Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son

SEQUENTIAL ESTIMATION OF STEADY-STATE QUANTILES: SOME NEW DEVELOPMENTS IN METHODS AND SOFTWARE

Christos Alexopoulos
David Goldsman

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

Anup C. Mokashi

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

James R. Wilson

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

ABSTRACT

Recent developments are summarized concerning Sequest and Sequem, sequential procedures for estimating nonextreme and extreme steady-state quantiles of a simulation output process. The procedures deliver point and confidence-interval (CI) estimators of a given quantile, where each CI approximately satisfies given requirements on its coverage probability and its absolute or relative precision. The public-domain Sequest software now includes both procedures. The software is applied to a user-supplied dataset exhibiting warm-up effects, autocorrelation, and a multimodal marginal distribution. For the simulation analysis method of standardized time series (STS), we also sketch an elementary proof of a functional central limit theorem (FCLT) that is needed to develop STS-based quantile-estimation procedures when the output process satisfies a conventional density-regularity condition and either (i) a geometric-moment contraction condition and an FCLT for a related binary process, or (ii) conventional strong-mixing conditions.

1 INTRODUCTION

To evaluate long-run performance or risk for complex systems, steady-state simulations play a fundamental role in a wide range of disciplines. On one hand, the steady-state expected value of an ergodic output process equals the long-run average of a time series of such outputs almost surely (a.s.). On the other hand, under broadly applicable conditions a steady-state quantile of the selected output can measure the long-run performance or risk for each individual output as well as overall system performance. For example, in a production-system simulation, let X_i denote the cycle time of the ith departing job (i.e., the job's time in the system), where $i \geq 1$. In the evaluation of an existing or proposed system design, an important performance measure may be $x_{0.95}$, the steady-state 0.95-quantile of each job's cycle-time distribution because as $i \to \infty$, the long-run probability is 95% that X_i does not exceed $x_{0.95}$.

To formalize the discussion, we assume that $\{X_i : i \geq 1\}$ is stationary with cumulative distribution function (c.d.f.) $F(x) \equiv \Pr\{X_i \leq x\}$ and probability density function (p.d.f.) $f(x)$ for all $x \in \mathbb{R}$, where $f(x)$ is continuous on its support. Given $p \in (0,1)$, the p-quantile of this distribution is $x_p \equiv F^{-1}(p) \equiv \inf\{x : F(x) \geq p\}$. If $\{X_i : i = 1, \ldots, n\}$ consists of independent and identically distributed (i.i.d.) outputs, then we can compute standard point and confidence-interval (CI) estimators of x_p (Serfling 1980, Section 2.3.3 and Section 2.6.1).

978-1-7281-3283-9/19/$31.00 ©2019 IEEE

If the simulation is not initialized in steady-state operation or $\{X_i : i \geq 1\}$ is autocorrelated, then the estimation of steady-state quantiles involves substantial challenges. In particular, successive responses generated by a simulation are rarely i.i.d. normal random variables (r.v.'s). For example, in a queueing network simulation that has the empty-and-idle initial condition but substantial long-run congestion, successive observations $\{X_i : i \geq 1\}$ of time in the system for departing customers are contaminated by warm-up effects that depend on the customer index i; hence those observations are neither independent nor identically distributed. As another example, successive losses or gains $\{X_i : i = 1,\ldots,n\}$ in the value of a financial portfolio over a given n-period time horizon are not mutually independent because of their joint stochastic dependence on the economic conditions prevailing over that time horizon. Moreover, in both of these examples the associated p.d.f. $f(x)$ often has highly nonnormal properties such as pronounced skewness or multiple modes. See Alexopoulos et al. (2017, p. 22:3) and Alexopoulos et al. (2019b, pp. 2–3) for a brief review of previous work on quantile estimation for steady-state simulation.

In this paper we summarize our recent work on two sequential procedures for estimating a steady-state quantile whose order p is given—namely, Sequest (Alexopoulos et al. 2019b), which is designed for estimating nonextreme quantiles (i.e., $0.05 \leq p \leq 0.95$); and Sequem (Alexopoulos et al. 2017), which is designed for estimating extreme upper quantiles (i.e., $0.95 < p \leq 0.999$). Section 2 provides an overview of our approach to delivering improved point and CI estimators of a given quantile based on batching and sectioning as well as the deletion of any warm-up period and the adjustment of each CI's half-length to compensate for any harmful effects of autocorrelation or nonnormality. In Section 3 the public-domain Sequest software (now including Sequem) is applied to a user-supplied dataset exhibiting warm-up effects, autocorrelation, and a multimodal marginal distribution. For the analysis method of standardized time series (STS), in Section 4 we sketch an elementary proof of a functional central limit theorem (FCLT) that is needed to develop STS-based quantile-estimation procedures when the output process satisfies a conventional density-regularity condition and either (i) a geometric-moment contraction condition and an FCLT for a related binary process; or (ii) conventional strong-mixing conditions, which are much harder to check than (i). In Section 5 we discuss our ongoing work on steady-state simulation quantile estimation.

2 OVERVIEW OF SEQUEST

Sequest uses the methods of batching and sectioning to estimate x_p. From the simulation-generated time series $\{Y_i = X_{w+i} : i = 1,\ldots,n\}$ of length $n = bm$ that has been observed beyond the end w of any warm-up period deleted by Sequest, we form b nonoverlapping batches each of size m so that for $j = 1,\ldots,b$, the jth batch consists of the subseries $\{Y_{(j-1)m+1},\ldots,Y_{jm}\}$. We sort the observations in the jth batch into ascending order to obtain the order statistics $Y_{j,(1)} \leq \cdots \leq Y_{j,(m)}$; and $\widehat{y}_p(j,m) \equiv Y_{j,(\lceil mp \rceil)}$ is the jth batch quantile estimator (BQE) of x_p. Similarly from the entire warmed-up sample $\{Y_1,\ldots,Y_n\}$, we compute the order statistics $Y_{(1)} \leq \cdots \leq Y_{(n)}$; and $\widetilde{y}_p(n) \equiv Y_{(\lceil np \rceil)}$ is the sectioning-based estimator of x_p. (Here we use the conventional definition of $\widetilde{y}_p(n)$ to simplify the discussion and save space; see Alexopoulos et al. (2019b, Equation (16)) for the refined version of $\widetilde{y}_p(n)$ that is used in Sequest and Sequem.) We also compute an estimator of the variance of the BQEs, $\widetilde{S}^2_{\widehat{y}_p} \equiv (1/b)\sum_{j=1}^{b}\big[\widehat{y}_p(j,m) - \widetilde{y}_p(n)\big]^2$. Let $\widehat{\varphi}_{\widehat{y}_p}(b,m)$ and $\widehat{\mathscr{B}}_{\widehat{y}_p}(b,m)$ respectively denote the sample lag-one correlation and sample skewness of the BQEs $\big\{\widehat{y}_p(j,m) : j = 1,\ldots,b\big\}$.

