage	79.0%	87.0%	88.0%	92.3%
avg. rel. precision	0.058	0.032	0.059	0.038
avg. CI half length	1.050	0.581		0.692
var. CI half length	0.028	0.033		0.044
±1.00% PRECISION				
# replications		400		400
avg. sample size		164,176		231,861
coverage		87.3%		95.5%
avg. rel. precision		0.010		0.010
avg. CI half length		0.176		0.177
var. CI half length		5.9E-5		5.3E-5

For the $M/M/1$/LIFO queue waiting time process, Steiger (1999) and Law and Carson (1979) report undercoverage for the cases of no precision and 15% precision; and Table 2 shows that ASAP3 outperformed ASAP and the LC procedure in these cases.

3.3 $M/H_2/1$ Queue

In this section we report results for the $M/H_2/1$ queueing system with hyperexponential service times and a server utilization $\tau = 0.8$. The mean interarrival time is 1, and no customers are present at time zero. The steady-state mean waiting time in the queue for this system is 8.0. Table 3 summarizes the results we obtained by applying ASAP3 and the LC procedure to the queue waiting times generated by this system

From Table 3 we concluded that ASAP3 significantly outperformed ASAP in virtually every respect. In the no precision case, ASAP delivered nominal 90% CIs whose empirical coverage was only 76%; moreover the variance of the CI half-length was 43.7. By contrast in the no precision case ASAP3 delivered nominal 90% CIs with empirical coverage of 87.8%; and the variance of the CI half-length was 0.5962. Moreover in the no precision case, the LC procedure delivered 88% CI coverage but required

Table 2: Performance of Batch Means Procedures for the $M/M/1$/LIFO Queue Waiting Time Process Based on Independent Replications of Nominal 90% and 95% CIs

Precision Requirement	Nominal 90% CIs ASAP	ASAP3	LC	95% CIs ASAP3
NO PRECISION				
# replications	100	400	100	400
avg. sample size	5,025	53,958	3,120	53,958
coverage	72.0%	87.0%	64.0%	92.5%
avg. rel. precision	0.210	0.082	0.236	0.098
avg. CI half length	0.652	0.261		0.312
var. CI half length	0.074	0.106		0.008
±15% PRECISION				
# replications	100	400	100	400
avg. sample size	14,317	54,017	13,944	54,265
coverage	77.0%	86.8%	76.0%	92.8%
avg. rel. precision	0.119	0.081	0.131	0.096
avg. CI half length	0.372	0.260		0.308
var. CI half length	0.004	0.004		0.005
±7.5% PRECISION				
# replications	100	400	100	400
avg. sample size	57,539	68,325	74,624	90,911
coverage	82.0%	87.5%	84.0%	92.5%
avg. rel. precision	0.062	0.069	0.064	0.071
avg. CI half length	0.196	0.219		0.226
var. CI half length	6.0E-4	5.1E-4		2.9E-4
±3.75% PRECISION				
# replications		400		400
avg. sample size		258,228		371,072
coverage		90.8%		95.5%
avg. rel. precision		0.036		0.036
avg. CI half length		0.115		0.115
var. CI half length		3.2E-5		4.1E-5

Table 3: Performance of Batch Means Procedures for the $M/H_2/1$ Queue Waiting Time Process with Traffic Intensity $\tau = 0.8$ Based on Independent Replications of Nominal 90% and 95% CIs

Precision Requirement	Nominal 90% CIs ASAP	ASAP3	LC	95% CIs ASAP3
NO PRECISION				
# replications	100	400	100	400
avg. sample size	16,716	42,022	86,144	42,022
coverage	76.0%	87.8%	88.0%	91.8%
avg. rel. precision	0.539	0.2026	0.106	0.2447
avg. CI half-length	4.760	1.614		1.9500
var. CI half-length	43.730	0.5962		0.9084
±15% PRECISION				
# replications	100	400	100	400
avg. sample size	148,820	76,214	86,144	96,706
coverage	88.0%	88.0%	88.0%	93.3%
avg. rel. precision	0.102	0.1308	0.106	0.1354
avg. CI half-length	0.802	1.0329		1.0687
var. CI half-length	0.055	0.0273		0.0165
±7.5% PRECISION				
# replications	100	400	100	400
avg. sample size	405,854	228,482	229,632	309,560
coverage	93.0%	90.0%	90.0%	94.5%
avg. rel. precision	0.053	0.07054	0.067	0.07084
avg. CI half-length	0.421	0.5623		0.5647
var. CI half-length	0.021	2.0E−3		1.8E−3
±3.75% PRECISION				
# replications		400		400
avg. sample size		798,234		1,115,986
coverage		90.0%		94.7%
avg. rel. precision		0.0359		0.0360
avg. CI half-length		0.2867		0.2878
var. CI half-length		2.5E−4		2.1E−4

an average sample size of 86,144 while ASAP3 required an average sample size of 42,022. At the 15% and 7.5% precision levels, all three procedures delivered acceptable coverage; and although within each precision level both ASAP3 and the LC procedure required comparable average sample sizes, ASAP's average sample sizes were nearly twice as large as for ASAP3 and the LC procedure.

