Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

CONSTRUCTING CONFIDENCE INTERVALS FOR A QUANTILE USING BATCHING AND SECTIONING WHEN APPLYING LATIN HYPERCUBE SAMPLING

Hui Dong

Supply Chain Management and
Marketing Sciences Department
Rutgers University
Newark, NJ 07102, USA

Marvin K. Nakayama

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

ABSTRACT

Quantiles are often used in risk evaluation of complex systems. In some situations, as in regulations regarding safety analyses of nuclear power plants, a confidence interval is required for the quantile of the simulation's output variable. In our current paper, we develop methods to construct confidence intervals for quantiles when applying Latin hypercube sampling, a variance reduction technique that extends stratification for sampling in higher dimensions. Our approaches employ the batching and sectioning methods when applying replicated Latin hypercube sampling, with a single Latin hypercube sample in each batch, and samples across batches are independent. We have established the asymptotic validity of the confidence intervals developed in this paper. Moreover, we have proven that quantile estimators from a single Latin hypercube sample and replicated Latin hypercube samples satisfy weak Bahadur representations. An advantage of sectioning over batching is that the sectioning confidence interval typically has better coverage, which we observe in numerical experiments.

1 INTRODUCTION

For a given constant $0 < p < 1$, the p-quantile of a continuous random variable Y is the smallest constant ξ_p such that $P(Y \le \xi_p) = p$. For example, the median is a special case of a quantile when $p = 0.5$. If Y has cumulative distribution function (CDF) F, then the p-quantile ξ_p can be calculated as the inverse of the CDF at value p, denoted as $F^{-1}(p)$. Quantiles have wide applications in risk assessment of complex systems. For example, in financial portfolio management, quantiles are known as values at risk (VaRs), and analysts often measure risk and capital adequacy in terms of 0.99-quantiles. Another important use of quantiles occurs in safety analyses of nuclear power plants. The U.S. Nuclear Regulatory Commission requires plant licensees to demonstrate that their facilities satisfy a 95/95 criterion (U.S. Nuclear Regulatory Commission 1989), which means that the licensee has to establish with at least 95% confidence that the 0.95-quantile of the regulated output variable (for example, peak cladding temperature) falls below a mandated threshold. Hence, not only do we want a point estimator of the 0.95-quantile, but also an upper one-sided 95% confidence interval (CI) of the quantile so that if the upper confidence limit lies below the given threshold, the 95/95 criterion is satisfied.

One simulation method to estimate a quantile is crude Monte Carlo (CMC). With CMC, a sample of n independent and identically distributed (i.i.d.) outputs is generated, through which their CDF is estimated, and then a point estimator of the quantile is obtained by inverting the CDF estimator (Section 2.3 in Serfling 1980). Let F' denote the derivative (when it exists) of F. If $F'(\xi_p) > 0$, then the quantile estimator from CMC obeys a central limit theorem (CLT) (Section 2.3.3 in Serfling 1980), from which an asymptotically valid CI can be constructed if we have a consistent estimator of the CLT's asymptotic variance. Section 2.6.2 in Serfling (1980) describes one approach, using a finite difference, for consistently estimating the asymptotic variance. Instead of utilizing a CLT to build a CI, an alternative method exploits a binomial property to obtain an exact distribution-free CI based on order statistics (Section 2.6.1 of Serfling 1980).

In addition to CMC, there has been abundant work conducted to obtain more efficient quantile estimators by applying variance-reduction techniques (VRTs); see Chapter 4 in Glasserman (2004) for an overview of VRTs to estimate a mean. VRTs that have been used for quantile estimation include importance sampling (IS) (Glynn 1996, Glasserman, Heidelberger, and Shahabuddin 2000, Liu and Yang 2012); control variates (CV) (Hsu and Nelson 1990, Hesterberg and Nelson 1998); and correlation-induction methods, including Latin hypercube sampling (LHS) and antithetic variates (AV) (Avramidis and Wilson 1998, Jin, Fu, and Xiong 2003). However, the proof techniques employed to establish the validity of quantile CIs when applying CMC do not work if VRTs are used. To address this issue, Chu and Nakayama (2012) develop a general approach for building a CI when using several VRTs, including IS, CV, AV and combined IS and stratified sampling. In this paper, we study LHS.

Latin hypercube sampling was developed by McKay, Conover, and Beckman (1979) as a way of applying stratification in higher dimensions. Stein (1987) analyzes the asymptotic variance of the LHS estimator of an expectation, and shows that it is no larger than its counterpart from CMC. Owen (1992) proves a CLT holds for an LHS estimator of a mean of a bounded output, and Loh (1996) extends the CLT for the LHS estimator of a mean when the output's absolute third moment is finite. Avramidis and Wilson (1998) use LHS to obtain a point estimator of a quantile, and prove it satisfies a CLT. Helton and Davis (2003) survey the extensive literature on LHS, discuss its applications, and compare LHS and CMC.

