Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

SEQUEST: A SEQUENTIAL PROCEDURE FOR ESTIMATING STEADY-STATE QUANTILES

Christos Alexopoulos
David Goldsman

H. Milton Stewart School of Industrial
and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, USA

Anup Mokashi

SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513-8617, USA

Rong Nie
Qing Sun
Kai-Wen Tien
James R. Wilson

Edward P. Fitts Department of Industrial and Systems Engineering
North Carolina State University
Raleigh, NC 27695-7906, USA

ABSTRACT

Sequest is a fully sequential procedure that delivers improved point and confidence-interval (CI) estimators for a designated steady-state quantile by exploiting a combination of ideas from batching and sectioning. Sequest incorporates effective methods to do the following: (a) eliminate bias in the sectioning-based point estimator that is caused by initialization of the simulation or an inadequate simulation run length (sample size); and (b) adjust the CI half-length for the effects of skewness or correlation in the batching-based point estimators of the designated quantile. Sequest delivers a CI designed to satisfy user-specified requirements concerning both the CI's coverage probability and its absolute or relative precision. We found that Sequest exhibited good small- and large-sample properties in a preliminary evaluation of the procedure's performance on a suite of test problems that includes some problems designed to "stress test" the procedure.

1 INTRODUCTION

Simulation is perhaps the most widely used tool in the fields of industrial and systems engineering, operations research, and the management sciences. Steady-state simulations play a fundamental role in system design, and they are particularly appropriate for evaluating long-run system performance or risk. For example, let X_i denote the loss in the value of a financial portfolio over the ith time interval of a fixed length (say, two weeks or even one trading day) so that $X_i > 0$ represents the magnitude of the loss while $X_i \le 0$ indicates a gain of magnitude $|X_i| = -X_i$ over the ith time interval for $i = 1, 2, \ldots$ (Glasserman 2004). For all cut-off values $x \in \mathbb{R}$, we let $F(x) \equiv \Pr\{X_i \le x\}$ and $f(x) = F'(x)$ respectively denote the cumulative distribution function (c.d.f.) and the probability density function (p.d.f.) of X_i for $i = 1, 2, \ldots$. With this setup, for $0 < p < 1$ the $100p\%$ Value at Risk (VaR) for the portfolio is usually defined in standard statistical terminology as the p-quantile $x_p \equiv F^{-1}(p) \equiv \inf\{x : F(x) \ge p\}$, so that if we take $p = 0.99$, then the probability is 99% that the loss X_i in the ith time interval will not exceed $x_{0.99}$.

Similarly, if in a call center simulation X_i denotes the waiting time spent on hold before the ith caller reaches a service representative, then call-center management may seek convincing evidence that the $p = 0.9$ quantile $x_{0.9}$ of call-waiting time does not exceed a critical threshold x^*. If we take $x^* = 3$

978-1-4799-7486-3/14/$31.00 ©2014 IEEE

minutes, then convincing evidence that 90% of all callers wait at most 3 minutes on hold might be that a simulation-generated one-sided upper 95% confidence interval (CI) for $x_{0.90}$ does not include 3 minutes.

In the development of effective steady-state simulation analysis procedures, the main obstacle is that generally the associated output processes do not even approximately satisfy the basic assumptions underlying conventional statistical methods. In particular, successive responses are rarely independent and identically distributed (i.i.d.) normal random variables—for example, consecutive waiting times in a heavily congested queueing simulation with the empty-and-idle initial condition; and similar considerations apply to successive losses or gains in the value of a financial portfolio over an extended time horizon. When the data are identically distributed but stochastically dependent (e.g., correlated), the point estimation of x_p is straightforward: sort the observations in order $X_{(1)} \leq \cdots \leq X_{(n)}$ to yield the estimator $\widehat{x}_p = X_{(\lceil np \rceil)}$, where $\lceil \cdot \rceil$ denotes the ceiling function. If the $\{X_i\}$ are also independent and $f(x_p) > 0$, then valid large-sample CIs for x_p can also be easily computed. In this situation, the variate $\sqrt{n}(\widehat{x}_p - x_p)$ is asymptotically normal with mean zero and variance $p(1-p)/[f(x_p)]^2$ (Serfling 1980, Section 2.3.3); thus for $0 < \alpha < 1$, an asymptotically valid $100(1-\alpha)\%$ CI for x_p has the form $\widehat{x}_p \pm z_{1-\alpha/2}\big[\widehat{\mathrm{Var}}(\widehat{x}_p)\big]^{1/2}$, where $\widehat{\mathrm{Var}}(\widehat{x}_p)$ is a suitable estimator of $\mathrm{Var}(\widehat{x}_p)$ and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution. When the $\{X_i\}$ are dependent and subject to initialization bias, then the problem of computing point and CI estimators of x_p that are free of initialization bias and asymptotically reliable becomes much more difficult.

If the X_i are dependent and possibly contaminated by an initial transient, then the quantile x_p can be estimated using the data observed on a single run using any of the methods described in Bekki et al. (2010), Chen and Kelton (2006, 2008), Iglehart (1976), Jain and Chlamtac (1985), Jin, Fu, and Xiong (2003), Raatikainen (1987, 1990), and Seila (1982a, 1982b). The relatively sparse simulation literature on this problem reflects the following difficulties: (a) lack of an adequate theoretical basis for some of the existing methods; (b) lack of effective guidelines for using the methods in practice; (c) poor performance of the estimators in industrial-strength applications; and (d) excessive computational or storage requirements.

