Confidence Intervals for Quantiles Using Sectioning When Applying Variance-Reduction Techniques

MARVIN K. NAKAYAMA, New Jersey Institute of Technology

We develop confidence intervals (CIs) for quantiles when applying variance-reduction techniques (VRTs) and sectioning. Similar to batching, sectioning partitions the independent and identically distributed (i.i.d.) outputs into nonoverlapping batches and computes a quantile estimator from each batch. But rather than centering the CI at the average of the quantile estimators across the batches, as in batching, sectioning centers the CI at the overall quantile estimator based on all the outputs. A similar modification is made to the sample variance, which is used to determine the width of the CI. We establish the asymptotic validity of the sectioning CI for importance sampling and control variates, and the proofs rely on first showing that the corresponding quantile estimators satisfy a Bahadur representation, which we have done in prior work. Here, we present some numerical results.

Categories and Subject Descriptors: I.6.6 [**Simulation and Modeling**]: Simulation Output Analysis; G.3 [**Probability and Statistics**]: Nonparametric statistics; G.3 [**Probability and Statistics**]: Probabilistic algorithms (including Monte Carlo)

General Terms: Performance, Theory

Additional Key Words and Phrases: control variates, importance sampling, quantile, value-at-risk, variance reduction

ACM Reference Format:
Marvin K. Nakayama. 2014. Confidence intervals for quantiles using sectioning when applying variance-reduction techniques. ACM Trans. Model. Comput. Simul. 24, 4, Article 19 (March 2014), 21 pages.
DOI: http://dx.doi.org/10.1145/2558328

1. INTRODUCTION

For a continuous random variable X having a strictly increasing cumulative distribution function (CDF) F, we define the p-quantile, $0 < p < 1$, as the constant ξ_p such that $F(\xi_p) = p$. Equivalently, $\xi_p = F^{-1}(p)$. For example, the median is the 0.5-quantile $\xi_{0.5}$.

In practice, the p-quantile with $p \approx 0$ or $p \approx 1$ is commonly used to assess risk. For example, in finance, where a quantile is also called a value-at-risk, the 0.99-quantile is often used to measure portfolio risk (e.g., see Duffie and Pan [1997]). The U.S. Nuclear Regulatory Commission (NRC) allows nuclear plant licensees to demonstrate their facilities are in compliance using a *95/95 criterion*. This requires showing that, with 95% confidence, the 0.95-quantile of a certain output variable lies below a mandated

This material is based on work supported in part by the National Science Foundation under Grant Numbers CMMI-0926949, CMMI-1200065, and DMS-1331010. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Author's address: M. K. Nakayama, Department of Computer Science, GITC Building, New Jersey Institute of Technology, Newark, NJ 07102.

DOI: http://dx.doi.org/10.1145/2558328

threshold [U.S. Nuclear Regulatory Commission 1989]. Thus, not only do we want a point estimate for a quantile, but there is also a need for a confidence interval (CI) for a quantile to provide a measure of the statistical error of the estimate. Until now, the simulations for the 95/95 analyses of nuclear power plants have only used crude Monte Carlo (CMC; i.e., without the use of any variance reduction; see p. 75 of the report of the Nuclear Energy Agency Committee on the Safety of Nuclear Installations [2007]).

Simulation-based estimation of quantiles typically entails first running the simulation model n times to get n outputs, which are then used to construct an estimator of the CDF F, and then inverting it to obtain a quantile estimator. There has been substantial work on variance-reduction techniques (VRTs) to provide point estimators for quantiles, and the VRTs can significantly increase statistical efficiency, especially when estimating extreme quantiles (i.e., when $p \approx 0$ or 1). VRTs for quantile estimation include importance sampling (IS) [Glynn 1996; Glasserman et al. 2000; Sun and Hong 2010], control variates (CV) [Hsu and Nelson 1990; Hesterberg and Nelson 1998], and correlation-induction methods, such as antithetic variates (AV) and Latin hypercube sampling (LHS) [Avramidis and Wilson 1998; Jin et al. 2003].

The construction of a CI for a quantile when applying VRTs has received much less attention. One approach for obtaining a CI with CMC is based on the binomial distribution (see Section 2.6.1 of Serfling [1980]), but this approach is not valid when using VRTs. However, Hsu and Nelson [1990] are able to generalize the binomial CI to a multinomial one for the case of control variates.

Chu and Nakayama [2012] develop a general approach to construct a CI for a quantile when applying a VRT, and they show the framework includes IS, combined IS and stratification, CV, and AV. Nakayama [2011b] demonstrates that replicated LHS also fits in this setting. The method of Chu and Nakayama [2012] utilizes a finite difference to estimate $1/f(\xi_p)$, which appears in the variance constant of the quantile estimator's central limit theorem (CLT), where f denotes the density function corresponding to F. Although they prove the asymptotic validity of their method, the method may not perform well in practice for small sample sizes n and extreme p.

Subsequently, other methods for constructing a CI for a quantile have been developed for IS. These include those by Liu and Yang [2012], who use the bootstrap, and by Nakayama [2011a], who applies a kernel method [Wand and Jones 1995].

In this article, we construct CIs for a quantile using a method known as *sectioning* when applying a VRT. Asmussen and Glynn [2007] (Section III.5a) develop sectioning for the case of CMC, and we now extend it to VRTs. Similar to batching, sectioning divides the n outputs into $b \geq 2$ batches, each of size $m = n/b$. From each batch, a quantile estimator is formed. For batching, we construct a CI by computing the sample average and variance of the b batch quantile estimators. In contrast, sectioning replaces the sample average with the overall quantile estimator from all n observations. Because quantile estimators are biased, with the bias converging to 0 as the sample size grows large, sectioning has the advantage that the CI is centered at a less-biased point estimator than the batching point estimator, whose bias is determined by the batch size $m < n$. This appears to lead to better coverage for sectioning, and we present some numerical results that support this claim.

The rest of this article is organized as follows: Section 2 reviews the use of CMC to estimate a quantile and construct CIs for it. We discuss in Section 3 estimating a quantile using VRTs, including IS and CV. Section 4 presents numerical results, and concluding remarks are in Section 5. All of the proofs are collected in an appendix. The material on the IS sectioning method has appeared previously without proof in Nakayama [2012], which also provides numerical results for a smaller stochastic model than that considered here.

2. QUANTILE ESTIMATION USING CMC

Consider a real-valued random variable X having CDF F. For a fixed $0 < p < 1$, define the p-quantile of F (or equivalently of X) as $\xi_p = F^{-1}(p) \equiv \inf\{x : F(x) \geq p\}$. Assume that F is differentiable at ξ_p and that $f(\xi_p) > 0$, where $f(\xi_p) = \frac{d}{dx}F(x)|_{x=\xi_p}$.

With CMC, we run our simulation model for n i.i.d. replications, resulting in i.i.d. outputs $X_1, X_2, \ldots, X_n$, each having CDF F. We then estimate the CDF via the empirical CDF F_n, defined by

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq y), \tag{1}$$

where $I(\,\cdot\,)$ denotes the indicator function, which assumes the value 1 (0, respectively) when its argument is true (false, respectively). The CMC p-quantile estimator is then $\xi_{p,n} = F_n^{-1}(p)$.

Several different techniques have been developed to construct a confidence interval for ξ_p when applying CMC. One approach exploits the fact that $nF_n(\xi_p)$ follows a binomial(n, p) distribution (e.g., see Section 2.6.1 of Serfling [1980]).

Another method first shows that when $f(\xi_p) > 0$, the CMC p-quantile estimator $\xi_{p,n}$ satisfies a central limit theorem (Section 2.3.3 of Serfling [1980])

$$\frac{\sqrt{n}}{\kappa_p}(\xi_{p,n} - \xi_p) \Rightarrow N(0, 1), \tag{2}$$

as $n \to \infty$, where $\kappa_p = \sqrt{p(1-p)}/f(\xi_p)$, $\Rightarrow$ denotes convergence in distribution (e.g., Section 1.2.4 of [Serfling 1980]), and $N(a, b^2)$ is a normal random variable with mean a and variance b^2. If we have a consistent estimator $\kappa_{p,n}$ of κ_p, then we can unfold the CLT to obtain an approximate-100$(1-\alpha)$% CI for ξ_p as

$$C_n' \equiv \left(\xi_{p,n} \pm z_\alpha \frac{\kappa_{p,n}}{\sqrt{n}}\right), \tag{3}$$

where $z_\alpha = \Phi^{-1}(1-\alpha/2)$ and Φ is the CDF of a $N(0, 1)$. The CI is asymptotically valid in the sense that its coverage $P(\xi_p \in C_n') \to 1-\alpha$ as $n \to \infty$.

A complication of this approach is that it is nontrivial to construct a consistent estimator $\kappa_{p,n}$ of κ_p. Some techniques for accomplishing this have been developed in the statistics literature, including using a finite difference [Bloch and Gastwirth 1968] and kernel methods [Falk 1986]. To describe the first approach, note that $\lambda_p \equiv 1/f(\xi_p) = \frac{d}{dp}F^{-1}(p) = \lim_{h\to 0}[F^{-1}(p+h) - F^{-1}(p-h)]/(2h)$, which is known as the *sparsity function* [Tukey 1965] or the *quantile density* [Parzen 1979]. This suggests the finite-difference estimator

$$\lambda_{p,n} = \frac{F_n^{-1}(p+h_n) - F_n^{-1}(p-h_n)}{2h_n}, \tag{4}$$

where $h_n > 0$ is a user-specified *bandwidth*. When $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, then $\lambda_{p,n} \Rightarrow \lambda_p$ as $n \to \infty$.

One issue with quantile estimators is that they are generally biased. To explain the bias, it helps to express the CMC p-quantile estimator as $\xi_{p,n} = F_n^{-1}(p) = X_{(\lceil np \rceil)}$, where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the order statistics of the n i.i.d. outputs and $\lceil\,\cdot\,\rceil$ is the ceiling function. Under certain conditions, Proposition 2 of Avramidis and Wilson [1998] establishes that

$$\text{Bias}[\xi_{p,n}] \equiv E[\xi_{p,n}] - \xi_p = \frac{1}{n}\left(\frac{\lceil np \rceil - np - p}{f(\xi_p)} - \frac{p(1-p)f'(\xi_p)}{2(f(\xi_p))^3}\right) + o(1/n),$$

where f' denotes the derivative of f and a function $g(n) = o(h(n))$ means that $g(n)/h(n) \to 0$ as $n \to \infty$. Thus, whereas the bias approaches 0 as $n \to \infty$, the convergence may not be monotonic because of the rounding up in the ceiling function. (One alternative quantile estimator comes from inverting a linearly interpolated version of F_n, and Avramidis and Wilson [1998] show that its bias differs from the preceding displayed equation in that the numerator in the first term in the large parentheses is instead $0.5 - p$.) Resampling schemes [Efron 1979] can sometimes be used to estimate or reduce the bias of certain estimators. But the lack of smoothness of the quantile estimator $\xi_{p,n}$ as a function of the outputs can cause issues in applying the jackknife (e.g., the jackknife quantile variance estimator is not consistent), although the problem can be remedied by modifying the jackknife to use all subsamples in which $d > 1$ outputs are deleted (instead of just $d = 1$ as in the standard jackknife), where $d \to \infty$ at a certain rate as $n \to \infty$ (e.g., see Section 2.3 of Shao and Tu [1995]).