However, the papers described above do not include methods to construct asymptotically valid CIs for quantiles under LHS. Recently, Nakayama (2011, 2012) proposed approaches to estimate CIs for quantiles by replicated LHS (rLHS). In this paper, we develop new methods for constructing CIs for quantiles through rLHS that give better results than those in Nakayama (2011, 2012). In the new approaches, we generate $b \geq 2$ independent LHS samples, each of size m, and we compute the rLHS quantile estimator by inverting the CDF estimator from all $n = bm$ observations. One method we consider is *batching* (Glasserman 2004, p. 491), which treats each replicated LHS sample as a batch. The other is *sectioning* (Section III.5a in Asmussen and Glynn 2007 and Nakayama 2014), which takes the batching CI and replaces the batching point estimator with the overall point estimator throughout. Since quantile estimators are generally biased with the bias vanishing as the sample size increases (Avramidis and Wilson 1998), one advantage of sectioning over batching is that sectioning's CI is centered at a typically less-biased point estimator, thus resulting in the sectioning CI having better coverage. The CIs are asymptotically valid as $m \to \infty$ with $b \geq 2$ fixed. Our proof of the asymptotic validity of the sectioning CI depends on showing that the quantile estimators from single-sample LHS (ssLHS) and rLHS each satisfy so-called (weak) Bahadur representations (Bahadur 1966, Ghosh 1971), which we have established in Dong and Nakayama (2014).

Nakayama (2012) proposes batching and sectioning methods for a different type of rLHS, in which the number r of independent LHS samples grows large; there are $b \geq 2$ batches, each batch contains independent r/b LHS samples, and each LHS sample has fixed size m. Thus, the total number of observations is $n = b[(r/b)m)] = rm$. Because they have multiple replicated LHS per batch, we refer to these as *mrLHS* methods. Nakayama (2012) establishes the asymptotic validity of batching and sectioning for mrLHS as $r \to \infty$ with m and $b \geq 2$ fixed. Compared with that, the techniques in our paper have just a single LHS sample in each batch, with LHS samples independent across the batches. Here each of the b batches' single LHS sample has size m, where m grows large, so the overall number of observations is $n = bm$. We call these *single replicated LHS (srLHS)* methods, and we establish the asymptotic validity of srLHS batching and sectioning as $m \to \infty$ with $b \geq 2$ fixed. For a fixed overall sample size n and number b of batches, srLHS allows the size m of each LHS sample to be larger than for mrLHS with $r > b$, which leads to srLHS reducing variance more. We present numerical results in Section 5 that demonstrate this.

The rest of the paper is organized as follows. In Section 2, methods for estimating a quantile and constructing a quantile CI under CMC are reviewed. Section 3 describes how to generate outputs using LHS. In Section 4, we present detailed steps of srLHS. In Section 5, we give numerical results of coverage experiments using different approaches to generate the outputs, including CMC, mrLHS and srLHS. Section 6 provides concluding remarks and subsequent research topics. The formal mathematical proofs are contained in Dong and Nakayama (2014).

2 BACKGROUND

Consider a random variable Y that can be expressed as

$$Y = g(U_1, U_2, \ldots, U_d) \tag{1}$$

where $g : \Re^d \to \Re^1$ is a deterministic function and inputs $U_1, U_2, \ldots, U_d$ are i.i.d. unif$[0,1]$ random variables. Let F be the CDF of Y, and define $\xi_p = F^{-1}(p) \equiv \inf\{y : F(y) \geq p\}$ as the p-quantile of F (or equivalently of Y). The goal is to estimate ξ_p and construct a $100(1-\alpha)\%$ CI for it through a sample of Y with size n. The framework in (1) covers the special setting of

$$Y = c(X_1, X_2, \ldots, X_d) \tag{2}$$

with $c : \Re^d \to \Re^1$ and $X_1, X_2, \ldots, X_d$ being arbitrary input random variables that are independent but not necessarily identically distributed. In this case, the random variables $X_1, X_2, \ldots, X_d$ can be transformed from i.i.d. unif$[0,1]$ random variables $U_1, U_2, \ldots, U_d$ through $X_j = H_j^{-1}(U_j)$ for $j = 1, 2, \ldots, d$, where H_j is the marginal CDF of X_j. Therefore, we have

$$Y = c(H^{-1}(U_1), H_2^{-1}(U_2), \ldots, H_d^{-1}(U_d)) = g(U_1, U_2, \ldots, U_d).$$

2.1 Constructing a Confidence Interval When Applying CMC

Let $\mathbf{U} = (U_1, U_2, \ldots, U_d)$ denote a vector of d i.i.d. unif[0,1] random variable inputs and $\mathbf{u} = (u_1, u_2, \ldots, u_d)$ be a realization of $\mathbf{U}$. In the case of CMC, we first generate n i.i.d. input vectors $\mathbf{U}_i = (U_{i,1}, U_{i,2}, \cdots, U_{i,d})$, $i = 1, 2, \ldots, n$, which are arranged as:

$$\begin{matrix} U_{1,1} & U_{1,2} & \cdots & U_{1,d} \\ U_{2,1} & U_{2,2} & \cdots & U_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ U_{n,1} & U_{n,2} & \cdots & U_{n,d} \end{matrix},$$

and then apply function g from (1) to each row to generate a sample of Y with size n:

$$\begin{aligned} Y_1 &= g(U_{1,1}, U_{1,2}, \ldots, U_{1,d}), \\ Y_2 &= g(U_{2,1}, U_{2,2}, \ldots, U_{2,d}), \\ &\vdots \\ Y_n &= g(U_{n,1}, U_{n,2}, \ldots, U_{n,d}). \end{aligned}$$