In this paper we develop Sequest, a fully sequential procedure that delivers improved point and CI estimators for a designated steady-state quantile based on a combination of ideas from batching (Tafazzoli and Wilson 2011) and sectioning (Asmussen and Glynn 2007, Section III.5a). Sequest incorporates effective methods to do the following: (a) eliminate bias in the sectioning-based point estimator that is caused by initialization of the simulation or an inadequate simulation run length (sample size); and (b) adjust the CI half-length for the effects of skewness or correlation in the batching-based point estimators of the designated quantile. Sequest delivers a CI designed to satisfy user-specified requirements concerning both the CI's coverage probability and its absolute or relative precision. The remainder of this article is organized as follows. In Section 2, we summarize the basic assumptions underlying the design of Sequest. Section 3 provides an overview of Sequest and a formal algorithmic statement of the procedure. Section 4 contains a summary of the results of a preliminary experimental performance evaluation of Sequest. Section 5 contains concluding remarks and an outline of the next steps in our work on Sequest. The slides for the oral presentation of this article are available online via www.ise.ncsu.edu/jwilson/wsc14sequest.pdf.

2 BASIC ASSUMPTIONS OF SEQUEST

To lay a sufficiently broad foundation for building point and CI estimators of the steady-state p-quantile x_p, we assume the stationary simulation output process $\{X_i : i = 0, 1, \ldots\}$ can be expressed as a (measurable, possibly nonlinear) function of a sequence of "shocks" $\{\varepsilon_i : i \in \mathbb{Z}\}$ that are i.i.d. random variables,

$$X_i = \mathbb{X}(\ldots, \varepsilon_{i-2}, \varepsilon_{i-1}, \varepsilon_i) \quad \text{for } i = 0, 1, \ldots, \tag{1}$$

so the $\{\varepsilon_i\}$ may be regarded as the stream of random numbers driving the simulation, and the function $\mathbb{X}(\cdot)$ represents the operations performed by the simulation model on its probabilistic inputs up to time i so as to generate the corresponding output response X_i. We assume that in a nonempty open interval $\mathscr{D}(x_p)$

containing the desired quantile x_p, the random variable X_i has a p.d.f. $f(x)$ with derivative $f'(x)$ such that

$$f(x_p) > 0 \quad \text{and} \quad \sup\big\{\, f(x) + |f'(x)| : x \in \mathscr{D}(x_p) \,\big\} < \infty \;. \tag{2}$$

We also assume that $\{X_i : i = 0, 1, \ldots\}$ satisfies the *geometric-moment contraction* (GMC) condition—i.e., there exist constants $\psi > 0$, $C > 0$, and $r \in (0,1)$ such that for the independent input processes $\{\varepsilon_j : j \in \mathbb{Z}\}$ and $\{\varepsilon^*_j : j \in \mathbb{Z}\}$ each consisting of i.i.d. variates, we have

$$\mathrm{E}\big[\,\big|\mathbb{X}(\ldots,\varepsilon_{-2},\varepsilon_{-1},\varepsilon_0,\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_i) - \mathbb{X}(\ldots,\varepsilon^*_{-2},\varepsilon^*_{-1},\varepsilon^*_0,\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_i)\big|^{\psi}\,\big] \le Cr^i \quad \text{for } i \ge 0\,. \tag{3}$$

The GMC condition (3) requires that if two paired replications of the simulation model associated with the function $\mathbb{X}(\cdot)$ are initialized independently but use common random numbers after the simulation starting time, then the difference $X_i - X^*_i$ between the matching output responses generated by the two simulations at time i will converge to zero in the mean of order ψ as the time index $i \to \infty$. If the GMC condition (3) holds, then the difference $X_i - X^*_i$ also converges in probability to zero as $i \to \infty$ (Bickel and Doksum 2007).

As noted by Wu (2005), condition (3) is easier to check than the usual strong mixing condition. The setup (1)–(3) applies to finite-order moving-average and autoregressive processes; and the latter class of processes forms the basis for the autoregressive method of steady-state simulation analysis (Law 2014). Moreover, conditions (1)–(3) are satisfied by a rich diversity of widely used linear and nonlinear processes, including conditional heteroscedastic (ARCH) processes, random coefficient autoregressive (RCA) processes, and threshold autoregressive (TAR) processes, as well as a broad class of Markov chains (Alexopoulos, Goldsman, and Wilson 2012).

Let $\{X_1, \ldots, X_n\}$ denote a data set from which we wish to build point and CI estimators of x_p using the Sequest procedure applied to b nonoverlapping batches each of size m. For $j = 1, \ldots, b$, we sort the jth batch of observations $\{X_{(j-1)m+1}, \ldots, X_{jm}\}$ in ascending order to obtain the order statistics $X_{j,(1)} \le X_{j,(2)} \le \cdots \le X_{j,(m)}$; and the associated batch quantile estimator (BQE) based on the jth batch of size m is

$$\widehat{x}_p(j,m) = X_{j,(\lceil mp \rceil)}\,. \tag{4}$$

Similarly from the entire data set $\{X_1, \ldots, X_n\}$ and its associated order statistics $X_{(1)} \le \cdots \le X_{(n)}$, we compute the overall point estimator of x_p,

$$\widetilde{x}_p(n) = X_{(\lceil np \rceil)}\,. \tag{5}$$

Using (4) and (5), we also compute a modified estimator of the variance of the BQEs,

$$\widetilde{S}^2_{\widehat{x}_p}(b,m) \equiv b^{-1} \sum_{j=1}^{b} \big[\widehat{x}_p(j,m) - \widetilde{x}_p(n)\big]^2. \tag{6}$$

Under the assumptions (1)–(3), it is straightforward to extend the analysis presented in Section 2 of Alexopoulos, Goldsman, and Wilson (2012) to show that as $m \to \infty$ with b fixed, an asymptotically valid $100(1-\alpha)\%$ CI for x_p has the form

$$\widetilde{x}_p(n) \pm t_{1-\alpha/2,b-1} \widetilde{S}_{\widehat{x}_p}(b,m) \big/ \sqrt{b}, \tag{7}$$

where $t_{q,\nu}$ is the q-quantile of Student's t distribution with ν degrees of freedom. For Markov processes, the validity of the CI (7) can be established under geometric ergodicity (Muñoz 2010).

