Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, eds.

USING SECTIONING TO CONSTRUCT CONFIDENCE INTERVALS FOR QUANTILES WHEN APPLYING IMPORTANCE SAMPLING

Marvin K. Nakayama

Computer Science Department
New Jersey Institute of Technology
Newark, NJ 07102, USA

ABSTRACT

Quantiles, which are known as values-at-risk in finance, are often used to measure risk. Confidence intervals provide a way of assessing the error of quantile estimators. When estimating extreme quantiles using crude Monte Carlo, the confidence intervals may have large half-widths, thus motivating the use of variance-reduction techniques (VRTs). This paper develops methods for constructing confidence intervals for quantiles when applying the VRT importance sampling. The confidence intervals, which are asymptotically valid as the number of samples grows large, are based on a technique known as sectioning. Empirical results seem to indicate that sectioning can lead to confidence intervals having better coverage than other existing methods.

1 INTRODUCTION

Consider a random variable X representing the (random) performance of a stochastic model over a finite time period. Let F be the cumulative distribution function (CDF) of X. For a fixed $0 < p < 1$, the p-quantile of F (or equivalently, of X) is a constant ξ_p such that $F(\xi_p) = p$; i.e., $\xi_p = F^{-1}(p)$. A well-known example is the *median*, which is the 0.5-quantile. We assume that F is too complicated to compute in closed form, but we have a simulation model that outputs samples from F.

Quantiles frequently arise in practice as risk measures. For example, the 0.99-quantile, which is also called the 99% (or 1%) *value-at-risk* in finance, is widely used to measure portfolio risk; e.g., see Duffie and Pan (1997). The Nuclear Regulatory Commission (NRC; U.S. Nuclear Regulatory Commission 1989) requires that nuclear-power-plant licensees estimate the 0.95-quantile of various output variables (e.g., peak cladding temperature) in simulations of loss-of-coolant accidents (LOCAs).

The estimation of quantiles via simulation is typically carried out in the following manner. Since the p-quantile is the inverse of the true CDF F, a natural way to estimate ξ_p is to first estimate the CDF from n independent and identically distributed (i.i.d.) samples from F, and then invert the estimated CDF to obtain a quantile estimator.

In addition to computing a point estimator for ξ_p, it is also important to provide a measure of the estimator's statistical error. This is typically done by constructing a confidence interval (CI) for ξ_p, and a CI provides error bounds in which one is highly confident that the true quantile lies. Indeed, the NRC requires that plant licensees compute an upper-one-sided 95% CI for a 0.95-quantile in LOCA simulations and show that the CI lies entirely below a mandated threshold. This is known as the *95/95 criterion*; e.g., see Section 24.9 of Lurie, Abramson, and Vail (2011) and U.S. Nuclear Regulatory Commission (2010). At present, nuclear engineers perform the simulation analysis only using crude Monte Carlo (CMC; i.e., sampling without the use of any variance reduction).

A commonly used approach for developing a CI for a quantile is to first establish that the quantile estimator satisfies a central limit theorem (CLT), and then unfold the CLT to obtain a CI. One difficulty with this approach is that the asymptotic variance in the CLT is nontrivial to estimate. For the case of

978-1-4673-4781-5/12/$31.00 ©2012 IEEE

CMC, there have been techniques developed in the statistics literature (e.g., Bloch and Gastwirth 1968; Falk 1986) to consistently estimate the asymptotic variance. Unfortunately, these methods require the user to specify some parameters, for which determining appropriate values can be difficult in practice.

One problem with CMC is that the half-width of the resulting CI may be quite large, especially when estimating extreme quantiles (i.e., when $p \approx 0$ or $p \approx 1$). To obtain more efficient quantile estimators, we can apply a variance-reduction technique (VRT); see Chapter V of Asmussen and Glynn (2007) for an overview of VRTs to estimate a mean. VRTs developed for estimating a quantile include importance sampling (IS; Glynn 1996), combined IS and stratified sampling (IS+SS; Glasserman, Heidelberger, and Shahabuddin 2000), control variates (CV; Hsu and Nelson 1990; Hesterberg and Nelson 1998), and correlation-induction methods, such as Latin hypercube sampling (LHS) and antithetic variates (AV) (Avramidis and Wilson 1998; Jin, Fu, and Xiong 2003). IS, which is the focus of the current paper, is especially well suited to study rare events; see Glynn and Iglehart (1989). When applying a VRT to estimate a quantile, the VRT is typically applied to obtain an estimate of the CDF, which is inverted to arrive at a quantile estimate.

While most of the above papers prove that the resulting VRT quantile estimator satisfies a CLT, none provides a method for constructing a CI for the quantile based on the CLT. To address this issue, Chu and Nakayama (2012) develop a general framework to construct CIs for quantiles when applying a wide spectrum of VRTs, including IS, IS+SS, CV and AV. This approach uses a finite-difference estimator to estimate the asymptotic variance constant in the CLT. Nakayama (2011b) extends the applicability of the method to a type of LHS. Nakayama (2011a) develops an alternative estimator for the asymptotic variance constant using kernel methods (Wand and Jones 1995) for the case of IS. One drawback of these methods is that while the resulting CIs are asymptotically valid, their performance can be poor for finite sample sizes when $p \approx 1$. Liu and Yang (2012) also develop a bootstrap estimator of the asymptotic variance of the IS quantile estimator, but it converges more slowly than the kernel estimator in Nakayama (2011a).