Define F_n to be the CMC estimator of the CDF F, where

$$F_n(y) = \frac{1}{n} \sum_{1 \leq i \leq n} I(Y_i \leq y)$$

and $I(\cdot)$ denotes the indicator function, which equals 1 (resp., 0) when its argument is true (resp., false). The CMC point estimator of the p-quantile $\xi_p = F^{-1}(p)$ is given by

$$\xi_{p,n} = F_n^{-1}(p). \tag{3}$$

An equivalent way to calculate $F_n^{-1}(p)$ is to sort the values $Y_1, Y_2, \ldots, Y_n$ in non-decreasing order. Let $Y_{n:1} \leq Y_{n:2} \leq \cdots \leq Y_{n:n}$ be the ordered statistics, and then $\xi_{p,n} = Y_{n:\lceil np \rceil}$, where $\lceil np \rceil$ is the smallest integer that is greater than or equal to np.

Section 2.3.3 in Serfling (1980) proves that if $F'(\xi_p) > 0$, where F' denotes the derivative (when it exists) of F, then the CMC p-quantile estimator $\xi_{p,n}$ satisfies a CLT

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

as $n \to \infty$, where $\tau_p = \sqrt{p(1-p)}/F'(\xi_p)$, $N(a, b^2)$ is a normal random variable with mean a and variance b^2, and $\Rightarrow$ denotes convergence in distribution (e.g., see Section 1.2.4 of Serfling 1980). From the CLT, an asymptotic $100(1-\alpha)\%$ confidence interval $C_{n,\mathrm{CMC}}$ can be constructed as follows:

$$C_{n,\mathrm{CMC}} = [\xi_{p,n} \pm z_{\alpha/2} \tau_p / \sqrt{n}],$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution, i.e., $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ with Φ the CDF of a $N(0,1)$ random variable.

For most cases, the value of $F'(\xi_p)$ is not known, so it has to be estimated as well. For CMC, the statistic literature contains several consistent estimators for τ_p. For example, Bloch and Gastwirth (1968) and Bofinger (1975) use a finite-difference method to estimate $\lambda_p \equiv 1/F'(\xi_p)$. By noting that

$$\frac{d}{dp}F^{-1}(p) = \lim_{h \to 0}[F^{-1}(p+h) - F^{-1}(p-h)]/(2h) = \frac{1}{F'(F^{-1}(p))} = \frac{1}{F'(\xi_p)} = \lambda_p,$$

we obtain a natural estimator of λ_p as

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

where $h_n > 0$ is a user-specified *bandwidth*. Then an estimator of τ_p is $\tau_{p,n} = \sqrt{p(1-p)}\lambda_{p,n}$, and when $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, we have $\tau_{p,n} \Rightarrow \tau_p$ as $n \to \infty$. Therefore, a two-sided $100(1-\alpha)\%$ confidence interval for ξ_p is

$$\hat{C}_{n,\mathrm{CMC}} = [\xi_{p,n} \pm z_{\alpha/2} \tau_{p,n} / \sqrt{n}],$$

which is asymptotically valid in the sense that $P(\xi_p \in \hat{C}_{n,\mathrm{CMC}}) \to 1-\alpha$ as $n \to \infty$.

2.2 Batching and Sectioning When Using CMC

The quality of the estimator $\tau_{p,n}$ depends on the choice of the bandwidth h_n, but specifying an appropriate value for h_n is nontrivial in practice; e.g., see Bofinger (1975), Hall and Sheather (1988) and Chu and Nakayama (2012). We instead consider alternative methods — batching and sectioning — that do not attempt to consistently estimate τ_p. While these approaches still require the user to specify one parameter, the number b of batches, selecting a reasonable choice for b appears simpler than determining a bandwidth that works well.

In batching, the overall sample $Y_1, Y_2, \ldots, Y_n$, of size n is first divided into $b \geq 2$ batches, each of which contains a sample of size $m = n/b$. Let $Y_{j,i} = Y_{(j-1)m+i}$ for $j = 1, 2, \ldots, b$, and $i = 1, 2, \ldots, m$, be the ith output in the jth batch, and define a CDF estimator for each batch j, $j = 1, 2, \ldots, b$, as

$$F_{j,m}(y) = \frac{1}{m}\sum_{1 \leq i \leq m} I(Y_{j,i} \leq y).$$

The corresponding p-quantile estimator for the jth batch is $\xi_{p,j,m} = F_{j,m}^{-1}(p)$. We define the CMC batching p-quantile estimator as

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

The sample variance of $\xi_{p,j,m}$, $j = 1,2,\ldots,b$, is

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

Define the critical value $t_{b-1,\alpha/2} = G^{-1}(1-\alpha/2)$, with G being the CDF of a Student t random variable with $b-1$ degrees of freedom. The CMC batching two-sided $100(1-\alpha)\%$ confidence interval for ξ_p is given by

$$C_{b,m,\mathrm{batch}} = [\bar{\xi}_{p,b,m} \pm t_{b-1,\alpha/2} S_{b,m,\mathrm{batch}}/\sqrt{b}].$$

Quantile estimators are generally biased, with the bias vanishing as the sample size grows large (Avramidis and Wilson 1998). The bias of the CMC batching p-quantile estimator $\bar{\xi}_{p,b,m}$ is determined by the batch size $m = n/b < n$, so the batching CI is centered at an estimator that can have considerable bias. Thus, the batching CI can suffer from poor coverage, especially when the overall sample size n is not very large.