Rather than trying to consistently estimate κ_p to construct a CI for ξ_p, we can instead apply a *cancellation method* [Glynn and Iglehart 1990], which cancels out κ_p in the relevant limit theorem in a manner analogous to the Student t-statistic. (To illustrate the cancellation in a Student t-statistic, suppose that $Z_1, Z_2, \ldots, Z_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables, and let $\bar{Z}_n = (1/n)\sum_{i=1}^{n} Z_i$ and $S_n^2 = (1/(n-1))\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2$. Then $\bar{Z}_n \sim N(\mu, \sigma^2/n)$ and $S_n \sim \sigma\sqrt{\chi^2_{n-1}/(n-1)}$, where $\sim$ means "is distributed as" and χ^2_d denotes a chi-squared random variable with d degrees of freedom (df). Hence, $T_n \equiv \sqrt{n}(\bar{Z}_n - \mu)/S_n$ has a Student t-distribution with $n-1$ df, which does not depend on σ since it was canceled out in the ratio defining T_n.) One cancellation approach is *batching*. Here, we divide the n outputs in $b \geq 2$ nonoverlapping batches, each of size $m = n/b$, which we assume is an integer. The jth batch, $j = 1, 2, \ldots, b$, consists of the outputs X_i for $(j-1)m+1 \leq i \leq jm$, and from these outputs, we construct an estimator $F_{m,j}$ of the CDF F as

$$F_{m,j}(y) = \frac{1}{m}\sum_{i=(j-1)m+1}^{jm} I(X_i \leq y). \tag{5}$$

Then compute the p-quantile estimator $\xi_{p,m,j} = F_{m,j}^{-1}(p)$ from the jth batch. We next compute the sample average

$$\bar{\xi}_{p,m,b} = \frac{1}{b}\sum_{j=1}^{b} \xi_{p,m,j} \tag{6}$$

of the b batch quantile estimators, which are i.i.d., and also their sample variance

$$S^2_{m,b,\text{batch}} = \frac{1}{b-1}\sum_{j=1}^{b}(\xi_{p,m,j} - \bar{\xi}_{p,m,b})^2.$$

The CMC batching approximate-$100(1-\alpha)$% CI for ξ_p is given by

$$\left(\bar{\xi}_{p,m,b} \pm t_{b-1,\alpha}\frac{S_{m,b,\text{batch}}}{\sqrt{b}}\right),$$

where $t_{b-1,\alpha}$ is the upper-$\alpha/2$ critical point of a Student t random variable T with $b-1$ df; if G is the CDF of T, then $t_{b-1,\alpha} = G^{-1}(1-\alpha/2)$.

In the context of applying batch means to construct a CI for a steady-state mean, Schmeiser [1982] recommends choosing $10 \leq b \leq 30$, which we follow. For a fixed total sample size n, the relatively small value for b leads to a larger batch size $m = n/b$.

Because the asymptotic validity of batching rests on the approximate normality of the batch estimators, larger m helps ensure this by virtue of a CLT.

The bias of quantile estimators can be problematic for the batching CI. Although the bias approaches 0 as the sample size n gets large, the bias can be significant for small n. The bias of the batching point estimator $\bar{\xi}_{p,m,b}$ is determined by the batch size m, which is smaller than the total number n of outputs. Hence, the batching CI is centered at a point that may have large bias when n is small, resulting in poor coverage.

Asmussen and Glynn [2007] (Section III.5a) develop *sectioning* as an alternative approach to construct a CI for ξ_p. Similar to batching, sectioning again works with the same $b \geq 2$ batches, each of size $m = n/b$, where the batches are sometimes instead called *sections*. But sectioning replaces the batching point estimator $\bar{\xi}_{p,m,b}$ with the overall quantile estimator $\xi_{p,n} = F_n^{-1}(p)$ throughout the formula for the batching CI. Specifically, let

$$S^2_{m,b,\mathrm{sect}} = \frac{1}{b-1}\sum_{j=1}^{b}(\xi_{p,m,j} - \xi_{p,n})^2,$$

which differs from the batching sample variance $S^2_{m,b,\mathrm{batch}}$ by instead subtracting the overall point estimate $\xi_{p,n}$. The CMC sectioning approximate-100$(1-\alpha)$% CI for ξ_p is then

$$\left(\xi_{p,n} \pm t_{b-1,\alpha}\frac{S_{m,b,\mathrm{sect}}}{\sqrt{b}}\right).$$

Note that the sectioning CI is centered at the overall quantile estimator $\xi_{p,n}$ instead of the batching point estimator $\bar{\xi}_{p,m,b}$. This should lead to better coverage since $\xi_{p,n}$ will typically be less biased than $\bar{\xi}_{p,m,b}$.

The asymptotic validity of the CMC sectioning CI can be established by first showing that the quantile estimator $\xi_{p,n}$ satisfies a *Bahadur representation*, for which we now provide a heuristic derivation. Because $F_n \approx F$ for large n, it is plausible that $\xi_{p,n} = F_n^{-1}(p) \approx F^{-1}(p) = \xi_p$. Thus, since $F(\xi_p) = p$, we see that

$$p \approx F(\xi_{p,n}) \approx F(\xi_p) + f(\xi_p)(\xi_{p,n} - \xi_p) \approx F_n(\xi_p) + f(\xi_p)(\xi_{p,n} - \xi_p),$$

where the second approximation follows from a Taylor expansion and the last since $F_n \approx F$. Rearranging terms then yields $\xi_{p,n} \approx \xi_p + [p - F_n(\xi_p)]/f(\xi_p)$.

We can make this rigorous by replacing the approximation with an equality and introducing an error term:

$$\xi_{p,n} = \xi_p + \frac{p - F_n(\xi_p)}{f(\xi_p)} + R'_n, \tag{7}$$

where R'_n vanishes at some rate as $n \to \infty$. Assuming that $f(\xi_p) > 0$ and the second derivative of F is bounded in a neighborhood of ξ_p, Bahadur [1966] first established an almost sure (a.s.) rate for the error term:

$$R'_n = O(n^{-3/4}(\log n)^{1/2}(\log\log n)^{1/4}) \text{ a.s.}, \tag{8}$$

as $n \to \infty$, where "$Y_n = O(g(n))$ a.s." means there is an event Ω_0 for which $P(\Omega_0) = 1$ and for each $\omega \in \Omega_0$, there exists a constant $B(\omega)$ such that $|Y_n(\omega)| \leq B(\omega)g(n)$ for all sufficiently large n. Ghosh [1971] proved a weaker form of the error rate under less stringent conditions: if $f(\xi_p) > 0$, then

$$\sqrt{n}R'_n \Rightarrow 0 \tag{9}$$

as $n \to \infty$. We call (7) and (8) ((9), respectively) a *strong* (*weak*, respectively) Bahadur representation. (Note that (8) implies (9).) In either case, a Bahadur representation shows that a quantile estimator can be approximated as a linear transformation of a CDF estimator. The weak form suffices for most applications, including ours. For example, the Bahadur representation ensures that $\xi_{p,n}$ satisfies the CLT in (2). To see why, rearrange terms in (7) and scale by $\sqrt{n}$ to get $\sqrt{n}(\xi_{p,n}-\xi_p) = \sqrt{n}[p-F_n(\xi_p)]/f(\xi_p)+\sqrt{n}R'_n$. Because $F_n(\xi_p)$ is the sample average of i.i.d. indicator functions, the first term on the right side converges weakly to $N(0, p(1-p)/f^2(\xi_p))$ as $n \to \infty$. The second term vanishes as $n \to \infty$ by (9), so Slutsky's theorem (p. 19 of Serfling [1980]) implies (2). A Bahadur representation thus gives insight into why a quantile estimator, which is *not* a sample average, obeys a CLT.

A (weak) Bahadur representation also provides the key that allows replacing the batching point estimator $\bar{\xi}_{p,m,b}$ with the overall point estimator $\xi_{p,n}$ to obtain the sectioning CI from the batching CI. To see why, first note that the linearity of the CDF estimator ensures that for equal-sized batches, the average of the batch CDF estimators (5) equals the overall CDF estimator (1). Each batch quantile estimator satisfies a Bahadur representation, so $\bar{\xi}_{p,m,b}$ can be approximated by the average of linear transformations of the batch CDF estimators. But the latter is just a linear transformation of the overall CDF estimator, which is roughly the overall quantile estimator, again by a Bahadur representation. Hence, $\bar{\xi}_{p,m,b} \approx \xi_{p,n}$, although the two are not equal in general, and (9) guarantees the approximation is fine enough to permit replacing $\bar{\xi}_{p,m,b}$ with $\xi_{p,n}$ in the batching CI. The proofs in the Appendix provide the full details when using variance reduction.

3. CONFIDENCE INTERVALS WHEN APPLYING VARIANCE REDUCTION

We now want to develop CIs for ξ_p when applying VRTs. We have a total computation budget n, which represents the number of outputs generated. Let $\hat{F}_n$ be the estimator of the CDF F from all n outputs, and the specific form of $\hat{F}_n$ depends on the VRT applied. The overall VRT p-quantile estimator is $\hat{\xi}_{p,n} = \hat{F}_n^{-1}(p)$.

Chu and Nakayama [2012] establish a set of conditions on $\hat{F}_n$ under which a VRT p-quantile estimator satisfies a weak Bahadur representation,

$$\hat{\xi}_{p,n} = \xi_p + \frac{p-\hat{F}_n(\xi_p)}{f(\xi_p)} + R_n, \text{ with } \sqrt{n}R_n \Rightarrow 0, \tag{10}$$

as $n \to \infty$. (They also prove a *perturbed* weak Bahadur representation for the p_n-quantile estimator, where $p_n \to p$ as $n \to \infty$, which they use to show the consistency of a VRT finite-difference estimator of the quantile density $\lambda_p = 1/f(\xi_p)$.) Their general framework covers a broad class of VRTs, including (under appropriate moment conditions) IS, stratified sampling (SS), combined IS+SS, AV, and CV. Nakayama [2011b] shows that replicated LHS also satisfies the conditions of Chu and Nakayama [2012].