Let $A = \max\big\{\big[1 + \widehat{\varphi}_{\widehat{y}_p}(b,m)\big]\big/\big[1 - \widehat{\varphi}_{\widehat{y}_p}(b,m)\big], 1\big\}$ denote the multiplicative adjustment to the half-length of the CI for x_p that is designed to compensate for correlation of the BQEs. From the sample skewness $\widehat{\mathscr{B}}_{\widehat{y}_p}(b,m)$ of the BQEs, compute the skewness-adjustment parameter $\beta = \widehat{\mathscr{B}}_{\widehat{y}_p}(b,m)\big/\big(6\sqrt{b}\big)$;

and define the skewness-adjustment function,

$$G(\zeta) \equiv \left\{ \begin{array}{ll} \zeta & \text{if } |\beta| \le \varepsilon_s, \\ \dfrac{[1+6\beta(\zeta-\beta)]^{1/3}-1}{2\beta} & \text{if } |\beta| > \varepsilon_s, \end{array} \right\} \quad \text{for all} \quad \zeta \in \mathbb{R},$$

where in general $\omega^{1/3} \equiv \operatorname{sign}(\omega)|\omega|^{1/3}$ for all real ω and ε_s is an arbitrarily small positive number. The correlation- and skewness-adjusted CI estimator of x_p has midpoint $\widetilde{y}_p(n)$ and half-length

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

Sequest progressively increases m until the termination condition $H \le H^*$ is satisfied, where $H^* = r^*\left|\widetilde{y}_p(n)\right|$ if the relative-precision level r^* is given, and $H^* = h^*$ if the absolute-precision level h^* is given. Let $n^\dagger$ denote the final sample size. See Alexopoulos et al. (2019b, Section 2) for a discussion of (i) the theoretical, heuristic, and practical considerations leading to our use of the point estimator $\widetilde{y}_p(n^\dagger)$ and the CI estimator $\widetilde{y}_p(n^\dagger) \pm H$ in the design of Sequest; and (ii) a formal algorithmic statement of Sequest and an explanation of the steps of the procedure. See Alexopoulos et al. (2017, Section 2) for a similar discussion of Sequem.

The Sequest software (now including Sequem) has a graphical user interface, enabling the user to do the following: (i) specify the parameters of any test process detailed in Alexopoulos et al. (2017, Section 4) or Alexopoulos et al. (2019b, Section 3), and apply either quantile-estimation algorithm automatically to a realization of the selected process that is generated by the software in real time; or (ii) apply Sequest or Sequem semiautomatically to a user-supplied dataset contained in a plain-text file. In both cases, the user has the ability to specify an upper bound on the total sample size. If in case (ii) the dataset is sufficiently large to allow normal termination of Sequest, then the selected algorithm delivers point and CI estimators of x_p that (approximately) satisfy the user-specified requirements on CI coverage and precision; otherwise, the selected algorithm terminates after providing an estimate of the sample size required to continue execution in the current step. The numerical example presented in Section 3 illustrates the use of Sequest and Sequem, including screenshots showing the results of running the software on a user-supplied dataset.

We developed stand-alone, public-domain software implementations of Sequest that include Sequem and can run under the Linux, MacOS, and Windows operating systems. Stable links to these implementations are provided in Alexopoulos et al. (2019a).

3 EXPERIMENTATION WITH THE SEQUEST SOFTWARE PACKAGE

The performance of the Sequest and Sequem procedures was evaluated using the aforementioned test processes. This evaluation was based on estimation of various performance metrics (e.g., the absolute bias of the point estimate, the CI coverage probability, the CI relative half-length, and the overall sample size) based on independent replications of each test process. In this section we demonstrate the software package using a single sample path generated from a simulation model for the aircraft maintenance facility in Problem 2.32 of Law (2015). Although this setting does not allow a thorough evaluation of the two procedures, it illustrates the functionality of the software with a model that is more “realistic” than the processes in Alexopoulos et al. (2017, 2019b), which were designed for stress-testing of the procedures.

An aircraft inspection/repair facility handles seven different types of jets, as described in Table 1 below. The times between successive plane arrivals of type ℓ ($\ell = 1,2,\ldots,7$) are exponentially distributed with mean a_ℓ; all times are in days. The facility uses $c = 12$ parallel (identical) service stations, each of which sequentially handles the inspection and (if necessary) repair of all the engines on a plane, but can deal with only one engine at a time. For example, a type-2 plane has three engines, so when it enters service, each engine must undergo an inspection-and-possible-repair process before the next engine on this plane can begin service; and all three engines must be inspected and (if necessary) repaired before the plane leaves the service station. Each service station is capable of dealing with any type of plane. A plane arriving to find an idle service station goes immediately into service, while an arriving plane finding all service stations busy must join a single FIFO queue.

Table 1: Model parameters from Problem 2.32 of Law (2015).

Plane type (ℓ)	Number of engines	a_ℓ	A_ℓ	B_ℓ	p_ℓ	r_ℓ
1	4	8.1	0.7	2.1	0.30	2.1
2	3	2.9	0.9	1.8	0.26	1.8
3	2	3.6	0.8	1.6	0.18	1.6
4*	4	8.4	1.9	2.8	0.12	3.1
5	4	10.9	0.7	2.2	0.36	2.2
6	2	6.7	0.9	1.7	0.14	1.7
7*	3	3.0	1.6	2.0	0.21	2.8

Two of the seven types of planes are classified as wide-body (denoted by an asterisk in Table 1), while the other five are classified as regular. For each engine on a type-ℓ plane, the following process takes place, where Erlang(k,μ) denotes the gamma distribution with integer shape parameter k and mean μ:

- The engine is initially inspected, taking an amount of time distributed uniformly on (A_ℓ, B_ℓ).
- A decision is made as to whether repair is needed; the probability that repair is needed is p_ℓ. If no repair is needed, inspection of the jet's next engine begins; or if this was the last engine, the jet leaves the facility.
- If repair is needed, it is carried out, taking an amount of time distributed as Erlang$(2, r_\ell)$.
- After repair, a second inspection is done, taking an amount of time distributed uniformly between $A_\ell/2$ and $B_\ell/2$. The probability that the engine needs further repair is $p_\ell/2$.
- If the initial repair is successful, the engine is done. If the engine fails the second inspection, it requires a second repair, taking an amount of time from the Erlang$(2, r_\ell/2)$ distribution. For each inspection-repair cycle after the first cycle, the inspection times are uniformly distributed between $A_\ell/2$ and $B_\ell/2$, the probability of failing inspection is $p_\ell/2$, and the repair time has the Erlang$(2, r_\ell/2)$ distribution. These cycles continue until the engine finally passes inspection.

While the full model in Law (2015) also involves operational costs and entails the evaluation of designs involving nonpreemtive priority assigned to wide-body jets in a single waiting line or allocation of a subset of stations to wide-body jets, we focus on the simplest setting and the estimation of various quantiles of the time-in-system of an arbitrary plane in steady state. We note that the basic facility is an M/G/c queueing system: the overall arrival process is Poisson with a rate of $\lambda = 1.439$ planes per day; and the service times (including the associated inspection times) are a mixture of seven distributions containing geometrically distributed random sums. To compute the first two moments of the service time conditional on the type of plane, we evaluated the event that initial repair is needed for each engine; and from the number of secondary repairs and inspections required for each engine, we obtained the mean service time for an arbitrary plane $\mathrm{E}[S] = 6.702$ and the respective variance $\mathrm{Var}[S] = 8.033$. Hence $\rho = \lambda \mathrm{E}[S]/c = 0.804$ is the long-run fraction of the facility's total service capacity that is busy.

Exact calculation of selected steady-state quantiles for the time-in-system process is difficult. Instead we computed nearly exact estimates of those parameters based on a single run starting in the empty-and-idle state, and we deleted the first 10,000 observations from the simulation-generated dataset $\left\{X_i : i = 1, \ldots, 5 \times 10^6\right\}$ to eliminate any warm-up effects. Figure 1 depicts a histogram of the truncated dataset. The sample minimum, mean, standard deviation, skewness, kurtosis, and maximum of the truncated dataset were 1.602, 7.458, 4.160, 1.085, 1.458, and 47.097, respectively. The sample lag-one autocorrelation in the warmed-up dataset was 0.1165. Using the exact values of λ, E$[S]$, and Var$[S]$ given above, we computed the standard approximation E$[Y_i] \approx 7.337$ days for the steady-state expected time in the M/G/c queueing system (Whitt 1993, Equation (2.14)); and the closeness of the sample mean $\overline{Y} = 7.458$ days to this approximation partially validates the simulation. From Figure 1 we concluded that the steady-state p.d.f. $f(x)$ of the process $\{X_i\}$ had four modes (near 2.5, 4, 6, and 9.5 days) and three antimodes (near 3, 5, and 8 days).