Equation (7) provides the foundation on which Sequest is built. In particular, Sequest is designed to achieve the following:

- Determine a data truncation point w (i.e., the end of the warm-up period) beyond which the truncated sample statistics (4)–(6) are approximately free of initialization bias;
- Adjust the CI half-length $t_{1-\alpha/2,b-1}\widetilde{S}_{\widehat{x}_p}(b,m)/\sqrt{b}$ to compensate for any skewness or correlation in the batch quantile estimators $\left\{\widehat{x}_p(j,m) : j = 1,\ldots,b\right\}$; and
- Determine sufficiently large values of the truncation point w, the batch size m, the batch count b, and the total sample size $n = w + bm$ so that the user-specified precision and coverage probability are achieved by the final CI estimator of x_p.

3 OVERVIEW OF SEQUEST

Sequest is an adaptation to steady-state quantile estimation of Skart, a sequential procedure for estimating steady-state means (Tafazzoli and Wilson 2011). The name Sequest is an abbreviation of the phrase "Sequential quantile estimation technique"; moreover, *sequest* is a now-obsolete English word with the meaning "to follow" (Simpson and Weiner 1989). The name Sequest seems appropriate because the procedure is designed to "follow" a simulation-generated process as closely as possible so as to deliver sufficiently precise and reliable point and CI estimators of user-specified quantiles of that process. Sequest addresses the start-up problem by iteratively applying the randomness test of von Neumann (1941) to BQEs with increasing sizes for each batch; and when the randomness test is finally passed, the warm-up period is taken to be the initial batch. Sequest addresses the nonnormality problem by exploiting a Cornish–Fisher expansion for the classical Student's t-ratio based on a random sample from a nonnormal (skewed) distribution; and analysis of this expansion leads to a modified t-ratio that incorporates terms due to Johnson (1978) and Willink (2005) so as to compensate for any skewness in the final set of "warmed up" (truncated) BQEs. Sequest addresses the correlation problem by using a first-order autoregressive model of the truncated BQEs to compensate for any residual correlation between those BQEs. To achieve the user-specified precision in the final CI for x_p, Sequest may request additional simulation-generated observations; and several iterations of Sequest may be performed until a CI satisfying the precision requirement is finally delivered. A formal algorithmic statement of Sequest is given below.

Algorithmic Statement of Sequest

[0] Set the initial sample size $n \leftarrow 4096$, batch size $m \leftarrow 64$, and batch count $b \leftarrow 64$. Set the initial absolute tolerance on the sample variance of the BQEs, $\varepsilon_a \leftarrow 10^{-10}$, and the associated relative tolerance, $\varepsilon_r \leftarrow 10^{-5}$. Set the tolerance on the skewness adjustment, $\varepsilon_s \leftarrow 10^{-3}$. Set the randomness test size, $\alpha_{\text{ran}} \leftarrow 0.25$. Set the parameters $\eta \leftarrow 2.82888$ and $\theta \leftarrow 2$ of the upper-bound function on absolute skewness of the BQEs,

$$\mathscr{B}^*(p) = \exp\left(-\eta|p-0.5|^{\theta}\right) \quad \text{for} \quad p \in (0,1).$$

Finally, set the upper bound on the number of iterations of the skewness-reducing batch-size adjustment step, $u^* \leftarrow 5$.

[1] From the initial time series $\{X_i : i = 1,\ldots,n\}$, form b batches of size m to compute the BQEs (4). Compute the sample mean and sample variance of the BQEs,

$$\bar{x}_p(b,m) \leftarrow \frac{1}{b}\sum_{j=1}^{b}\widehat{x}_p(j,m) \quad \text{and} \quad S^2_{\widehat{x}_p}(b,m) \leftarrow \frac{1}{b-1}\sum_{j=1}^{b}\left[\widehat{x}_p(j,m)-\bar{x}_p(b,m)\right]^2. \tag{8}$$

[a] If

$$S_{\widehat{x}_p}(b,m) \le \min\{\varepsilon_a, \varepsilon_r|\bar{x}_p(b,m)|\},$$

then go to step **[1b]**; otherwise go to step **[2]**.

[b] Update the batch size and the total sample size according to $m \leftarrow 2m$ and $n \leftarrow 2n$; obtain the required additional observations by restarting the simulation if necessary; update the BQEs (4) and the sample statistics (8); and return to step **[1a]**.

[2] Apply von Neumann's test for randomness to the current set of BQEs $\{\widehat{x}_p(j,m) : j = 1,\ldots,b\}$ by computing the test statistic

$$C_b \leftarrow 1 - \frac{\sum_{j=1}^{b-1}[\widehat{x}_p(j,m) - \widehat{x}_p(j+1,m)]^2}{2(b-1)S^2_{\widehat{x}_p}(b,m)}\,. \tag{9}$$

[a] If

$$|C_b| \le z_{1-\alpha_{\text{ran}}/2}\sqrt{(b-2)/(b^2-1)}\,,$$

then go to step **[3]**; otherwise proceed to step **[2b]**.

[b] Update the batch size and sample size according to $m \leftarrow 2m$ and $n \leftarrow 2n$; obtain the required additional observations by restarting the simulation if necessary; update the BQEs (4), the sample statistics (8), and the randomness test statistic (9); and return to step **[2a]**.

[3] Set the length of the warm-up period according to $w \leftarrow m$. Initialize the skewness-reduction iteration counter, $u \leftarrow 0$.