An alternative approach to produce a CI for a quantile is to use *batching* (Schmeiser 1982). In batching, the n i.i.d. samples are divided into $b \geq 2$ nonoverlapping batches of equal size $m = n/b$. We compute from each batch an estimate of the CDF, which is inverted to obtain a quantile estimate. The resulting b quantile estimates from the b batches are averaged to obtain a point estimate, at which the batching CI is centered. The half-width of the batching CI is determined by the sample standard deviation of the b batch quantile estimates.

One drawback of the batching CI is that it can have poor coverage when the sample size n is small, especially for extreme quantiles. The problem arises because quantile estimators are biased, where the bias decreases to 0 as the sample size increases; e.g., see Avramidis and Wilson (1998). With batching, the amount of bias is determined by $m = n/b$, the size of each batch, which is smaller than the total sample size n. Thus, the point estimate in batching can suffer from large bias, especially for small n, and this potentially leads to poor coverage; e.g., see Nakayama (2011b).

To address this issue, we now consider alternative approaches to construct CIs for ξ_p based on the idea of *sectioning*. Section III.5a of Asmussen and Glynn 2007 develops sectioning for the case of CMC, and we now extend it to IS. Similar to batching, sectioning instead centers the CI at the overall quantile estimate obtained from inverting the CDF estimate based on all n samples. Also, for the half-width of the CI, we replace the average of the batch quantile estimates by the overall quantile estimate in the formula for the sample standard deviation of the batch quantile estimates.

An advantage of sectioning over batching is that the sectioning point estimate is computed by inverting the estimated CDF from all n samples. This results in the sectioning point estimator of the quantile having smaller bias than that for batching, and the reduced bias can improve coverage for small sample size n, as our experiments in this paper show.

The rest of the paper has the following organization. Sections 2 and 3 review quantile estimation when applying CMC and IS, respectively. Section 4 describes how to construct CIs using sectioning and batching with IS. We present in Section 5 numerical results from running experiments on a small model. Concluding remarks are provided in Section 6.

2 REVIEW OF QUANTILE ESTIMATION USING CMC

Let X be a random variable having CDF F; i.e., $F(y) = P(X \le y)$. For a fixed $0 < q < 1$ and real-valued function G, let $G^{-1}(q) = \inf\{x : G(x) \ge q\}$. Now fix $0 < p < 1$. The goal is to estimate and construct a CI for the p-quantile $\xi_p \equiv F^{-1}(p)$ via simulation.

When applying crude Monte Carlo, we accomplish this by generating independent and identically distributed (i.i.d.) samples $X_1, X_2, \ldots, X_n$ from F. Estimate F using the *empirical CDF* F_n, defined by

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

with $I(\cdot)$ denoting the indicator function, which takes on the value 1 (resp., 0) when the argument is true (resp., false). A point estimator for ξ_p is then

$$\xi_{p,n} = F_n^{-1}(p). \tag{2}$$

To build a CI for ξ_p, we first note that if F is differentiable at ξ_p with $f(\xi_p) > 0$, where f denotes the derivative of F when it exists, the p-quantile estimator $\xi_{p,n}$ satisfies the CLT (Section 2.3.3 of Serfling 1980)

$$\sqrt{n}(\xi_{p,n} - \xi_p) \Rightarrow N(0, \tau_p^2)$$

as $n \to \infty$, where $N(a, c^2)$ is a normal random variable with mean a and variance c^2, $\tau_p^2 = p(1-p)/f^2(\xi_p)$, and $\Rightarrow$ denotes convergence in distribution (Section 3 of Billingsley 1999). If one has a consistent estimator $\tau_{p,n}$ of τ_p, then an asymptotically valid $100(1-\alpha)\%$ CI for ξ_p is $(\xi_{p,n} \pm z_\alpha \tau_{p,n}/\sqrt{n})$, where $z_\alpha = \Phi^{-1}(1-\alpha/2)$ and Φ is the CDF of a $N(0,1)$ random variable. However, constructing a consistent estimator for τ_p is a delicate matter because estimating $f(\xi_p)$ is nontrivial. There have been different methods developed in the statistics literature to estimate $f(\xi_p)$, including finite-difference estimators (Bloch and Gastwirth 1968, Bofinger 1975) and kernel estimators (Falk 1986, Jones 1992).