Similar to batching, sectioning addresses this issue by replacing the batching point estimator $\bar{\xi}_{p,b,m}$ in the batching CI with the overall quantile estimator $\xi_{p,n}$ using all n outputs, as defined in (3). Asmussen and Glynn (2007) (Section III.5a) develop sectioning for the CMC case and Nakayama (2014) extends it for importance sampling and control variates. Define

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

The CMC sectioning two-sided $100(1-\alpha)\%$ CI for ξ_p is then

$$C_{b,m,\mathrm{sec}} = [\xi_{p,n} \pm t_{b-1,\alpha/2} S_{b,m,\mathrm{sec}}/\sqrt{b}].$$

Both $C_{b,m,\mathrm{batch}}$ and $C_{b,m,\mathrm{sec}}$ are asymptotically valid CIs, i.e., $P(\xi_p \in C) \to 1-\alpha$ as $m \to \infty$ with $b \ge 2$ fixed, where $C \in \{C_{b,m,\mathrm{batch}}, C_{b,m,\mathrm{sec}}\}$. But the sectioning CI usually has better coverage when the overall sample size $n = bm$ is small. For both batching and sectioning, Section III.5a of Asmussen and Glynn (2007) suggests choosing the number of batches satisfying $b \le 30$. Nakayama (2014) provides numerical results for sectioning and batching with $b = 10$ and $b = 20$ when applying each of CMC, importance sampling and control variates, and found that $b = 10$ resulted in significantly better coverage when n is small, especially for $p \approx 1$.

3 LATIN HYPERCUBE SAMPLING

In this section, we describe the method to generate a single Latin hypercube sample, and establish important properties of it as described in Theorem 1 below. Let

$$\begin{matrix} U_{1,1} & U_{1,2} & \cdots & U_{1,d} \\ U_{2,1} & U_{2,2} & \cdots & U_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ U_{s,1} & U_{s,2} & \cdots & U_{s,d} \end{matrix},$$

be $s \times d$ i.i.d. unif$[0,1]$ random variables. Choose d independent random permutations of $(1,2,\dots,s)$, denoted as $\pi_1,\pi_2,\dots,\pi_d$, where $\pi_j = (\pi_j(1),\pi_j(2),\dots,\pi_j(s))$. For $i = 1,2,\dots,s$, and $j = 1,2,\dots,d$, define

$$V_{i,j} = \frac{\pi_j(i) - 1 + U_{i,j}}{s},$$

which we arrange as

$$\begin{matrix} V_{1,1} & V_{1,2} & \cdots & V_{1,d} \\ V_{2,1} & V_{2,2} & \cdots & V_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ V_{s,1} & V_{s,2} & \cdots & V_{s,d} \end{matrix} . \tag{5}$$

Then, by applying function $g : \Re^d \to \Re^1$ from (1) to each row of (5), we have

$$\begin{aligned} \hat{Y}_1 &= g(V_{1,1}, V_{1,2}, \cdots, V_{1,d}), \\ \hat{Y}_2 &= g(V_{2,1}, V_{2,2}, \cdots, V_{2,d}), \\ &\vdots \\ \hat{Y}_s &= g(V_{n,1}, V_{n,2}, \cdots, V_{n,d}). \end{aligned} \tag{6}$$

Outputs $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_s$ are called a Latin hypercube sample of size s. Because each row i in (5) has d i.i.d. unif[0,1] random variables, we have $\hat{Y}_i \sim F$ for $i = 1,2,\dots,s$. But since each column j in (5) uses the same permutation π_j, the rows in (5) are dependent, thus making $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_s$ dependent as well.

Compared with CMC, LHS yields smaller asymptotic variances for an expectation estimator (Stein 1987) and also for a quantile estimator (Avramidis and Wilson 1998). McKay, Conover, and Beckman (1979) give an explicit form of the covariance among the outputs of a Latin hypercube sample. Under the conditions given below, Avramidis and Wilson (1998) prove a CLT for a quantile estimator using a single-sample Latin hypercube sample (ssLHS). In this paper, we further establish that the ssLHS quantile estimator satisfies a weak Bahadur (1966) representation (Ghosh 1971), as stated in Theorem 1 below. A Bahadur representation shows that a quantile estimator can be approximated as the true quantile plus a linear transformation of a CDF estimator evaluated at the true quantile, with a remainder term that vanishes at some rate as the sample size grows large.

To obtain our results, we first impose two continuity conditions from Avramidis and Wilson (1998):

CC1 *The function $g(\cdot)$ in (1) has a finite set of discontinuities $\mathscr{D}$.*

CC2 *There exists a neighborhood $\mathscr{N}(\xi_p)$ of ξ_p such that for each $x \in \mathscr{N}(\xi_p)$ and for each $j = 1,\dots,d$, there exists a finite set $\mathscr{Q}_j(x)$ such that*

$$P(g(\mathbf{U}) = x | U_j = u_j) = 0$$

for every $u_j \in [0,1] - \mathscr{Q}_j(x)$.

