Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

SEQUEM: ESTIMATING EXTREME STEADY-STATE QUANTILES VIA THE MAXIMUM TRANSFORMATION

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

Kai-Wen Tien

Harold and Inge Marcus Department of
Industrial and Manufacturing Engineering
Pennsylvania State University
University Park, PA 16802, USA

James R. Wilson

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

ABSTRACT

This article presents Sequem, a fully sequential procedure for computing point estimators and confidence intervals (CIs) for extreme steady-state quantiles of a simulation output process. The method is an enhancement of the Sequest procedure proposed by Alexopoulos et al. in 2014 for estimating nonextreme steady-state quantiles. Sequem exploits a combination of batching, sectioning, and the maximum transformation technique to achieve the following: (a) reduction in point-estimator bias arising from initial conditions or inadequate simulation run length; and (b) adjustment of the CI half-length to compensate for the effects of skewness or correlation in the corresponding quantile point estimators obtained from nonoverlapping batches. The CIs delivered by Sequem satisfy user-specified requirements related to coverage probability and absolute or relative precision. A preliminary evaluation based on three "stress-testing" processes revealed that Sequem exhibited good performance when used in challenging settings.

1 INTRODUCTION

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 in a call center simulation, let X_k denote the waiting time spent on hold before the kth caller reaches a service representative for $k = 1, 2, \ldots$. Call center management may seek convincing evidence that in the long run, at least 99% of all call waiting times do not exceed a critical threshold x^*, say $x^* = 2$ minutes. For each possible threshold $x \in \mathbb{R}$, we let $F_X(x) \equiv \Pr\{X_k \le x\}$ and $f_X(x) = F'_X(x)$ respectively denote the cumulative distribution function (c.d.f.) and the probability density function (p.d.f.) of the steady-state distribution of X_k that is achieved as $k \to \infty$. With this setup, for $0 < p < 1$ we define the p-quantile $x_p \equiv F_X^{-1}(p) \equiv \min\{x : F_X(x) \ge p\}$ of the steady-state distribution of call waiting times, so that if we take $p = 0.99$, then the long-run probability is 0.99 that X_k does not exceed $x_{0.99}$. As convincing evidence that in the long run at least 99% of all call waiting times do not exceed 2 minutes, management might require that an asymptotically valid 95% confidence interval (CI) for $x_{0.99}$ lies to the left of the point $x^* = 2$ minutes. Many reliability and risk simulations are similarly based on point and CI estimators of a selected extreme quantile x_p, where p is taken sufficiently close to one to reflect the desired likelihood of achieving an acceptable level of system performance.

978-1-4673-9743-8/15/$31.00 ©2015 IEEE

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). When a simulation-generated time series $\{X_k : k = 1,\ldots,n\}$ of length n is composed of identically distributed but stochastically dependent (e.g., correlated) observations, the point estimation of x_p is straightforward: sort the observations in ascending order $X_{(1)} \le \cdots \le X_{(n)}$ to yield the estimator $\widehat{x}_p = X_{(\lceil np \rceil)}$, where $\lceil \cdot \rceil$ denotes the ceiling function. If the observations $\{X_k : k = 1,\ldots,n\}$ are also independent and $f_X(x_p) > 0$, then a valid large-sample CI 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 $p(1-p)\big/\big[f_X(x_p)\big]^2$ as $n \to \infty$ (Serfling 1980, Section 2.3.3); therefore as $n \to \infty$ with a fixed value of $\alpha \in (0,1)$, 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 $\mathrm{Var}(\widehat{x}_p)$ computed from the data set $\{X_k : k = 1,\ldots,n\}$ and z_γ is the γ quantile of the standard normal distribution.

If the $\{X_k\}$ 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. The relatively sparse simulation literature on this problem prior to 2014—including 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)—reflects the following difficulties: (i) lack of an adequate theoretical basis for some of the existing methods; (ii) lack of effective guidelines for using the methods in practice; (iii) poor performance of the estimators in industrial-strength applications; and (iv) excessive computational or storage requirements.

The Sequest method proposed by Alexopoulos et al. (2014) is the first fully sequential procedure in the literature that delivers an improved CI for a designated steady-state quantile and satisfies user-specified requirements on the CI's coverage probability and its absolute or relative precision. The improvement over existing methods is with respect to the CI's coverage probability and the average required sample size. Sequest is based on a combination of ideas from batching (Tafazzoli and Wilson 2011) and sectioning (Asmussen and Glynn 2007, Section III.5a). In particular, Sequest incorporates effective methods to do the following: (i) eliminate bias in the sectioning-based point estimator that is caused by an atypical initial condition for the simulation or by an inadequate simulation run length (sample size); and (ii) adjust the CI half-length for the effects of skewness or correlation in the batching-based point estimators of the designated quantile.

Substantial experimentation with Sequest revealed that in the estimation of extreme quantiles (that is, x_p for $p \in [0.95,1)$ or $p \in (0,0.05]$) with no precision requirement, Sequest may deliver CI coverage substantially below the nominal level, or it may require excessive sample sizes. The Sequem method proposed in this paper addresses the problem of estimating extreme quantiles by adopting and extending the maximum transformation method of Heidelberger and Lewis (1984). For simplicity we only consider estimating x_p for $p \in [0.95,1)$. The name Sequem is an abbreviation of the phrase "Sequential extreme quantile estimation via the maximum transformation."