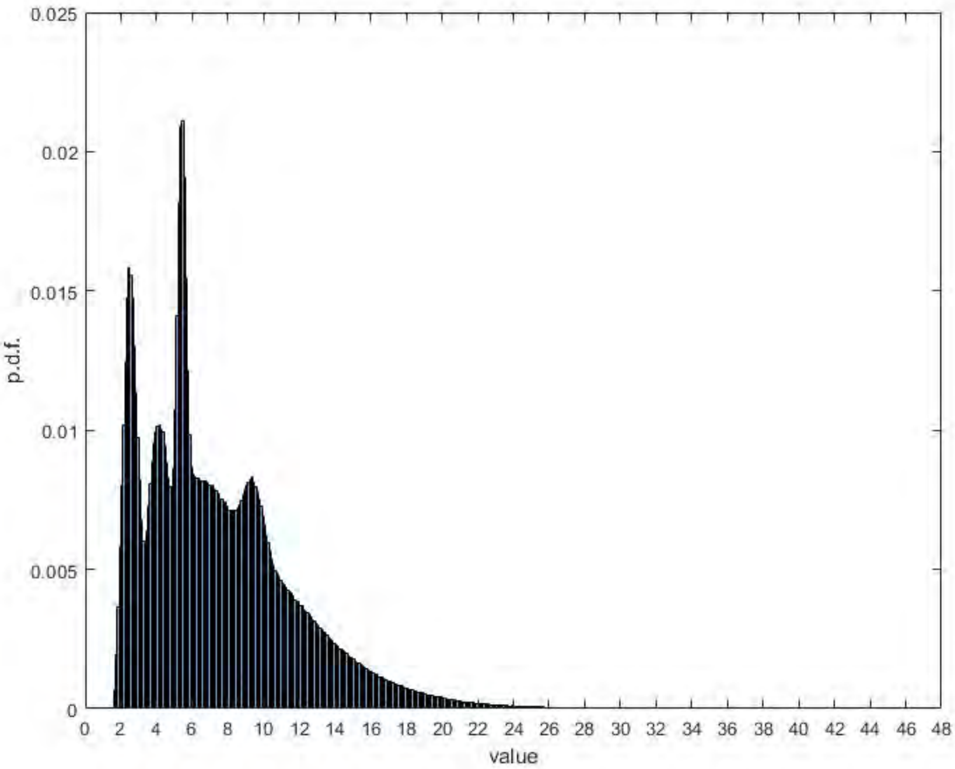


Figure 1: Histogram of times in system from the model in Section 3.

Table 2 displays experimental results from Sequest (for all values of p) and Sequem (in bold typeface for $p \geq 0.95$) in the absence of a precision requirement. For each value of p in the first column, each estimation procedure was supplied with the entire dataset of size 5 million and stopped as soon as a sufficiently large sample size was identified. Column 2 lists the nearly exact sample quantiles obtained from the truncated data set $\{X_i : i = 10^4 + 1, \ldots, 5 \times 10^6\}$ that was used to create the histogram in Figure 1. Columns 7–9 list the truncation point w, the batch size m, and the overall sample size $n = w + bm$. Column 3 lists the point estimate obtained from Sequest based on the truncated sample of size $n^\dagger = n - w$ after the removal of the initial w observations, and column 4 displays the absolute bias of the point estimate $\widetilde{y}_p(n^\dagger)$. Columns 5 and 6 list the absolute and relative half-lengths of the 95% CIs for x_p, where each CI's relative half-length is expressed as a percentage of its absolute midpoint.

An examination of Table 2 reveals that both methods performed very well with regard to this output process: (a) in all cases the length of the warm-up period w was practically negligible relative to the overall sample size n, and it was much smaller than the truncation point used for the computation of the point estimates in column 2; (b) the estimates of the absolute bias and the CI relative precision in columns 4 and 6 were small; and (c) the overall sample sizes required to obtain such accurate 95% CIs were also relatively small. The variability of the sample size in the neighborhoods of the modes is partially attributable to the skewness of the quantile estimators obtained from the nonoverlapping batches; a heuristic explanation of this phenomenon is given in Section 4.2 of Alexopoulos et al. (2018) and in Section EC.3 of the e-companion of Alexopoulos et al. (2019b). As p increased from 0.98 to 0.995, Sequem required progressively larger sample sizes than Sequest (from one to two orders of magnitude); this is a reasonable compromise for the superior CI coverage probability delivered by Sequem under no CI precision requirements (Alexopoulos et al. 2017).

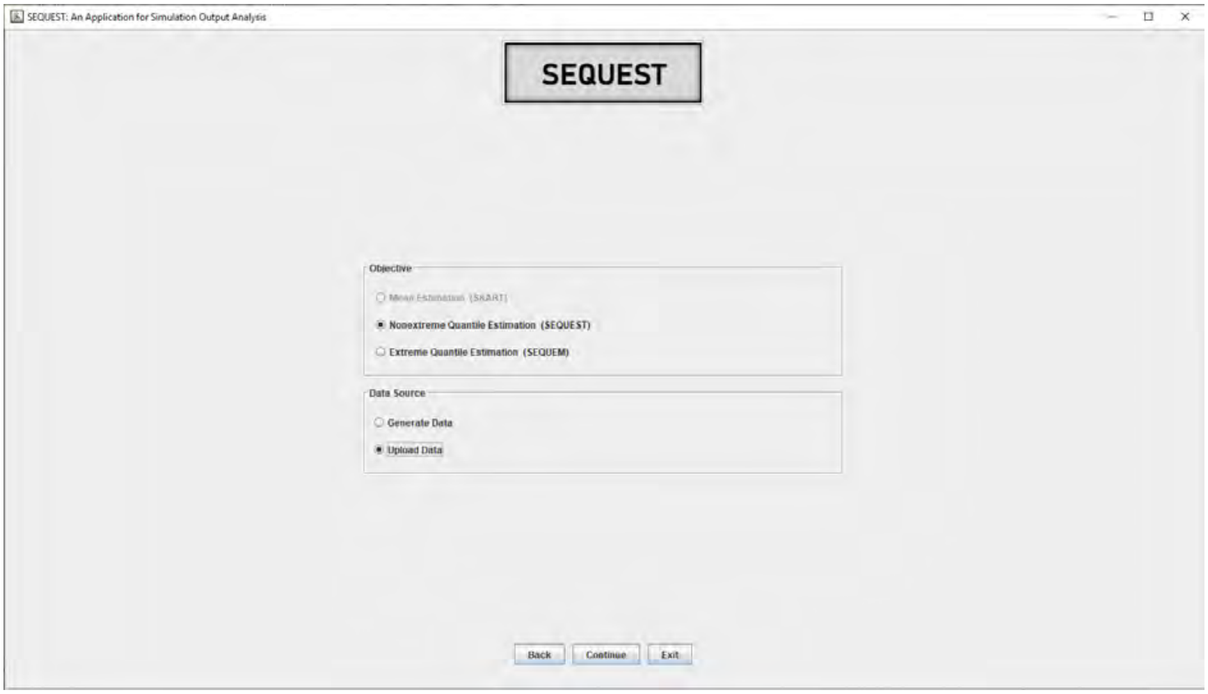
Table 2: Experimental results for point and 95% CI estimators of the p-quantile x_p of the time-in-system process from Problem 2.32 of Law (2015).

