A Sequential Procedure for Estimating Steady-State Quantiles

Christos Alexopoulos\textsuperscript{1}  David Goldsman\textsuperscript{1}  Anup C. Mokashi\textsuperscript{2}
Rong Nie\textsuperscript{3}  Qing Sun\textsuperscript{3}  Kai-Wen Tien\textsuperscript{3}  James R. Wilson\textsuperscript{3}

\textsuperscript{1}Georgia Tech  \textsuperscript{2}SAS Institute  \textsuperscript{3}North Carolina State University

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

November 9, 2014
Outline

1 Introduction
   - Setup for Quantile Estimation
   - Basis for Our Approach

2 Sequest: A Sequential Procedure for Quantile Estimation
   - Main Objectives of the Procedure
   - Main Steps of the Procedure

3 Experimental Performance Evaluation
   - Results for AR(1) Process
   - Results for $M/M/1$ Queue-Waiting-Time Process
   - Results for $M/M/1/LIFO$ Queue-Waiting-Time Process

4 Conclusions
Notation and Assumptions

- We have a simulation output process \( \{X_i : i = 1, 2, \ldots \} \) with steady-state c.d.f. \( F(x) = \Pr\{X_i \leq x\} \) 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 \( x_p \equiv F^{-1}(p) \equiv \inf\{x : F(x) \geq p\} \) based on a simulation-generated time series \( \{X_i : i = 1, \ldots, n\} \) of sufficient length \( n \), where the CI has the form

\[
\tilde{x}_p(n) \pm H, \quad (1)
\]

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

\[
H \leq H^* = \begin{cases} 
  r^* |\tilde{x}_p(n)|, & \text{for relative precision level } r^*, \\
  h^*, & \text{for absolute precision level } h^*. 
\end{cases}
\]
We organize \( \{X_i : i = 1, \ldots, n\} \) into \( b \) nonoverlapping batches of size \( m \) \((n = bm)\). From the \( j \)th batch \( \{X_{(j-1)m+1}, \ldots, X_{jm}\} \), we obtain the order statistics \( X_{j,(1)} \leq X_{j,(2)} \leq \cdots \leq X_{j,(m)} \) and the associated batch quantile estimator (BQE)

\[
\hat{x}_p(j, m) = X_{j, ([mp])} \quad \text{for} \quad j = 1, \ldots, b. \tag{2}
\]

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

\[
\tilde{x}_p(n) = X_{([np])}. \tag{3}
\]
Using (2) and (3), we also compute a modified estimator of the variance of the BQEs,

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

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

\[ \tilde{x}_p(n) \pm H, \text{ where } H = t_{1-\alpha/2,b-1} \tilde{S}_{\hat{x}_p}(b, m) / \sqrt{b}, \]  

where \( t_{q,\nu} \) is the \( q \)-quantile of Student’s \( t \) distribution with \( \nu \) 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 $\{\hat{x}_p(j, m) : j = 1, \ldots, b\}$; 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,

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

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 $\bar{x}_p(b, m)$, 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 $|\bar{x}_p(b, m)|$), 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 $\alpha_{\text{ran}}$, 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 \( j \)th warmed-up batch \( \{Y_{(j-1)m+i} : i = 1, \ldots, m\} \), compute \( j \)th warmed-up BQE \( \hat{\gamma}_p(j, m) \). Compute the sample mean \( \bar{\gamma}_p(b, m) \), sample variance \( S^2_{\bar{\gamma}_p}(b, m) \), and sample skewness \( \hat{B}_{\bar{\gamma}_p}(b, m) \) of the warmed-up BQEs.

[b] If \( \mid \hat{B}_{\bar{\gamma}_p}(b, m) \mid \leq \mathcal{B}^*(p) \) or \( u = u^* \), then go to step [4]; otherwise increase the batch size according to

\[
m \leftarrow \left\lfloor m \cdot \text{mid} \left\{ \sqrt{2}, \left[ \frac{\hat{B}_{\bar{\gamma}_p}(b, m)}{\mathcal{B}^*(p)} \right]^2, 16 \right\} \right\rfloor ,
\]

where \( \text{mid}\{u_1, u_2, u_3\} \equiv u(2) \), 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 $\bar{y}_p(b, m)$, sample variance $S^2_{\hat{y}_p}(b, m)$, and sample skewness $\hat{B}_{\hat{y}_p}(b, m)$.

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

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

that will be applied to the half-length of the CI estimator for $x_p$. 
From the warmed-up time series of length $n'$, compute the order statistics $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n')}$; then compute the overall sectioning-based point estimator $\tilde{y}_p(n')$ of $x_p$ as

$$\tilde{y}_p(n') \leftarrow Y_{([n'p])}.$$  \hfill (6)

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

$$\beta \leftarrow \hat{B}_{\tilde{y}_p}(b, m) / (6\sqrt{b}),$$

and define the skewness-adjustment function

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

for all real $\zeta$, where $\sqrt[3]{\zeta} \equiv \text{sign}(\zeta) \sqrt[3]{|\zeta|}$. 
Compute the modified sample variance of the warmed-up BQEs,

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

based on the overall quantile point estimator (6).