[a] Update the total sample size,

$$n \leftarrow w + bm\,,$$

and obtain the additional observations by restarting the simulation if necessary. Skip the first w observations in the overall time series of length n so that we have the "warmed-up" (truncated) time series of length $n' \leftarrow n - w$,

$$\{Y_i = X_{w+i} : i = 1,\ldots,n'\}\,. \tag{10}$$

For $j = 1,\ldots,b$, define the jth "warmed-up" batch by

$$\{Y_{(j-1)m+i} : i = 1,\ldots,m\}, \tag{11}$$

with associated order statistics

$$Y_{j,(1)} \le Y_{i,(2)} \le \cdots \le Y_{j,(m)} \tag{12}$$

so that the jth "warmed-up" BQE is

$$\widehat{y}_p(j,m) \leftarrow Y_{j,(\lceil mp \rceil)}\,. \tag{13}$$

Compute the sample mean $\bar{y}_p(b,m)$ and sample variance $S^2_{\widehat{y}_p}(b,m)$ of the BQEs in (13). Compute the sample skewness of the BQEs,

$$\widehat{\mathscr{B}}_{\widehat{y}_p}(b,m) \leftarrow \frac{b}{(b-1)(b-2)}\sum_{j=1}^{b}\left[\frac{\widehat{y}_p(j,m) - \bar{y}_p(b,m)}{S_{\widehat{y}_p}(b,m)}\right]^3.$$

[b] If

$$\left|\widehat{\mathscr{B}}_{\widehat{y}_p}(b,m)\right| \le \mathscr{B}^*(p) \quad \text{or} \quad u = u^*\,,$$

then go to step **[4]**; otherwise increase the batch size according to

$$m \leftarrow \left\lceil m \cdot \text{mid}\left\{\sqrt{2}, \left[\widehat{\mathscr{B}}_{\widehat{y}_p}(b,m)|/\mathscr{B}^*(p)\right]^2, 16\right\}\right\rceil\,,$$

where $\mathrm{mid}\{u_1, u_2, u_3\} \equiv u_{(2)}$, and return to step **[3a]**.

[4] Update the batch count, batch size, and total sample size according to

$$b \leftarrow b/2, \quad m \leftarrow 2m, \quad \text{and} \quad n \leftarrow w + bm;$$

and obtain the required additional observations by restarting the simulation if necessary.

[5] Update the warmed-up BQEs (13), their sample mean $\bar{y}_p(b,m)$, sample variance $S^2_{\widehat{y}_p}(b,m)$, and sample skewness $\widehat{\mathscr{B}}_{\widehat{y}_p}(b,m)$; then compute the sample lag-one correlation of the BQEs,

$$\widehat{\varphi}_{\widehat{y}_p}(b,m) \leftarrow \frac{1}{b-1} \sum_{j=1}^{b-1} \frac{[\widehat{y}_p(j,m) - \bar{y}_p(b,m)][\widehat{y}_p(j+1,m) - \bar{y}_p(b,m)]}{S^2_{\widehat{y}_p}(b,m)},$$

and the associated correlation adjustment

$$A \leftarrow \max\left\{ \left[1 + \widehat{\varphi}_{\widehat{y}_p}(b,m)\right] \big/ \left[1 - \widehat{\varphi}_{\widehat{y}_p}(b,m)\right], 1 \right\}$$

that will be applied to the half-length of the CI estimator for x_p.

From the warmed-up time series (10), compute the associated the order statistics $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n')}$; then compute the overall sectioning-based point estimator $\widetilde{y}_p(n')$ of x_p as follows:

$$\widetilde{y}_p(n') \leftarrow Y_{(\lceil n'p \rceil)}. \tag{14}$$

From the updated sample skewness $\widehat{\mathscr{B}}_{\widehat{y}_p}(b,m)$ compute the associated skewness-adjustment parameter,

$$\beta \leftarrow \widehat{\mathscr{B}}_{\widehat{y}_p}(b,m) \big/ \left(6\sqrt{b}\right),$$

and define the skewness-adjustment function

$$G(\zeta) = \begin{cases} \zeta, & \text{if } |\beta| \le \varepsilon_s, \\ \dfrac{\sqrt[3]{1 + 6\beta(\zeta - \beta)} - 1}{2\beta}, & \text{if } |\beta| > \varepsilon_s, \end{cases}$$

for all real ζ, where $\sqrt[3]{\zeta} \equiv \mathrm{sign}(\zeta)\sqrt[3]{|\zeta|}$ (Tafazzoli and Wilson 2011). Compute the modified sample variance of the BQEs,

$$\widetilde{S}^2_{\widehat{y}_p}(b,m) \leftarrow \frac{1}{b} \sum_{j=1}^{b} [\widehat{y}_p(j,m) - \widetilde{y}_p(n')]^2$$

based on the overall quantile point estimator (14).

[6] Compute the "half-length" of the bias-, correlation-, and skewness-adjusted $100(1-\alpha)\%$ CI for the p-quantile x_p,

$$H \leftarrow \max\left\{ G(t_{1-\alpha/2,b-1}), G(t_{\alpha/2,b-1}) \right\} \left[A\widetilde{S}^2_{\widehat{y}_p}(b,m)/b \right]^{1/2},$$

and the associated CI,

$$\widetilde{y}_p(n') \pm H. \tag{15}$$

If no precision level is specified, then deliver the CI (15) and stop; otherwise proceed to step **[7]**.

[7] Apply the appropriate absolute- or relative-precision stopping rule.

[a] If the half-length H of the current CI (15) satisfies the user-specified precision requirement

$$H \le H^*, \tag{16}$$

where

$$H^* = \begin{cases} r^* \left|\widetilde{y}_p(n')\right|, & \text{for a relative precision level } r^*, \\ h^*, & \text{for an absolute precision level } h^*, \end{cases} \tag{17}$$

then deliver the CI (15) and stop; otherwise proceed to step **[7b]**.

[b] For the fixed batch count b, estimate the batch size m required to satisfy (16)–(17),

$$m \leftarrow \left\lceil m \cdot \text{mid}\left\{1.02, (H/H^*)^2, 2\right\} \right\rceil .$$

Update the length of the warmed-up time series to $n' \leftarrow bm$. Obtain the required additional observations by restarting the simulation if necessary, and return to step **[5]**.