No CI Precision Requirement								
p	x_p	$\widetilde{y}_p(n^\dagger)$	$\left\|\text{Bias}\left[\widetilde{y}_p(n^\dagger)\right]\right\|$	H	CI Rel. Prec. (%)	w	m	n
0.05	2.351	2.354	0.003	0.013	0.533	256	4,180	134,016
0.10	2.674	2.671	0.003	0.018	0.678	437	5,844	187,445
0.15	3.156	3.194	0.038	0.094	2.938	437	3,152	101,301
0.20	3.850	3.876	0.026	0.068	1.745	437	2,452	78,901
0.25	4.349	4.346	0.003	0.073	1.765	437	3,996	128,309
0.30	4.932	4.961	0.029	0.080	1,614	437	2,452	78,901
0.35	5.305	5.305	0.000	0.042	0.794	437	4,128	132,533
0.40	5.545	5.530	0.015	0.020	0.354	437	13,916	445,749
0.45	5.928	5.895	0.033	0.039	0.665	437	13,916	445,749
0.50	6.528	6.515	0.013	0.079	1.214	437	6,952	222,901
0.55	7.147	7.135	0.012	0.068	0.950	437	7,880	252,597
0.60	7.798	7.784	0.014	0.077	0.990	437	8,268	265,013
0.65	8.495	8.483	0.012	0.079	0.935	437	7,836	251,189
0.70	9.150	9.121	0.029	0.068	0.742	437	9,368	300,213
0.75	9.772	9.745	0.027	0.048	0.495	437	13,916	445,749
0.80	10.576	10.534	0.042	0.073	0.694	437	13,916	445,749
0.85	11.690	11.644	0.046	0.108	0.924	437	9,836	315,189
0.90	13.138	13.079	0.059	0.055	0.417	437	25,312	810,421
0.92	13.891	13.843	0.048	0.063	0.455	437	24,584	787,125
0.94	14.834	14.802	0.032	0.125	0.848	384	12,432	398,208
0.95	15.416	15.377	0.039	0.119	0.776	437	15,320	490,677
		15.408	**0.008**	**0.204**	**1.327**	**512**	**4,978**	**319,104**
0.96	16.116	16.082	0.034	0.172	1.071	437	9,836	315,189
		16.195	**0.113**	**0.329**	**2.030**	**512**	**2,234**	**143,488**
0.98	18.184	18.732	0.548	0.642	3.428	256	512	16,640
		18.248	**0.064**	**0.252**	**1.379**	**768**	**2,048**	**328,448**
0.99	20.145	20.862	0.711	0.667	3.196	256	512	16,640
		20.218	**0.073**	**0.239**	**1.182**	**512**	**1,450**	**464,512**
0.995	22.028	22.739	0.717	0.532	2.338	256	612	19,840
		22.001	**0.027**	**0.183**	**0.830**	**1,536**	**2,048**	**1,377,792**

Below we illustrate the functionality of the Sequest method with regard to the estimation of the 99th percentile of our time-in-system process with a relative precision of 1% for the 95% CI for $x_{0.99}$. Figure 2 displays the initial (experimental) screen of the application, where the user specifies the method choice (in this case, Sequest) and the fact that a dataset will be read from a text file. Figure 3 depicts a portion of the next screen, where we have inserted the location of the input file and we have imposed an artificial upper bound on the sample size (to mimic a more-realistic situation where the user has a smaller dataset).

Figure 4 depicts a portion of the subsequent screen where we specified the estimation of the 99th percentile with a relative precision of 1% for the relative half-length of the 95% CI and then hit the **Finalize Inputs** button. The user proceeds by hitting the **Continue** button at the bottom of this screen and then hitting the **Start** button at the next screen. Figure 5 displays a caption of the first outcome, where Sequest points out that the supplied sample size is insufficient for computing a 95% for $x_{0.99}$ with a relative precision of 1%.

At this junction we hit the **Back** button twice, raise the upper bound on the sample size to 500,000, and repeated the steps in the previous paragraph. (If we used the sample sizes recommended by Sequest, we would have to perform this cycle three times.) Figure 6 displays the final output of the method. Notice that the overall sample size required to obtain the 95% CI, namely $x_{0.99} \in 20.065 \pm 0.151$, with an estimated relative precision of 0.750% increased from 16,640 from the case of no precision requirement (Table 2) to


SEQUEST: An Application for Simulation Output Analysis

SEQUEST

Objective

○ Mean Estimation (SKART)

● Nonextreme Quantile Estimation (SEQUEST)

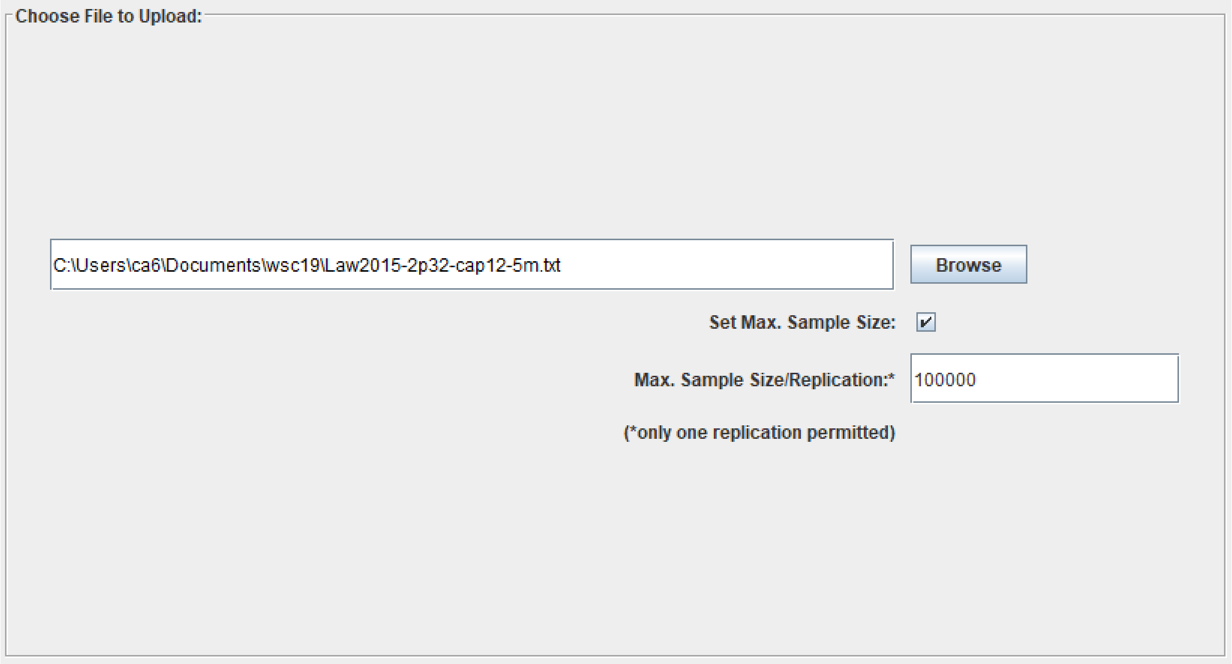
○ Extreme Quantile Estimation (SEQUEM)

Data Source

○ Generate Data

● Upload Data

Figure 2: Initial experimental screen of Sequest.