Batching is an alternative approach to construct a CI for ξ_p that avoids consistently estimating τ_p and $f(\xi_p)$. To apply the method, partition the n samples into a fixed number $b \ge 2$ of (nonoverlapping) batches, each of size $m = n/b$. Thus, for $j = 1, 2, \ldots, b$, the jth batch consists of samples X_i, $i = (j-1)m+1, (j-1)m+2, \ldots, jm$. Let $F_{m,j}$ be the CDF estimator from the jth batch (of size m), so

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

From the jth batch, we compute a quantile estimator $\xi_{p,m,j} = F_{m,j}^{-1}(p)$. Then let

$$S_{m,b,\text{batch}}^2 = \frac{1}{b-1}\sum_{j=1}^{b}(\xi_{p,m,j} - \bar{\xi}_{p,m,b})^2, \tag{3}$$

where

$$\bar{\xi}_{p,m,b} = \frac{1}{b}\sum_{j=1}^{b}\xi_{p,m,j}. \tag{4}$$

The batching $100(1-\alpha)\%$ CI for ξ_p when applying CMC is then

$$\left(\bar{\xi}_{p,m,b} \pm t_{b-1,\alpha}\frac{S_{m,b,\text{batch}}}{\sqrt{b}}\right), \tag{5}$$

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

Sectioning (Section III.5a of Asmussen and Glynn 2007) is a method similar to batching for constructing a CI for ξ_p. The n samples are again divided into $b \geq 2$ batches, each of size $m = n/b$, but now the batches are instead called *sections*. (Asmussen and Glynn 2007 suggest that b should be chosen no greater than 30.) Then compute

$$S^2_{m,b,\text{sect}} = \frac{1}{b-1} \sum_{j=1}^{b} (\xi_{p,m,j} - \xi_{p,n})^2, \tag{6}$$

which differs from the batching variance estimator (3) because $S^2_{m,b,\text{sect}}$ subtracts the overall point estimator $\xi_{p,n}$ in (2) from each summand rather than the average (4) of the b batch quantile estimates. The sectioning $100(1-\alpha)\%$ CI for ξ_p is then

$$\left(\xi_{p,n} \pm t_{b-1,\alpha} \frac{S_{m,b,\text{sect}}}{\sqrt{b}} \right). \tag{7}$$

Both (5) and (7) are asymptotically valid CIs as $n \to \infty$ with $b \geq 2$ fixed.

The sectioning CI in (7) and the batching CI in (5) differ in the following ways. The half-widths of the batching and sectioning CI are determined by (3) and (6), respectively. Moreover, the sectioning CI is centered at $\xi_{p,n}$ from (2), whereas the batching CI is centered at $\bar{\xi}_{p,m,b}$ in (4), which is the average of the b batch estimates $\xi_{p,m,j}$, $j = 1,2,\ldots,b$. Note that $\xi_{p,n}$ is obtained by inverting the CDF estimator based on all n samples, whereas each $\xi_{p,m,j}$ is computed by inverting the CDF estimator from the m samples from the jth section (or batch). Since $m = n/b < n$, each batch's quantile estimator is based on a smaller number of samples than is used to compute $\xi_{p,n}$. But quantile estimators are known to be biased, with the bias decreasing in the sample size; e.g., see Avramidis and Wilson (1998). This leads to $\bar{\xi}_{p,m,b}$ typically being more biased than $\xi_{p,n}$. Thus, since the batching CI is centered at $\bar{\xi}_{p,m,b}$, the batching CI can have poorer coverage than the sectioning CI in (7), especially when n is small.

3 REVIEW OF QUANTILE ESTIMATION USING IMPORTANCE SAMPLING

Glynn (1996) shows how to apply IS to estimate quantiles, as we now explain. First suppose that the CDF F is absolutely continuous with density function f, and let g be another density function satisfying $g(x) > 0$ whenever $f(x) > 0$. Let $L(x) = f(x)/g(x)$, which is known as the *likelihood ratio*, and let E_* denote expectation when X has density g. Then we can write

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

This expression suggests the following approach to estimate $F(y)$ and ξ_p. First use density g to generate n i.i.d. samples $(X_1, L_1), (X_2, L_2), \ldots, (X_n, L_n)$ of (X, L), where $L \equiv L(X)$. Then estimate $F(y)$ by

$$\hat{F}_n(y) = 1 - \frac{1}{n} \sum_{i=1}^{n} I(X_i > y) L_i. \tag{8}$$

(We use a hat to denote IS estimators.) The IS estimator of the p-quantile of F is finally

$$\hat{\xi}_{p,n} = \hat{F}_n^{-1}(p). \tag{9}$$

Our development of this IS quantile estimator is for the simple case when the random variable X has a CDF F with a density f, but IS can be generalized to much broader settings; e.g., see Glynn and Iglehart (1989).

Glynn (1996) also develops an alternative IS quantile estimator by computing a different CDF estimator

$$\hat{F}'_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq y) L_i, \tag{10}$$

which is based on the representation $F(y) = E[I(X \le y)] = E_*[I(X \le y)L(X)]$. Inverting this CDF estimator leads to another IS quantile estimator

$$\hat{\xi}'_{p,n} = \hat{F}'^{-1}_n(p). \tag{11}$$

As Glynn (1996) notes, the IS quantile estimator in (9) is more appropriate when $p \approx 1$, whereas the one in (11) should be applied for $p \approx 0$. To simplify the discussion from here on, we only focus on the quantile estimator (9), but the following results also apply to the estimator in (11) with minor modifications.