3.4 $M/M/1$ Queue

Table 4 summarizes our results for the $M/M/1$ queueing system with exponential service times, a server utilization $\tau = 0.8$, and a first-in-first-out queueing discipline. The mean interarrival time is 1, and no customers are present at time zero. The steady-state mean waiting time in the queue for this system is 3.2.

From Table 4 we concluded that in comparison with ASAP, ASAP3 delivered improved small-sample performance, with a reduction in the variance of the CI half-length from 5.21 to 0.031 in the no precision case. Furthermore, while all three methods delivered comparable results in terms of CI coverage and average relative precision for the ±7.5% precision case, ASAP3's average required sample size of 72,060 was 47% smaller than the average sample size of 136,491 required by ASAP.

4 EFFICIENCY ANALYSIS

If we have the "ideal" situation in which the target output process $\{X_j : j = 1, 2, \ldots\}$ is stationary and Gaussian with the known steady-state variance parameter (SSVP)

$$\gamma_X \equiv \lim_{n\to\infty} n\mathrm{Var}\big[\,\overline{X}(n)\,\big] = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(X_j, X_{j+\ell}), \qquad (7)$$

and if the series on the far right-hand side of (7) is absolutely convergent so that γ_X is well defined, then the nominal $100(1-\alpha)\%$ CI for μ_X,

$$\overline{X}(n) \pm z_{1-\alpha/2}\sqrt{\gamma_X/n}, \qquad (8)$$

is asymptotically valid in the sense that

$$\lim_{n\to\infty} \Pr\left\{ \mu_X \in \overline{X}(n) \pm z_{1-\alpha/2}\sqrt{\gamma_X/n} \right\} = 1-\alpha. \qquad (9)$$

It follows from (7)–(9) that in this ideal situation, an efficient procedure (in the sense of Chow and Robbins (1965) and Nádas (1969)) for computing a $100(1-\alpha)\%$ CI for μ_X

Table 4: Performance of Batch Means Procedures for the $M/M/1$ Queue Waiting Time Process with Traffic Intensity $\tau = 0.8$ Based on Independent Replications of Nominal 90% and 95% CIs

Precision Requirement	Nominal 90% CIs ASAP	ASAP3	LC	95% CIs ASAP3
NO PRECISION				
# replications	100	400	100	400
avg. sample size	8,482	41,326	32,960	41,326
coverage	85.0%	88.8%	85.0%	93.3%
avg. rel. precision	0.286	0.110	0.096	0.132
avg. CI half-length	1.310	0.348		0.417
var. CI half-length	5.210	0.031		0.046
±15% PRECISION				
# replications	100	400	100	400
avg. sample size	38,911	43,796	32,960	46,106
coverage	87.0%	88.5%	85.0%	93.0%
avg. rel. precision	0.112	0.098	0.096	0.110
avg. CI half-length	0.441	0.310		0.349
var. CI half-length	0.009	0.009		0.007
±7.5% PRECISION				
# replications	100	400	100	400
avg. sample size	136,491	72,060	75,648	97,643
coverage	90.0%	86.8%	87.0%	93.3%
avg. rel. precision	0.056	0.070	0.065	0.071
avg. CI half-length	0.222	0.220		0.224
var. CI half-length	0.002	4.2E−4		2.8E−4
±3.75% PRECISION				
# replications		400		400
avg. sample size		256,186		365,353
coverage		89.5%		93%
avg. rel. precision		0.036		0.036
avg. CI half-length		0.114		0.115
var. CI half-length		4.1E−5		4.0E−5

with relative precision r^* will require a sample size of

$$n^* = n^*(r^*) = z^2_{1-\alpha/2}\gamma_X \big/ (r^*\mu_X)^2; \tag{10}$$

and as $r^* \to 0$, the resulting confidence interval, $\overline{X}[n^*(r^*)] \pm z_{1-\alpha/2}\sqrt{\gamma_X / n^*(r^*)}$, will have a coverage probability that approaches the limiting value $1-\alpha$. In the preceding sections of this article, the experimental evidence suggested that the CIs generated by ASAP3 were asymptotically valid to a reasonable approximation in the sense that some degree of convergence to the coverage probability $1-\alpha$ was achieved in nearly all cases as we considered progressively smaller values of the relative precision r^*.