As with CMC, the VRT weak Bahadur representation in (10) implies that $\hat{\xi}_{p,n}$ satisfies the CLT

$$\frac{\sqrt{n}}{\kappa_p}(\hat{\xi}_{p,n} - \xi_p) \Rightarrow N(0,1) \tag{11}$$

as $n \to \infty$, where

$$\kappa_p = \frac{\psi_p}{f(\xi_p)}. \tag{12}$$

The value of ψ_p depends on the VRT applied, but the denominator $f(\xi_p)$ is independent of the VRT. For the VRTs mentioned in the previous paragraph, Chu and Nakayama [2012] and Nakayama [2011b] develop consistent estimators $\hat{\psi}_{p,n}$ of ψ_p. Also, they show that the finite difference

$$\hat{\lambda}_{p,n} = \frac{\hat{F}_n^{-1}(p+h_n) - \hat{F}_n^{-1}(p-h_n)}{2h_n} \tag{13}$$

consistently estimates $\lambda_p = 1/f(\xi_p)$ as $n \to \infty$ for bandwidth $h_n \to 0$ and $\sqrt{n}h_n \to b \in (0,\infty]$ when f is continuous at ξ_p and a (perturbed) weak Bahadur representation holds for the VRT p_n-quantile estimator with $p_n \to p$ as $n \to \infty$. Hence, $\hat{\kappa}_{p,n} = \hat{\psi}_{p,n}\hat{\lambda}_{p,n}$ consistently estimates κ_p, leading to the VRT finite-difference approximate-100$(1-\alpha)$% CI for ξ_p as

$$\left(\hat{\xi}_{p,n} \pm z_\alpha \frac{\hat{\kappa}_{p,n}}{\sqrt{n}}\right),$$

which is asymptotically valid as $n \to \infty$.

To apply batching or sectioning to construct a CI, we divide the n outputs into $b \geq 2$ batches, each of size $m = n/b$. From each batch j, let $\hat{F}_{m,j}$ be the CDF estimator constructed from the outputs in the jth batch, and compute the jth batch's quantile estimator $\hat{\xi}_{p,m,j} = \hat{F}_{m,j}^{-1}(p)$. Compute the batching point estimator

$$\bar{\xi}_{p,m,b} = \frac{1}{b}\sum_{j=1}^{b} \hat{\xi}_{p,m,j} \tag{14}$$

and the batching sample variance

$$\hat{S}^2_{m,b,\text{batch}} = \frac{1}{b-1}\sum_{j=1}^{b}\left(\hat{\xi}_{p,m,j} - \bar{\xi}_{p,m,b}\right)^2. \tag{15}$$

The VRT batching approximate-100$(1-\alpha)$% CI for ξ_p is then

$$C_{m,b,\text{batch}} = \left(\bar{\xi}_{p,m,b} \pm t_{b-1,\alpha}\frac{\hat{S}_{m,b,\text{batch}}}{\sqrt{b}}\right). \tag{16}$$

For sectioning, we replace the batching point estimator $\bar{\xi}_{p,m,b}$ with the overall quantile estimator $\hat{\xi}_{p,n}$ throughout the formula for the batching CI. Specifically, let

$$\hat{S}^2_{m,b,\text{sect}} = \frac{1}{b-1}\sum_{j=1}^{b}\left(\hat{\xi}_{p,m,j} - \hat{\xi}_{p,n}\right)^2, \tag{17}$$

and the VRT sectioning approximate-100$(1-\alpha)$% CI for ξ_p is

$$C_{m,b,\text{sect}} = \left(\hat{\xi}_{p,n} \pm t_{b-1,\alpha}\frac{\hat{S}_{m,b,\text{sect}}}{\sqrt{b}}\right). \tag{18}$$

We also define another CI that combines sectioning and batching by centering the CI at the overall point estimator $\hat{\xi}_{p,n}$ as in sectioning, but using the batching half-width. Specifically, the VRT combined sectioning-batching approximate-100$(1-\alpha)$% CI for ξ_p is

$$C_{m,b,\text{sb}} = \left(\hat{\xi}_{p,n} \pm t_{b-1,\alpha}\frac{\hat{S}_{m,b,\text{batch}}}{\sqrt{b}}\right). \tag{19}$$

When the batches are i.i.d., the asymptotic validity of the batching, sectioning, and combined sectioning-batching CIs can be established by exploiting the weak Bahadur representation in (10).

The following subsections examine some specific VRTs. To simplify notation, we continue to use the same notation ($\hat{F}_n$ for the VRT estimator of the CDF F, $\hat{\xi}_{p,n}$ for the VRT estimator of the p-quantile ξ_p, etc.) rather than introduce new variables for each special case.

3.1. Importance Sampling

We now describe how to apply the VRT IS to estimate a quantile, as developed by Glynn [1996].

Let F_* be another CDF such that F is absolutely continuous with respect to F_*; that is, $\int_A dF(x) = 0$ for every set $A \subset \Re$ such that $\int_A dF_*(x) = 0$. By the Radon-Nikodym Theorem [Billingsley 1995, Theorem 32.2], this property ensures the existence of a "density" $L(\cdot)$ such that

$$F(x) = \int_{-\infty}^{x} L(u)\, dF_*(u)$$

for all $x \in \Re$. (When F and F_* have respective densities f and f_*, then $L(u) = f(u)/f_*(u)$.) Let E_* (P_*, respectively) denote the expectation operator (probability measure, repectively) when X has CDF F_*. By Theorem 16.11 of Billingsley [1995], for every (integrable) function $h(X)$, we have

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)\, dF(x) = \int_{-\infty}^{\infty} h(x)L(x)\, dF_*(x) = E_*[h(X)L(X)].$$

Then we can express the original CDF F as

$$\begin{aligned} F(y) &= 1 - E[I(X > y)] = 1 - \int I(x > y)\, F(dx) = 1 - \int I(x > y)\, L(x)\, F_*(dx) \\ &= 1 - E_*[I(X > y)\, L(X)]. \end{aligned}$$

This suggests that we can estimate $F(y)$ via IS by generating i.i.d. outputs $X_1, X_2, \ldots, X_n$ using F_* and then forming an IS estimator of $F(y)$ as

$$\hat{F}_n(y) = 1 - \frac{1}{n}\sum_{i=1}^{n} I(X_i > y)L(X_i). \tag{20}$$

Inverting this leads to an IS p-quantile estimator

$$\hat{\xi}_{p,n} = \hat{F}_n^{-1}(p). \tag{21}$$

Glynn [1996] develops another IS p-quantile estimator by inverting an alternative CDF estimator

$$\hat{F}'_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le y)L(X_i),$$

which is based on the fact that $F(y) = E[I(X \le y)] = E_*[I(X \le y)L(X)]$. We can invert $\hat{F}'_n$ to obtain the corresponding p-quantile estimator $\hat{\xi}'_{p,n} = \hat{F}'^{-1}_n(p)$. As noted by Glynn [1996], the IS p-quantile estimator $\hat{\xi}_{p,n}$ from (21) is more appropriate when $p \approx 1$, and the other $\hat{\xi}'_{p,n}$ should be used for $p \approx 0$. To simplify the discussion, we focus on the first estimator, but our results also hold for the second (under appropriate modifications).

Glynn [1996] proves that if $f(\xi_p) > 0$ and $E_*[L^3(X)] < \infty$, then the IS p-quantile estimator $\hat{\xi}_{p,n}$ in (21) obeys the CLT in (11), with

$$\psi_p^2 = E_*[I(X > \xi_p)L^2(X)] - (1-p)^2 \tag{22}$$

in (12), which Chu and Nakayama [2012] prove can be consistently estimated by $(1/n)\sum_{i=1}^{n} I(X_i > \hat{\xi}_{p,n})L^2(X_i) - (1-p)^2$. Chu and Nakayama [2012] further relax the third-moment condition to

$$E_*[I(X > \xi_p - \delta)L^{2+\epsilon}(X)] < \infty \text{ for some } \epsilon > 0 \text{ and } \delta > 0 \tag{23}$$

and show that it also suffices to establish that $\hat{\xi}_{p,n}$ satisfies a (perturbed) weak Bahadur representation. For the IS p-quantile estimator $\hat{\xi}'_{p,n}$, the moment condition for the CLT and (perturbed) weak Bahadur representation is $E_*[I(X < \xi_p + \delta)L^{2+\epsilon}(X)] < \infty$ for some $\epsilon > 0$ and $\delta > 0$. Sun and Hong [2010] prove that the IS p-quantile estimator $\hat{\xi}'_{p,n}$ satisfies a strong Bahadur representation for fixed p under more stringent conditions.

For batching and sectioning with IS, we partition the n i.i.d. output pairs (X_i, L_i), $1 \le i \le n$, where $L_i = L(X_i)$, into $b \ge 2$ nonoverlapping batches, each of size $m = n/b$. Thus, for each $j = 1, 2, \ldots, b$, the outputs (X_i, L_i), $(j-1)m + 1 \le i \le jm$, make up the jth batch. For each batch j, we compute an estimator of the CDF F as

$$\hat{F}_{m,j}(y) = 1 - \frac{1}{m} \sum_{i=(j-1)m+1}^{jm} I(X_i > y)L_i, \tag{24}$$

which we invert to obtain the jth IS batch quantile estimator

$$\hat{\xi}_{p,m,j} = \hat{F}_{m,j}^{-1}(p). \tag{25}$$

We can then form the IS CIs (16), (18), and (19) for ξ_p using batching, sectioning, and combined sectioning-batching, respectively, by substituting (21) based on (20) for $\hat{\xi}_{p,n}$ and (25) based on (24) for $\hat{\xi}_{p,m,j}$. The following result establishes the asymptotic validity of these IS CIs for ξ_p. (Choosing $F_* \equiv F$ reduces IS to CMC, so the theorem also covers CMC as a special case.)

THEOREM 3.1. *Suppose $f(\xi_p) > 0$ and (23) holds. Then, when applying IS,*

$$P_*(\xi_p \in C) \to 1 - \alpha$$

as $m \to \infty$ with b fixed for $C = C_{m,b,sect}$, $C_{m,b,batch}$, or $C_{m,b,sb}$ in (16), (18), and (19), respectively.

3.2. Control Variates

Hsu and Nelson [1990] and Hesterberg and Nelson [1998] develop methods for applying CV to estimate quantiles. For simplicity, we start by describing the case of a single control, then later expand to multiple controls. Let Q be another output random variable that is correlated with X, and suppose we know the mean $\nu = E[Q]$. For example, in a G/G/1 queue, Q may represent the sum Y of the first k customers interarrival times, so $\nu = k\eta$, where η is the (known) mean of the interarrival time distribution. Another possibility for the CV when the interarrival times are i.i.d. exponential is $Q = I(Y \le y_p)$, where $y_p = G^{-1}(p)$ and G is the CDF of an Erlang-k random variable.

