A Sequential Procedure for Estimating Steady-State Quantiles

Christos Alexopoulos[1] David Goldsman[1] Anup C. Mokashi[2]
Rong Nie[3] Qing Sun[3] Kai-Wen Tien[3] James R. Wilson[3]

[1]Georgia Tech
[2]SAS Institute
[3]North Carolina State University

www.ise.ncsu.edu/jwilson/informs14-sequest.pdf

November 9, 2014

Outline

Notation and Assumptions

- We have a simulation output process $\{X_i : i = 1, 2, \ldots\}$ with steady-state c.d.f. $F(x) = \Pr\{X_i \le x\}$ and p.d.f. $f(x) = F'(x)$.
- Given $p \in (0, 1)$, we seek both point and confidence interval (CI) estimators of the p-quantile $x_p \equiv F^{-1}(p) \equiv \inf\{x : F(x) \ge p\}$ based on a simulation-generated time series $\{X_i : i = 1, \ldots, n\}$ of sufficient length n, where the CI has the form

$$\widetilde{x}_p(n) \pm H, \tag{1}$$

with a user-specified confidence coefficient $1 - \alpha \in (0, 1)$ and a user-specified precision specification of the form

$$H \le H^* = \begin{cases} r^* \left|\widetilde{x}_p(n)\right|, & \text{for relative precision level } r^*, \\ h^*, & \text{for absolute precision level } h^*. \end{cases}$$

Notation and Assumptions (Cont'd)

- We organize $\{X_i : i = 1, \ldots, n\}$ into b nonoverlapping batches of size m ($n = bm$). From the jth batch $\{X_{(j-1)m+1}, \ldots, X_{jm}\}$, we obtain the order statistics $X_{j,(1)} \le X_{j,(2)} \le \cdots \le X_{j,(m)}$ and the associated batch quantile estimator (BQE)

$$\widehat{x}_p(j,m) = X_{j,(\lceil mp \rceil)} \quad \text{for} \quad j = 1, \ldots, b. \tag{2}$$

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

$$\widetilde{x}_p(n) = X_{(\lceil np \rceil)}\,. \tag{3}$$

Notation and Assumptions (Cont'd)

- Using (2) and (3), we also compute a modified estimator of the variance of the BQEs,

$$\widetilde{S}^2_{\widehat{x}_p}(b,m) \equiv b^{-1} \sum_{j=1}^{b} \big[\, \widehat{x}_p(j,m) - \widetilde{x}_p(n) \,\big]^2. \tag{4}$$

- Under certain mild assumptions, we have shown that as $m \to \infty$ with b fixed, an asymptotically valid $100(1-\alpha)\%$ CI for x_p has the form

$$\widetilde{x}_p(n) \pm H, \text{ where } H = t_{1-\alpha/2,b-1}\widetilde{S}_{\widehat{x}_p}(b,m)\big/\sqrt{b}, \tag{5}$$

where $t_{q,\nu}$ is the q-quantile of Student's t distribution with ν degrees of freedom.

Sequest: A Sequential Procedure for Estimating Steady-State Quantiles

Sequest is a sequential procedure that delivers improved point and CI estimators for x_p by exploiting a combination of ideas from batching and sectioning to do the following:

- determine the length w of the warm-up period beyond which the truncated sample statistics (2)–(4) are approximately free of initialization bias;
- adjust the CI half-length H in (5) to compensate for any skewness or correlation in the BQEs $\{\widehat{x}_p(j,m) : j = 1, \ldots, b\}$; and
- determine sufficiently large values of w, m, b, and the total sample size $n = w + bm$ so that the user-specified precision and coverage probability are achieved by the final CI estimator (1) of x_p.

Initialization

[0] Set the initial sample size $n \leftarrow 4{,}096$, batch size $m \leftarrow 64$, and batch count $b \leftarrow 64$. Set the randomness test size, $\alpha_{\text{ran}} \leftarrow 0.25$. Set the parameters $\eta \leftarrow 2.82888$ and $\theta \leftarrow 2$ of the upper-bound function on absolute skewness of the BQEs,

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

Set the upper bound $u^* \leftarrow 5$ on the iterations of the skewness-reducing batch-size adjustment step [3].

Determining the Length w of the Warm-up Period

[1] Compute the BQEs $\{\hat{x}_p(j,m) : j = 1,\ldots b\}$, their sample mean $\overline{x}_p(b,m)$, and sample variance $S^2_{\hat{x}_p}(b,m)$.