4 EXPERIMENTAL PERFORMANCE EVALUATION OF SEQUEST

4.1 First-Order Autoregressive (AR(1)) Process

Table 1 shows the results of applying Sequest to a first-order autoregressive (AR(1)) process with the initial condition $X_0 = 0$, the autoregressive parameter $\rho = 0.995$, and the steady-state mean $\mu_X = 100$. This process is generated via the relation $X_i = \mu_X + \rho(X_{i-1} - \mu_X) + \varepsilon_i$, for $i = 1, 2, \ldots$, where $\{\varepsilon_i : i = 1, 2, \ldots\}$ are i.i.d. $N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon^2 = 1$. We applied Sequest to 1000 replications of this process.

Table 1: Performance of Sequest-delivered point and 95% CI estimators of the p-quantile x_p of the AR(1) process described in Section 4.1 based on 1000 replications.

No CI Precision Requirement								
p	x_p	Avg. $\widetilde{y}_p(n')$	$\left\lvert\text{Bias}\left[\widetilde{y}_p(n')\right]\right\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\overline{m}$	$\bar{n}$
0.3	94.7494	94.7606	0.0112	1.4025	1.4801	94.30%	5139	167014
0.5	100	100.0280	0.0290	1.4646	1.4642	94.60%	4107	133468
0.7	105.2506	105.2467	0.0039	1.5552	1.4777	94.90%	3866	125638
0.9	112.8316	112.7742	0.0574	1.7782	1.5768	93.40%	3912	127009
0.95	116.4691	116.3653	0.1038	1.9177	1.6480	94.30%	4328	140325
CI Relative Precision = 1.3%								
p	x_p	Avg. $\widetilde{y}_p(n')$	$\left\lvert\text{Bias}\left[\widetilde{y}_p(n')\right]\right\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\overline{m}$	$\bar{n}$
0.3	94.7494	94.8064	0.0570	1.0577	1.1156	94.30%	6785	219673
0.5	100	100.0245	0.0245	1.1138	1.1136	95.30%	5463	176863
0.7	105.2506	105.2691	0.0185	1.1786	1.1196	93.50%	5177	167596
0.9	112.8316	112.8059	0.0257	1.2807	1.1354	93.80%	5718	184808
0.95	116.4691	116.4293	0.0398	1.3314	1.1434	94.90%	6781	218818
CI Relative Precision = 1.0%								
p	x_p	Avg. $\widetilde{y}_p(n')$	$\left\lvert\text{Bias}\left[\widetilde{y}_p(n')\right]\right\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\overline{m}$	$\bar{n}$
0.3	94.7494	94.7919	0.0425	0.8435	0.8899	94.50%	9829	317085
0.5	100	100.0254	0.0254	0.8883	0.8880	95.70%	8087	260821
0.7	105.2506	105.2553	0.0047	0.9352	0.8885	94.40%	7727	249201
0.9	112.8316	112.8234	0.0082	1.0055	0.8912	94.50%	8859	285313
0.95	116.4691	116.4423	0.0268	1.0327	0.8869	94.90%	10652	342688

The high correlation between successive observations in this process makes it a severe test of Sequest's ability to handle correlated observations and to deliver an approximately valid correlation-adjusted CI. The steady-state marginal standard deviation of this test process is $\sigma_X = \sigma_\varepsilon \big/ \sqrt{1-\rho^2} = 10.0125$; therefore this process starts approximately ten standard deviations below its steady-state mean. The magnitude and duration of the initial transient in simulation-generated realizations of the AR(1) process under study was purposely designed to "stress-test" Sequest's ability to eliminate initialization bias as well as to compensate effectively for pronounced correlation among successive observations of a target process.

Table 1 shows that for all precision levels, Sequest's sampling efficiency was good. For $p \in [0.3, 0.95]$ in the no precision case, Sequest delivered the point estimator $\widetilde{y}_p(n')$ of x_p with average bias (labeled $\overline{\text{Bias}}\big[\widetilde{y}_p(n')\big]$ in the table) ranging in magnitude from 0.0039 to 0.104, while the corresponding values of x_p ranged from 94.7494 to 116.4691. Moreover in the no precision case, Sequest delivered nominal 95% CIs with coverages ranging from 93.4% to 94.9% and average values of the CI relative precision $100 \times \big|H/\widetilde{y}_p(n')\big|$ ranging from 1.46% to 1.65%. The results for relative precision levels $r^* = 0.013$ and $r^* = 0.01$ were judged to be similarly good—especially with respect to the increase in sample size required to satisfy the precision requirement relative to the no precision case.

4.2 *M/M/*1 Queue-Waiting-Time Process

Consider an $M/M/1$ queueing system with interarrival rate $\lambda = 0.8$ and service rate $\omega = 1$, and let X_i be the time spent in queue by entity i prior to receiving service. Let $\rho = \lambda/\omega = 0.8$ denote the traffic intensity. It is well known that the steady-state c.d.f. of X_i is

$$F(x) = \begin{cases} 0, & \text{for } x < 0, \\ 1 - \rho, & \text{for } x = 0, \\ 1 - \rho e^{-\omega(1-\rho)x}, & \text{for } x > 0; \end{cases} \tag{18}$$

hence the response X_i has steady-state mean $\mu_X = 4$ and the steady-state quantiles can be evaluated analytically by inverting the c.d.f. (18). We assume that the system starts empty ($X_1 = 0$). Table 2 shows the results of applying Sequest to 1000 replications of this process.

Table 2: Performance of Sequest-delivered point and 95% CI estimators of the p-quantile x_p of the $M/M/1$ queue waiting-time-process described in Section 4.2 based on 1000 replications.