Assuming that $f(\xi_p) > 0$ and $E_*[L^3] < \infty$, Glynn (1996) proves the CLT

$$\sqrt{n}(\hat{\xi}_{p,n} - \xi_p) \Rightarrow N(0, \kappa_p^2)$$

as $n \to \infty$, where

$$\kappa_p^2 = \frac{E_*[I(X > \xi_p)L^2] - (1-p)^2}{f^2(\xi_p)} \equiv \frac{\psi_p^2}{f^2(\xi_p)}. \tag{12}$$

Chu and Nakayama (2012) relax Glynn's moment condition for the CLT to

$$E_*[I(X > \xi_p - \delta)L^{2+\varepsilon}] < \infty \text{ for some } \varepsilon > 0 \text{ and } \delta > 0. \tag{13}$$

(For the quantile estimator $\hat{\xi}'_{p,n}$ in (11), the required moment condition is instead $E_*[I(X < \xi_p + \delta)L^{2+\varepsilon}] < \infty$ for some $\varepsilon > 0$ and $\delta > 0$.) If we have a consistent estimator $\hat{\kappa}_{p,n}$ of κ_p from (12), then we can use it to construct an asymptotically valid $100(1-\alpha)\%$ CI for ξ_p as

$$\left(\hat{\xi}_{p,n} \pm z_\alpha \frac{\hat{\kappa}_{p,n}}{\sqrt{n}}\right). \tag{14}$$

The numerator ψ_p^2 in the expression for κ_p^2 in (12) can be estimated by

$$\hat{\psi}_{p,n}^2 = \frac{1}{n}\sum_{i=1}^{n} I(X_i > \hat{\xi}_{p,n})L_i^2 - (1-p)^2, \tag{15}$$

which Chu and Nakayama (2012) prove is consistent. (The consistency proof is complicated by the fact that the summands $I(X_i > \hat{\xi}_{p,n})L_i^2$, $i = 1, 2, \ldots, n$, are not independent since they all depend on $\hat{\xi}_{p,n}$, which is a function of all n samples.) Chu and Nakayama (2012) also develop a consistent estimator for $\lambda_p \equiv 1/f(\xi_p)$ using a finite difference (Section 7.1 of Glasserman 2004). Nakayama (2011a) uses kernel methods to consistently estimate $f(\xi_p)$.

We now describe the finite-difference estimator for λ_p from Chu and Nakayama (2012). By the chain rule of calculus, we have that $\frac{d}{dp}F^{-1}(p) = 1/f(F^{-1}(p)) = \lambda_p$. Since $\frac{d}{dp}F^{-1}(p) = \lim_{h\to 0}[F^{-1}(p+h) - F^{-1}(p-h)]/(2h)$, a natural estimator of λ_p is

$$\hat{\lambda}_{p,n} = \frac{\hat{F}_n^{-1}(p+h_n) - \hat{F}_n^{-1}(p-h_n)}{2h_n}, \tag{16}$$

where $h_n > 0$ is a user-specified (small) *bandwidth*. Assuming (13) holds and f is continuous in a neighborhood in ξ_p, Chu and Nakayama (2012) show that $\hat{\lambda}_{p,n} \Rightarrow \lambda_p$ as $n \to \infty$ when $h_n \to 0$ and $\sqrt{n}h_n \to a \in (0, \infty]$ as $n \to \infty$. For example, we can choose the bandwidth $h_n = cn^{-d}$ for constants $c > 0$ and $0 < d \le 1/2$.

The kernel estimator of $f(\xi_p)$ from Nakayama (2011a) is the plug-in estimator

$$\hat{f}_n(\hat{\xi}_{p,n}), \tag{17}$$

where

$$\hat{f}_n(y) = \frac{1}{n}\sum_{i=1}^{n} k_h(y - X_i)L_i$$

is the IS kernel density estimator, $k_h(x) = \frac{1}{h}k\left(\frac{x}{h}\right)$, $k(\cdot)$ is a *kernel* function (often taken to be a symmetric density function; see Chapter 2 of Wand and Jones 1995), and $h = h_n > 0$ is a user-specified bandwidth. Under certain conditions, Nakayama (2011a) shows that $\hat{f}_n(\hat{\xi}_{p,n}) \Rightarrow f(\xi_p)$ as $n \to \infty$ when $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. Also, it is shown that $\hat{f}_n(\hat{\xi}_{p,n})$ satisfies a CLT, which establishes its rate of convergence as $(nh_n)^{-1/2}$ rather than the canonical $n^{-1/2}$.