Running the simulation n i.i.d. times results in outputs (X_i, Q_i), $i = 1, 2, \ldots, n$, as i.i.d. replicates of (X, Q), where X has CDF F. For any constant β, we define an estimate of the CDF F of X as

$$\hat{F}_{n,\beta}(y) = F_n(y) - \beta(\bar{Q}_n - \nu),$$

where F_n is from (1) and $\bar{Q}_n = (1/n)\sum_{i=1}^{n} Q_i$. Note that $\hat{F}_{n,\beta}(y)$ is an unbiased estimator of $F(y)$ for any constant β. The choice of β that minimizes the variance of $\hat{F}_{n,\beta}(y)$ is $\beta_*(y) = \text{Cov}[I(X \leq y), Q]/\text{Var}[Q]$ (e.g., see p. 138 of Asmussen and Glynn [2007]), which is typically unknown and must be estimated. (Cov and Var denote the covariance and variance operators, respectively.) A consistent estimator of $\beta_*(y)$ is

$$\hat{\beta}_n(y) = \frac{[(1/n)\sum_{i=1}^{n} I(X_i \leq y)Q_i] - F_n(y)\bar{Q}_n}{(1/n)\sum_{j=1}^{n}(Q_j - \bar{Q}_n)^2}.$$

Substituting this into $\hat{F}_{n,\beta}$ then yields the CV estimator of the CDF F as

$$\hat{F}_n(y) = F_n(y) - \hat{\beta}_n(y)(\bar{Q}_n - \nu),$$

which is typically no longer an unbiased estimator of $F(y)$ because of the dependence between $\hat{\beta}_n(y)$ and $\bar{Q}_n$. We invert $\hat{F}_n$ to obtain the CV p-quantile estimator

$$\hat{\xi}_{p,n} = \hat{F}_n^{-1}(p).$$

Hesterberg and Nelson [1998] discuss how to efficiently invert $\hat{F}_n$. Chu and Nakayama [2012] prove that if $f(\xi_p) > 0$ and $0 < \text{Var}[Q] < \infty$, then $\hat{\xi}_{p,n}$ satisfies the weak Bahadur representation in (10), as well as the perturbed version. In (12), $\psi_p^2 = p(1 - p) - (\text{Cov}[I(X \leq \xi_p), Q])^2/\text{Var}[Q]$ for CV, which can be consistently estimated by $p(1 - p) - [(1/n)\sum_{i=1}^{n} I(X_i \leq \hat{\xi}_{p,n})Q_i - F_n(\hat{\xi}_{p,n})\bar{Q}_n]^2/[(1/n)\sum_{j=1}^{n}(Q_j - \bar{Q}_n)^2]$.

We now describe how to extend CV for quantile estimation to exploit multiple controls [Hesterberg and Nelson 1998]. Suppose $\boldsymbol{Q} = (Q^{(1)}, Q^{(2)}, \ldots, Q^{(r)})^\top$ is a vector of r random variables (i.e., controls) correlated with X and with known mean $\boldsymbol{\nu} = (\nu^{(1)}, \nu^{(2)}, \ldots, \nu^{(r)})^\top$, where superscript $\top$ denotes transpose. Assume that the $r \times r$ covariance matrix $\Sigma_{\boldsymbol{Q}}$ of $\boldsymbol{Q}$ is nonsingular. Define the r-vector $\Sigma_{\boldsymbol{Q},X}(y) = (\text{Cov}[Q^{(1)}, I(X \leq y)], \text{Cov}[Q^{(2)}, I(X \leq y)], \ldots, \text{Cov}[Q^{(r)}, I(X \leq y)])^\top$. Let $\boldsymbol{Q}_i = (Q_i^{(1)}, Q_i^{(2)}, \ldots, Q_i^{(r)})^\top$ be the r-vector of controls on the ith replication. Then, the multiple-CV estimator of the CDF F from n i.i.d. replicates $(X_i, \boldsymbol{Q}_i)$, $i = 1, 2, \ldots, n$, is defined as

$$\hat{F}_n(y) = F_n(y) - \hat{\boldsymbol{\beta}}_n(y)^\top(\bar{\boldsymbol{Q}}_n - \boldsymbol{\nu}), \tag{26}$$

where F_n is from (1), $\bar{\boldsymbol{Q}}_n = (1/n)\sum_{i=1}^{n} \boldsymbol{Q}_i$, the r-vector $\hat{\boldsymbol{\beta}}_n(y) = \hat{\Sigma}_{\boldsymbol{Q},n}^{-1}\hat{\Sigma}_{\boldsymbol{Q},X,n}(y)$ is an estimator of the optimal multiplier $\boldsymbol{\beta}_*(y) = \Sigma_{\boldsymbol{Q}}^{-1}\Sigma_{\boldsymbol{Q},X}(y)$, the $r \times r$ matrix $\hat{\Sigma}_{\boldsymbol{Q},n} = (\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Q}_i\boldsymbol{Q}_i^\top) - \bar{\boldsymbol{Q}}_n\bar{\boldsymbol{Q}}_n^\top$ estimates $\Sigma_{\boldsymbol{Q}}$, and the r-vector $\hat{\Sigma}_{\boldsymbol{Q},X,n}(y) = [\frac{1}{n}\sum_{i=1}^{n} I(X_i \leq y)\boldsymbol{Q}_i] - F_n(y)\bar{\boldsymbol{Q}}_n$ estimates $\Sigma_{\boldsymbol{Q},X}(y)$. The multiple-CV p-quantile estimator is then $\hat{\xi}_{p,n} = \hat{F}_n^{-1}(p)$, which Hesterberg and Nelson [1998] describe how to compute efficiently.

When applying multiple controls with batching or sectioning with b batches, each of size $m = n/b$, we define for each batch $j = 1, 2, \ldots, b$, the CDF estimator

$$\hat{F}_{m,j}(y) = F_{m,j}(y) - \hat{\boldsymbol{\beta}}_{m,j}(y)^\top(\bar{\boldsymbol{Q}}_{m,j} - \boldsymbol{\nu}), \tag{27}$$

where $F_{m,j}$ is defined in (5), $\bar{\boldsymbol{Q}}_{m,j} = (1/m)\sum_{i=(j-1)m+1}^{jm} \boldsymbol{Q}_i$, the r-vector $\hat{\boldsymbol{\beta}}_{m,j}(y) = \tilde{\Sigma}_{\boldsymbol{Q},m,j}^{-1}\tilde{\Sigma}_{\boldsymbol{Q},X,m,j}(y)$, the $r \times r$ matrix $\tilde{\Sigma}_{\boldsymbol{Q},m,j} = (\frac{1}{m}\sum_{i=(j-1)m+1}^{jm} \boldsymbol{Q}_i\boldsymbol{Q}_i^\top) - \bar{\boldsymbol{Q}}_{m,j}\bar{\boldsymbol{Q}}_{m,j}^\top$, and the r-vector $\tilde{\Sigma}_{\boldsymbol{Q},X,m,j}(y) = [\frac{1}{m}\sum_{i=(j-1)m+1}^{jm} I(X_i \leq y)\boldsymbol{Q}_i] - F_{m,j}(y)\bar{\boldsymbol{Q}}_{m,j}$. We then obtain the

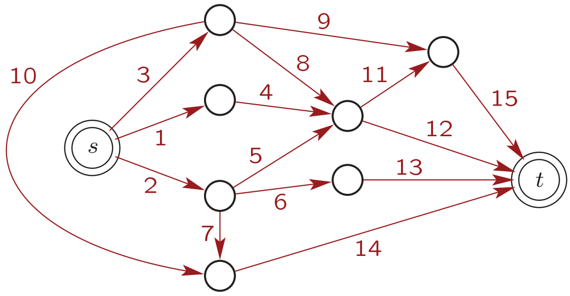

Fig. 1. A stochastic activity network.

multiple-CV batch-j p-quantile estimator $\hat{\xi}_{p,m,j} = \hat{F}^{-1}_{m,j}(p)$. We subsequently form the multiple-CV CI for ξ_p using batching by substituting the multiple-CV batch-j quantile estimator $\hat{\xi}_{p,m,j}$ into (14)–(16). Moreover, the multiple-CV sectioning and combined sectioning-batching CIs come from substituting the multiple-CV p-quantile estimator $\hat{\xi}_{p,n}$ into (18) and (19), respectively. The following result establishes the asymptotic validity of these multiple-CV CIs for ξ_p, which include the single-CV intervals as special cases ($r = 1$).

THEOREM 3.2. *Suppose $f(\xi_p) > 0$ and the covariance matrix $\Sigma_{\boldsymbol{Q}}$ of the r-vector $\boldsymbol{Q}$ of controls, $r \geq 1$, is nonsingular. Then, when applying CV,*

$$P(\xi_p \in C) \to 1 - \alpha$$

as $m \to \infty$ with b fixed for $C = C_{m,b,\mathrm{sect}}$, $C_{m,b,\mathrm{batch}}$ or $C_{m,b,\mathrm{sb}}$ in (16), (18), *and* (19), *respectively.*

We can also define other CV batching and sectioning CIs by replacing $\hat{\boldsymbol{\beta}}_{m,j}(y)$ with $\hat{\boldsymbol{\beta}}_n(y)$ in (27). Specifically, for each batch j, define another CDF estimator $\hat{F}'_{m,j}(y) = F_{m,j}(y) - \hat{\boldsymbol{\beta}}_n(y)^\top(\bar{\boldsymbol{Q}}_{m,j} - \boldsymbol{\nu})$ and the corresponding p-quantile estimator $\hat{\xi}'_{p,m,j} = \hat{F}'^{-1}_{m,j}(p)$. Then we obtain new batching, sectioning, and combined sectioning-batching CIs (16), (18), and (19), respectively, by replacing each $\hat{\xi}_{p,m,j}$ with $\hat{\xi}'_{p,m,j}$ in (14), (15), and (17). The resulting CIs are also asymptotically valid, which can be established by slightly modifying the proof of Theorem 3.2. Because it is based on a larger sample size, $\hat{\boldsymbol{\beta}}_n(y)$ is generally a better estimator of $\boldsymbol{\beta}_*(y)$ than each $\hat{\boldsymbol{\beta}}_{m,j}(y)$, but using $\hat{\boldsymbol{\beta}}_n(y)$ in constructing the multiple-CV batch quantile estimators $\hat{\xi}'_{p,m,j}$, $1 \leq j \leq b$ also induces dependence among them, whereas the $\hat{\xi}_{p,m,j}$, $1 \leq j \leq b$, are independent.