Choose File to Upload:

C:\Users\ca6\Documents\wsc19\Law2015-2p32-cap12-5m.txt

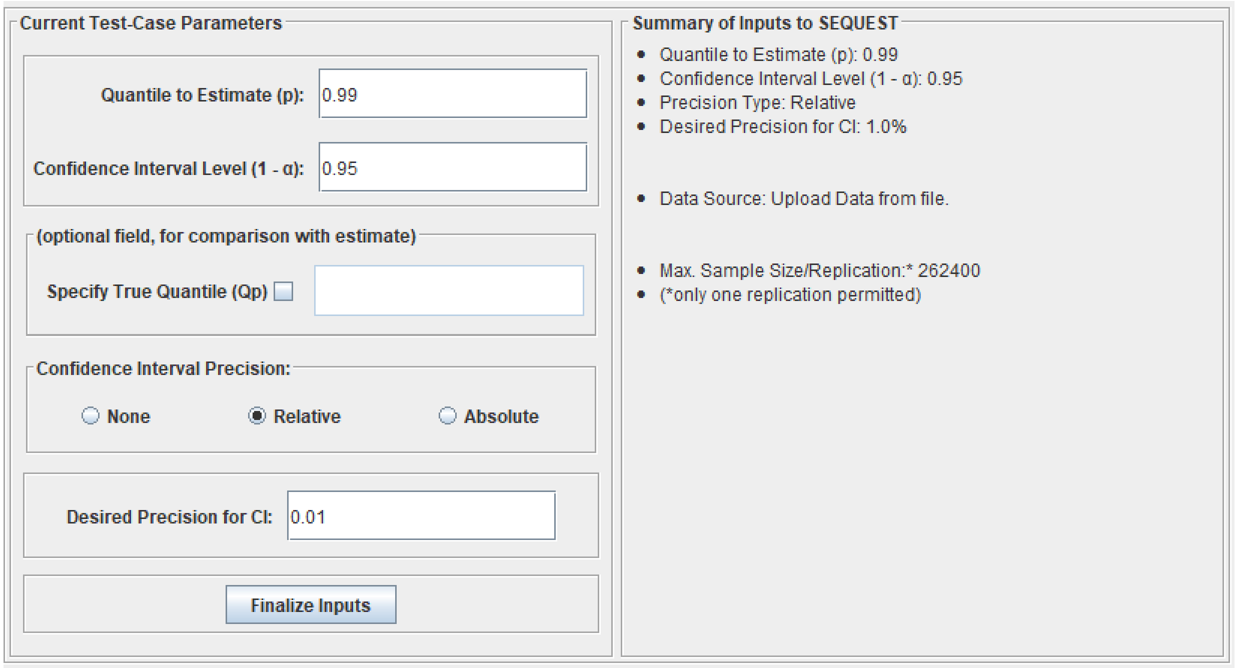
Set Max. Sample Size: ☑

Max. Sample Size/Replication:* 100000

(*only one replication permitted)

Figure 3: Partial screen of Sequest with the file path containing the data set in Section 3 and an upper bound of 100,000 on the sample size.

433,504; this increase is reasonably close to the well-known ratio $(3.196/0.750)^2 \approx 18.160$. Note that the latter performance metrics are on par with the output of Sequest under no CI precision requirement.



Current Test-Case Parameters

Quantile to Estimate (p): 0.99

Confidence Interval Level (1 - α): 0.95

(optional field, for comparison with estimate)

Specify True Quantile (Qp) ☐

Confidence Interval Precision:

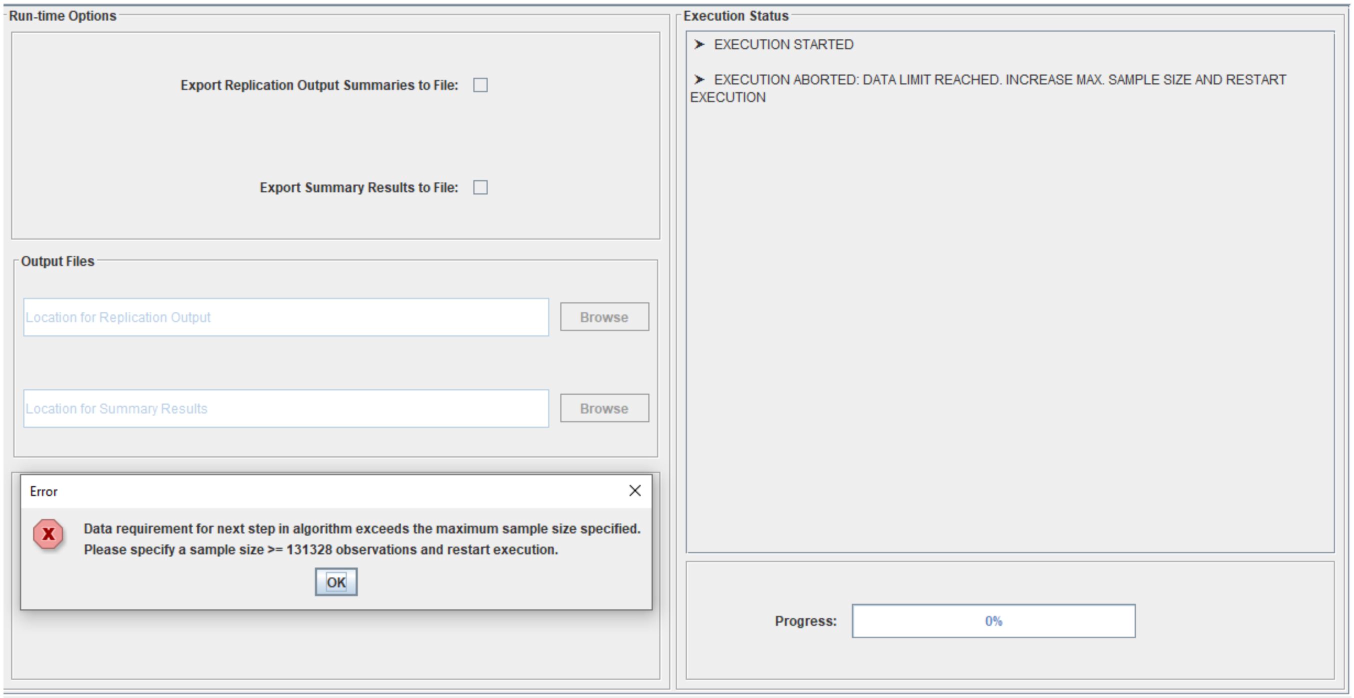
○ None ● Relative ○ Absolute

Desired Precision for CI: 0.01

Summary of Inputs to SEQUEST

- Quantile to Estimate (p): 0.99
- Confidence Interval Level (1 - α): 0.95
- Precision Type: Relative
- Desired Precision for CI: 1.0%
- Data Source: Upload Data from file.
- Max. Sample Size/Replication:* 262400
- (*only one replication permitted)

Figure 4: Partial screen of Sequest with the entries for p, confidence level $(1-\alpha)$ for the CI for x_p, and a relative precision requirement of 0.01.



Run-time Options

Export Replication Output Summaries to File: ☐

Export Summary Results to File: ☐

Output Files

Error

Data requirement for next step in algorithm exceeds the maximum sample size specified.
Please specify a sample size >= 131328 observations and restart execution.