4 BATCHING AND SECTIONING WITH IMPORTANCE SAMPLING

To apply batching with IS to construct a CI for ξ_p, we partition the n samples into $b \geq 2$ (nonoverlapping) batches, each of size m, where we assume that $n = bm$. Thus, for each $j = 1, 2, \ldots, b$, the samples (X_i, L_i), $i = (j-1)m+1, (j-1)m+2, \ldots, jm$, form the jth batch. We assume that as $n \to \infty$, the number b of batches remains fixed and the batch size $m \to \infty$. Let $\hat{F}_{m,j}$ denote the IS estimate of the CDF from the jth batch:

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

from which we obtain the jth *IS batch quantile estimate* $\hat{\xi}_{p,m,j} = \hat{F}^{-1}_{m,j}(p)$. Then compute the IS batching variance estimator

$$\hat{S}^2_{m,b,\text{batch}} = \frac{1}{b-1}\sum_{j=1}^{b}(\hat{\xi}_{p,m,j} - \tilde{\xi}_{p,m,b})^2, \tag{18}$$

where

$$\tilde{\xi}_{p,m,b} = \frac{1}{b}\sum_{j=1}^{b}\hat{\xi}_{p,m,j}. \tag{19}$$

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

$$C_{m,b,\text{batch}} \equiv \left(\tilde{\xi}_{p,m,b} \pm t_{b-1,\alpha}\frac{\hat{S}_{m,b,\text{batch}}}{\sqrt{b}}\right). \tag{20}$$

Now consider applying *sectioning* to construct a confidence interval for ξ_p when using IS. When sectioning is used, the b batches are instead called sections. We compute the IS sectioning variance estimator

$$\hat{S}^2_{m,b,\text{sect}} = \frac{1}{b-1}\sum_{j=1}^{b}(\hat{\xi}_{p,m,j} - \hat{\xi}_{p,n})^2, \tag{21}$$

which we note is similar to the IS batching variance estimator (18) except that in each summand we subtract $\hat{\xi}_{p,n}$ given in (9) rather than the sample average $\tilde{\xi}_{p,m,b}$ from (19). Finally we define the IS sectioning $100(1-\alpha)\%$ confidence interval for ξ_p as

$$C_{m,b,\text{sect}} \equiv \left(\hat{\xi}_{p,n} \pm t_{b-1,\alpha}\frac{\hat{S}_{m,b,\text{sect}}}{\sqrt{b}}\right). \tag{22}$$

We can also define a combined sectioning-batching CI as

$$C_{m,b,\text{sb}} \equiv \left(\hat{\xi}_{p,n} \pm t_{b-1,\alpha}\frac{\hat{S}_{m,b,\text{batch}}}{\sqrt{b}}\right), \tag{23}$$

which is centered at the overall quantile estimate $\hat{\xi}_{p,n}$ but whose half-width is the same as that for batching.

Let P_* denote the probability measure under IS, and the following theorem, whose proof appears in Nakayama (2012), establishes the asymptotic validity of the CIs. Choosing the IS change of measure P_* to be the same as the original measure P results in the likelihood ratio $L \equiv 1$, so the theorem also covers CMC as a special case.

Theorem 1 Suppose that F is differentiable at ξ_p, with $f(\xi_p) > 0$. Also, suppose that (13) holds. Then for CI $C = C_{m,b,\text{sect}}$, $C_{m,b,\text{batch}}$ or $C_{m,b,\text{sb}}$,

$$P_*\{\xi_p \in C\} \to 1-\alpha$$

as $m \to \infty$ with b fixed.

The assumptions in Theorem 1 on the original CDF F and its derivative f are standard conditions used to establish a CLT for a quantile estimator; e.g., see Section 2.3 of Serfling (1980). The proof of the asymptotic validity of the CIs relies on showing that the IS quantile estimator satisfies a so-called *Bahadur representation* (Bahadur 1966), which Chu and Nakayama (2012) prove when (13) further holds. Specifically, Chu and Nakayama (2012) show that

$$\hat{\xi}_{p,n} = \xi_p + \frac{p - \hat{F}_n(\xi_p)}{f(\xi_p)} + R_n \quad \text{with} \quad \sqrt{n}R_n \Rightarrow 0 \tag{24}$$

as $n \to \infty$. Under a stronger set of assumptions, Sun and Hong (2010) prove a stronger Bahadur representation for the IS quantile estimator by establishing an almost-sure rate at which R_n in (24) vanishes. In either case, a Bahadur representation establishes a type of first-order Taylor approximation, showing that a quantile estimator can be approximated by a linear transformation of a CDF estimator. CDF estimators, as in (1) for CMC and (8) and (10) for IS, are often sample averages of i.i.d. quantities, and as such, they satisfy CLTs (under appropriate moment conditions). Hence, a Bahadur representation provides insight into why a quantile estimator, which is not a sample average, satisfies a CLT.

We now describe the main ideas underlying the proof of Theorem 1; the full details are in Nakayama (2012). Note that (24) implies that for each batch (section) $j = 1,2,\ldots,b$, the corresponding quantile estimator satisfies a Bahadur representation:

$$\hat{\xi}_{p,m,j} = \xi_p + \frac{p - \hat{F}_{m,j}(\xi_p)}{f(\xi_p)} + R_{m,j} \quad \text{with} \quad \sqrt{m}R_{m,j} \Rightarrow 0 \tag{25}$$

as $m \to \infty$. As a consequence, $\sqrt{m}(\hat{\xi}_{p,m,j} - \xi_p) = \sqrt{m}[p - \hat{F}_{m,j}(\xi_p)]/f(\xi_p) + \sqrt{m}R_{m,j}$ converges weakly to a normal as $m \to \infty$ since (13) implies $\hat{F}_{m,j}(\xi_p)$ satisfies a CLT and $\sqrt{m}R_{m,j} \Rightarrow 0$. Because the $\hat{\xi}_{p,m,j}$, $j = 1,2,\ldots,b$, are i.i.d., it follows that the batching CI in (20) is asymptotically valid. For sectioning, (25) ensures that the average in (19) of the b batch quantile estimators can be expressed as