The remainder of this paper is organized as follows. Section 2 provides an overview of Sequem and a formal algorithmic statement of the procedure. Section 3 contains a summary of the results of a preliminary experimental performance evaluation of Sequem. Section 4 contains concluding remarks and an outline of the next steps in our work on Sequem. The slides for the oral presentation of this article are available online via www.ise.ncsu.edu/jwilson/wsc15sequem.pdf.

2 OVERVIEW OF SEQUEM

From the simulation-generated time series $\{X_1,\ldots,X_n\}$ of length $n = bm$, we form b nonoverlapping batches each of size m, so that the jth batch consists of the observations $\{X_{(j-1)m+1},\ldots,X_{jm}\}$. For $j = 1,\ldots,b$, we sort the jth batch in ascending order to obtain the order statistics $X_{j,(1)} \le X_{j,(2)} \le \cdots \le X_{j,(m)}$ and define the associated batch quantile estimator (BQE) by

$$\widehat{x}_p(j,m) \equiv \begin{cases} X_{j,(1)} & \text{if } p \le 0.5/m\,, \\ \delta_{p,m} X_{j,(\lceil mp+0.5 \rceil -1)} + (1-\delta_{p,m}) X_{j,(\lceil mp+0.5 \rceil)} & \text{if } 0.5/m < p < (m-0.5)/m\,, \\ X_{j,(m)} & \text{if } (m-0.5)/m \le p\,, \end{cases} \tag{1}$$

where

$$\delta_{p,m} \equiv \lceil mp + 0.5 \rceil - (mp + 0.5) \quad \text{for } m = 1,2,\ldots. \tag{2}$$

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

$$\widetilde{x}_p(n) \equiv \begin{cases} X_{(1)} & \text{if } p \le 0.5/n\,, \\ \delta_{p,n} X_{(\lceil np+0.5 \rceil -1)} + (1-\delta_{p,n}) X_{(\lceil np+0.5 \rceil)} & \text{if } 0.5/n < p < (n-0.5)/n, \\ X_{(n)} & \text{if } (n-0.5)/n \le p. \end{cases} \tag{3}$$

Using (1) and (3), we also compute a modified estimator of the variance of the BQEs,

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

Finally we compute the following $100(1-\alpha)\%$ CI for x_p,

$$\widetilde{x}_p(n) \pm t_{1-\alpha/2,b-1} \widetilde{S}_{\widehat{x}_p}(b,m) / \sqrt{b}, \tag{4}$$

where $t_{r,\nu}$ is the r-quantile of Student's t-distribution with ν degrees of freedom for $r \in (0,1)$. As $m \to \infty$ with b fixed, the asymptotic validity of the CI (4) can be established under any of the following conditions on the underlying process $\{X_k\}$: (i) ϕ-mixing (Sen 1972); (ii) geometric ergodicity for Markov chains (Muñoz 2010); or (iii) a geometric moment contraction condition (Alexopoulos, Goldsman, and Wilson 2012; Wu 2005). Condition (iii) is satisfied by a rich diversity of widely used linear and nonlinear processes, including autoregressive–moving average (ARMA) processes, generalized conditional heteroscedastic (GARCH) processes, random coefficient autoregressive (RCA) processes, and threshold autoregressive (TAR) processes as well as a broad class of Markov chains.

The maximum transformation technique of Heidelberger and Lewis (1984) converts the problem of estimating an extreme quantile to the problem of estimating a quantile closer to the median. If $X_1^*, X_2^*, \ldots, X_c^*$ are i.i.d. random variables with c.d.f. $F_X(\cdot)$ and if we define the random variable $Y = \max\{X_1^*, X_2^*, \ldots, X_c^*\}$ with c.d.f. $F_Y(y) = \Pr\{Y \le y\} = [F_X(y)]^c$ for all y, then $F_Y(x_p) = [F_X(x_p)]^c = p^c \equiv q$; hence the estimation of x_p reduces to the estimation of the q-quantile of the distribution of Y. Heidelberger and Lewis (1984) considered values of c such that $p^c \approx 0.5$, since estimators of the median typically have smaller mean squared error than estimators for extreme quantiles. However, such an assignment can lead to sample size explosion because for $p = 0.99$ one has $c \approx \ln(0.5)/\ln(0.99) \approx 69$; and the implications of this are explained below.

Heidelberger and Lewis (1984) applied the maximum transformation to an autocorrelated stationary process $\{X_k : k = 1, \ldots, n\}$ of length n by forming L contiguous groups of data, each consisting of cm consecutive observations so that $n = cmL$. Conceptually each group is arranged in a $c \times m$ matrix whose rows are formed from the consecutive nonoverlapping batches of size m within the group — that is, the first batch of m observations in the group forms the first row of the associated matrix, the second batch of m observations in the group forms the second row of the matrix, etc. This arrangement ensures that each column of the matrix consists of observations of the underlying process $\{X_k\}$ separated by lag m; and if m is sufficiently large, then the observations in each column of that matrix are approximately i.i.d. with the c.d.f. $F_X(\cdot)$ so that the maximum of the observations in each column is a random variable with the c.d.f. $F_Y(\cdot)$. For $\ell = 1,\ldots,L$, let $Y_{i,\ell}$ denote the maximum of the observations in column i of the matrix for group ℓ, where $i = 1,\ldots,m$. If m is large enough, then from the associated order statistics $Y_{(1),\ell} \leq \cdots \leq Y_{(m),\ell}$ we can compute the ℓth group quantile estimator (GQE) $Y_{(\lceil mq \rceil),\ell}$ of the q-quantile of Y for $\ell = 1,\ldots,L$. If for any convenient value of q we take $c = \lfloor \ln(q)/\ln(p) \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function, and if m is large enough, then the GQEs $\{Y_{(\lceil mq \rceil),\ell} : \ell = 1,\ldots,L\}$ are approximately i.i.d. normal unbiased estimators of x_p so an approximate CI for x_p similar to (4) can be based on the sample mean and sample variance of the GQEs. Heidelberger and Lewis (1984) demonstrated the potential of the maximum transformation approach, but they provided no guidelines for choosing the number of groups L and the batch size m.