				No CI Precision Requirement				
p	x_p	$\widetilde{y}_p(n')$	$\big\lvert\overline{\text{Bias}}\big[\widetilde{y}_p(n')\big]\big\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\overline{m}$	$\bar{n}$
0.3	0.6676	0.6673	3.20E-04	0.0645	9.6629	97.00%	9383	300439
0.5	2.35	2.3485	0.0015	0.1582	6.7352	96.50%	7829	250762
0.7	4.9041	4.8991	0.0050	0.2794	5.7038	95.50%	10880	348437
0.9	10.3972	10.3559	0.0413	0.3861	3.7288	93.50%	39730	1272035
0.95	13.8629	13.7678	0.0951	0.4529	3.2897	93.70%	80001	2560469
				CI Relative Precision = 3.0%				
p	x_p	$\widetilde{y}_p(n')$	$\big\lvert\overline{\text{Bias}}\big[\widetilde{y}_p(n')\big]\big\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\overline{m}$	$\bar{n}$
0.3	0.6676	0.6672	3.84E-04	0.0178	2.6718	96.50%	66786	2137343
0.5	2.35	2.3498	0.0002	0.0622	2.6486	95.90%	23577	754701
0.7	4.9041	4.9026	0.0015	0.1274	2.5987	94.90%	21705	694852
0.9	10.3972	10.3874	0.0098	0.2382	2.2933	95.40%	47338	1515501
0.95	13.8629	13.8546	0.0083	0.2933	2.1167	95.80%	88303	2826131
				CI Relative Precision = 2.5%				
p	x_p	$\widetilde{y}_p(n')$	$\big\lvert\overline{\text{Bias}}\big[\widetilde{y}_p(n')\big]\big\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\overline{m}$	$\bar{n}$
0.3	0.6676	0.6672	3.79E-04	0.0150	2.2477	95.50%	93812	3002160
0.5	2.35	2.3488	0.0012	0.0524	2.2292	95.00%	32330	1034797
0.7	4.9041	4.9022	0.0019	0.1079	2.2007	95.70%	28921	925752
0.9	10.3972	10.3885	0.0087	0.2107	2.0278	94.60%	54055	1730439
0.95	13.8629	13.8549	0.0080	0.2642	1.9066	95.50%	95675	3062058

The warm-up period for this process is relatively short, and consequently the effects of initialization bias on our quantile estimators are much less than for the AR(1) process in Section 4.1. However, the marginal c.d.f. (18) of the $M/M/1$ queue waiting times is markedly nonnormal, having an atom at zero (that is, a nonzero probability mass at zero) and an exponential tail. This characteristic induces a positive skewness in the batch quantile estimators (13) that significantly distorts the behavior of the conventional sectioning-based CI given by Equation (7), resulting in coverage probabilities significantly below the nominal level $1-\alpha$.

Table 2 shows that for all precision levels, Sequest's sampling efficiency was good. For $p \in [0.3, 0.95]$ in the no precision case, Sequest delivered the point estimator $\widetilde{y}_p(n')$ of x_p with average bias ranging in magnitude from 3.20×10^{-4} (for $x_{0.3} = 0.6676$) to 0.0951 (for $x_{0.95} = 13.8629$). Moreover in the no precision case, Sequest delivered nominal 95% CIs with coverages ranging from 97.0% (for $p = 0.3$) to 93.5% (for $p = 0.9$) together with average values of the CI relative precision ranging from 3.29% (for $p = 0.95$) to 9.66% (for $p = 0.3$). The results for relative precision levels $r^* = 0.03$ and $r^* = 0.025$ were judged to be similarly good—especially with respect to the increase in sample size required to satisfy the precision requirement relative to the no precision case.

4.3 *M/M/*1/LIFO Queue-Waiting-Time Process

The next test process was the sequence of queue waiting times for the $M/M/1/\text{LIFO}$ queue, with customers in the queue being served in last-in-first-out (LIFO) order, an empty-and-idle initial condition, arrival rate $\lambda = 1.0$, and service rate $\mu = 1.25$. In steady-state operation this system has a server utilization of $\rho = 0.8$ and a mean queue waiting time of $\mu_X = 3.2$. The $M/M/1/\text{LIFO}$ queue-waiting-time process was selected for two reasons: (a) unlike the two previous test processes, the autocorrelation function for this process does not decline in magnitude geometrically fast with increasing lags; and (b) the process has a highly nonnormal marginal distribution that significantly distorts the behavior of the conventional sectioning-based CI (7), resulting in coverage probabilities significantly below the nominal level $1-\alpha$. Table 3 shows the results of applying Sequest to 1000 replications of this process. We computed the "exact" value of each selected quantile $x_p = F^{-1}(p)$ as follows: (a) we numerically inverted the Laplace transform of the steady-state marginal c.d.f. $F_{B_{\text{FIFO}}}(\cdot)$ of a busy period in the $M/M/1$ queue with the same arrival rate λ and service rate μ (Kleinrock 1975, Equation (5.144)) using the Euler algorithm of Abate and Whitt (2006); (b) we combined the relation

$$F(x) = \begin{cases} 0, & \text{if } x < 0, \\ (1-\rho) + \rho F_{B_{\text{FIFO}}}(x), & \text{if } x \geq 0, \end{cases}$$

with the result of (a) to compute a piecewise-linear approximation to $F(x)$ for $0 \leq x \leq 75$ in increments of size $\Delta x = 10^{-3}$; and (c) we inverted the result of (b) to yield an estimate of x_p with high accuracy.