4. NUMERICAL RESULTS

We ran numerical experiments on a stochastic activity network (SAN) previously studied in Juneja et al. [2007] and Chu and Nakayama [2012]. (Nakayama [2012] provides empirical results with IS for a smaller SAN.) Also known as a stochastic PERT, a SAN models the time to complete a project consisting of a collection of activities with precedence constraints and random durations [Adlakha and Kulkarni 1989]. We consider the SAN in Figure 1, which has $d = 15$ activities, corresponding to the directed edges in the network. Each edge ℓ has a random length A_ℓ, which is the time to complete activity ℓ. We assume A_ℓ is exponentially distributed with mean 2 for $1 \leq \ell \leq 8$, and mean 1 for $9 \leq \ell \leq 15$. All the A_ℓ are mutually independent. There are $q = 10$ paths through the network, and we let B_j denote the set of activities on the jth path. We have $B_1 = \{1, 4, 11, 15\}$, $B_2 = \{1, 4, 12\}$, $B_3 = \{2, 5, 11, 15\}$, $B_4 = \{2, 5, 12\}$, $B_5 = \{2, 6, 13\}$,

Table I. Coverages (with Average Half-Widths) and Relative Bias for CMC

			$b = 10$			$b = 20$				Relative Bias (%)		
n	FD	Kernel	Batch	Section	SB	Batch	Section	SB	Exact	$b = 1$	$b = 10$	$b = 20$
					$p = 0.8$							
100	0.900 (1.029)	0.860 (0.877)	0.661 (0.885)	0.890 (0.966)	0.859 (0.885)	0.247 (0.770)	0.873 (0.886)	0.820 (0.770)	0.907 (0.934)	−0.49	−5.13	−9.45
400	0.884 (0.461)	0.891 (0.453)	0.841 (0.485)	0.904 (0.500)	0.894 (0.485)	0.665 (0.450)	0.896 (0.470)	0.880 (0.450)	0.907 (0.467)	−0.13	−1.36	−2.69
1,600	0.876 (0.226)	0.889 (0.229)	0.878 (0.248)	0.898 (0.251)	0.894 (0.248)	0.834 (0.236)	0.897 (0.239)	0.892 (0.236)	0.898 (0.233)	−0.03	−0.34	−0.68
6,400	0.899 (0.115)	0.904 (0.115)	0.905 (0.125)	0.908 (0.126)	0.906 (0.125)	0.888 (0.119)	0.902 (0.120)	0.901 (0.119)	0.909 (0.117)	−0.01	−0.09	−0.18
					$p = 0.95$							
100	0.949 (2.663)	0.707 (1.362)	0.856 (1.676)	0.865 (1.724)	0.852 (1.676)	0.170 (1.166)	0.787 (1.367)	0.714 (1.166)	0.907 (1.741)	−1.58	−1.77	−12.71
400	0.900 (0.928)	0.813 (0.844)	0.679 (0.842)	0.893 (0.915)	0.862 (0.842)	0.265 (0.731)	0.876 (0.837)	0.821 (0.731)	0.902 (0.871)	−0.35	−3.63	−6.71
1,600	0.891 (0.443)	0.856 (0.433)	0.833 (0.457)	0.897 (0.471)	0.889 (0.457)	0.676 (0.425)	0.890 (0.443)	0.876 (0.425)	0.898 (0.435)	−0.10	−0.98	−1.92
6,400	0.897 (0.219)	0.882 (0.218)	0.882 (0.235)	0.901 (0.238)	0.897 (0.235)	0.841 (0.223)	0.901 (0.226)	0.898 (0.223)	0.900 (0.218)	−0.03	−0.25	−0.49
					$p = 0.99$							
100	0.508 (2.063)	0.525 (1.653)	0.042 (1.676)	0.700 (2.555)	0.540 (1.676)	0.000 (1.166)	0.658 (2.231)	0.399 (1.166)	0.941 (3.686)	−5.04	−21.17	−29.96
400	0.925 (2.649)	0.734 (1.425)	0.740 (1.614)	0.842 (1.697)	0.822 (1.614)	0.060 (1.118)	0.782 (1.423)	0.678 (1.118)	0.912 (1.843)	−1.32	−4.12	−12.57
1,600	0.980 (1.541)	0.834 (0.856)	0.898 (1.020)	0.906 (1.047)	0.898 (1.020)	0.711 (1.091)	0.943 (1.153)	0.930 (1.091)	0.898 (0.922)	−0.35	0.55	4.13
6,400	0.939 (0.540)	0.878 (0.453)	0.894 (0.499)	0.898 (0.506)	0.893 (0.499)	0.779 (0.461)	0.892 (0.472)	0.881 (0.461)	0.897 (0.461)	−0.10	−0.23	−1.15

$B_6 = \{2, 7, 14\}$, $B_7 = \{3, 8, 11, 15\}$, $B_8 = \{3, 8, 12\}$, $B_9 = \{3, 9, 15\}$, and $B_{10} = \{3, 10, 14\}$. Let $X = \max_{1 \le j \le 10} \sum_{\ell \in B_j} A_\ell$, which is the length of the longest path from s to t in the SAN and corresponds to the time to complete the project. Let F denote the CDF of X. We are interested in estimating the p-quantile ξ_p of F and constructing CIs for ξ_p for different values of p when applying various simulation methods (CMC, CV, and IS). The goal of the experiments is to study the *coverage* of different nominal 90% CIs, where we estimated the coverage as the proportion of the constructed CIs that contain the "true value" of ξ_p from 10^4 independent experiments. The asymptotic validity of our CIs ensures that the coverage converges to the nomimal level 0.9 as the sample size n grows large, but it can be off for small n.

Because computing the CDF F is extremely tedious, we estimated the "true value" of ξ_p and $\lambda_p = 1/f(\xi_p)$ using a single CMC simulation with $n = 5 \times 10^7$. This gave $\xi_{0.8} = 11.7655$, $\xi_{0.95} = 15.3478$, $\xi_{0.99} = 19.1259$, $\xi_{0.999} = 24.28996$, $\lambda_{0.8} = 14.1895$, $\lambda_{0.95} = 48.572$, $\lambda_{0.99} = 225.225$, and $\lambda_{0.999} = 2236.371$. The "true value" of λ_p was estimated using the finite-difference estimator (4) with bandwidth $h_n = 0.5n^{-1/2}$.

For CMC, we generate n i.i.d. replicates of X, where, in each independent replication, we independently generate $A_1, A_2, \ldots, A_{15}$ having their original exponential distributions and then compute X from the lengths A_ℓ obtained. Table I contains the results for CMC. The first column gives the sample size n. The next several columns present the coverage (and average half-widths) of nominal 90% CIs constructed using different methods. The column labeled FD is for the CI in (3) with $\lambda_p = 1/f(\xi_p)$ from $\kappa_p = \sqrt{p(1-p)}/f(\xi_p)$ estimated via the finite-difference estimator (4) of Bloch and Gastwirth [1968] with bandwidth $h_n = 0.5n^{-1/2}$. As noted by Chu and Nakayama [2012], when $p \approx 1$ and n is small, we may have $p + h_n > 1$, which would lead to F_n^{-1} in the first term in the numerator of (4) being evaluated outside its domain $(0, 1]$. This issue must be addressed in some way, and we make the following adjustments, which Chu and Nakayama [2012] also applied. When $p + h_n > 1$, we replace $p + h_n$, $p - h_n$, and $2h_n$ in the FD with $1 - (1-p)/10$, $2p - 1 + (1-p)/10$, and $9(1-p)/5$, respectively, the second chosen so the two arguments to F_n^{-1} in (4) are symmetric about p, and the last is the difference of the first two. The next column contains results for the CI in (3) with a plug-in kernel estimator [Wand and Jones 1995] for $f(\xi_p)$. The plug-in kernel estimator is given by $\hat{f}_n(\xi_{p,n})$, where $\hat{f}_n$ is the kernel density estimator

$$\hat{f}_n(y) = (1/n)\sum_{i=1}^{n} \frac{1}{h_n} k\left(\frac{y - X_i}{h_n}\right),$$

k is the Gaussian kernel (i.e., the density function of a $N(0, 1)$), and the bandwidth $h_n = cn^{-v}$ with $c = 2$ and $v = 1/5$. (We also ran additional experiments [not shown] with $c = 0.5$ and 1, $v = 1/2$ and $1/3$, and the uniform and Epanechnikov kernels [Wand and Jones 1995], but $c = 2$, $v = 1/5$ and the Gaussian kernel generally gave the best results. A "good" choice for h_n depends on the distribution F of the outputs and the level p of the quantile, and determining an appropriate bandwidth in practice can be tricky, which is a drawback of the kernel method, as well as FD.) The columns labeled "Batch," "Section," and "SB" are for batching, sectioning, and combined sectioning-batching, respectively, with either $b = 10$ or $b = 20$ batches. The column with heading "Exact" is for the CI in (3) with the "true value" of λ_p. The last three columns of Table I give the point estimators' percent relative biases, estimated from the 10^4 independent experiments. The column for $b = 1$ is for the overall point estimator $\xi_{p,n}$ (i.e., only a single batch), so the percent relative bias is $100(E[\xi_{p,n}] - \xi_p)/\xi_p$, which uses the "true value" of ξ_p. The columns for $b = 10$ and $b = 20$ are for the batching point estimator (6) with the corresponding number of batches.

As n gets large, the coverage for all methods approaches the nominal level 0.9, but the small-sample performance is not as good, especially for $p = 0.99$. In general, sectioning appears to lead to better coverage than the other methods (except "Exact," which is not implementable in practice since it requires knowledge of the value of the unknown λ_p). Batching does extremely poorly for $p = 0.99$ and $n = 100$; this is likely due to the significant bias (see the last three columns of Table I) of the batching point estimator caused by the small batch size $m = n/b$. The coverage and bias for batching with $b = 20$ are much worse than when $b = 10$ because of the former's larger bias. For sectioning and SB, $b = 20$ is somewhat inferior to $b = 10$, with the deterioration increasing as p grows for small n.

We also experimented with CVs using multiple controls. We wanted to choose CVs that are strongly correlated with $I(X \le \xi_p)$. We attempt to do this by using all paths B_j with the largest expected length. There are $r = 3$ paths, B_1, B_3, B_7, that are tied with the largest expectation, which is 6, and let $Y^{(1)} = \sum_{\ell \in B_1} A_\ell$, $Y^{(2)} = \sum_{\ell \in B_3} A_\ell$, and $Y^{(3)} = \sum_{\ell \in B_7} A_\ell$. For each $j = 1, 2, 3$, $Y^{(j)}$ is the sum of four independent (but not identically distributed) exponentials, two with mean 2 and two with unit mean, so all the $Y^{(j)}$ have the same CDF G, which is worked out in Chu and Nakayama [2012]. We then define the CVs $Q^{(j)} = I(Y^{(j)} \le G^{-1}(p))$, $j = 1, 2, 3$. Table II gives the results for CV. For the FD column, we applied the approach of Chu and Nakayama [2012], in which

Table II. Coverages (Average Half-Widths) and Relative Bias for CV
The FD column contains results when applying the estimator from Chu and Nakayama [2012].