Sequem addresses the latter two fundamental problems and extends the group quantile method of Heidelberger and Lewis (1984) in two key respects.

- Given the effectiveness of the Sequest procedure (Alexopoulos et al. 2014) for estimating nonextreme quantiles ($0.1 \leq p \leq 0.9$), Sequem estimates the $q = 0.9$ quantile of the max-transformed data so that in Sequem we take $c = \lfloor \ln(0.9)/\ln(p) \rfloor$. This assignment results in substantially smaller values of c and limits the aforementioned sample size explosion.
- Sequem uses a sectioning mechanism based on applying the maximum transformation to the entire simulation-generated time series $\{X_k : k = 1,\ldots,n\}$ of length n by conceptually arranging that time series into a $c \times (mL)$ matrix so that the first subseries of mL consecutive observations form the first row of the matrix, the second subseries of mL consecutive observations form the second row of the matrix, and so on. This ensures that each column of the matrix consists of observations that are separated by lag mL, where $mL \gg m$ so that from the maximum values in each column $\{\widetilde{Y}_i : i = 1,\ldots,mL\}$, we compute a point estimator of x_p with substantially reduced bias and variance.

A formal algorithmic statement of Sequem is given in Figure 1. Step **[0]** of Sequem initializes various experimental parameters of Sequem. Steps **[1]** and **[2]** compute an initial estimate of the length w of the warm-up period that is sufficiently large to include any pronounced (primarily deterministic) initial transient so that beyond observation w, the underlying process $\{X_k : k > w\}$ exhibits approximately stationary stochastic behavior. In particular, the first loop in steps **[1a–b]** starts with $b = 64$ batches of size $m_0 = 256$ and progressively increases the batch size by the factor of $\tau_{\mathrm{wrm}} = 2$ until the sample standard deviation of the BQEs exceeds the value $\varepsilon_a = 10^{-10}$ and the estimated coefficient of variation of the BQEs exceeds the threshold $\varepsilon_r = 10^{-5}$. The second loop in steps **[1c–d]** applies von Neumann's randomness test with a significance level that decreases gradually from the "aggressive" value $\alpha_{\mathrm{wrmi}} = 0.25$ to about $\alpha_{\mathrm{wrmf}} = 0.001$. Each time the BQEs fail the randomness test, the batch size is doubled ($\tau_{\mathrm{wrm}} = 2$) and the test is repeated.

Step **[2]** performs a more stringent randomness test to obtain a batch size m large enough so that observations X_k and X_{k+m} of the underlying process separated by lag m are approximately i.i.d., which ensures the approximate validity of the maximum transformation. The level of significance α_{mxt} decreases at the same rate as in steps **[1c–d]**, and the batch size is doubled ($\tau_{\mathrm{mxt}} = 2$) after each failed test.

Step **[3]** starts by skipping the initial w observations of the underlying process, where the truncation point w is taken to be the sum of the final batch sizes from steps **[1]** and **[2]**. Step **[3b]** obtains additional data and forms $L = 64$ groups of size cm as described in Heidelberger and Lewis (1984). For the ℓth group ($\ell = 1,\ldots,L$), we compute the group quantile estimator $Y_{(\lceil mq \rceil),\ell}$ from the $c \times m$ matrix associated with that group. (This grouping is retained in steps **[3]–[5]**.) The loop in steps **[3c–d]** sequentially increases the batch size m until the estimated absolute skewness of the GQEs falls below the threshold $\mathscr{B}^*(p) = 0.6$. To avoid an explosion of the batch size m, we limit the number of iterations of step **[3]** to $u^* = 50$, we impose temporary upper bounds n^* and m^* on n and m respectively that have no effect beyond step **[3]**, and we exploit Equation (11) to increase the batch size in a less pronounced fashion than in steps **[1]** and **[2]**.

[0] Set the value $p \in [0.95, 1)$, $\alpha \in (0,1)$, and the absolute (h^*) or relative (r^*) precision requirement on the half-length of the approximate $100(1-\alpha)\%$ CI for x_p. Initialize the remaining experimental parameters as follows.

- **[a]** Set the "baseline" batch size $m_0 \leftarrow 256$. Set the initial batch size $m \leftarrow m_0$, the initial batch count $b \leftarrow 64$, and the initial sample size $n \leftarrow mb$. For the initial phase of determining the length of the warm-up period in step **[1]**, set the absolute tolerance on the sample variance of the BQEs, $\varepsilon_a \leftarrow 10^{-10}$, and the associated relative tolerance, $\varepsilon_r \leftarrow 10^{-5}$. For computing the CI skewness adjustment in step **[4]**, set the associated tolerance, $\varepsilon_s \leftarrow 10^{-3}$.

- **[b]** For determining the preliminary length of the warm-up period in step **[1]**, set the starting randomness test size $\alpha_{\text{wrmi}} \leftarrow 0.25$, the final randomness test size $\alpha_{\text{wrmf}} \leftarrow 0.001$, the associated batch-size inflation factor $\tau_{\text{wrm}} \leftarrow 2$, and $\ell^* \leftarrow 15$, the maximum number of iterations of the randomness tests allowed in each of steps **[1c]** and **[2c]**. For determining in step **[2]** a batch size sufficiently large to ensure proper operation of the maximum transform, set the starting randomness test size $\alpha_{\text{mxti}} \leftarrow 0.25$, the final randomness test size $\alpha_{\text{mxtf}} \leftarrow 0.001$, and the associated batch-size inflation factor $\tau_{\text{mxt}} \leftarrow 2$.