Table 3 shows that for all precision levels, Sequest's sampling efficiency was good. For $p \in [0.3, 0.95]$ in the no precision case, Sequest delivered the point estimator $\widetilde{y}_p(n')$ of x_p with average bias ranging in magnitude from 2.80×10^{-5} (for $x_{0.5} = 0.4692$) to 0.0084 (for $x_{0.95} = 14.4052$). Moreover in the no precision case, Sequest delivered nominal 95% CIs with coverages ranging from 97.2% (for $p = 0.5$) to 95.8% (for $p = 0.3$) together with average values of the CI relative precision ranging from 3.27% (for $p = 0.95$) to 27.86% (for $p = 0.3$). The results for relative precision levels $r^* = 0.03$ and $r^* = 0.025$ were judged to be similarly good—especially with respect to the amount of additional sampling compared with the no precision case that was required to satisfy the precision requirement. It is particularly noteworthy that with the imposition of the precision requirement $r^* = 0.03$, the average relative precision of the CIs delivered by Sequest for $p = 0.3$ declined from 27.86% (for the no precision case with average batch size $\overline{m} = 462$ and average total sample size $\bar{n} = 13940$) to 2.71% (with $\overline{m} = 32011$ and $\bar{n} = 1024513$ when $r^* = 0.03$); and at the same time, the CI coverages for these two cases were 95.8% and 95.5%, respectively.

5 CONCLUSIONS

In this article we describe Sequest, a fully sequential procedure for delivering improved point and CI estimators of steady-state quantiles of a simulation output process. Sequest delivers a CI designed to satisfy user-specified requirements concerning both the CI's coverage probability and its absolute or relative precision. We found that Sequest exhibited good small- and large-sample properties in a preliminary evaluation of the procedure's performance in a suite of test problems that includes some problems designed

Table 3: Performance of Sequest-delivered point and 95% CI estimators of the p-quantile x_p of the $M/M/1/$LIFO queue-waiting-time process described in Section 4.3 based on 1000 replications.

No CI Precision Requirement

p	x_p	$\widetilde{y}_p(n')$	$\left\lvert\overline{\text{Bias}}\left[\widetilde{y}_p(n')\right]\right\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\overline{m}$	$\overline{n}$
0.3	0.1129	0.1146	0.0017	0.0319	27.8575	95.80%	462	14940
0.5	0.4692	0.4692	2.80E-05	0.0625	13.3175	97.20%	433	14000
0.7	1.3579	1.3581	1.51E-04	0.0872	6.4228	97.10%	1704	54654
0.9	6.718	6.7200	0.0020	0.2666	3.9675	95.90%	9807	313911
0.95	14.4052	14.3968	0.0084	0.4710	3.2714	96.00%	24949	798473

CI Relative Precision = 3.0%

p	x_p	$\widetilde{y}_p(n')$	$\left\lvert\overline{\text{Bias}}\left[\widetilde{y}_p(n')\right]\right\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\overline{m}$	$\overline{n}$
0.3	0.1129	0.1129	1.57E-06	0.0031	2.7109	95.50%	32011	1024513
0.5	0.4692	0.4691	9.09E-05	0.0126	2.6933	94.90%	5834	186825
0.7	1.3579	1.3580	7.81E-05	0.0362	2.6665	95.20%	5030	161097
0.9	6.718	6.7168	0.0012	0.1678	2.4975	95.40%	14353	459375
0.95	14.4052	14.3978	0.0074	0.3413	2.3705	95.60%	29962	958892

CI Relative Precision = 2.5%

p	x_p	$\widetilde{y}_p(n')$	$\left\lvert\overline{\text{Bias}}\left[\widetilde{y}_p(n')\right]\right\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\overline{m}$	$\overline{n}$
0.3	0.1129	0.1129	8.28E-06	0.0025	2.2423	95.80%	46588	1490978
0.5	0.4692	0.4691	1.13E-04	0.0105	2.2353	94.30%	8420	269587
0.7	1.3579	1.3580	9.14E-05	0.0304	2.2355	95.00%	7094	227142
0.9	6.718	6.7165	0.0015	0.1444	2.1499	95.60%	18139	580526
0.95	14.4052	14.3974	0.0078	0.3002	2.0848	95.60%	35195	1126361

to "stress test" the procedure. In all the test problems to which we have applied Sequest so far (including all the processes that Tafazzoli et al. (2011) used to evaluate the performance of Skart), we have found that Sequest was competitive with previously developed methods for estimating steady-state quantiles of simulation output processes (Bekki et al. 2010, Chen and Kelton 2006, Jain and Chlamtac 1985, Moore 1980, Raatikainen 1987, Raatikainen 1990, Seila 1982a, Seila 1982b).

Future work on Sequest will include adaptation of the maximum transformation (Heidelberger and Lewis 1984) to reduce the samples sizes required for estimating p-quantiles when p is close to 0 or 1, especially in the case that $p \in \{0.9, 0.95, 0.99\}$. We will also perform a thorough sensitivity analysis of the performance of Sequest with respect to variation in the procedure's numerous parameters ("magic numbers"), with the ultimate objective of achieving substantial improvements in the performance of the procedure. A critical aspect of future work will be to gain a better understanding of the effect of noninitialization bias on the performance of point and CI estimators of steady-state quantiles, and to exploit this understanding in formulating more effective methods for estimating the required batch and sample sizes at each step of Sequest.

ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation grants CMMI-1233141/1232998.

REFERENCES

Abate, J., and W. Whitt. 2006. "A Unified Framework for Numerically Inverting Laplace Transforms." *INFORMS Journal on Computing* 18 (4): 408–421.

Alexopoulos, C., D. Goldsman, and J. R. Wilson. 2012. "A New Perspective on Batched Quantile Estimation." In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 190–200. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. New York: Springer Science+Business Media.

Bekki, J. M., J. W. Fowler, G. T. Mackulak, and B. L. Nelson. 2010. "Indirect Cycle Time Quantile Estimation Using the Cornish–Fisher Expansion." *IIE Transactions* 42 (1): 31–44.

Bickel, P. J., and K. A. Doksum. 2007. *Mathematical Statistics: Basic Ideas and Selected Topics*. 2nd ed. Upper Saddle River, NJ: Pearson Prentice Hall.

Chen, E. J., and W. D. Kelton. 2006. "Quantile and Tolerance-Interval Estimation in Simulation." *European Journal of Operational Research* 168:520–540.