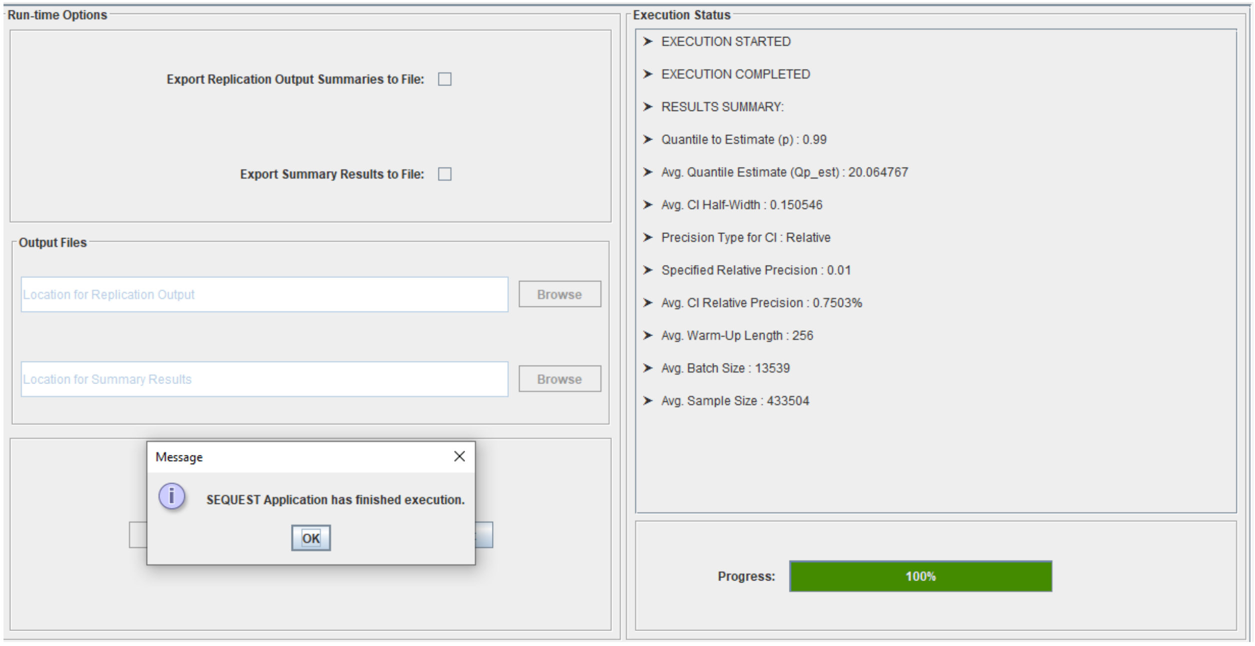
Execution Status

➤ EXECUTION STARTED

➤ EXECUTION ABORTED: DATA LIMIT REACHED. INCREASE MAX. SAMPLE SIZE AND RESTART EXECUTION

Progress: 0%

Figure 5: Partial screen of Sequest noting the inadequacy of the sample size with regard to the experimental setup in Figure 4.



Export Replication Output Summaries to File: □

Export Summary Results to File: □

SEQUEST Application has finished execution.

- EXECUTION STARTED
- EXECUTION COMPLETED
- RESULTS SUMMARY:
- Quantile to Estimate (p) : 0.99
- Avg. Quantile Estimate (Qp_est) : 20.064767
- Avg. CI Half-Width : 0.150546
- Precision Type for CI : Relative
- Specified Relative Precision : 0.01
- Avg. CI Relative Precision : 0.7503%
- Avg. Warm-Up Length : 256
- Avg. Batch Size : 13539
- Avg. Sample Size : 433504

Progress: 100%

Figure 6: Final partial screen of Sequest for the estimation of $x_{0.99}$ with the 95% CI precision requirement of 0.01.

4 AN APPROACH TO QUANTILE ESTIMATION USING STANDARDIZED TIME SERIES

To formulate point and CI estimators of x_p based on the method of standardized time series using the warmed-up process $\{Y_i : i \geq 1\}$, we let $\mathbb{Z} \equiv \{0, \pm 1, \pm 2, \ldots\}$; and for each $x \in \mathbb{R}$ and $i \geq 1$, we let $I_i(x) \equiv 1$ if $Y_i \leq x$, and $I_i(x) \equiv 0$ otherwise. We assume that $\{Y_i : i \geq 1\}$ and the associated binary process $\{I_i(x_p) : i \geq 1\}$ satisfy the following conditions.

Geometric-Moment Contraction (GMC) Condition: The process $\{Y_i : i \geq 0\}$ is expressed in terms of a function $\xi(\cdot)$ of a sequence of i.i.d. random variables $\{\varepsilon_j : j \in \mathbb{Z}\}$ such that $Y_i = \xi(\ldots, \varepsilon_{i-1}, \varepsilon_i)$ for $i \geq 0$; moreover, there exist constants $\psi > 0$, $C > 0$, and $r \in (0,1)$ such that for two independent sequences $\{\varepsilon_j : j \in \mathbb{Z}\}$ and $\{\varepsilon_j^{\dagger} : j \in \mathbb{Z}\}$ each consisting of i.i.d. random variables with the same distribution as ε_0,

$$\mathrm{E}\Big[\big|\xi(\ldots, \varepsilon_{-1}, \varepsilon_0, \varepsilon_1, \ldots, \varepsilon_i) - \xi\big(\ldots, \varepsilon_{-1}^{\dagger}, \varepsilon_0^{\dagger}, \varepsilon_1, \ldots, \varepsilon_i\big)\big|^{\psi}\Big] \leq Cr^i \quad \text{for} \quad i \geq 0.$$

Density-Regularity (DR) Condition: The p.d.f. $f(x)$ is continuous at every $x \in \mathbb{R}$ with $\sup_{x \in \mathbb{R}} f(x) < \infty$; and at the quantile x_p to be estimated, $f(x_p) > 0$ and the derivative $f'(x_p)$ exists.

Functional Central Limit Theorem for $\{I_i(x_p)\}$: Let $\bar{I}(x_p, n) \equiv (1/n)\sum_{i=1}^{n} I_i(x_p)$ for $n \geq 1$. We define the autocorrelation function $\rho_{I(x_p)}(\ell) \equiv \mathrm{Corr}\big[I_i(x_p), I_{i+\ell}(x_p)\big]$ at each lag $\ell \in \mathbb{Z}$; and we assume that $\sum_{\ell \in \mathbb{Z}} \big|\rho_{I(x_p)}(\ell)\big| < \infty$ so we have the corresponding variance parameter $\sigma_{I(x_p)}^2 \equiv \lim_{n \to \infty} n\mathrm{Var}\big[\bar{I}(x_p, n)\big] = p(1-p)\sum_{\ell \in \mathbb{Z}} \rho_{I(x_p)}(\ell) \in (0, \infty)$. Finally we assume that the sequence of random functions

$$\mathscr{I}_n(t; p, x_p) \equiv \frac{1}{\sigma_{I(x_p)} n^{1/2}} \sum_{i=1}^{\lfloor nt \rfloor} [I_i(x_p) - p] \quad \text{for} \quad t \in [0,1] \quad \text{and} \quad n \geq 1$$

satisfies the functional central limit theorem (FCLT)

$$\mathscr{I}_n(\cdot; p, x_p) \underset{n \to \infty}{\Longrightarrow} \mathscr{W}(\cdot), \tag{1}$$

where $\lfloor\cdot\rfloor$ denotes the floor function; $\mathscr{W}(\cdot)$ denotes standard Brownian motion on $[0,1]$; and $\underset{n\to\infty}{\Longrightarrow}$ denotes weak convergence as $n\to\infty$ in the space D of real-valued functions on $[0,1]$ that are right-continuous with left-hand limits (Billingsley 1999, Section 12).