- **[c]** For controlling the batch size in the skewness-reduction step **[3]**, set the upper-bound function on absolute skewness of the BQEs, $\mathscr{B}^*(\delta) = 0.60$, $\delta \in (0,1)$. Set the constant for the maximum transformation to $c \leftarrow \lfloor \ln(0.9)/\ln(p) \rfloor$, the corresponding max-transformed quantile $q \leftarrow p^c$, and the number of groups $L \leftarrow 64$ that are used to compute the final point and CI estimators of x_p. Set the upper bound $u^* \leftarrow 50$ on the number of iterations of step **[3]**. Set the upper bound $\tau_{\text{skw}} \leftarrow 2$ on the batch-size adjustment factor to be used when testing the BQEs for excessive absolute skewness in step **[3]**. Finally set the (temporary) upper bound $n^* \leftarrow 3.0 \times 10^8$ on the total sample size to be considered when assigning the corresponding maximum batch size m^* in step **[3]**.

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

$$\bar{x}_p(b,m) \leftarrow \frac{1}{b}\sum_{j=1}^{b} \widehat{x}_p(j,m) \quad \text{and} \quad S^2_{\widehat{x}_p}(b,m) \leftarrow \frac{1}{b-1}\sum_{j=1}^{b}\left[\widehat{x}_p(j,m) - \bar{x}_p(b,m)\right]^2 . \tag{5}$$

Initialize the iteration counter $\ell \leftarrow 1$ for the iterations of the warm-up randomness test in step **[1]**.

- **[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 **[1c]**.

- **[b]** Update the batch size and the total sample size according to $m \leftarrow \lfloor m\tau_{\text{wrm}} \rfloor$ and $n \leftarrow bm$; obtain the required additional observations by restarting the simulation if necessary; update the BQEs (1) and the sample statistics (5); and return to step **[1a]**.

- **[c]** 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

  $$C_b \leftarrow 1 - \frac{\sum_{j=1}^{b-1}[\widehat{x}_p(j,m) - \widehat{x}_p(j+1,m)]^2}{2(b-1)S^2_{\widehat{x}_p}(b,m)} . \tag{6}$$

  Compute the size of the current randomness test, $\alpha_{\text{wrm}} \leftarrow \alpha_{\text{wrmi}}\big(0.60^{\ell-1}\big) + \alpha_{\text{wrmf}}\big(1 - 0.60^{\ell-1}\big)$. If

  $$|C_b| \le z_{1-\alpha_{\text{wrm}}/2}\sqrt{(b-2)/(b^2-1)} \quad \text{or} \quad \ell \ge \ell^*,$$

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

Figure 1: Algorithmic statement of Sequem.

[d] Update the iteration counter, batch size, and sample size according to $\ell \leftarrow \ell + 1$, $m \leftarrow \lfloor m\tau_{\text{wrm}} \rfloor$, and $n \leftarrow bm$, respectively; obtain the required additional observations by restarting the simulation if necessary; update the BQEs (1) and the sample statistics (5); and return to step **[1c]**.

[2] Determine a batch size m sufficiently large so that the BQEs pass a more stringent test of randomness that is needed to ensure the validity of the maximum transform as follows.

[a] Set the length of the warm-up period according to $w \leftarrow m$, the current batch size. Initialize the iteration counter $\ell \leftarrow 1$ for the maximum-transform randomness test. Update the batch count according to $b \leftarrow \min\{cL, 256\}$ and reset the batch size $m \leftarrow m_0$.

[b] Update the total sample size, $n \leftarrow w + bm$, and obtain the additional observations needed 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" time series of length $n' \leftarrow n - w = bm$ with the following indexing scheme: $\{X'_k = X_{w+k} : k = 1, \ldots, n'\}$. From the latter time series, form b batches of size m to compute the associated BQEs (1) and the associated sample statistics (5) and (6). Compute the size of the current maximum-transform randomness test, $\alpha_{\text{mxt}} \leftarrow \alpha_{\text{mxti}}\big(0.60^{\ell-1}\big) + \alpha_{\text{mxtf}}\big(1 - 0.60^{\ell-1}\big)$. If

$$|C_b| \le z_{1-\alpha_{\text{mxt}}/2}\sqrt{(b-2)/(b^2-1)} \quad \text{or} \quad \ell \ge \ell^*,$$

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

[c] Update the iteration counter, batch size, and sample size according to $\ell \leftarrow \ell + 1$, $m \leftarrow \lfloor m\tau_{\text{mxt}} \rfloor$, and $n \leftarrow bm$, respectively; and return to step **[2b]**.

[3] Determine a batch size m sufficiently large so that the GQEs have manageable skewness as follows.

[a] Initialize the skewness-reduction iteration counter, $u \leftarrow 1$. Set the final length of the warm-up period by incrementing the initial value of w from step **[2a]** by the final value of m from step **[2]**, $w \leftarrow w + m$. Set the upper limit

$$m^* \leftarrow \lfloor (n^* - w)/(cL) \rfloor$$

on the batch size allowed on any iteration of the following substeps **[3b–d]**.