Define the ssLHS CDF estimator based on a single LHS sample of size $s = n$ as

$$\hat{F}_n(y) = \frac{1}{n} \sum_{1 \le i \le n} I(\hat{Y}_i \le y) \tag{7}$$

and let $\hat{\xi}_{p,n} = \hat{F}_n^{-1}(p)$ be the ssLHS p-quantile estimator. We then have the following result:

Theorem 1 Suppose continuity conditions CC1 and CC2 hold, CDF F has a bounded second derivative in a neighborhood of ξ_p, and $F'(\xi_p) > 0$. Then the ssLHS quantile estimator $\hat{\xi}_{p,n}$ with sample size n satisfies the weak Bahadur representation

$$\hat{\xi}_{p,n} = \xi_p - \frac{\hat{F}_n(\xi_p) - p}{F'(\xi_p)} + \hat{R}_n, \quad \text{with } n^{1/2}\hat{R}_n \Rightarrow 0 \text{ as } n \to \infty, \tag{8}$$

where $\hat{F}_n(\cdot)$ is defined in (7).

Chu and Nakayama (2012) establish a set of three technical conditions on any CDF estimator that ensure the corresponding quantile estimator satisfies a weak Bahadur representation holds, and our proof of Theorem 1 shows the three conditions hold for the ssLHS CDF estimator $\hat{F}_n$ in (7). Establishing the three conditions for $\hat{F}_n$ is complicated by the fact that $\hat{Y}_i$, $i = 1, 2, \ldots, n$, are dependent, and we prove Theorem 1 by modifying and extending some of the arguments that Avramidis and Wilson (1998) develop to prove the ssLHS quantile estimator $\hat{\xi}_{p,n}$ obeys a CLT. The full details are provided in Dong and Nakayama (2014).

4 CONFIDENCE INTERVAL ESTIMATION FOR A QUANTILE WHEN APPLYING LHS

Avramidis and Wilson (1998) prove that under the conditions in Theorem 1, the ssLHS quantile estimator $\hat{\xi}_{p,n}$ satisfies the CLT

$$\frac{\sqrt{n}}{\eta_p}(\hat{\xi}_{p,n} - \xi_p) \Rightarrow N(0,1) \tag{9}$$

as $n \to \infty$, where $0 < \eta_p < \infty$ is a constant. Define $0 < \psi_p < \infty$ such that $\sqrt{n}[\hat{F}_n(\xi_p) - p] \Rightarrow N(0, \psi_p^2)$ as $n \to \infty$. Hence, by (8), we can see that

$$\eta_p = \psi_p / F'(\xi_p). \tag{10}$$

From the CLT (9), we can obtain an ssLHS CI for ξ_p as

$$C_{n,\mathrm{ssLHS}} = [\hat{\xi}_{p,n} \pm z_{\alpha/2}\eta_p/\sqrt{n}].$$

However, the asymptotic variance η_p^2 in (9) is not known nor easily estimated (Glasserman 2004, p. 242), making it difficult to construct a confidence interval directly using ssLHS. In this paper we instead develop CIs for ξ_p using replicated LHS with batching or sectioning.

Let $b \geq 2$ be the number of batches. For each batch, a single LHS sample of size m is generated by letting $s = m$ in Section 3, and LHS samples across batches are generated independently. The overall sample size across all batches is $n = bm$. Because a single LHS sample is used in each batch, we refer to this method as single rLHS method (srLHS) to distinguish it from the approach proposed by Nakayama (2012), where the total number r of independent LHS samples across all batches grows large with a fixed size m for each LHS sample and a fixed number b of batches. We refer to the latter as the multiple rLHS (mrLHS) method, and batching and sectioning with mrLHS are asymptotically valid when $r \to \infty$ with b and m fixed.

For srLHS, generate b independent Latin hypercube samples as in (6), each of which has size $s = m$, and let $\tilde{Y}_{j,i}$, $i = 1, 2, \ldots, m$, be the m outputs in the jth batch, $j = 1, 2, \ldots, b$. Since the batches are independent, $\tilde{Y}_{j,i}$ and $\tilde{Y}_{j',i'}$ are independent for $j \neq j'$, but $\tilde{Y}_{j,i}$ and $\tilde{Y}_{j,i'}$ are dependent since they are in the same batch j. Then we define $\check{F}_{b,m}$ as the srLHS CDF estimator, with

$$\check{F}_{b,m}(y) = \frac{1}{b}\sum_{1 \leq j \leq b} \frac{1}{m} \sum_{1 \leq i \leq m} I(\tilde{Y}_{j,i} \leq y), \tag{11}$$

which is based on all $n = bm$ outputs, and let

$$\check{\xi}_{p,b,m} = \check{F}_{b,m}^{-1}(p)$$

be the srLHS overall p-quantile estimator. Let $\tilde{F}_{j,m}$ be the CDF estimator from batch j, $j = 1, 2, \ldots, b$, i.e., $\tilde{F}_{j,m}(y) = \frac{1}{m}\sum_{1 \leq i \leq m} I(\tilde{Y}_{j,i} \leq y)$. The corresponding p-quantile estimator from each batch j is $\tilde{\xi}_{p,j,m} = \tilde{F}_{j,m}^{-1}(p)$. Similar to the CMC batching method, define the srLHS batching p-quantile estimator

$$\bar{\xi}_{p,b,m} = \frac{1}{b}\sum_{1 \leq j \leq b} \tilde{\xi}_{p,j,m}.$$

The sample variance of $\tilde{\xi}_{p,j,m},\ j=1,2,\ldots,b$, is

$$\check{S}^2_{b,m,\text{batch}} = \frac{1}{b-1}\sum_{1\le j\le b}(\tilde{\xi}_{p,j,m} - \acute{\xi}_{p,b,m})^2. \tag{12}$$