Compute the “half-length” of the bias-, correlation-, and skewness-adjusted 100(1 − α)% CI estimator of \( x_p \),

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

and the associated CI,

\[
\bar{y}_p(n') \pm H. \quad (7)
\]

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

   \[ H \leq H^* = \begin{cases} 
   r^*|\bar{y}_p(n')|, & \text{for relative precision level } r^*, \\
   h^*, & \text{for absolute precision level } h^*, 
   \end{cases} \tag{8} \]

   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.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$x_p$</th>
<th>Avg. $\bar{y}_p(n')$</th>
<th>$\bar{H}$</th>
<th>Avg. CI Rel. Prec. (%)</th>
<th>CI Coverage</th>
<th>$\bar{m}$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>94.7494</td>
<td>94.7606</td>
<td>1.4025</td>
<td>1.4801</td>
<td>94.3%</td>
<td>5,139</td>
<td>167,014</td>
</tr>
<tr>
<td>0.5</td>
<td>100</td>
<td>100.0280</td>
<td>1.4646</td>
<td>1.4642</td>
<td>94.6%</td>
<td>4,107</td>
<td>133,468</td>
</tr>
<tr>
<td>0.7</td>
<td>105.2506</td>
<td>105.2467</td>
<td>1.4777</td>
<td>94.9%</td>
<td>3,866</td>
<td>125,638</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>112.8316</td>
<td>112.7742</td>
<td>1.5768</td>
<td>93.4%</td>
<td>3,912</td>
<td>127,009</td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>116.4691</td>
<td>116.3653</td>
<td>1.6480</td>
<td>94.3%</td>
<td>4,328</td>
<td>140,325</td>
<td></td>
</tr>
</tbody>
</table>

CI Relative Precision = 1.0%

<table>
<thead>
<tr>
<th>$p$</th>
<th>$x_p$</th>
<th>Avg. $\bar{y}_p(n')$</th>
<th>$\bar{H}$</th>
<th>Avg. CI Rel. Prec. (%)</th>
<th>CI Coverage</th>
<th>$\bar{m}$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>94.7494</td>
<td>94.7919</td>
<td>0.8435</td>
<td>0.8899</td>
<td>94.5%</td>
<td>9,829</td>
<td>317,085</td>
</tr>
<tr>
<td>0.5</td>
<td>100</td>
<td>100.0254</td>
<td>0.8883</td>
<td>0.8880</td>
<td>95.7%</td>
<td>8,087</td>
<td>260,821</td>
</tr>
<tr>
<td>0.7</td>
<td>105.2506</td>
<td>105.2553</td>
<td>0.9352</td>
<td>0.8885</td>
<td>94.4%</td>
<td>7,727</td>
<td>249,201</td>
</tr>
<tr>
<td>0.9</td>
<td>112.8316</td>
<td>112.8234</td>
<td>1.0055</td>
<td>0.8912</td>
<td>94.5%</td>
<td>8,859</td>
<td>285,313</td>
</tr>
<tr>
<td>0.95</td>
<td>116.4691</td>
<td>116.4423</td>
<td>1.0327</td>
<td>0.8869</td>
<td>94.9%</td>
<td>10,652</td>
<td>342,688</td>
</tr>
</tbody>
</table>
$M/M/1$ 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.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$x_p$</th>
<th>$\tilde{y}_p(n')$</th>
<th>$\tilde{H}$</th>
<th>Avg. CI Rel. Prec. (%)</th>
<th>CI Coverage</th>
<th>$\bar{m}$</th>
<th>$\bar{n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.6676</td>
<td>0.6673</td>
<td>0.0645</td>
<td>9.6629</td>
<td>97.0%</td>
<td>9,383</td>
<td>300,439</td>
</tr>
<tr>
<td>0.5</td>
<td>2.35</td>
<td>2.3485</td>
<td>0.1582</td>
<td>6.7352</td>
<td>96.5%</td>
<td>7,829</td>
<td>250,762</td>
</tr>
<tr>
<td>0.7</td>
<td>4.9041</td>
<td>4.8991</td>
<td>0.2794</td>
<td>5.7038</td>
<td>95.5%</td>
<td>10,880</td>
<td>348,437</td>
</tr>
<tr>
<td>0.9</td>
<td>10.3972</td>
<td>10.3559</td>
<td>0.3861</td>
<td>3.7288</td>
<td>93.5%</td>
<td>39,730</td>
<td>1,272,035</td>
</tr>
<tr>
<td>0.95</td>
<td>13.8629</td>
<td>13.7678</td>
<td>0.4529</td>
<td>3.2897</td>
<td>93.7%</td>
<td>80,001</td>
<td>2,560,469</td>
</tr>
</tbody>
</table>

CI Relative Precision $= 2.5\%$

<table>
<thead>
<tr>
<th>$p$</th>
<th>$x_p$</th>
<th>$\tilde{y}_p(n')$</th>
<th>$\tilde{H}$</th>
<th>Avg. CI Rel. Prec. (%)</th>
<th>CI Coverage</th>
<th>$\bar{m}$</th>
<th>$\bar{n}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0.6676</td>
<td>0.6672</td>
<td>0.0150</td>
<td>2.2477</td>
<td>95.5%</td>
<td>93,812</td>
<td>3,002,160</td>
</tr>
<tr>
<td>0.5</td>
<td>2.35</td>
<td>2.3488</td>
<td>0.0524</td>
<td>2.2292</td>
<td>95.0%</td>
<td>32,330</td>
<td>1,034,797</td>
</tr>
<tr>
<td>0.7</td>
<td>4.9041</td>
<td>4.9022</td>
<td>0.1079</td>
<td>2.2007</td>
<td>95.7%</td>
<td>28,921</td>
<td>925,752</td>
</tr>
<tr>
<td>0.9</td>
<td>10.3972</td>
<td>10.3885</td>
<td>0.2107</td>
<td>2.0278</td>
<td>94.6%</td>
<td>54,055</td>
<td>1,730,439</td>
</tr>
<tr>
<td>0.95</td>
<td>13.8629</td>
<td>13.8549</td>
<td>0.2642</td>
<td>1.9066</td>
<td>95.5%</td>
<td>95,675</td>
<td>3,062,058</td>
</tr>
</tbody>
</table>

Alexopoulos et al. Sequential Steady-State Quantile Estimation November 9, 2014 15 / 18
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.
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.