To complement the preceding experimental results, in this section we perform an empirical efficiency analysis of ASAP3 on the test problems for which γ_X can be evaluated exactly (or at least to the limits of machine accuracy); and we will take $n^* = n^*(r^*)$ as a benchmark sample size to compare with $\bar{n} = \bar{n}(r^*)$, the average sample size required by ASAP3 to deliver CIs with relative precision r^*. In particular we use the sample-size ratio $\bar{n}/n^*$ to gauge the efficiency of ASAP3 in the given test problems.

Tables 5 and 6 give the values of n^* for the $M/H_2/1$ and $M/M/1$ queueing systems. In the Appendix of Steiger et al. (2005b), we summarize the numerical methods used to evaluate the SSVP for each test problem reported in Tables 5 and 6.

Table 5: Comparison of the Average Sample Size $\bar{n}$ Required by ASAP3 with the Theoretical n^* Required for Efficiently Computing a Valid 90% CI for μ with Relative Precision r^*

Output Process	r^*	n^*	ASAP3 $\bar{n}$	ASAP3 $\bar{n}/n^*$
$M/M/1$ Queue Waiting Times	15%	53,306	103,742	1.946
$\tau = 0.9$, $\mu_X = 9$, $\gamma_X = 35{,}901$	7.5%	213,222	287,568	1.349
(Steiger et al., 2005a)	3.75%	852,886	969,011	1.136
$M/M/1$ Queue Waiting Times	15%	14,853	43,796	2.949
$\tau = 0.8$, $\mu_X = 3.2$,	7.5%	59,412	72,060	1.213
$\gamma_X = 1{,}264.64$	3.75%	237,650	256,186	1.078
$M/H_2/1$ Queue Waiting Times	15%	45,486	76,214	1.676
$\tau = 0.8$, $\mu_X = 8$,	7.5%	181,942	228,482	1.256
$\gamma_X = 24{,}204.8$	3.75%	727,765	798,234	1.097

Table 6: Comparison of the Average Sample Size $\bar{n}$ Required by ASAP3 with the Theoretical n^* Required for Efficiently Computing a Valid 95% CI for μ with Relative Precision r^*

Output Process	r^*	n^*	ASAP3 $\bar{n}$	ASAP3 $\bar{n}/n^*$
$M/M/1$ Queue Waiting Times	15%	75,675	140,052	1.851
$\tau = 0.9$, $\mu_X = 9$, $\gamma_X = 35{,}901$	7.5%	302,670	382,958	1.265
(Steiger et al., 2005a)	3.75%	1,210,797	1,341,522	1.108
$M/M/1$ Queue Waiting Times	15%	21,086	46,106	2.187
$\tau = 0.8$, $\mu_X = 3.2$,	7.5%	84,344	97,643	1.158
$\gamma_X = 1{,}264.64$	3.75%	337,376	365,353	1.083
$M/H_2/1$ Queue Waiting Times	15%	64,574	96,706	1.498
$\tau = 0.8$, $\mu_X = 8$,	7.5%	258,293	309,560	1.198
$\gamma_X = 24{,}204.8$	3.75%	1,033,169	1,115,986	1.080

From Tables 5 and 6 we concluded that for progressively smaller values of r^*, the ratio $\bar{n}/n^*$ tended to 1. Since Tables 5 and 6 are based on test processes that exhibit pronounced effects due to initialization bias, nonnormality, and stochastic dependence (correlation), it is not surprising that the ratio $\bar{n}/n^*$ may be substantially larger than 1 in some cases simply because of the large warm-up period and batch size that may be required to achieve some semblance of the "ideal" situation in which the truncated batch means constitute a stationary Gaussian process. All in all, we judged ASAP3's required sample sizes to be reasonable for the test problems used in this study.

5 CONCLUSIONS

The undercoverage problem encountered with ASAP was virtually eliminated in the design of ASAP2 and ASAP3 by the elimination of the test for independence of the batch means from the latter two procedures. Both ASAP2 and ASAP3 test only for stationary multivariate normality of the batch means and always deliver a CI adjusted for correlation, if any, among the final batch means. Excessive variabilities seen with ASAP in the final sample sizes, and to some extent in the final CI half-lengths, were partially resolved in ASAP2 by decreasing the significance level of the test for stationary multivariate normality on each iteration of that test. Moreover, the means and variances of the final CI half-lengths delivered by ASAP3 were greatly reduced in comparison with the corresponding quantities delivered by ASAP and ASAP2; and ASAP3 has achieved this performance improvement by progressively increasing the batch size until we can conclude that the correlation between adjacent batch means does not significantly exceed 0.8 in the sense that a one-sided upper 99% confidence interval for this correlation lies entirely below 0.8.