The two-sided $100(1-\alpha)\%$ CI for ξ_p using srLHS batching is

$$\check{C}_{b,m,\text{batch}} = [\acute{\xi}_{p,b,m} \pm t_{b-1,\alpha/2}\check{S}_{b,m,\text{batch}}/\sqrt{b}]. \tag{13}$$

As with CMC, when the overall sample size $n=bm$ is not very large, the srLHS batching p-quantile estimator $\acute{\xi}_{p,b,m}$ can be significantly biased as the bias is determined by the batch size $m=n/b<n$. Hence, the srLHS batching CI in (13) may suffer from poor coverage since it is centered at a poor point estimator.

We can again address this issue by applying sectioning. In this approach, the batching estimator $\acute{\xi}_{p,b,m}$ is replaced with the overall quantile estimator $\check{\xi}_{p,b,m} = \check{F}^{-1}_{b,m}(p)$ for $\check{F}_{b,m}$ defined in (11), and instead of (12), we compute

$$\check{S}^2_{b,m,\text{sec}} = \frac{1}{b-1}\sum_{1\le j\le b}(\tilde{\xi}_{p,j,m} - \check{\xi}_{p,b,m})^2.$$

Then the two-sided $100(1-\alpha)\%$ CI for ξ_p using srLHS sectioning is

$$\check{C}_{b,m,\text{sec}} = [\check{\xi}_{p,b,m} \pm t_{b-1,\alpha/2}\check{S}_{b,m,\text{sec}}/\sqrt{b}]. \tag{14}$$

Since the srLHS sectioning CI is centered at a typically less-biased point estimator than the srLHS batching CI in (13), the srLHS sectioning CI should have better coverage when $n=bm$ is small.

To establish the asymptotic validity of the srLHS sectioning method, we first prove that the srLHS overall p-quantile estimator $\check{\xi}_{p,b,m}$ satisfies a weak Bahadur representation, which is given next. The result is established by exploiting Theorem 1 and the independence of the LHS samples across batches.

Theorem 2 Under the same conditions as in Theorem 1, the srLHS overall quantile estimator $\check{\xi}_{p,b,m}$ based on rLHS with $b\ge 2$ independent LHS samples, each of size m, satisfies

$$\check{\xi}_{p,b,m} = \xi_p - \frac{\check{F}_{b,m}(\xi_p)-p}{F'(\xi_p)} + \check{R}_{b,m} \quad \text{with } n^{1/2}\check{R}_{b,m} = b^{1/2}m^{1/2}\check{R}_{b,m} \Rightarrow 0$$

as $m\to\infty$ with $b\ge 2$ fixed.

The next theorem further establishes the asymptotic validity of the srLHS batching and sectioning CIs.

Theorem 3 Under the same conditions as in Theorem 1,

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

as $m\to\infty$ with $b\ge 2$ fixed, where $C\in\{\check{C}_{b,m,\text{batch}},\check{C}_{b,m,\text{sec}}\}$.

The asymptotic validity of the srLHS batching CI can be established as follows. For each batch $j=1,2,\ldots,b$, its quantile estimator $\tilde{\xi}_{p,j,m}$ is approximately normally distributed since it is based on a single LHS sample of size $m\to\infty$ and Avramidis and Wilson (1998) prove that ssLHS quantile estimators satisfy a CLT. The independence of the batches then ensures the validity of the batching CI. The keys to proving the asymptotic validity of the srLHS sectioning CI are the Bahadur representations in Theorems 1 and 2. These enable us to show that the srLHS batching point estimator $\acute{\xi}_{p,b,m}$ and the srLHS overall quantile estimator $\check{\xi}_{p,b,m}$ are sufficiently close to allow replacing $\acute{\xi}_{p,b,m}$ in the batching CI with $\check{\xi}_{p,b,m}$ and still retain an asymptotically valid CI. The full details are provided in Dong and Nakayama (2014).

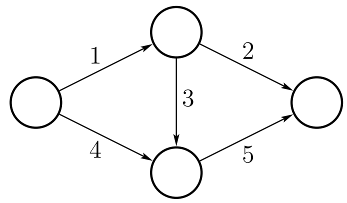

Figure 1: A stochastic activity network.