From the jth batch $\{Y_{(j-1)m+1},\ldots,Y_{jm}\}$, we compute the jth batch mean of the associated binary r.v.'s, $\bar{I}_j(x_p,m)\equiv(1/m)\sum_{\ell=1}^{m}I_{(j-1)m+\ell}(x_p)$ for $j=1,\ldots,b$, where the batch count b is fixed. Let $\mathbf{Z}_b$ denote a $b\times 1$ standard normal random vector. If $\{\mathscr{Q}_m : m\geq 1\}$ is a sequence of r.v.'s and $\{r_m : m\geq 1\}$ is a sequence of positive constants, then the notation $\mathscr{Q}_m = O_{\text{a.s.}}(r_m)$ means there is a (bounded) r.v. U such that $|\mathscr{Q}_m/r_m|\leq U$ for $m\geq 1$ a.s. Based on this setup, Alexopoulos et al. (2019b) prove the following result.

Theorem 1 If $\{Y_i : i\geq 1\}$ satisfies the GMC and DR conditions as well as the FCLT (1), then

$$\widehat{y}_p(j,m) = x_p - \frac{\bar{I}_j(x_p,m)-p}{f(x_p)} + O_{\text{a.s.}}\left\{\frac{[\log(m)]^{3/2}}{m^{3/4}}\right\} \quad \text{as} \quad m\to\infty$$

for $j=1,\ldots,b$; and $m^{1/2}\big[\widehat{y}_p(1,m)-x_p,\ldots,\widehat{y}_p(b,m)-x_p\big]^{\mathsf{T}} \underset{m\to\infty}{\Longrightarrow} \big[\sigma_{I(x_p)}/f(x_p)\big]\mathbf{Z}_b$.

As in Section 2, we let $\widetilde{y}_p(n)$ denote the conventional sectioning-based estimator of x_p based on the warmed-up time series $\{Y_i : i=1,\ldots,n\}$ of length n; and we consider the asymptotic behavior as $n\to\infty$ of the centered-and-scaled quantile-estimation process

$$\mathfrak{X}_n(t;p,x_p) \equiv \frac{\lfloor nt\rfloor\{\widetilde{y}_p(\lfloor nt\rfloor)-x_p\}}{n^{1/2}} = -\frac{\sigma_{I(x_p)}\mathscr{I}_n(t;p,x_p)}{f(x_p)} + O_{\text{a.s.}}\left\{\frac{[\log(n)]^{3/2}}{n^{1/4}}\right\} \text{ for } t\in[0,1],\ n\geq 1. \quad (2)$$

Theorem 2 If $\{Y_i : i\geq 1\}$ satisfies the assumptions of Theorem 1, then

$$\big[f(x_p)/\sigma_{I(x_p)}\big]\mathfrak{X}_n(\cdot;p,x_p) \underset{n\to\infty}{\Longrightarrow} \mathscr{W}(\cdot). \quad (3)$$

Proof. Let U^* denote a (bounded) r.v. such that in Equation (2), we have

$$\big|O_{\text{a.s.}}\{[\log(n)]^{3/2}/n^{1/4}\}\big| \leq U^*[\log(n)]^{3/2}/n^{1/4} \quad \text{for} \quad n\geq 1 \quad \text{a.s.} \quad (4)$$

For $x\in D$, let $\|x\|\equiv\sup\{|x(t)| : t\in[0,1]\}$; and let Λ denote the class of strictly increasing, continuous mappings of $[0,1]$ onto itself, where $I\in\Lambda$ denotes the identity map. For $x,y\in D$, let $d(x,y)\equiv \inf_{\lambda\in\Lambda}\big\{\max\big[\|I-\lambda\|,\|x-y\circ\lambda\|\big]\big\}$ denote the distance between x and y in the Skorohod metric on D (Billingsley 1999, Equation(12.13)). From Equations (2) and (4), we see that

$$d\left[\frac{f(x_p)\mathfrak{X}_n(\cdot;p,x_p)}{\sigma_{I(x_p)}}, -\mathscr{I}_n(\cdot;p,x_p)\right] \leq \frac{f(x_p)U^*[\log(n)]^{3/2}}{\sigma_{I(x_p)}n^{1/4}} \underset{n\to\infty}{\Longrightarrow} 0; \quad (5)$$

and Equation (3) follows from the FCLT (1), Equation (5), and the convergence-together theorem (Billingsley 1999, Theorem 3.1). ∎

Finally we obtain the FCLT required for an STS-based quantile-estimation procedure. Let

$$\mathfrak{A}_n(t)\equiv\big(\lfloor nt\rfloor/n^{1/2}\big)\big[\widetilde{y}_p(\lfloor nt\rfloor)-\widetilde{y}_p(n)\big] \quad \text{for} \quad t\in[0,1],\ n\geq 1$$

denote the STS quantile-estimation process; and let $\mathscr{B}(t)\equiv\mathscr{W}(t)-t\mathscr{W}(1)$ for $t\in[0,1]$ denote a standard Brownian bridge that is independent of $\mathscr{W}(1)$ (Billingsley 1999, pp. 101–104).

Theorem 3 If $\{Y_i : i\geq 1\}$ satisfies the assumptions of Theorem 1, then

$$\big[f(x_p)/\sigma_{I(x_p)}\big]\big\{n^{1/2}\big[\widetilde{y}_p(n)-x_p\big],\mathfrak{A}_n(\cdot)\big]\big\} \underset{n\to\infty}{\Longrightarrow} [\mathscr{W}(1),\mathscr{B}(\cdot)]. \quad (6)$$

Proof. Define the map $\Theta : D \mapsto \mathbb{R} \times D$ as $\Theta(y) \equiv [y(1), y(t) - ty(1)]$ for $y \in D$, where $\mathbb{R} \times D$ has the metric $\tau(\cdot,\cdot)$ defined as follows: for $\mathfrak{s}_i = (r_i, x_i) \in \mathbb{R} \times D$ $(i = 1,2)$, we have $\tau(\mathfrak{s}_1, \mathfrak{s}_2) \equiv \big\{(r_1 - r_2)^2 + [d(x_1, x_2)]^2\big\}^{1/2}$. Let $\text{Disc}(\Theta)$ denote the discontinuity points of Θ. We show that $\Theta(\cdot)$ is continuous at every $y \in C$ so $\Pr\{\mathscr{W} \in \text{Disc}(\Theta)\} = 0$. By the continuous-mapping theorem (Whitt 2002, Theorem 3.4.3), $[f(x_p)/\sigma_{I(x_p)}]\Theta(\mathfrak{X}_n) \underset{n\to\infty}{\Longrightarrow} [\mathscr{W}(1), \mathscr{B}(\cdot)]$. Let $\mathfrak{B}_n(t) \equiv \mathfrak{X}_n(t; p, x_p) - t\mathfrak{X}_n(1; p, x_p)$ for $t \in [0,1]$. Since $n^{1/2}\big[\widetilde{y}_p(n) - x_p\big] \underset{n\to\infty}{\Longrightarrow} \big[\sigma_{I(x_p)}/f(x_p)\big]\mathscr{W}(1)$, we see that $\big\|\mathfrak{A}_n(\cdot) - \mathfrak{B}_n(\cdot)\big\| \underset{n\to\infty}{\Longrightarrow} 0$ by Slutsky's Theorem. Equation (6) follows by the convergence-together theorem. ∎