- **[a]** If the BQEs exhibit no significant variation (i.e., $S^2_{\hat{x}_p}(b,m)$ is too close to 0 or too small relative to $|\overline{x}_p(b,m)|$), then go to step **[1b]**; otherwise go to step **[2]**
- **[b]** Perform the updates $m \leftarrow 2m$ and $n \leftarrow 2n$; obtain the required additional observations by restarting the simulation if necessary; update the BQEs and their sample statistics; and return to step **[1a]**.

[2] Apply von Neumann's test for randomness to the current BQEs.

- **[a]** If the randomness test is passed at the significance level α_{ran}, then go to step **[3]**; otherwise go to step **[2b]**
- **[b]** Perform the updates $m \leftarrow 2m$ and $n \leftarrow 2n$; obtain the required additional observations by restarting the simulation if necessary; update the BQEs and their sample statistics; and return to step **[2a]**.

Reducing the Skewness of the BQEs

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

[a] Update the total sample size, $n \leftarrow w + bm$, and obtain the additional observations by restarting the simulation if necessary. Skip the first w observations to obtain the "warmed-up" series of length $n' = n - w$, $\{Y_i = X_{w+i} : i = 1, \ldots, n'\}$. From the jth warmed-up batch $\{Y_{(j-1)m+i} : i = 1, \ldots, m\}$, compute jth warmed-up BQE $\widehat{y}_p(j, m)$. Compute the sample mean $\overline{y}_p(b, m)$, sample variance $S^2_{\widehat{y}_p}(b, m)$, and sample skewness $\widehat{\mathscr{B}}_{\widehat{y}_p}(b, m)$ of the warmed-up BQEs.

[b] If $\left|\widehat{\mathscr{B}}_{\widehat{y}_p}(b, m)\right| \leq \mathscr{B}^*(p)$ or $u = u^*$, then go to step **[4]**; otherwise increase the batch size according to

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

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

Computing the Point and CI Quantile Estimators

[4] Perform the updates $b \leftarrow b/2$, $m \leftarrow 2m$, and $n \leftarrow w + bm$; and obtain the required additional observations by restarting the simulation if necessary.

[5] Update the warmed-up BQEs, their sample mean $\overline{y}_p(b,m)$, sample variance $S^2_{\widehat{y}_p}(b,m)$, and sample skewness $\widehat{\mathscr{B}}_{\widehat{y}_p}(b,m)$.

- **[a]** Compute the sample lag-one correlation $\widehat{\varphi}_{\widehat{y}_p}(b,m)$ of the warmed-up BQEs and the associated correlation adjustment

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

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

Computing the Point and CI Quantile Estimators (Cont'd)

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

$$\widetilde{y}_p(n') \leftarrow Y_{(\lceil n' p \rceil)} \,. \tag{6}$$

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

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

and define the skewness-adjustment function

$$G(\zeta) = \begin{cases} \zeta \,, & \text{if } |\beta| \le 10^{-3} \,, \\ \dfrac{\sqrt[3]{1 + 6\beta(\zeta - \beta)} - 1}{2\beta} \,, & \text{if } |\beta| > 10^{-3} \,, \end{cases}$$

for all real ζ, where $\sqrt[3]{\zeta} \equiv \operatorname{sign}(\zeta)\,\sqrt[3]{|\zeta|}$.

Computing the Point and CI Quantile Estimators (Cont'd)

[c] Compute the modified sample variance of the warmed-up BQEs,

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

based on the overall quantile point estimator (6).

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

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

and the associated CI,

$$\widetilde{y}_p(n') \;\pm\; H\,. \tag{7}$$

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

Satisfying the Precision Requirement

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

[a] If the half-length H of the current CI (7) satisfies the user-specified precision requirement

$$H \le H^* = \begin{cases} r^* \left| \widetilde{y}_p(n') \right|, & \text{for relative precision level } r^*, \\ h^*, & \text{for absolute precision level } h^*, \end{cases} \tag{8}$$

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

[b] For the fixed batch count b, estimate the batch size m required to satisfy (8),

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

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

First-Order Autoregressive (AR(1)) Process

The table below shows the results of applying Sequest to an AR(1) process with the initial condition $X_0 = 0$, the autoregressive parameter $\rho = 0.995$, steady-state mean $\mu_X = 100$, and steady-state standard deviation $\sigma_X = 10.01$.