The main difference of our srLHS methods and the mrLHS approaches in Nakayama (2012) is that the mrLHS techniques require the number of LHS samples per batch to grow to infinity with each LHS sample size remaining fixed, while the srLHS methods require only one Latin hypercube sample that grows large in each batch. If the total sample size n is fixed, the size of each independent LHS in srLHS can be much bigger than that in mrLHS, and this can result in better variance reduction. The reason for this is that Stein (1987) shows that a ssLHS estimator has asymptotic variance that decreases at rate c/m, where m is the size of the single Latin hypercube sample and c is a constant. Thus, the larger LHS sample sizes in srLHS result in smaller variance. We will compare these two approaches through numerical experiments in Section 5.

5 NUMERICAL RESULTS

The asymptotic validity of the confidence intervals estimated by the batching and sectioning methods is tested on a stochastic activity network (SAN), which has been previously studied in Hsu and Nelson (1990). The network models the completion time of a project consisting of a group of activities with precedence constraints and random durations. Let $d = 5$ be the number of activities in the project, whose graph is shown in Figure 1. There are three paths in the network, and the completion time of the project is the longest one. For $1 \leq j \leq d$, let A_j be the duration of the jth activity of the project. All A_j, $1 \leq j \leq d$, follow an exponential distribution with mean 1, and are mutually independent. The network has $q = 3$ paths, denoted as $B_1 = \{1,2\}$, $B_2 = \{1,3,5\}$, $B_3 = \{4,5\}$. Let $Y = \max_{1\leq \ell \leq q} \sum_{j\in B_\ell} A_j$ be the output of the experiment, i.e., the completion time of the project. Our experiments estimate the p-quantile of Y for $p = 0.8$ and $p = 0.95$ and construct nominal 90% confidence intervals using either batching or sectioning.

We used $b = 10$ independent batches in our experiment. (Experiments in Nakayama 2014 with $b = 10$ and $b = 20$ using CMC, importance sampling and control variates showed $b = 20$ often results in poorer coverage than $b = 10$ for the same overall sample size n, especially when n is small or $p \approx 1$.) We compare three simulation approaches to generate outputs: CMC (Section 2.2), mrLHS (Nakayama 2012), and srLHS (Section 4). In srLHS, we generate a single LHS sample of size m for each of the b batches, so the total size across all batches is $n = bm$. We varied the total sample size $n = 4^k \times 100$ for $k = 0,1,2,3$. As the total size n grows large, the size of each LHS m grows large since b is fixed. In the experiments with mrLHS, we fixed the size of each LHS at $m = 10$, so as n grows large, the number of LHS samples in each batch grows large but the size of each LHS sample remains unchanged at $m = 10$. Thus, when $n = 100$, the srLHS and mrLHS approaches are identical, but they differ as n increases. The CDF of the project-completion time Y is (Hsu and Nelson 1990)

$$F(y) = 1 + (3 - 3y - y^2/2)e^{-y} + (-3 - 3y + y^2/2)e^{-2y} - e^{-3y} \tag{15}$$

for $y \geq 0$. Its density $F'(y)$ is positive and its second derivative is bounded. Using (15) we numerically calculate the true 0.8-quantile and 0.95-quantile as $\xi_{0.8} = 4.7145$ and $\xi_{0.95} = 6.6645$, and use them in the coverage experiments. We ran 10^3 independent replications to estimate the coverage $P(\xi_p \in C)$ of a CI C.

From the results in Table 1, we see that the srLHS method has the smallest average half-width, followed by the mrLHS approach, and then CMC. When $n = 100$, the performance of srLHS and mrLHS

Table 1: Coverages (and average half-widths) of batching and sectioning methods for nominal 90% CIs using CMC, mrLHS ($b = 10$, $m = 10$) and srLHS ($b = 10$, $m = n/b$) are given.

n	p = 0.8					
	CMC		mrLHS		srLHS	
	Batching	Sectioning	Batching	Sectioning	Batching	Sectioning
100	0.644 (0.477)	0.885 (0.517)	0.534 (0.337)	0.886 (0.383)	0.559 (0.340)	0.915 (0.386)
400	0.835 (0.260)	0.910 (0.267)	0.807 (0.189)	0.897 (0.197)	0.745 (0.159)	0.903 (0.168)
1600	0.887 (0.134)	0.898 (0.136)	0.900 (0.097)	0.919 (0.099)	0.871 (0.081)	0.903 (0.083)
6400	0.892 (0.067)	0.896 (0.068)	0.904 (0.049)	0.903 (0.049)	0.895 (0.041)	0.903 (0.041)

n	p = 0.95					
	CMC		mrLHS		srLHS	
	Batching	Sectioning	Batching	Sectioning	Batching	Sectioning
100	0.887 (0.926)	0.893 (0.952)	0.912 (0.825)	0.883 (0.852)	0.882 (0.834)	0.881 (0.862)
400	0.699 (0.461)	0.901 (0.500)	0.667 (0.402)	0.889 (0.436)	0.638 (0.347)	0.880 (0.382)
1600	0.823 (0.253)	0.887 (0.260)	0.844 (0.222)	0.901 (0.229)	0.814 (0.176)	0.902 (0.184)
6400	0.878 (0.128)	0.886 (0.129)	0.874 (0.114)	0.894 (0.115)	0.867 (0.088)	0.901 (0.089)

are comparable as the two techniques are identical in this case, and as n grows, the advantage of srLHS over mrLHS starts to show and increases with n. For example, when $n = 400$, for the case of $p = 0.95$, the average half width from srLHS with sectioning is about 12% smaller than that from mrLHS; when $n = 1600$, the improvement is approximately 20%, and when $n = 6400$, the difference is roughly 23%. The widening gap of srLHS over mrLHS results from the LHS sample size in srLHS growing as n increases, whereas the LHS sample size in mrLHS remains fixed. We also notice that in general, sectioning leads to better coverage than batching, which occurs because the sectioning CI is centered at a less-biased point estimator. The coverage of all three approaches under both batching and sectioning converge to about 90% when the overall sample size n gets large.