[b] Update the total sample size, $n \leftarrow w + cmL$, and obtain the additional observations by restarting the simulation if necessary. Skip the first w observations in the time series of length n so that we have the warmed-up time series of length $n'' \leftarrow n - w = cmL$ with the indexing scheme

$$\{X''_{k,i,\ell} = X_{w+(\ell-1)mc+(k-1)m+i} : \; k = 1, \ldots, c; \; i = 1, \ldots, m; \; \ell = 1, \ldots, L\}.$$

Therefore within ℓth group consisting of c adjacent batches of size m, we let $X''_{k,i,\ell}$ denote the ith observation in the kth batch ("data row") constituting that group. Within the ℓth group whose associated $c \times m$ matrix is obtained conceptually by vertically concatenating (stacking) the group's successive data rows each of length m, compute the maximum $Y_{i,\ell}$ of the observations in the ith column of the resulting matrix so that we have

$$Y_{i,\ell} = \max\big\{X''_{k,i,\ell} : k = 1, \ldots, c\big\} \quad \text{for } i = 1, \ldots, m \text{ and } \ell = 1, \ldots, L.$$

Within each group ℓ, compute the associated order statistics $Y_{(1),\ell} \le Y_{(2),\ell} \le \cdots \le Y_{(m),\ell}$ so that the ℓth warmed-up GQE based on c batches ("data rows") of size m is

$$\widehat{y}_p(c, m, \ell) \leftarrow \begin{cases} Y_{(1)\ell} & \text{if } q \le 0.5/m, \\ \delta_{q,m} Y_{(\lceil mq+0.5 \rceil - 1),\ell} + (1 - \delta_{q,m}) X_{(\lceil mq+0.5 \rceil),\ell} & \text{if } 0.5/m < q < (m-0.5)/m, \\ X_{(m),\ell} & \text{if } (m-0.5)/m \le q, \end{cases} \tag{7}$$

Figure 1 (Continued): Algorithmic statement of Sequem.

for $\ell = 1, \ldots, L$. From the L GQEs in (7), compute the sample mean, variance, and skewness,

$$\bar{y}_p(c,m,L) \leftarrow \frac{1}{L} \sum_{\ell=1}^{L} \widehat{y}_p(c,m,\ell)\,, \tag{8}$$

$$S^2_{\widehat{y}_p}(c,m,L) \leftarrow \frac{1}{L-1} \sum_{\ell=1}^{L} \left[\widehat{y}_p(c,m,\ell) - \bar{y}_p(c,m,L)\right]^2, \tag{9}$$

$$\widehat{\mathscr{B}}_{\widehat{y}_p}(c,m,L) \leftarrow \frac{L}{(L-1)(L-2)} \sum_{\ell=1}^{L} \left[\frac{\widehat{y}_p(c,m,\ell) - \bar{y}_p(c,m,L)}{S_{\widehat{y}_p}(c,m,L)}\right]^3. \tag{10}$$

[c] If

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

then go to step **[4]**; otherwise compute $\psi(u) \equiv \max\left\{1.10, \tau_{\text{skw}}/\sqrt{u}\right\}$, the upper limit on the batch-size inflation factor, increase the batch size according to

$$m \leftarrow \left\lceil m \cdot \text{mid}\left\{1.05, \left[\widehat{\mathscr{B}}_{\widehat{y}_p}(c,m,L)/\mathscr{B}^*(p)\right]^2, \ \psi(u)\right\}\right\rceil, \tag{11}$$

where $\text{mid}\{\zeta_1, \zeta_2, \zeta_3\} \equiv \zeta_{(2)}$, and increment the skewness-reduction iteration counter $u \leftarrow u + 1$.

[d] If $m > m^*$, then set $m \leftarrow m^*$ and $u \leftarrow u^*$. Return to step **[3b]**.

[4] Update the group count and batch size according to $L \leftarrow L/2$ and $m \leftarrow 2m$.

[5] With the updated values of m and L, recompute the warmed-up GQEs (7), their sample mean (8), sample variance (9), and sample skewness (10); then compute the sample lag-one correlation of the GQEs,

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

and the associated correlation adjustment

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

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

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

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

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 $\sqrt[3]{\zeta} \equiv \text{sign}(\zeta)\sqrt[3]{|\zeta|}$ (Tafazzoli and Wilson 2011).

Figure 1 (Continued): Algorithmic statement of Sequem.

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

$$H \leftarrow \max\big\{G(t_{1-\alpha/2,L-1}), G(t_{\alpha/2,L-1})\big\}\Big[A S^2_{\widetilde{y}_p}(c,m,L)/L\Big]^{1/2}. \tag{12}$$

To compute the "sectioning-based" point estimator of x_p that has been adapted to the maximum transform method, we use the following modified indexing scheme for the warmed-up time series $\{X_{w+j} : j = 1,\ldots,cmL\}$: let $X^{\#}_{k,i} = X_{w+(k-1)mL+i}$ for $k = 1,\ldots,c$ and $i = 1,\ldots,mL$ so that within the kth subseries ("data row") of length mL formed from L adjacent batches of size m, $X^{\#}_{k,i}$ denotes the ith observation in that subseries. From the $c \times (mL)$ matrix that is conceptually obtained by vertically concatenating (stacking) the successive data rows each of length mL, compute the maximum $\widetilde{Y}_i$ of the observations in the ith column of the resulting matrix so that we have

$$\widetilde{Y}_i = \max\big\{X^{\#}_{k,i} : k = 1,\ldots,c\big\} \text{ for } i = 1,\ldots,mL.$$