Chen, E. J., and W. D. Kelton. 2008. "Estimating Steady-State Distributions via Simulation-Generated Histograms." *Computers and Operations Research* 35 (4): 1003–1016.

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. Heidelberg, Germany: Springer-Verlag.

Heidelberger, P., and P. A. W. Lewis. 1984. "Quantile Estimation in Dependent Sequences." *Operations Research* 32:185–209.

Iglehart, D. L. 1976. "Simulating Stable Stochastic Systems, VI: Quantile Estimation." *Journal of the Association for Computing Machinery* 23:347–360.

Jain, R., and I. Chlamtac. 1985. "The P^2 Algorithm for Dynamic Calculation of Quantiles and Histograms without Storing Observations." *Communications of the ACM* 28 (10): 1076–1085.

Jin, X., M. C. Fu, and X. Xiong. 2003. "Probabilistic Error Bounds for Simulation Quantile Estimators." *Management Science* 49:230–246.

Johnson, N. J. 1978. "Modified t Tests and Confidence Intervals for Asymmetrical Populations." *Journal of the American Statistical Association* 73 (363):536–544.

Kleinrock, L. 1975. *Queueing Systems, Volume I: Theory*. New York: Wiley.

Law, A. M. 2014. *Simulation Modeling and Analysis*. 5th ed. New York: McGraw-Hill.

Moore, L. W. 1980. *Quantile Estimation Methods in Regenerative Processes*. Ph. D. thesis, Department of Statistics, University of North Carolina, Chapel Hill, NC.

Muñoz, D. F. 2010. "On the Validity of the Batch Quantile Method for Markov Chains." *Operations Research Letters* 38 (3): 223–226.

Raatikainen, K. E. E. 1987. "Simultaneous Estimation of Several Percentiles." *Simulation* 49:159–163.

Raatikainen, K. E. E. 1990. "Sequential Procedure for Simultaneous Estimation of Several Percentiles." *Transactions of the Society for Computer Simulation* 7 (1): 21–44.

Seila, A. F. 1982a. "A Batching Approach to Quantile Estimation in Regenerative Simulations." *Management Science* 28 (5): 573–581.

Seila, A. F. 1982b. "Estimation of Percentiles in Discrete Event Simulation." *Simulation* 6:193–200.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons.

Simpson, J. A., and E. S. C. Weiner. (Eds.) 1989. *Oxford English Dictionary*. Oxford, UK: Oxford University Press.

Tafazzoli, A., and J. R. Wilson. 2011. "Skart: A Skewness- and Autoregression-Adjusted Batch Means Procedure for Simulation Analysis." *IIE Transactions* 43 (2): 110–128.

Tafazzoli, A., J. R. Wilson, E. K. Lada, and N. M. Steiger. 2011. "Performance of Skart: A Skewness- and Autoregression-Adjusted Batch-Means Procedure for Simulation Analysis." *INFORMS Journal on Computing* 23:297–314.

von Neumann, J. 1941. "Distribution of the Ratio of the Mean Square Successive Difference to the Variance." *Annals of Mathematical Statistics* 12 (4):367–395.

Willink, R. 2005. "A Confidence Interval and Test for the Mean of an Asymmetric Distribution." *Communications in Statistics—Theory and Methods* 34:753–766.

Wu, W. B. 2005. "On the Bahadur Representation of Sample Quantiles for Dependent Sequences." *Annals of Statistics* 33 (4): 1924–1963.

AUTHOR BIOGRAPHIES

CHRISTOS ALEXOPOULOS is a professor in the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. His research interests are in the areas of simulation, statistics, and optimization of stochastic systems. He is a member of INFORMS and an active participant in the Winter Simulation Conference, having been *Proceedings* Co-Editor in 1995, Associate Program Chair in 2006, and a member of the Board of Directors since 2008. He is also an Area Editor of the *ACM Transactions on Modeling and Computer Simulation*. His e-mail address is christos@isye.gatech.edu, and his Web page is www.isye.gatech.edu/∼christos.

DAVID GOLDSMAN is a professor in the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. His research interests include simulation output analysis, ranking and selection, and healthcare simulation. He was Program Chair of the Winter Simulation Conference in 1995 and a member of the WSC Board of Directors between 2001–2009. He is currently a trustee of the WSC Foundation. His e-mail address is sman@gatech.edu, and his Web page is www.isye.gatech.edu/∼sman.

ANUP C. MOKASHI is an operations research development tester for SAS Simulation Studio at the SAS Institute. He is a member of IIE and INFORMS. His e-mail address is Anup.Mokashi@sas.com.

RONG NIE is a second-year Ph.D. student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. Her current research interests are focused on applied probability and statistical modeling, with applications to healthcare and production systems. Before coming to NCSU, she obtained a Master of Science degree in Statistics from Virginia Tech (2012). Her email address is rnie@ncsu.edu.

QING SUN is a first-year Ph.D. student in the Graduate Program in Operations Research at North Carolina State University. Her current research interests are focused on simulation analysis and probability theory. Her email address is qsun3@ncsu.edu.

KAI-WEN TIEN received his Master of Industrial Engineering degree from North Carolina State University in 2014. Currently he is a Ph.D. student in the Harold and Igne Marcus Department of Industrial and Manufacturing Engineering at Penn State University. His research interests are focused on probability theory, queueing theory, and simulation analysis. His email is tkaiwen@ncsu.edu.

JAMES R. WILSON is a professor in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. His current research interests are focused on probabilistic and statistical issues in the design and analysis of simulation experiments, with special emphasis on applications in healthcare and production. As a WSC participant, he served as *Proceedings* Editor (1986), Associate Program Chair (1991), and Program Chair (1992). During the period 1997–2004, he was a member of the WSC Board of Directors. He is a member of ACM, ASA, and ASEE; and he is a Fellow of IIE and INFORMS. His e-mail address is jwilson@ncsu.edu, and his Web page is www.ise.ncsu.edu/jwilson.