6 CONCLUDING REMARKS

In this paper, we develop methods to construct CIs for quantiles when applying rLHS. We employ batching and sectioning to build the CIs. Because of the bias of quantile estimators, the sectioning CI typically has better coverage than the batching CI since the former is centered at a less-biased point estimator. The proposed srLHS methods improve on previous mrLHS methods of batching and sectioning with multiple replicated Latin hypercube samples by Nakayama (2012), and use only one Latin hypercube sample in each batch. To establish the asymptotic validity of the srLHS methods, we have proven weak Bahadur representations hold for quantile estimators for both a single Latin hypercube sample and replicated Latin hypercube samples, where the sizes of the LHS samples increase.

The asymptotic regimes of srLHS and mrLHS differ: srLHS has the batch size $m \to \infty$ with the number $b \geq 2$ of batches fixed, whereas mrLHS requires the number r of independent LHS samples to satisfy

$r \to \infty$, with fixed batch size m and fixed number b of batches. Thus, the asymptotic validity of srLHS covers the setting of large LHS sample sizes, whereas mrLHS only allows fixed LHS sample sizes but the number of LHS samples increases. As a consequence, for a given overall sample size, srLHS permits larger LHS sample sizes than mrLHS, which results in srLHS reducing variance more, as we observed in the numerical results.

We are currently investigating methods for constructing a CI for a quantile when applying ssLHS. One approach for doing this is to develop a consistent estimator for the asymptotic variance η_p^2 in the ssLHS CLT in (9) by separately estimating the numerator and denominator in (10). We can consistently estimate $\lambda_p = 1/F'(\xi_p)$ via a finite difference as in (4) using the results of Chu and Nakayama (2012). The square of the numerator ψ_p in (10) is the asymptotic variance of the ssLHS CDF estimator evaluated at ξ_p. Owen (1992) develops a consistent estimator for the variance of an ssLHS estimator of a mean, and we are working on extending this to estimate ψ_p^2.

Acknowledgments

This work has been supported in part by the National Science Foundation under Grants No. 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.

REFERENCES

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

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).

Dong, H., and M. K. Nakayama. 2014. "Quantile Estimation Using Latin Hypercube Sampling". working draft.

Ghosh, J. K. 1971. "A New Proof of the Bahadur Representation of Quantiles and an Application". *Annals of Mathematical Statistics* 42:1957–1961.

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*, 180–185: Publishing House of St. Petersburg University, St. Petersburg, Russia.

Hall, P., and S. J. Sheather. 1988. "On the Distribution of a Studentized Quantile". *Journal of the Royal Statistical Society B* 50:381–391.

Helton, J. C., and F. J. Davis. 2003. "Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems". *Reliability Engineering & System Safety* 81:23–69.

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.

Liu, J., and X. Yang. 2012. "The Convergence Rate and Asymptotic Distribution of Bootstrap Quantile Variance Estimator for Importance Sampling". *Advances in Applied Probability* 44:815–841.

Loh, W.-L. 1996. "On Latin Hypercube Sampling". *Annals of Statistics* 24:2058–2080.

McKay, M. D., W. J. Conover, and R. J. Beckman. 1979. "A Comparison of Three Methods for Selecting Input Variables in the Analysis of Output from a Computer Code". *Technometrics* 21:239–245.

Nakayama, M. K. 2011. "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 for Quantiles When Applying Replicated Latin Hypercube Sampling and Sectioning". In *Proceedings of the 2012 Autumn Simulation Conference*.

Nakayama, M. K. 2014. "Confidence Intervals Using Sectioning for Quantiles When Applying Variance-Reduction Techniques". *ACM Transactions on Modeling and Computer Simulation*. To appear.

Owen, A. B. 1992. "A Central Limit Theorem for Latin Hypercube Sampling". *Journal of the Royal Statistical Society B* 54:541–551.

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

Stein, M. 1987. "Large Sample Properties of Simulations Using Latin Hypercube Sampling". *Technometrics* 29:143–151. correction 32:367.

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.

AUTHOR BIOGRAPHY

Hui Dong is currently a research scientist in operations at Amazon Corporate LLC. She received her Ph.D. in Management from Rutgers University. Her research interests include simulation, optimization and their applications in risk evaluation, supply chain and general operations management. She has served as a reviewer for *Annals of Operations Research* and *Winter Simulation Conference*. Her email address is huidong@rutgers.edu.

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 Simulation Area Editor for the *INFORMS Journal on Computing* and an associate editor for *ACM Transactions on Modeling and Computer Simulation*. His research interests include simulation, modeling, statistics, risk analysis, and energy. His email address is marvin@njit.edu.