Then compute the associated order statistics $\widetilde{Y}_{(1)} \le \widetilde{Y}_{(2)} \le \cdots \le \widetilde{Y}_{(mL)}$ and set the point estimator for x_p as

$$\widetilde{y}_p(c,m,L) \leftarrow \begin{cases} \widetilde{Y}_{(1)} & \text{if } q \le 0.5/(mL)\,, \\ \delta_{q,mL}\widetilde{Y}_{(\lceil mLq+0.5\rceil-1)} + (1-\delta_{q,mL})\widetilde{Y}_{(\lceil mLq+0.5\rceil)} & \text{if } 0.5/(mL) < q < (mL-0.5)/(mL), \\ \widetilde{Y}_{(mL)} & \text{if } (mL-0.5)/(mL) \le q. \end{cases} \tag{13}$$

The associated CI estimator for x_p has the form

$$\widetilde{y}_p(c,m,L) \,\pm\, H. \tag{14}$$

If no precision level is specified, then deliver the CI (14) 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 (14) satisfies the user-specified precision requirement

$$H \le H^*, \tag{15}$$

where

$$H^* = \begin{cases} r^*\big|\widetilde{y}_p(c,m,L)\big|, & \text{for a relative precision level } r^*, \\ h^*, & \text{for an absolute precision level } h^*, \end{cases} \tag{16}$$

then deliver the CI (14) and stop; otherwise proceed to step **[7b]**.

[b] For the fixed batch count b, estimate the batch size m required to satisfy (15)–(16),

$$m \leftarrow \big\lceil m \cdot \mathrm{mid}\big\{1.02, (H/H^*)^2, 1.2\big\}\big\rceil.$$

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

Figure 1 (Continued): Algorithmic statement of Sequem.

Step **[4]** halves the number of groups L and doubles the batch size m in an attempt to improve the coverage of the CI (14) in step **[6]**. This action neither increases the total sample size nor changes the bias of the sectioning-based point estimator of x_p. However, in general the absolute skewness of the GQEs decreases and the CI half-length increases because of the increase in m and the decrease in L.

Step **[5]** obtains adjustments for excess autocorrelation or skewness of the GQEs that will be used in the computation of the CI half-length (12) in step **[6]**. The latter step computes the approximate CI for x_p based on the sectioning technique applied to the entire set of max-transformed observations $\{\widetilde{Y}_i : i = 1,\ldots,mL\}$ to yield the final point estimator (13). The CI (14) is symmetric about (13) to guard against undercoverage.

Finally, step **[7]** sequentially increases the batch size m until the absolute or relative precision requirement on the CI half-length is satisfied as stipulated by (15) and (16). The assignment in step **[7b]** is based on substantial experimentation with sequential procedures for estimating steady-state means (Tafazzoli and Wilson 2011) and the Sequest approach for estimating steady-state quantiles (Alexopoulos et al. 2014).

3 EXPERIMENTAL PERFORMANCE EVALUATION OF SEQUEM

In this section we conduct a preliminary performance evaluation of Sequem based on three test processes presenting various statistical challenges. Each table below contains performance statistics for two levels of CI relative precision: $r^* = \infty$ (i.e., no precision requirement) and a problem-dependent value of r^* which is significantly smaller than the average CI relative precision obtained when $r^* = \infty$. The second precision requirement is intended to assess the effectiveness of step **[7]** of Sequem.

Sequem has been implemented in Java, and the code will be freely available soon after the completion of a more-detailed performance evaluation and the development of a graphical user interface.

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

Table 1 shows the results of applying Sequem 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 $X_k = \mu_X + \rho(X_{k-1} - \mu_X) + \varepsilon_k$, for $k = 1,2,\ldots$, where $\{\varepsilon_k : k = 1,2,\ldots\}$ are i.i.d. $N(0,\sigma_\varepsilon^2)$ with $\sigma_\varepsilon^2 = 1$. We applied Sequem to 1000 replications of this process.

Table 1: Performance of Sequem-delivered point and 95% CI estimators of the p-quantile x_p of the AR(1) process described in Section 3.1 based on 1000 replications.

No CI Precision Requirement								
p	x_p	Avg. $\widetilde{y}_p(n')$	Avg. $\left\lvert \text{Bias}\left[\widetilde{y}_p(n')\right] \right\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Cover.	$\overline{m}$	$\overline{n}$
0.95	116.4691	116.4	0.614	1.59	1.37	93.4%	3207	207766
0.99	123.2926	123.3	0.385	0.975	0.791	95.2%	5583	1789741
0.995	125.7906	125.8	0.308	0.772	0.614	94.1%	6430	4324081
CI Relative Precision = 0.5%								
p	x_p	Avg. $\widetilde{y}_p(n')$	Avg. $\left\lvert \text{Bias}\left[\widetilde{y}_p(n')\right] \right\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Cover.	$\overline{m}$	$\overline{n}$
0.95	116.4691	116.5	0.212	0.530	0.455	94.7%	22176	1421778
0.99	123.2926	123.3	0.229	0.545	0.442	93.6%	11661	3734704
0.995	125.7906	125.8	0.225	0.540	0.429	93.7%	9190	6178909

The high correlation between successive observations in this process makes it a severe test of Sequem'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" Sequem's ability to eliminate initialization bias as well as to compensate effectively for pronounced correlation between successive observations of a target process. For both precision requirements considered, the average lengths of the warm-up period for estimating the 0.95, 0.99, and 0.995 quantiles were 2,507 2,993, and 3,030, respectively.