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

No CI Precision Requirement							
p	x_p	Avg. $\tilde{y}_p(n')$	$\bar{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\bar{m}$	$\bar{n}$
0.3	94.7494	94.7606	1.4025	1.4801	94.3%	5,139	167,014
0.5	100	100.0280	1.4646	1.4642	94.6%	4,107	133,468
0.7	105.2506	105.2467	1.5552	1.4777	94.9%	3,866	125,638
0.9	112.8316	112.7742	1.7782	1.5768	93.4%	3,912	127,009
0.95	116.4691	116.3653	1.9177	1.6480	94.3%	4,328	140,325
CI Relative Precision = 1.0%							
p	x_p	Avg. $\tilde{y}_p(n')$	$\bar{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\bar{m}$	$\bar{n}$
0.3	94.7494	94.7919	0.8435	0.8899	94.5%	9,829	317,085
0.5	100	100.0254	0.8883	0.8880	95.7%	8,087	260,821
0.7	105.2506	105.2553	0.9352	0.8885	94.4%	7,727	249,201
0.9	112.8316	112.8234	1.0055	0.8912	94.5%	8,859	285,313
0.95	116.4691	116.4423	1.0327	0.8869	94.9%	10,652	342,688

$M/M/1$ Queue-Waiting-Time Process

The table below shows the results of applying Sequest to an $M/M/1$ queueing system with interarrival rate $\lambda = 0.8$, service rate $\omega = 1$, and utilization $\rho = \lambda/\omega = 0.8$.

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

No CI Precision Requirement							
p	x_p	$\widetilde{y}_p(n')$	$\bar{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\bar{m}$	$\bar{n}$
0.3	0.6676	0.6673	0.0645	9.6629	97.0%	9,383	300,439
0.5	2.35	2.3485	0.1582	6.7352	96.5%	7,829	250,762
0.7	4.9041	4.8991	0.2794	5.7038	95.5%	10,880	348,437
0.9	10.3972	10.3559	0.3861	3.7288	93.5%	39,730	1,272,035
0.95	13.8629	13.7678	0.4529	3.2897	93.7%	80,001	2,560,469
CI Relative Precision = 2.5%							
p	x_p	$\widetilde{y}_p(n')$	$\bar{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\bar{m}$	$\bar{n}$
0.3	0.6676	0.6672	0.0150	2.2477	95.5%	93,812	3,002,160
0.5	2.35	2.3488	0.0524	2.2292	95.0%	32,330	1,034,797
0.7	4.9041	4.9022	0.1079	2.2007	95.7%	28,921	925,752
0.9	10.3972	10.3885	0.2107	2.0278	94.6%	54,055	1,730,439
0.95	13.8629	13.8549	0.2642	1.9066	95.5%	95,675	3,062,058

$M/M/1$/LIFO Queue-Waiting-Time Process

The table below shows the results of applying Sequest to an $M/M/1$/LIFO queueing system with interarrival rate $\lambda = 1.0$, service rate $\omega = 1.25$, and utilization $\rho = \lambda/\omega = 0.8$.

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

No CI Precision Requirement							
p	x_p	$\widetilde{y}_p(n')$	$\bar{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\bar{m}$	$\bar{n}$
0.3	0.1129	0.1146	0.0319	27.8575	95.8%	462	14,940
0.5	0.4692	0.4692	0.0625	13.3175	97.2%	433	14,000
0.7	1.3579	1.3581	0.0872	6.4228	97.1%	1,704	54,654
0.9	6.718	6.7200	0.2666	3.9675	95.9%	9,807	313,911
0.95	14.4052	14.3968	0.4710	3.2714	96.0%	24,949	798,473
CI Relative Precision = 2.5%							
p	x_p	$\widetilde{y}_p(n')$	$\bar{H}$	Avg. CI Rel. Prec. (%)	CI Coverage	$\bar{m}$	$\bar{n}$
0.3	0.1129	0.1129	0.0025	2.2423	95.8%	46,588	1,490,978
0.5	0.4692	0.4691	0.0105	2.2353	94.3%	8,420	269,587
0.7	1.3579	1.3580	0.0304	2.2355	95.0%	7,094	227,142
0.9	6.718	6.7165	0.1444	2.1499	95.6%	18,139	580,526
0.95	14.4052	14.3974	0.3002	2.0848	95.6%	35,195	1,126,361

Conclusions

1. In a performance evaluation that includes some problems with characteristics typical of routine applications as well as problems designed to "stress test" the procedure, we have found that Sequest was competitive with previous methods.

2. We are continuing to refine Sequest to improve its execution time, memory requirements, and statistical estimation efficiency relative to its competitors.

3. We have also developed Sequem, an extension of Sequest that incorporates a modification of the maximum transformation (Heidelberger and Lewis 1984) to
    - reduce the sample sizes required for estimating extreme p-quantiles, in particular for $p \in [0.9, 0.995)$, and
    - resolve the CI undercoverage issues.

   We have also found Sequem's performance to be competitive.

Acknowledgment

Thanks to the NSF for grants CMMI–1233141/1232998.