		$b = 10$			$b = 20$				Relative Bias (%)		
n	FD	Batch	Section	SB	Batch	Section	SB	Exact	$b = 1$	$b = 10$	$b = 20$
					$p = 0.8$						
100	0.852 (0.795)	0.645 (0.990)	0.955 (1.122)	0.922 (0.990)	0.027 (0.787)	0.968 (1.133)	0.866 (0.787)	0.886 (0.795)	0.12	−5.87	−16.82
400	0.879 (0.402)	0.902 (0.480)	0.926 (0.492)	0.920 (0.480)	0.895 (0.506)	0.948 (0.515)	0.944 (0.506)	0.899 (0.404)	0.03	0.34	−0.03
1,600	0.886 (0.201)	0.896 (0.222)	0.904 (0.225)	0.901 (0.222)	0.895 (0.218)	0.909 (0.219)	0.906 (0.218)	0.898 (0.203)	0.00	0.07	0.15
6,400	0.900 (0.101)	0.907 (0.110)	0.908 (0.110)	0.907 (0.110)	0.906 (0.105)	0.909 (0.106)	0.909 (0.105)	0.909 (0.110)	0.00	0.01	0.03
					$p = 0.95$						
100	0.884 (2.093)	0.075 (1.162)	0.925 (2.054)	0.707 (1.162)	0.000 (0.793)	0.901 (1.799)	0.558 (0.793)	0.820 (1.422)	0.41	−16.92	−25.81
400	0.900 (0.823)	0.839 (1.020)	0.949 (1.075)	0.931 (1.020)	0.124 (0.754)	0.950 (0.962)	0.868 (0.754)	0.886 (0.750)	0.12	−1.18	−9.19
1,600	0.894 (0.387)	0.902 (0.450)	0.926 (0.460)	0.918 (0.450)	0.894 (0.504)	0.947 (0.513)	0.944 (0.504)	0.894 (0.378)	0.01	0.25	0.59
6,400	0.896 (0.191)	0.894 (0.210)	0.903 (0.212)	0.898 (0.210)	0.897 (0.205)	0.908 (0.206)	0.905 (0.205)	0.895 (0.190)	0.00	0.05	0.10
					$p = 0.99$						
100	0.284 (1.087)	0.001 (1.176)	0.696 (2.863)	0.348 (1.176)	0.000 (0.881)	0.634 (2.385)	0.260 (0.881)	0.409 (1.498)	−5.41	−27.45	−34.43
400	0.853 (2.140)	0.424 (1.344)	0.853 (1.733)	0.731 (1.344)	0.000 (0.771)	0.871 (1.804)	0.498 (0.771)	0.815 (1.518)	0.46	−7.58	−20.79
1,600	0.974 (1.321)	0.825 (1.080)	0.945 (1.144)	0.925 (1.080)	0.346 (0.857)	0.934 (0.984)	0.887 (0.857)	0.886 (0.800)	0.09	−1.17	−5.58
6,400	0.940 (0.479)	0.899 (0.498)	0.927 (0.510)	0.920 (0.498)	0.889 (0.518)	0.945 (0.526)	0.940 (0.518)	0.896 (0.404)	0.01	0.20	−0.01

the FD estimator (13) uses the CV CDF estimator (26). Again, sectioning generally seems to have better coverage than the other implementable methods for small n.

We apply IS using a method from Chu and Nakayama [2012], which combines ideas from Juneja et al. [2007] and Glynn [1996]. Juneja et al. [2007] examine estimating a tail probability of the longest path in a SAN, and they define the IS CDF F_* as a mixture of q distributions. Each distribution in the mixture exponentially tilts the activities along one path and leaves the other edges with their original distributions, where all edges are independent. The details on how to determine the mixture weights are given in Chu and Nakayama [2012], and we choose the value for the tilting parameter using an approach from Glynn [1996]. Full details are available in Chu and Nakayama [2012].

Table III presents the results for IS. We chose more extreme values for p than in the previous tables since IS is more suited to use for extreme quantiles, and CMC and CV do not work well for $p \approx 1$. For FD, we apply the approach of Chu and Nakayama [2012], in which the FD estimator (13) uses the IS CDF estimator (20). The column headed "Kernel" uses a plug-in kernel estimator of $f(\xi_p)$ developed in Nakayama [2011a], with a Gaussian kernel and bandwidth $h_n = 2n^{-1/5}$. As before, sectioning appears to provide better coverage than the other implementable methods for small n, especially as p approaches 1.

Table III. Coverages (with Average Half-Widths) and Relative Bias for IS
The FD and Kernel columns contain results when applying the estimators from Chu and Nakayama [2012] and Nakayama [2011a], respectively.

			$b = 10$			$b = 20$				Relative Bias (%)		
n	FD	Kernel	Batch	Section	SB	Batch	Section	SB	Exact	$b = 1$	$b = 10$	$b = 20$
						$p = 0.95$						
100	0.983 (1.232)	0.823 (0.719)	0.851 (0.972)	0.932 (1.028)	0.922 (0.972)	0.644 (1.005)	0.953 (1.065)	0.943 (1.005)	0.876 (0.777)	−0.27	−2.74	−5.28
400	0.923 (0.439)	0.878 (0.389)	0.886 (0.453)	0.913 (0.467)	0.904 (0.453)	0.822 (0.444)	0.923 (0.457)	0.915 (0.444)	0.897 (0.402)	−0.09	−0.68	−1.34
1,600	0.902 (0.207)	0.890 (0.200)	0.900 (0.222)	0.910 (0.225)	0.904 (0.222)	0.880 (0.214)	0.910 (0.216)	0.906 (0.214)	0.897 (0.203)	−0.01	−0.16	−0.33
6,400	0.899 (0.102)	0.895 (0.101)	0.901 (0.111)	0.906 (0.112)	0.904 (0.111)	0.895 (0.106)	0.901 (0.107)	0.898 (0.106)	0.899 (0.102)	−0.01	−0.05	−0.09
						$p = 0.99$						
100	0.980 (1.432)	0.794 (0.795)	0.790 (1.259)	0.954 (1.362)	0.945 (1.259)	0.394 (1.330)	0.977 (1.469)	0.969 (1.330)	0.862 (0.876)	−0.34	−3.96	−8.03
400	0.987 (0.752)	0.864 (0.445)	0.875 (0.543)	0.924 (0.564)	0.913 (0.543)	0.764 (0.549)	0.936 (0.571)	0.928 (0.549)	0.889 (0.460)	−0.06	−0.84	−1.78
1,600	0.991 (0.381)	0.887 (0.231)	0.901 (0.261)	0.917 (0.266)	0.910 (0.261)	0.869 (0.252)	0.915 (0.257)	0.911 (0.252)	0.894 (0.232)	0.01	−0.18	−0.39
6,400	0.944 (0.138)	0.892 (0.117)	0.903 (0.129)	0.905 (0.131)	0.902 (0.129)	0.901 (0.124)	0.905 (0.125)	0.903 (0.124)	0.889 (0.117)	0.02	−0.02	−0.07
						$p = 0.999$						
100	0.972 (1.616)	0.768 (0.868)	0.707 (1.674)	0.972 (1.861)	0.962 (1.674)	0.186 (1.785)	0.988 (2.067)	0.980 (1.785)	0.858 (1.014)	−0.39	−5.30	−10.82
400	0.987 (0.869)	0.852 (0.505)	0.864 (0.653)	0.928 (0.684)	0.919 (0.653)	0.662 (0.693)	0.953 (0.733)	0.944 (0.693)	0.893 (0.544)	−0.08	−1.02	−2.26
1,600	0.993 (0.441)	0.885 (0.264)	0.896 (0.304)	0.915 (0.310)	0.909 (0.304)	0.847 (0.298)	0.921 (0.304)	0.915 (0.298)	0.906 (0.276)	−0.01	−0.22	−0.48
6,400	0.993 (0.222)	0.894 (0.134)	0.901 (0.149)	0.905 (0.151)	0.900 (0.149)	0.892 (0.144)	0.907 (0.145)	0.905 (0.144)	0.904 (0.138)	0.00	−0.05	−0.11

In terms of the amount of variance reduction obtained, CV in our examples somewhat decreases the average half-widths (AHWs) of the CIs compared to CMC when p is not too extreme and n is large. The shrinkage is not extremely large because $I(X \leq \xi_p)$ is not very highly correlated with any linear transformation of the controls $\boldsymbol{Q}^{(j)} = I(Y^{(j)} \leq G^{-1}(p))$, $j = 1, 2, 3$. This is especially the case when $p = 0.99$. Also, when the sample size or batch size is small, CV may not reduce variance, and this may be due to the noise in the estimate of the optimal multiplier $\boldsymbol{\beta}_*(y) = \Sigma_{\boldsymbol{Q}}^{-1}\Sigma_{\boldsymbol{Q},X}(y)$.

On the other hand, IS significantly reduces variance, with approximately a factor of 2 (4, respectively) decrease in the AHW compared to CMC for $p = 0.95$ ($p = 0.99$, respectively). For $p = 0.999$, IS leads to even more variance reduction, with the AHW reduced by about a factor of 20 compared to CMC (not shown).

In all our tables, sectioning always has larger AHWs than batching and SB, which results in sectioning having the highest coverage of the three methods. To see why this seems to occur, note that the batching variance estimator (15) is the ordinary sample variance of the batch quantile estimates, and the batching point estimator $\tilde{\xi}_{p,m,b}$ has the same bias as each batch's estimator $\hat{\xi}_{p,m,j}$. But in the sectioning variance estimator (17), the overall point estimator $\hat{\xi}_{p,n}$, which is subtracted from each batch's estimate instead of the batching point estimate, is typically less biased than each $\hat{\xi}_{p,m,j}$. This contributes to $\hat{S}^2_{m,b,\mathrm{sect}}$ being inflated compared to $\hat{S}^2_{m,b,\mathrm{batch}}$, which is reflected in sectioning having larger AHWs than batching and SB. The discrepancy is most pronounced when n is small or $p \approx 1$, especially for CMC and CV. Combining this with sectioning centering the CI at a less-biased point estimator leads to the increased coverage for sectioning over batching. Sectioning's and SB's CIs are centered at the same point, but the wider AHW of sectioning seems to result in higher coverage.