Table 1 shows that for both precision levels, Sequem's sampling efficiency was good. Moreover in the no precision case, Sequem delivered nominal 95% CIs with coverages ranging from 93.4% to 95.2% and with average values of the CI relative precision $100 \times \left\lvert H/\widetilde{y}_p(n') \right\rvert$ ranging from 0.6% to 1.4%. The results for 0.5% relative precision ($r^* = 0.005$) 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.

For example, Sequem delivered CI estimators of the 0.95, 0.99, and 0.995 quantiles with the following properties: (i) respective coverage probabilities of 94.7%, 93.6%, and 93.7%; (ii) respective average relative precisions of about 0.46%, 0.44%, and 0.43%; and (iii) respective average sample sizes of about 1.4, 3.7, and 6.2 million. These results are competitive with all the results reported in the literature for significantly less challenging versions of the AR(1) process (Chen and Kelton 2006).

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

Consider an $M/M/1$ queueing system with interarrival rate $\lambda = 0.9$ and service rate $\omega = 1$, and let X_k be the time spent in queue by customer k prior to receiving service. Let $\rho = \lambda/\omega = 0.9$ denote the traffic intensity. It is well known that the steady-state c.d.f. of X_k is defined as follows: $F_X(x) = 0$ for $x < 0$; $F_X(x) = 1 - \rho$ for $x = 0$; and $F_X(x) = 1 - \rho\exp[-\omega(1-\rho)x]$ for $x > 0$. Hence the response X_k has steady-state mean $\mu_X = 9$, and the steady-state quantiles can be evaluated analytically by inverting $F_X(\cdot)$. We assume that the system starts with 113 customers initially in the system, and we record successive queue waiting times only for the customers arriving after the beginning of the simulation. Table 2 shows the results of applying Sequem to 1000 replications of this process.

Table 2: Performance of Sequem-delivered point and 95% CI estimators of the p-quantile x_p of the $M/M/1$ queue waiting-time-process described in Section 3.2 based on 1000 replications.

No CI Precision Requirement								
p	x_p	Avg. $\widetilde{y}_p(n')$	Avg. $\left|\text{Bias}\left[\widetilde{y}_p(n')\right]\right|$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Cover.	$\overline{m}$	$\bar{n}$
0.95	28.90372	28.8	0.583	1.54	5.36	94.2%	46154	2961218
0.99	44.998097	45.0	0.683	1.73	3.85	95.1%	46936	15027284
0.995	51.929568	51.8	0.718	1.78	3.44	95.1%	44013	29584593
CI Relative Precision = 3%								
p	x_p	Avg. $\widetilde{y}_p(n')$	Avg. $\left|\text{Bias}\left[\widetilde{y}_p(n')\right]\right|$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Cover.	$\overline{m}$	$\bar{n}$
0.95	28.90372	28.9	0.295	0.761	2.64	96.0%	87874	5631307
0.99	44.998097	45.0	0.445	1.127	2.50	95.2%	65422	20942986
0.995	51.929568	51.9	0.499	1.255	2.42	95.3%	56976	38295547

As for the AR(1) process, the warm-up period for this process is also pronounced. For all three precision requirements, the average lengths of the warm-up period for estimating the 0.95, 0.99, and 0.995 quantiles were 7,333, 7,649, and 7,669, respectively.

The marginal c.d.f. $F_X(\cdot)$ 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 (1) that significantly distorts the behavior of the conventional sectioning-based CI given by Equation (4), resulting in coverage probabilities significantly below the nominal level $1 - \alpha$.

Table 2 shows that for both precision levels, Sequem's sampling efficiency was good. In the case of no precision requirement, Sequem delivered the point estimator $\widetilde{y}_p(n')$ of x_p with average absolute biases of about 0.6, 0.7, and 0.7 for the 0.95, 0.99, and 0.995 quantiles, respectively. Moreover in the case of no precision requirement, Sequem delivered nominal 95% CIs for the 0.95, 0.99, and 0.995 quantiles with respective coverages of 94.2%, 95.1%, and 95.1%, respective average relative precisions of 5.4%, 3.8%, and 3.4%, and respective average sample sizes of about 3 million, 15 million, and 30 million. From the recent technical report of Alexopoulos et al. (2015) containing the formal algorithmic statement of the latest version of Sequest as well as some relevant numerical results, we see that to deliver a 95% CI for the 0.95, 0.99, and 0.995 quantiles in the case of no precision requirement, Sequest respectively required on the average sample sizes of 10 million, 32 million, and 47 million, with corresponding coverage probabilities of 95%, 90%, and 86%, respectively. Therefore in this especially difficult test process, Sequem achieved a substantial reduction in total sample size while maintaining much closer conformance to the nominal coverage probability compared with Sequest.

With a relative precision requirement of 3%, Sequem's point estimators of the 0.95, 0.99, and 0.995 quantiles had average absolute biases of about 0.29, 0.44, and 0.50, respectively. Moreover, Sequem delivered nominal 95% CIs for the 0.95, 0.99, and 0.995 quantiles with respective coverages of 96.0%, 95.2%, and 95.3%, respective average relative precisions of about 2.6%, 2.5%, and 2.4%, and respective average sample sizes of about 5.6 million, 21 million, and 38 million.

3.3 *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 $\omega = 3.2$. The $M/M/1/\text{LIFO}$ queue-waiting-time process was selected for two reasons: (i) unlike the two previous test processes, the autocorrelation function for this process does not decline in magnitude geometrically fast with increasing lags; and (ii) the process has a highly nonnormal marginal distribution that significantly distorts the behavior of the conventional sectioning-based CI (4), resulting in coverage probabilities significantly below the nominal level $1-\alpha$.