Remark 1 Comparable results hold if the GMC condition is replaced by an appropriate ϕ-mixing condition (Bradley 2005, Equations (1.2) and (2.2)). For example, Sen (1972) derives a Bahadur representation with a remainder term of the form $O_{\text{a.s.}}[\log(n)/n^{1/8}]$ under the assumption that (i) $f(x)$ satisfies the DR condition and $f'(x)$ is positive and bounded in some neighborhood of x_p; and (ii) $\{Y_i : i \geq 1\}$ is ϕ-mixing with $\sum_{\ell=1}^{\infty}[\phi(\ell)]^{1/2} < \infty$. In this situation Billingsley (1968, Theorem 21.1) ensures that the FCLT (1) holds, and hence Theorems 2 and 3 hold. The ϕ-mixing condition (ii) is much harder to verify compared with the GMC condition, which can at least be checked empirically as discussed in Alexopoulos et al. (2012). ◀

Remark 2 We can show comparable results if the GMC condition is replaced by an appropriate ρ-mixing condition (Bradley 2005, Equations (1.2) and (2.2)); but it is unclear such results hold for any α-mixing condition (Bradley 2005, Equations (1.1) and (2.2)). Wang et al. (2011) derive a Bahadur representation with a remainder term of the form $O_{\text{a.s.}}\{[\log(n)]^{3/4}/n^{1/2}\}$ by assuming (i) $f'(x)$ is bounded in some neighborhood of x_p; and (ii) $\{Y_i : i \geq 1\}$ is α-mixing with $\alpha(\ell) \leq \mathfrak{C}/\ell^{\mathfrak{h}}$ for $\ell \geq 1$ and some constants $\mathfrak{C} > 0$ and $\mathfrak{h} > 11$. We are unaware of any result for α-mixing processes comparable to Theorem 21.1 of Billingsley (1968) (for ϕ-mixing processes) or Theorem 19.3 of Billingsley (1999) (for ρ-mixing processes) that ensures either FCLT (1) or FCLT (6) holds. In any case, the relevant α- and ρ-mixing conditions are hard to verify. ◀

Remark 3 Using the FCLT (6), we can construct an STS-based sequential procedure for estimating steady-state quantiles with the following batching techniques: nonoverlapping batch means, batched STS area estimators, and overlapping batch means (Alexopoulos et al. 2016). The development in this section parallels to some extent the approach of Calvin and Nakayama (2013) for deriving STS-based point and CI estimators of x_p. However, the latter approach assumes the $\{Y_i : i \geq 1\}$ are i.i.d. and requires deeper analysis than the approach in this section. ◀

5 CONCLUSIONS

In this article we reviewed some new developments concerning the Sequest and Sequem sequential procedures for estimating nonextreme and extreme steady-state quantiles, respectively. We presented an example illustrating how the associated software is applied to a simulation output process exhibiting warm-up effects, autocorrelation, and nonnormality. Finally we established a readily-accessible theoretical foundation for STS-based quantile-estimation procedures when the output process satisfies the usual DR condition and either the GMC condition or certain mixing conditions. Ongoing work includes (i) further improvements to the Sequest and Sequem procedures and their associated implementations in a public-domain software package; and (ii) sequential quantile-estimation procedures based on the STS method.

ACKNOWLEDGMENTS

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

REFERENCES

Alexopoulos, C., D. Goldsman, A. C. Mokashi, K.-W. Tien, and J. R. Wilson. 2019a. "Online Availability of the Sequest Software for Linux, MacOS, and Windows". https://people.engr.ncsu.edu/jwilson/files/sequest-availability.pdf, accessed 13th July 2019.

Alexopoulos, C., D. Goldsman, A. C. Mokashi, K.-W. Tien, and J. R. Wilson. 2019b. "Sequest: A Sequential Procedure for Estimating Quantiles in Steady-State Simulations". *Operations Research* 67(4):1162–1183. https://people.engr.ncsu.edu/jwilson/files/sequest19or.pdf, accessed 7$^{\text{th}}$ September 2019.

Alexopoulos, C., D. Goldsman, A. C. Mokashi, and J. R. Wilson. 2017. "Automated Estimation of Extreme Steady-State Quantiles via the Maximum Transformation". *ACM Transactions on Modeling and Computer Simulation* 27(4):22:1–22:29.

Alexopoulos, C., D. Goldsman, A. C. Mokashi, and J. R. Wilson. 2018. "Sequential Estimation of Steady-State Quantiles: Lessons Learned and Future Directions". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1814–1825. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Alexopoulos, C., D. Goldsman, P. Tang, and J. R. Wilson. 2016. "SPSTS: A Sequential Procedure for Estimating the Steady-State Mean Using Standardized Time Series". *IIE Transactions* 48(9):864–880.

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, New Jersey: Institute of Electrical and Electronics Engineers.

Billingsley, P. 1968. *Convergence of Probability Measures*. New York: John Wiley & Sons.

Billingsley, P. 1999. *Convergence of Probability Measures*. 2nd ed. New York: John Wiley & Sons.

Bradley, R. C. 2005. "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions". *Probability Surveys* 2:107–144.

Calvin, J. M., and M. K. Nakayama. 2013. "Confidence Intervals for Quantiles with Standardized Time Series". In *Proceedings of the 2013 Winter Simlation Conference*, edited by R. Pasupathy, S.-H. Kim, R. H. A. Tolk, and M. E. Kuhl, 601–612. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Law, A. M. 2015. *Simulation Modeling and Analysis*. 5th ed. New York: McGraw-Hill.

Sen, P. K. 1972. "On the Bahadur Representation of Sample Quantiles for Sequences of ϕ-Mixing Random Variables". *Journal of Multivariate Analysis* 2(1):77–95.

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

Wang, X., S. Hu, and W. Yang. 2011. "The Bahadur Represenation for Sample Quantiles under Strongly Mixing Sequence". *Journal of Statistical Planning and Inference* 141:655–662.

Whitt, W. 1993. "Approximations for the GI/G/m Queue". *Production and Operations Management* 2(2):114–161.

Whitt, W. 2002. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues*. New York: Springer.

AUTHOR BIOGRAPHIES

CHRISTOS ALEXOPOULOS is a Professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech. His main research interests are in the areas of applied probability and computer simulation. His research has been recognized with three awards, including the INFORMS Simulation Society 2007 Outstanding Simulation Publication Award, the 2010 Best Paper Award in Operations Engineering and Analysis of *IIE Transactions*, and the Best Paper Award at the 24th Workshop on Principles of Advanced and Distributed Simulation (PADS). His e-mail address is christos@gatech.edu, and his Web page is www.isye.gatech.edu/∼christos.

DAVID GOLDSMAN is a Professor and the Director of Master's Programs in the H. Milton Stewart School of Industrial and Systems Engineering at the Georgia Institute of Technology. His research interests include simulation output analysis, statistical ranking and selection methods, and medical and humanitarian applications of operations research. Dave is a recipient of the INFORMS Simulation Society's Distinguished Service Award, and is a Fulbright Fellow and a Fellow of the Institute of Industrial and Systems Engineers (IISE). 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 Analyst at SAS Institute Inc. in Cary, North Carolina. He holds an MS in Industrial Engineering from North Carolina State University. His research interests include design and implementation of algorithms related to statistical aspects of discrete-event simulation. His career interests include applying simulation and other Operations Research techniques to large scale industrial problems. He is a member of IISE and INFORMS. His e-mail address is Anup.Mokashi@sas.com.

JAMES R. WILSON is a Professor Emeritus in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. His research interests concern modeling, analysis, and simulation of stochastic systems, especially as applied in healthcare, production, and quality systems engineering. He is a Fellow of INFORMS and IISE. His e-mail address is jwilson@ncsu.edu, and his Web page is www.ise.ncsu.edu/jwilson.