When applying IS, sectioning and SB appear to lead to overcoverage for small n, whereas batching has undercoverage. It is arguably better to have overcoverage than undercoverage, especially when evaluating risk, where being overly confident can have disasterous effects. Although sectioning and SB may not always have overcoverage with all stochastic models, for our SAN example, the apparent overcoverage may be explained as follows. The IS quantile estimators seem to have small bias—for the same p and n, the IS quantile estimators are usually less biased than those for CMC and CV. Also, when n is large, the AHWs for sectioning and SB are only slightly larger than that for the Exact column, where the relative increase roughly corresponds to the difference in the Student t critical point with $b - 1$ df and the normal critical point. But when n is small, AHW for sectioning and SB are much larger than that for Exact, which may indicate that the sectioning and batching variance estimates are "too large" in this case. Combining this with the small bias for the IS overall point estimator appears to lead to the overcoverage for sectioning and SB when n is small.

5. CONCLUSIONS

Sectioning is a cancellation method for constructing a CI for a quantile. Originally developed in Asmussen and Glynn [2007] for CMC, the technique is extended here to VRTs. The approach is similar to batching but with a key difference being that the batching point estimator is replaced by the overall point estimator. Because quantile estimators are biased, with the bias converging to 0 as the sample size increases, the sectioning CI is centered at a less-biased point than the batching point estimator, whose bias is determined by the batch size $m = n/b$. This seems to lead to sectioning having better coverage, which the numerical results confirm. The asymptotic validity of sectioning follows from the VRT quantile estimator satisfying a weak Bahadur representation and the independence of the batches. Our experiments seem to indicate choosing a small number b of batches works better than larger b, especially when $p \approx 1$, and we recommend setting $b = 10$. Although increasing b can lead to slightly smaller CIs (because of the larger number of df) when n is large, it also can hurt the coverage for smaller n because the asymptotic validity of sectioning (and batching) relies on a CLT to ensure each batch's point estimator is approximately normal.

Sectioning and other CI methods for quantiles require the user to specify parameters to implement the techniques. But sectioning's parameter b seems easier to choose than the bandwidth h_n for FD and kernel estimators. Also, our numerical experiments indicate that the FD and the kernel approaches can require larger sample sizes for the CI coverage to be close to the nominal level. But asymptotically the FD and kernel methods will have slight narrower CIs because they use a normal critical point instead of one from a Student t-distribution, which sectioning and batching require. (This difference is small for $b \geq 10$.) Liu and Yang [2012] consider a bootstrap variance estimator when applying IS to estimate a quantile, but Hall and Martin [1988] show that, for the CMC setting, the bootstrap estimator converges at a slower rate than the FD and kernel estimators. Another well-known approach for CMC [Serfling 1980, Section 2.6.1] exploits the fact that $nF_n(\xi_p)$ has a binomial (n, p) distribution, which can be used to construct an *exact* CI for ξ_p with finite n. However, this approach does not seem to extend to general VRTs. (An exception is CV, for which Hsu and Nelson [1990] develop an asymptotic multinomial CI.)

It would be interesting to investigate applying sectioning to construct a CI for a steady-state quantile when applying a VRT. Munoz [2010] establishes the asymptotic validity of the combined sectioning-batching method (although he calls it simply batching) for quantiles of the steady-state distribution of a function of a Markov chain. He requires that the estimator of the steady-state CDF satisfies a functional central limit theorem (FCLT) and the quantile estimator satisfies a Bahadur representation. To extend the proof of Munoz [2010] to incorporate CV for steady-state quantile estimation, one would need to assume his two conditions hold when applying CV.

APPENDIX

A.1. Proof of Theorem 3.1

We only consider the case when $C = C_{m,s,\mathrm{sect}}$, because the other two cases are similar. Chu and Nakayama [2012] show that $f(\xi_p) > 0$ and (23) ensure the weak Bahadur representation in (10) holds for IS. Thus, each batch j's IS quantile estimator also satisfies a weak Bahadur representation, which we write as

$$\hat{\xi}_{p,m,j} - \xi_p = Z_{m,j} + R_{m,j} \quad \text{with} \quad \sqrt{m}R_{m,j} \Rightarrow 0, \tag{28}$$

as $m = n/b \to \infty$ with fixed $b \geq 2$, where

$$Z_{m,j} = \frac{p - \hat{F}_{m,j}(\xi_p)}{f(\xi_p)}. \tag{29}$$

In (10), we can express

$$\begin{aligned} \frac{p - \hat{F}_n(\xi_p)}{f(\xi_p)} &= \frac{1}{f(\xi_p)}\left[p - \left(1 - \frac{1}{n}\sum_{i=1}^{n} I(X_i > \xi_p)L_i\right)\right] \\ &= \frac{1}{b}\sum_{j=1}^{b} \frac{1}{f(\xi_p)}\left[p - \left(1 - \frac{1}{m}\sum_{i=(j-1)m+1}^{jm} I(X_i > \xi_p)L_i\right)\right] \\ &= \frac{1}{b}\sum_{j=1}^{b} Z_{m,j} \equiv \bar{Z}_{m,b}; \end{aligned} \tag{30}$$

that is, the average of the linear transformation (29) of the batch CDF estimators equals the linear transformation of the overall CDF estimator. We then rewrite (10) as

$$\hat{\xi}_{p,n} - \xi_p = \bar{Z}_{m,b} + R_{bm} \quad \text{with} \quad \sqrt{bm}R_{bm} \Rightarrow 0, \tag{31}$$

as $n = bm \to \infty$ with b fixed. Furthermore, Chu and Nakayama [2012] show that $f(\xi_p) > 0$ and (23) ensure

$$\sqrt{m}Z_{m,j} \Rightarrow N\big(0, \kappa_p^2\big), \tag{32}$$

as $m \to \infty$, where κ_p^2 is defined in (12) and (22). Thus, (28) and (31) imply

$$\begin{aligned}
A_{m,b} &\equiv \frac{\sqrt{b}(\hat{\xi}_{p,n} - \xi_p)}{\hat{S}_{m,b,\mathrm{sect}}} = \frac{\sqrt{b}(\hat{\xi}_{p,n} - \xi_p)}{\left[\frac{1}{b-1}\sum_{j=1}^{b}\left((\hat{\xi}_{p,m,j} - \xi_p) - (\hat{\xi}_{p,n} - \xi_p)\right)^2\right]^{1/2}} \\
&= \left(\frac{\sqrt{m}}{\sqrt{m}}\right)\frac{\sqrt{b}\left(\bar{Z}_{m,b} + R_{bm}\right)}{\left[\frac{1}{b-1}\sum_{j=1}^{b}\left(Z_{m,j} - \bar{Z}_{m,b} + R_{m,j} - R_{bm}\right)^2\right]^{1/2}} \\
&= \left[\sqrt{b}\left(\sqrt{m}\bar{Z}_{m,b} + \sqrt{m}R_{bm}\right)\right]\Bigg[\frac{1}{(b-1)}\sum_{j=1}^{b}\Big(\left(\sqrt{m}Z_{m,j} - \sqrt{m}\bar{Z}_{m,b}\right)^2 \\
&\qquad + 2\left(\sqrt{m}Z_{m,j} - \sqrt{m}\bar{Z}_{m,b}\right)\left(\sqrt{m}R_{m,j} - \sqrt{m}R_{bm}\right) + \left(\sqrt{m}R_{m,j} - \sqrt{m}R_{bm}\right)^2\Big)\Bigg]^{-1/2} \\
&\equiv w\Big(\sqrt{m}R_{bm}, \left(\sqrt{m}Z_{m,j}, \sqrt{m}R_{m,j}\right) : j = 1, 2, \ldots, b\Big),
\end{aligned} \tag{33}$$

where $w : \Re^{2b+1} \to \Re$ is the function defined such that (33) holds, which is possible by (30). The b batches are independent, so (32) and Example 3.2 of Billingsley [1999] imply $(\sqrt{m}Z_{m,j} : j = 1, 2, \ldots, b) \Rightarrow (N_j : j = 1, 2, \ldots, b)$ as $m \to \infty$, where N_j, $j = 1, 2, \ldots, b$, are i.i.d. $N(0, \kappa_p^2)$. Moreover, the Bahadur representations in (28) and (31) ensure $(\sqrt{m}R_{bm}, \sqrt{m}R_{m,j} : j = 1, 2, \ldots, b) \Rightarrow (0, 0, \ldots, 0)$ as $n \to \infty$ by Theorem 3.9 of Billingsley [1999]. Another application of the same theorem then implies

$$\Big(\sqrt{m}R_{bm}, \left(\sqrt{m}Z_{m,j}, \sqrt{m}R_{m,j}\right) : j = 1, 2, \ldots, b\Big) \Rightarrow (0, (N_j, 0) : j = 1, 2, \ldots, b), \tag{34}$$

as $m \to \infty$ with b fixed. Because w in (33) is continuous with probability 1 at the limit in (34), the continuous-mapping theorem (Theorem 3.4.3 of Whitt 2002) guarantees that

$$A_{m,b} \Rightarrow \frac{\sqrt{b}(\frac{1}{b}\sum_{j=1}^{b} N_j)}{\sqrt{\frac{1}{b-1}\sum_{j=1}^{b}(N_j - \frac{1}{b}\sum_{\ell=1}^{b} N_\ell)^2}}, \tag{35}$$

as $m \to \infty$ with b fixed because of (30), where the limit has a Student t-distribution with $b-1$ df. Thus, the Portmanteau theorem (Theorem 2.1 of Billingsley 1999) implies $P_*(\xi_p \in C_{m,b,\mathrm{sect}}) = P_*(-t_{b-1,\alpha} < A_{m,b} < t_{b-1,\alpha}) \to 1 - \alpha$ as $m \to \infty$ by the continuity of the Student t CDF, where we recall that P_* denotes the probability measure under IS.