$$\begin{aligned}
\tilde{\xi}_{p,m,b} &= \frac{1}{b}\sum_{j=1}^{b}\left[\xi_p + \frac{p - \hat{F}_{m,j}(\xi_p)}{f(\xi_p)} + R_{m,j}\right] \\
&= \xi_p + \frac{p - \frac{1}{b}\sum_{j=1}^{b}\hat{F}_{m,j}(\xi_p)}{f(\xi_p)} + \frac{1}{b}\sum_{j=1}^{b}R_{m,j} \\
&= \xi_p + \frac{p - \hat{F}_n(\xi_p)}{f(\xi_p)} + \frac{1}{b}\sum_{j=1}^{b}R_{m,j}.
\end{aligned}$$

Combining this and (24) leads to

$$\sqrt{n}(\hat{\xi}_{p,n} - \tilde{\xi}_{p,m,b}) = \sqrt{n}\left(R_n - \frac{1}{b}\sum_{j=1}^{b}R_{m,j}\right) \Rightarrow 0$$

as $n = mb \to \infty$ with b fixed. This fact can then be used to justify replacing the estimator $\tilde{\xi}_{p,m,b}$ in the batching CI by $\hat{\xi}_{p,n}$ to obtain the sectioning CI in (22) and the combined sectioning-batching CI in (23).

5 NUMERICAL RESULTS

We ran experiments on a simple stochastic activity network (SAN), which was previously studied in Hsu and Nelson (1990) and Chu and Nakayama (2012). Also known as stochastic PERT networks, SANs are often used in practice to model the time to complete a project consisting of activities having precedence relations and random durations (Adlakha and Kulkarni 1989). We consider a SAN with $d = 5$ activities, which correspond to edges, labeled $1, 2, \ldots, 5$, in the SAN. The length A_ℓ of edge ℓ denotes the time to complete activity ℓ, and $A_1, A_2, \ldots, A_5$ are i.i.d. exponential with mean 1. The SAN has $q = 3$ paths, with $B_1 = \{1,2\}$, $B_2 = \{1,3,5\}$, and $B_3 = \{4,5\}$ as the sets of edges on the paths. Let $X = \max_{j=1,2,3} \sum_{\ell \in B_j} A_\ell$ denote the length of the longest path, and its CDF F is given by $F(x) = 1 + (3 - 3x - x^2/2)e^{-x} + (-3 - 3x + x^2/2)e^{-2x} - e^{-3x}$ for $x \geq 0$, and $F(x) = 0$ for $x < 0$. The goal is to estimate and construct CIs for the p-quantile ξ_p of F for different values of p. Differentiating $F(x)$ leads to its density $f(x)$, which is continuous and positive for all $x \geq 0$. In particular $f(\xi_p) > 0$ for any $0 < p < 1$, as required by Theorem 1.

We simulated the SAN by generating samples of $A_1, A_2, \ldots, A_d$ using IS. We applied an IS scheme from Chu and Nakayama (2012), which is based on ideas from Juneja, Karandikar, and Shahabuddin (2007) and Glynn (1996). The method samples from a mixture of distributions, where each distribution in the mixture exponentially tilts the durations of the activities on one path, leaving the activities not on that path with their original (exponential) distributions; see Chu and Nakayama (2012) for details. In our experiments, we chose different values of $p \approx 1$, so we use the quantile estimator from (9) rather than (11). Table 1 presents the results from constructing nominal 90% CIs for ξ_p for different sample sizes n. We estimated the coverage (and average half-widths) from running 10^4 independent replications.

We constructed CIs for ξ_p using six different methods. These include the batching CI in (20), the sectioning CI in (22), and the combined sectioning-batching CI in (23). The results for these three methods are in the columns in Table 1 labeled "Batch," "Section" and "SB," respectively. For all these CIs, we used $b = 10$ sections or batches.

The other three sets of CIs have the form in (14). One CI estimates $1/f(\xi_p)$ in (12) via the finite difference in (16) (column "FD"). Another employs the kernel estimator in (17) to estimate $f(\xi_p)$ (column "Kernel"). The last (column "Exact") uses the exact value of $f(\xi_p)$, which was computed numerically. All three of these CIs apply (15) to estimate the numerator in (12).

For the finite-difference estimator, we used bandwidth $h_n = 0.5n^{-1/2}$. In our experiments, when $p + h_n \geq 1$, which would result in $\hat{F}_n^{-1}$ being evaluated at a point outside of its domain $(0,1)$, we instead compute the finite-difference estimator by replacing $p + h_n$ and $p - h_n$ in (16) with $q_{1,n} = 1 - (1-p)/10$ and $q_{2,n} = 2p - 1 + (1-p)/10$, respectively ($q_{1,n}$ and $q_{2,n}$ are chosen to be symmetric around p); see Chu and Nakayama (2012) for more details. For the kernel estimator, we chose k to be the Gaussian kernel (i.e., k is the density of a $N(0,1)$) and the bandwidth $h_n = 0.5n^{-1/5}$.