ASAP3 is primarily designed for use in conjunction with a user-specified absolute or relative precision requirement on the final CI; and when it is used in this way, ASAP3 generally delivers CIs whose coverage probability is close to the nominal level. On the basis of all the experimentation we have performed with the procedure, ASAP3 appears to deliver CIs whose coverage probability is reasonably close to the nominal level even in the absence of a precision requirement; but in such cases there is of course no guarantee that the resulting CIs will be narrow enough to be useful in practice. Although ASAP3 does not provide a definitive resolution of all problems associated with the batch means method for steady-state simulation output analysis, many of the undesirable behaviors of its predecessors ASAP and ASAP2 have been eliminated; and there is good evidence to show that ASAP3's performance in practice compares favorably with other well-known batch means procedures. We believe the basic approach of ASAP3 has the potential to lead to new developments in the method of batch means.

Additional experimental results, follow-up papers and revised software, will be available on the website <www.ie.ncsu.edu/jwilson>.

REFERENCES

Alexopoulos, C., and D. Goldsman. 2004. To batch or not to batch? *ACM Transactions on Modeling and Computer Simulation* 14 (1): 76–114.

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

Chow, Y. S., and H. Robbins. 1965. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics* 36: 457–462.

Fishman, G. S. 1998. LABATCH.2: Software for statistical analysis of simulation sample path data. In *Proceedings of the 1998 Winter Simulation Conference*, ed. D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, 131–139. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online via <www.informs-cs.org/wsc98papers/017.PDF> [accessed July 15, 2005].

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

Lada, E. K., J. R. Wilson, and N. M. Steiger. 2003. A wavelet-based spectral method for steady-state simulation analysis. In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 422–430. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online via <www.informs-cs.org/wsc03papers/052.pdf> [accessed July 15, 2005].

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

Nádas, A. 1969. An extension of a theorem of Chow and Robbins on sequential confidence intervals for the mean. *The Annals of Mathematical Statistics* 40 (2): 667–671.

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 online via <www.lib.ncsu.edu/etd/public/etd-19231992992670/etd.pdf> [accessed July 15, 2005].

Steiger, N. M., E. K. Lada, J. R. Wilson, C. Alexopoulos, D. Goldsman, and F. Zouaoui. 2002. ASAP2: An improved batch means procedure for simulation output analysis. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, 336–344. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online via <www.informs-cs.org/wsc02papers/043.pdf> [accessed July 15, 2005].

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, and D. Goldsman. 2003. ASAP3 software and user's manual. Department of Industrial Engineering, North Carolina State University, Raleigh, North Carolina. Available online via <ftp.ncsu.edu/pub/eos/pub/jwilson/installasap3.exe> [accessed July 15, 2005].

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, and D. Goldsman. 2005a. ASAP3: A batch means procedure for steady-state simulation analysis. *ACM Transactions on Modeling and Computer Simulation* 15 (1): 39–73.

Steiger, N. M., E. K. Lada, and J. R. Wilson. 2005b. Performance evaluation of ASAP3 for steady-state simulation analysis. *IIE Transactions*, in review. Available online via <ftp.ncsu.edu/pub/eos/pub/jwilson/iiev15t.pdf> [accessed July 15, 2005].

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 <www.informs-cs.org/wsc99papers/061.PDF> [accessed July 15, 2005].

Steiger, N. M., and J. R. Wilson. 2000. Experimental performance evaluation of batch-means procedures for simulation output analysis. In *Proceedings of the 2000 Winter Simulation Conference*, ed. R. R. Barton, J. A. Joines, K. Kang, and P. A. Fishwick, 627–636. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online via <www.informs-cs.org/wsc00papers/084.PDF> [accessed July 15, 2005].

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

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

Steiger, N. M., and J. R. Wilson. 2002b. ASAP software and user's manual. Department of Industrial Engineering, North Carolina State University, Raleigh, North Carolina. Available online via <ftp.ncsu.edu/pub/eos/pub/jwilson/installasap.exe> [accessed July 15, 2005].

Steiger, N. M., E. K. Lada, J. R. Wilson, J. A. Joines, C. Alexopoulos, and D. Goldsman. 2004. Steady-state simulation analysis using ASAP3. In *Proceedings of the 2004 Winter Simulation Conference*, ed. R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, 672–680. Available online via <www.informs-sim.org/wsc04papers/081.pdf> [accessed July 15, 2005].

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

AUTHOR BIOGRAPHIES

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

EMILY K. LADA is an operations research development tester at the SAS Institute. She is a member of IIE and INFORMS. Her e-mail address is <eklada@eos.ncsu.edu>.

JAMES R. WILSON is professor and head of the Department of Industrial Engineering at North Carolina State University. He is a member of AAUW, ACM, ASA and INFORMS, and he is a Fellow of IIE. His e-mail address is <jwilson@eos.ncsu.edu>, and his web page is <www.ie.ncsu.edu/jwilson>.