A.2. Proof of Theorem 3.2

We only show that $\sqrt{m}(\tilde{\hat{\xi}}_{p,m,b} - \hat{\xi}_{p,n}) \Rightarrow 0$, which ensures that the difference between the batching and overall point estimators vanishes sufficiently fast to allow substituting the overall point estimator in the batching CI, and the rest of the proof is similar to that of Theorem 3.1. Chu and Nakayama [2012] prove that the overall single-CV p-quantile estimator $\hat{\xi}_{p,n}$ satisfies a Bahadur representation in (10) when $f(\xi_p) > 0$ and the single CV Q satisfies $0 < \mathrm{Var}[Q] < \infty$, and it is straightforward to modify their proof to handle the multiple-CV case under the conditions assumed here. Thus,

$$\hat{\xi}_{p,n} = \xi_p + \frac{1}{f(\xi_p)}\left[p - F_n(\xi_p) + \hat{\boldsymbol{\beta}}_n(\xi_p)^\top(\bar{\boldsymbol{Q}}_n - \boldsymbol{\nu})\right] + R_n \quad \text{with} \quad \sqrt{n}R_n \Rightarrow 0, \tag{36}$$

as $n \to \infty$. Similarly, for each batch j, we have

$$\hat{\xi}_{p,m,j} = \xi_p + \frac{1}{f(\xi_p)}\left[p - F_{m,j}(\xi_p) + \hat{\boldsymbol{\beta}}_{m,j}(\xi_p)^\top(\bar{\boldsymbol{Q}}_{m,j} - \boldsymbol{\nu})\right] + R_{m,j} \quad \text{with} \quad \sqrt{m}R_{m,j} \Rightarrow 0, \tag{37}$$

as $m = n/b \to \infty$ with b fixed. Hence, the batching point estimator in (14) satisfies

$$\begin{aligned}
\tilde{\xi}_{p,m,b} &= \xi_p + \frac{1}{f(\xi_p)}\left[p - \frac{1}{b}\sum_{j=1}^{b}\left(F_{m,j}(\xi_p) - \hat{\boldsymbol{\beta}}_{m,j}(\xi_p)^\top(\bar{\boldsymbol{Q}}_{m,j} - \boldsymbol{\nu})\right)\right] + \frac{1}{b}\sum_{j=1}^{b} R_{m,j} \\
&= \xi_p + \frac{1}{f(\xi_p)}\left[p - F_n(\xi_p) + \frac{1}{b}\sum_{j=1}^{b}\left(\hat{\boldsymbol{\beta}}_{m,j}(\xi_p)^\top(\bar{\boldsymbol{Q}}_{m,j} - \boldsymbol{\nu})\right)\right] + \frac{1}{b}\sum_{j=1}^{b} R_{m,j}
\end{aligned}$$

by the linearity of $F_{m,j}$ and F_n. It then follows from (36) that

$$\begin{aligned}
\sqrt{m}(\tilde{\xi}_{p,m,b} - \hat{\xi}_{p,n}) &= \frac{\sqrt{m}}{f(\xi_p)}\left[\frac{1}{b}\sum_{j=1}^{b}\left(\hat{\boldsymbol{\beta}}_{m,j}(\xi_p)^\top(\bar{\boldsymbol{Q}}_{m,j} - \boldsymbol{\nu})\right) - \hat{\boldsymbol{\beta}}_n(\xi_p)^\top(\bar{\boldsymbol{Q}}_n - \boldsymbol{\nu})\right] \\
&\quad + \sqrt{m}\left[\frac{1}{b}\sum_{j=1}^{b} R_{m,j} - R_n\right] \\
&= \frac{1}{f(\xi_p)}\left[\frac{1}{b}\sum_{j=1}^{b}\left([\hat{\boldsymbol{\beta}}_{m,j}(\xi_p) - \hat{\boldsymbol{\beta}}_n(\xi_p)]^\top\sqrt{m}(\bar{\boldsymbol{Q}}_{m,j} - \boldsymbol{\nu})\right)\right] \\
&\quad + \frac{1}{b}\sum_{j=1}^{b}\sqrt{m}R_{m,j} - \sqrt{m}R_n
\end{aligned}$$

because $(1/b)\sum_{j=1}^{b}\bar{\boldsymbol{Q}}_{m,j} = \bar{\boldsymbol{Q}}_n$ by the linearity of averaging. As $m = n/b \to \infty$ with b fixed, $\hat{\Sigma}_{\boldsymbol{Q},n}$ and $\tilde{\Sigma}_{\boldsymbol{Q},m,j}$ consistently estimate $\Sigma_{\boldsymbol{Q}}$, which we assumed to be nonsingular. Also, $\hat{\Sigma}_{\boldsymbol{Q},X,n}(\xi_p)$ and $\tilde{\Sigma}_{\boldsymbol{Q},X,m,j}(\xi_p)$ consistently estimate $\Sigma_{\boldsymbol{Q},X}(\xi_p)$, so the continuous-mapping theorem ensures $\hat{\boldsymbol{\beta}}_{m,j}(\xi_p) - \hat{\boldsymbol{\beta}}_n(\xi_p) = \tilde{\Sigma}^{-1}_{\boldsymbol{Q},m,j}\tilde{\Sigma}_{\boldsymbol{Q},X,m,j}(\xi_p) - \hat{\Sigma}^{-1}_{\boldsymbol{Q},n}\hat{\Sigma}_{\boldsymbol{Q},X,n}(\xi_p) \Rightarrow \mathbf{0}$, the r-vector of zeros, as $m = n/b \to \infty$ for fixed $b \geq 2$. Moreover, for each batch j, we have $\sqrt{m}(\bar{\boldsymbol{Q}}_{m,j} - \boldsymbol{\nu}) \Rightarrow N_r(\mathbf{0}, \Sigma_{\boldsymbol{Q}})$ as $m \to \infty$ by the ordinary multivariate CLT, where $N_r(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes an r-dimensional normal random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Therefore, $[\hat{\boldsymbol{\beta}}_{m,j}(\xi_p) - \hat{\boldsymbol{\beta}}_n(\xi_p)]^\top\sqrt{m}(\bar{\boldsymbol{Q}}_{m,j} - \boldsymbol{\nu}) \Rightarrow 0$ for each batch j as $m \to \infty$. In addition, $\sqrt{m}R_{m,j} \Rightarrow 0$ and $\sqrt{m}R_n \Rightarrow 0$ as $m = n/b \to \infty$ by (36) and (37). Consequently, Slutsky's theorem implies $\sqrt{m}(\tilde{\xi}_{p,m,b} - \hat{\xi}_{p,n}) \Rightarrow 0$ as $m \to \infty$, as desired.

ACKNOWLEDGMENTS

The author would like to thank Prof. Peter Glynn of Stanford University for suggesting to investigate this topic. The author also thanks an associate editor and referees for helpful comments that ultimately led to a better article.

REFERENCES

V. G. Adlakha and V. G. Kulkarni. 1989. A classified bibliography of research on stochastic PERT networks. *INFOR* 27 (1989), 272–296.

S. Asmussen and P. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. Springer, New York.

A. N. Avramidis and J. R. Wilson. 1998. Correlation-induction techniques for estimating quantiles in simulation. *Operations Research* 46 (1998), 574–591.

R. R. Bahadur. 1966. A note on quantiles in large samples. *Annals of Mathematical Statistics* 37 (1966), 577–580.

P. Billingsley. 1995. *Probability and Measure* (3rd ed.). John Wiley & Sons, New York.

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

D. A. Bloch and J. L. Gastwirth. 1968. On a simple estimate of the reciprocal of the density function. *Annals of Mathematical Statistics* 39 (1968), 1083–1085.

F. Chu and M. K. Nakayama. 2012. Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions On Modeling and Computer Simulation* 36, 2 (2012), Article 7 (25 pages plus 12–page online–only appendix).

D. Duffie and J. Pan. 1997. An overview of value at risk. *Journal of Derivatives* 4 (1997), 7–49.

B. Efron. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7 (1979), 1–26.

M. Falk. 1986. On the estimation of the quantile density function. *Statistics & Probability Letters* 4 (1986), 69–73.

J. K. Ghosh. 1971. A new proof of the Bahadur representation of quantiles and an application. *Annals of Mathematical Statistics* 42 (1971), 1957–1961.

P. Glasserman, P. Heidelberger, and P. Shahabuddin. 2000. Variance reduction techniques for estimating value-at-risk. *Management Science* 46 (2000), 1349–1364.

P. W. Glynn. 1996. Importance sampling for Monte Carlo estimation of quantiles. In *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation*. Publishing House of St. Petersburg University, St. Petersburg, Russia, 180–185.

P. W. Glynn and D. L. Iglehart. 1990. Simulation output analysis using standardized time series. *Mathematics of Operations Research* 15 (1990), 1–16.

P. Hall and M. A. Martin. 1988. Exact convergence rate of bootstrap quantile variance estimator. *Probability Theory and Related Fields* 80 (1988), 261–268.

T. C. Hesterberg and B. L. Nelson. 1998. Control variates for probability and quantile estimation. *Management Science* 44 (1998), 1295–1312.

J. C. Hsu and Barry L. Nelson. 1990. Control variates for quantile estimation. *Management Science* 36 (1990), 835–851.

X. Jin, M. C. Fu, and X. Xiong. 2003. Probabilistic error bounds for simulation quantile estimation. *Management Science* 49 (2: 2003), 230–246.

S. Juneja, R. Karandikar, and P. Shahabuddin. 2007. Asymptotics and fast simulation for tail probabilities of maximum of sums of few random variables. *ACM Transactions on Modeling and Computer Simulation* 17 (2007), article 2, 35 pages.

J. Liu and X. Yang. 2012. The convergence rate and asymptotic distribution of bootstrap quantile variance estimator for importance sampling. *Advances in Applied Probability* 44 (2012), 815–841.

D. F. Munoz. 2010. On the validity of the batch quantile method for Markov chains. *Operations Research Letters* 38 (2010), 223–226.

M. K. Nakayama. 2011a. Asymptotic properties of kernel density estimators when applying importance sampling. In *Proceedings of the 2011 Winter Simulation Conference*, S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu (Eds.). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, 556–568.

M. K. Nakayama. 2011b. Asymptotically valid confidence intervals for quantiles and values-at-risk when applying Latin hypercube sampling. *International Journal on Advances in Systems and Measurements* 4 (2011), 86–94.

M. K. Nakayama. 2012. Using sectioning to construct confidence intervals for quantiles when applying importance sampling. In *Proceedings of the 2012 Winter Simulation Conference*, C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher (Eds.). Institute of Electrical and Electronics Engineers, Piscataway, New Jersey, to appear.

Nuclear Energy Agency Committee on the Safety of Nuclear Installations. 2007. *BEMUSE Phase III Report: Uncertainty and Sensitivity Analysis of the LOFT L2-5 Test*. Organization for Economic Co-operation and Development Report NEA/CSNI/R(2007)4. Paris, France.

E. Parzen. 1979. Density quantile estimation approach to statistical data modelling. In *Smoothing Techniques for Curve Estimation*, T. Gasser and M. Rosenblatt (Eds.). Springer, Berlin.

B. W. Schmeiser. 1982. Batch size effects in the analysis of simulation output. *Operations Research* 30 (1982), 556–568.

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

J. Shao and D. Tu. 1995. *The Jackknife and Bootstrap*. Springer, New York.

L. Sun and L. J. Hong. 2010. Asymptotic representations for importance-sampling estimators of value-at-risk and conditional value-at-risk. *Operations Research Letters* 38 (2010), 246–251.

J. W. Tukey. 1965. Which part of the sample contains the information? *Proceedings of the National Academy of Sciences of the USA* 53 (1965), 127–134.

U.S. Nuclear Regulatory Commission. 1989. *Best-Estimate Calculations of Emergency Core Cooling Performance*. Nuclear Regulatory Commission Regulatory Guide 1.157. Washington, DC.

M. P. Wand and M. C. Jones. 1995. *Kernel Smoothing*. Chapman & Hall, London.

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

Received January 2013; revised August 2013; accepted November 2013