The results for the finite-difference estimator show that for $p = 0.95$, the coverage converges to the nominal level 0.9 as n increases. But for the more extreme values of p, the coverage does not seem to converge as n grows. The theory in Chu and Nakayama (2012) proves that the coverage will eventually converge to 0.9, but this requires huge sample sizes when $p \approx 1$. The finite-difference CIs consistently have over-coverage. This is due to the finite-difference estimator overestimating $1/f(\xi_p)$, as can be seen by the average half-width being larger than those for CIs using the exact value of $f(\xi_p)$.

The CIs based on the kernel estimator show that for each p, coverage converges to 0.9 as n grows. But the coverage is consistently below the nominal level. This can lead to the user being overconfident, which may be especially problematic when evaluating risk.

For each p, the coverage of the batching CIs converge to 0.9 from below as n grows. But for extreme p, the coverage for $n = 100$ is quite poor. The reason for the poor coverage for small n appears to be the bias of quantile estimators. The size $m = n/b$ of each batch is much smaller than n, leading to the batched

quantile estimator (19) being significantly biased. The CI in (20) is thus centered at a quite biased point estimator, negatively affecting coverage.

Sectioning leads to coverage converging to 0.9 as n increases. For all sample sizes except $n = 100$, the absolute difference from the nominal coverage for sectioning is about the same as that for batching, but batching approaches 0.9 from below, while sectioning converges from above. As we noted earlier, it is (arguably) more desirable to have overcoverage than undercoverage. The results for combined sectioning-batching (SB) are similar to plain sectioning, but SB has slightly smaller average half-widths when n is small. This seems to arise from the fact that for the batching variance estimate $\hat{S}^2_{m,b,\text{batch}}$ in (18), $\tilde{\xi}_{p,m,b}$ has the same bias as each $\hat{\xi}_{p,m,j}$, whereas for the sectioning variance estimate $\hat{S}^2_{m,b,\text{sect}}$ in (21), $\hat{\xi}_{p,n}$ has smaller bias than each $\hat{\xi}_{p,m,j}$. This leads to $\hat{S}^2_{m,b,\text{batch}}$ typically being smaller than $\hat{S}^2_{m,b,\text{sect}}$.

Finally, when the coverages are close to the nominal level, the CIs for batching, sectioning, and combined sectioning-batching are slightly wider on average than those for the finite-difference and kernel estimators and for the exact $f(\xi_p)$. The reason for this is that the former CIs use the Student-t critical point instead of the smaller normal critical point of the latter.

6 CONCLUDING REMARKS

We used sectioning to develop confidence intervals for a quantile when applying importance sampling. The CIs, which are asymptotically valid, are similar to those obtained with batching. The sectioning CI is centered at the quantile estimate based on inverting the CDF estimate from all n samples. In contrast, the batching CI is centered at the sample average of the b quantile estimates from the b batches, each of size $m = n/b$. Because quantile estimators are biased, the larger effective sample size of the sectioning point estimate leads to its CI being centered at a less biased estimator than the batching CI. This seems to lead to sectioning and combined sectioning-batching having better coverage than batching when n is small. The CIs for the combined sectioning-batching are slightly smaller than for sectioning when n is small.

ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation under Grants No. CMMI-0926949 and CMMI-1200065. 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. Also, the author would like to thank Peter Glynn for some suggesting to investigate this topic.

REFERENCES

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

Asmussen, S., and P. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. 1st ed. New York: Springer.

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

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

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

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

Bofinger, E. 1975. "Estimation of a density function using order statistics". *Australian Journal of Statistics* 17:1–7.

Chu, F., and M. K. Nakayama. 2012. "Confidence Intervals for Quantiles When Applying Variance-Reduction Techniques". *ACM Transactions On Modeling and Computer Simulation* 36:Article 7 (25 pages plus 12–page online–only appendix).

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

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

Glasserman, P. 2004. *Monte Carlo Methods in Financial Engineering*. New York: Springer.

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

Glynn, P. W. 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*, edited by S. M. Ermakov and V. B. Melas, 180–185: Publishing House of St. Petersburg University, St. Petersburg, Russia.

Glynn, P. W., and D. L. Iglehart. 1989. "Importance sampling for stochastic systems". *Management Science* 35:1367–1393.

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

Hsu, J. C., and B. L. Nelson. 1990. "Control variates for quantile estimation". *Management Science* 36:835–851.

Jin, X., M. C. Fu, and X. Xiong. 2003. "Probabilistic Error Bounds for Simulation Quantile Estimation". *Management Science* 49:230–246.

Jones, M. C. 1992. "Estimating Densities, Quantiles, Quantile Densities and Density Quantiles". *Ann. Inst. Statist. Math.* 44:721–727.

Juneja, S., 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:article 2, 35 pages.

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

Lurie, D., L. Abramson, and J. Vail. 2011. "Applying Statistics". U.S. Nuclear Regulatory Commission Report NUREG-1475, Rev. 1, U.S. Nuclear Regulatory Commission, Washington, DC.