Table 3 shows the results of applying Sequem to 1000 replications of this process. We computed the "exact" value of each selected quantile $x_p = F_X^{-1}(p)$ as follows: (i) 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 ω (Kleinrock 1975, Equation (5.144)) using the Euler algorithm of Abate and Whitt (2006); (ii) we combined the relation $F_X(x) = (1-\rho) + \rho F_{B_{\text{FIFO}}}(x)$ for $x \geq 0$ with the result of (i) to compute a piecewise-linear approximation to $F_X(x)$ for $0 \leq x \leq 75$ in increments of size $\Delta x = 10^{-3}$; and (iii) we inverted the result of (ii) to yield an estimate of x_p with high accuracy.

Table 3: Performance of Sequem-delivered point and 95% CI estimators of the p-quantile x_p of the $M/M/1/\text{LIFO}$ queue-waiting-time process described in Section 3.3 based on 1000 replications.

No CI Precision Requirement								
p	x_p	Avg. $\widetilde{y}_p(n')$	Avg. $\lvert\text{Bias}[\widetilde{y}_p(n')]\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Cover.	$\overline{m}$	$\overline{n}$
0.95	14.4052	14.4	0.307	0.871	6.06	96.4%	3240	208045
0.99	49.5819	49.6	0.902	2.51	5.07	97.0%	3242	1038381
0.995	71.8438	71.8	1.166	3.26	4.54	97.2%	3105	2088031
CI Relative Precision = 2%								
p	x_p	Avg. $\widetilde{y}_p(n')$	Avg. $\lvert\text{Bias}[\widetilde{y}_p(n')]\rvert$	$\overline{H}$	Avg. CI Rel. Prec. (%)	CI Cover.	$\overline{m}$	$\overline{n}$
0.95	14.4052	14.4	0.104	0.261	1.81	95.4%	23645	1513981
0.99	49.5819	49.6	0.370	0.888	1.79	94.1%	17230	5514620
0.995	71.8438	71.8	0.537	1.29	1.79	94.3%	13134	8827362

Table 3 shows that for both precision levels, Sequem's sampling efficiency was good. In the case of no precision requirement, Sequem delivered the point estimator $\widetilde{y}_{0.95}(n')$ of the 0.95 quantile with average absolute bias of about 0.3 and average sample size of about 208,000 when estimating the true value $x_{0.95} \approx 14.4$; on the other hand, the Sequest method in Alexopoulos et al. (2015) required on the average about 500,000 observations to deliver a 95% CI for $x_{0.95}$ from this process. Therefore in this case Sequem achieved about a 58% reduction in total sample size compared with Sequest.

Similarly in the case of no precision requirement, Sequem delivered the point estimator $\widetilde{y}_{0.99}(n')$ of the 0.99 quantile with average absolute bias of about 0.9 and average sample size of about 1 million when estimating the true value $x_{0.99} \approx 49.6$. In the no precision case, Sequem delivered CIs for the 0.95, 0.99, and 0.995 quantiles with respective coverages of 96.4%, 97.0%, and 97.2%, and with respective average relative precisions of 6%, 5%, and 4.5%. For the case of nominal relative precision of 2%, Sequem delivered CIs for the 0.95, 0.99, and 0.995 quantiles with respective average relative precisions of 1.8%, 1.8% and 1.8%, and with respective coverages of 95.4%, 94.1%, and 94.3%. We judged the corresponding average sample sizes of about 1.5 million, 5.5 million, and 8.8 million to be reasonable. For this test process, some current quantile-estimation procedures fail to deliver useful estimators of x_p for any value of $p \in (0,1)$ (Bekki et al. 2010).

4 CONCLUSIONS

This article describes Sequem, a fully sequential procedure for computing improved point estimators and CIs for steady-state quantiles of a simulation output process. The CIs are designed to meet user-specified criteria related to their coverage probability and absolute or relative precision. A preliminary evaluation of Sequem based on three output processes designed to "stress-test" the procedure revealed that Sequem was competitive with existing methods for estimating steady-state quantiles, including Bekki et al. (2010), Chen and Kelton (2006), and Heidelberger and Lewis (1984).

Future work on Sequem will focus on the following threads: (a) a thorough sensitivity analysis of the performance of Sequem with respect to the procedure's parameters; (b) a detailed performance evaluation based on an expanded suite of problems that includes all the processes in Tafazzoli et al. (2011); (c) the incorporation of effective data management techniques; and (d) the simultaneous estimation of multiple quantiles.

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, A. Mokashi, R. Nie, Q. Sun, K.-W. Tien, and J. R. Wilson. 2014. "Sequest: A Sequential Procedure for Estimating Steady-State Quantiles." In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 662–673. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Alexopoulos, C., D. Goldsman, A. Mokashi, Q. Sun, Y. Zhong, and J. R. Wilson. 2015. "Sequest10: A Sequential Procedure for Estimating Steady-State Quantiles." Technical Report. Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC. Available online via http://www.ise.ncsu.edu/jwilson/files/sequest10-4-14.pdf [accessed April 14, 2015].

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.

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.

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.

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

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.

Sen, P. K. 1972. "On the Bahadur Representation of Sample Quantiles for Sequences of ϕ-Mixing Random Variables." *Journal of Multivariate Analysis* 2:77–95.

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.

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. 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. 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.

KAI-WEN TIEN 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 e-mail is power751124@gmail.com.

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. His e-mail address is jwilson@ncsu.edu, and his Web page is www.ise.ncsu.edu/jwilson.