Nakayama, M. K. 2011a, December. "Asymptotic Properties of Kernel Density Estimators When Applying Importance Sampling". In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu, 556–568. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Nakayama, M. K. 2011b. "Asymptotically Valid Confidence Intervals for Quantiles and Values-at-Risk When Applying Latin Hypercube Sampling". *International Journal on Advances in Systems and Measurements* 4:86–94.

Nakayama, M. K. 2012. "Confidence Intervals Using Sectioning for Quantiles When Applying Variance-Reduction Techniques". In preparation.

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

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

Sun, L., and L. J. Hong. 2010. "Asymptotic Representations for Importance-Sampling Estimators of Value-at-Risk and Conditional Value-at-Risk". *Operations Research Letters* 38:246–251.

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

U.S. Nuclear Regulatory Commission 2010. "Acceptance criteria for emergency core cooling systems for light-water nuclear power reactors". Title 10, Code of Federal Regulations Section 50.46 (10CFR50.46), U.S. Nuclear Regulatory Commission, Washington, DC.

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

AUTHOR BIOGRAPHY

MARVIN K. NAKAYAMA is a professor in the Department of Computer Science at the New Jersey Institute of Technology. He received his Ph.D. in operations research from Stanford University. He won second prize in the 1992 George E. Nicholson Student Paper Competition sponsored by INFORMS and is a recipient of a CAREER Award from the National Science Foundation. He is the Stochastic Models Area Editor for *ACM Transactions on Modeling and Computer Simulation* and the Simulation Area Editor for *INFORMS Journal on Computing*.

Table 1: CIs were constructed using different methods, and the coverage levels (and average half-widths) were estimated from 10^4 independent replications.

p	n	FD	Kernel	Batch	Section	SB	Exact
0.95	100	0.984	0.797	0.841	0.945	0.936	0.879
		(0.633)	(0.362)	(0.532)	(0.565)	(0.532)	(0.401)
	400	0.922	0.865	0.888	0.917	0.908	0.895
		(0.226)	(0.200)	(0.236)	(0.243)	(0.236)	(0.207)
	1600	0.904	0.891	0.897	0.910	0.904	0.901
		(0.106)	(0.103)	(0.114)	(0.116)	(0.114)	(0.104)
	6400	0.898	0.894	0.900	0.903	0.901	0.898
		(0.052)	(0.052)	(0.057)	(0.057)	(0.057)	(0.052)
$1-10^{-2}$	100	0.981	0.777	0.803	0.959	0.952	0.873
		(0.712)	(0.390)	(0.661)	(0.714)	(0.661)	(0.445)
	400	0.989	0.864	0.879	0.924	0.916	0.897
		(0.372)	(0.223)	(0.271)	(0.281)	(0.271)	(0.232)
	1600	0.991	0.883	0.892	0.908	0.903	0.901
		(0.188)	(0.115)	(0.129)	(0.131)	(0.129)	(0.117)
	6400	0.941	0.890	0.895	0.904	0.901	0.897
		(0.068)	(0.058)	(0.064)	(0.064)	(0.064)	(0.059)
$1-10^{-3}$	100	0.975	0.743	0.748	0.969	0.960	0.861
		(0.793)	(0.405)	(0.838)	(0.924)	(0.838)	(0.492)
	400	0.990	0.844	0.874	0.932	0.922	0.895
		(0.420)	(0.246)	(0.316)	(0.331)	(0.316)	(0.260)
	1600	0.991	0.880	0.892	0.910	0.902	0.897
		(0.213)	(0.130)	(0.147)	(0.150)	(0.147)	(0.132)
	6400	0.994	0.894	0.899	0.902	0.900	0.897
		(0.107)	(0.066)	(0.072)	(0.073)	(0.072)	(0.066)
$1-10^{-4}$	100	0.971	0.712	0.693	0.977	0.971	0.851
		(0.853)	(0.413)	(1.011)	(1.134)	(1.011)	(0.528)
	400	0.990	0.840	0.863	0.933	0.925	0.898
		(0.460)	(0.266)	(0.357)	(0.375)	(0.357)	(0.283)
	1600	0.993	0.874	0.889	0.913	0.906	0.893
		(0.233)	(0.141)	(0.162)	(0.166)	(0.162)	(0.144)
	6400	0.991	0.893	0.899	0.906	0.902	0.898
		(0.117)	(0.072)	(0.079)	(0.080)	(0.079)	(0.072)
$1-10^{-5}$	100	0.960	0.683	0.626	0.981	0.974	0.836
		(0.899)	(0.415)	(1.172)	(1.338)	(1.172)	(0.557)
	400	0.990	0.831	0.855	0.940	0.932	0.893
		(0.494)	(0.280)	(0.397)	(0.420)	(0.397)	(0.304)
	1600	0.991	0.874	0.889	0.916	0.911	0.897
		(0.251)	(0.151)	(0.175)	(0.180)	(0.175)	(0.155)
	6400	0.992	0.893	0.900	0.906	0.902	0.903
		(0.126)	(0.077)	(0.085)	(0.086)	(0.085)	(0.078)