A Tutorial on Quantile Estimation via Monte Carlo

Hui Dong and Marvin K. Nakayama

Abstract Quantiles are frequently used to assess risk in a wide spectrum of application areas, such as finance, nuclear engineering, and service industries. This tutorial discusses Monte Carlo simulation methods for estimating a quantile, also known as a percentile or value-at-risk, where p of a distribution's mass lies below its p-quantile. We describe a general approach that is often followed to construct quantile estimators, and show how it applies when employing naive Monte Carlo or variance-reduction techniques. We review some large-sample properties of quantile estimators. We also describe procedures for building a confidence interval for a quantile, which provides a measure of the sampling error.

1 Introduction

Numerous application settings have adopted quantiles as a way of measuring risk. For a fixed constant $0 < p < 1$, the p-quantile of a continuous random variable is a constant ξ such that p of the distribution's mass lies below ξ. For example, the median is the 0.5-quantile. In finance, a quantile is called a *value-at-risk*, and risk managers commonly employ p-quantiles for $p \approx 1$ (e.g., $p = 0.99$ or $p = 0.999$) to help determine capital levels needed to be able to cover future large losses with high probability; e.g., see [33].

Nuclear engineers use 0.95-quantiles in *probabilistic safety assessments* (PSAs) of nuclear power plants. PSAs are often performed with Monte Carlo, and the U.S. Nuclear Regulatory Commission (NRC) further requires that a PSA accounts for the

Hui Dong
Amazon.com Corporate LLC*, Seattle, WA 98109, USA e-mail: huidong@amazon.com *This work is not related to Amazon, regardless of the affiliation.

Marvin K. Nakayama
New Jersey Institute of Technology, Computer Science Department, Newark, NJ 07102, USA e-mail: marvin@njit.edu

Monte Carlo sampling error; e.g., see [50], Section 3.2 of [49], and Section 24.9 of [51]. This can be accomplished by providing a confidence interval for ξ.

Quantiles also arise as risk measures in service industries. For out-of-hospital patient care, a 0.9-quantile is commonly employed to assess response times of emergency vehicles and times to transport patients to hospitals [5]. In addition, [20] examines the 0.9-quantile of customer waiting times at a call center.

This tutorial discusses various Monte Carlo methods for estimating a quantile. Section 2 lays out the mathematical setting. In Section 3 we outline a general approach for quantile estimation via Monte Carlo, and illustrate it for the special case of naive Monte Carlo (NMC). We examine large-sample properties of quantile estimators in Section 4. Section 5 shows how the basic procedure in Section 3 can also be used when employing *variance-reduction techniques* (VRTs), which can produce quantile estimators with smaller sampling error than when NMC is applied. We describe different methods for constructing confidence intervals for ξ in Section 6.

2 Mathematical Framework

Consider the following example, which we will revisit throughout the paper to help illustrate ideas and notation. The particular stochastic model in the example turns out to be simple enough that it can actually be solved through a combination of analytical and numerical methods, making Monte Carlo simulation unnecessary. But the tractability allows us to compute exact quantiles, which are useful for our numerical studies in Sections 5.7 and 6.4 comparing different Monte Carlo methods. Larger, more complicated versions of the model are usually analytically intractable.

Example 1 (Stochastic activity network (SAN)). A contractor is preparing a bid to work on a project, such as developing a software product, or constructing a building. She wants to determine a time ξ to use as the bid's promised completion date so that there is a high probability of finishing the project by ξ to avoid incurring a penalty. To try to figure out such a ξ, she builds a stochastic model of the project's duration.

The project consists of d activities, numbered $1,2,\ldots,d$. Certain activities must be completed before others can start, e.g., building permits must be secured prior to laying the foundation. Figure 1, which has been previously studied in [47, 29, 13, 15], presents a directed graph that specifies the precedence constraints of a project with $d = 5$ activities. The nodes in the graph represent particular epochs in time, and edges denote activities. For a given node v, all activities corresponding to edges into v must be completed before starting any of the activities for edges out of v. Hence, activity 1 must finish before beginning activities 2 and 3. Also, activity 5 can commence only after activities 3 and 4 are done.

For each $j = 1,2,\ldots,d$, activity j has a random duration X_j, which is the length of edge j and has marginal *cumulative distribution function* (CDF) G_j, where each G_j is an exponential distribution with mean 1; i.e., $G_j(x) = \mathsf{P}(X_j \le x) = 1 - e^{-x}$ for $x \ge 0$, and $G_j(x) = 0$ for $x < 0$. We further assume that $X_1, X_2, \ldots, X_d$ are mutually independent. The (random) time Y to complete the project is then the length of the

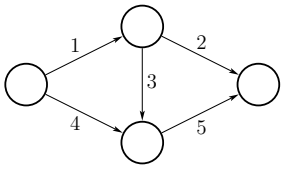

Fig. 1 A stochastic activity network with $d = 5$ activities.

longest path from the source, which is the leftmost node in Figure 1, to the sink, the rightmost node. The graph has $r = 3$ paths from source to sink,

$$\mathscr{P}_1 = \{1,2\}, \quad \mathscr{P}_2 = \{4,5\}, \quad \mathscr{P}_3 = \{1,3,5\}; \tag{1}$$

e.g., path $\mathscr{P}_3$ consists of activities 1, 3, and 5. For each $k = 1,2,\ldots,r$, let $T_k = \sum_{j\in\mathscr{P}_k} X_j$ be the (random) length of path $\mathscr{P}_k$. Thus,

$$Y = \max_{k=1,2,\ldots,r} T_k = \max(X_1+X_2,\, X_4+X_5,\, X_1+X_3+X_5) \tag{2}$$

represents the project's completion time, and we denote its CDF by F. □

More generally, consider a (complicated) stochastic model, and define P and E as the probability measure and expectation operator, respectively, induced by the model. Let Y be an $\mathfrak{R}$-valued output of the model representing its random performance or behavior, and define F as the CDF of Y, i.e.,

$$F(y) = \mathsf{P}(Y \le y) = \mathsf{E}[I(Y \le y)] \text{ for each } y \in \mathfrak{R}, \tag{3}$$

where $I(\cdot)$ denotes the indicator function, which takes value 1 (resp., 0) when its argument is true (resp., false). For a fixed constant $0 < p < 1$, define the p-quantile ξ of F as the generalized inverse of F; i.e.,

$$\xi = F^{-1}(p) \equiv \inf\{y : F(y) \ge p\}. \tag{4}$$

If F is continuous at ξ, then $F(\xi) = p$, but $F(\xi) \ge p$ in general.

Example 1 (continued). In her bid for the project, the contractor may specify the 0.95-quantile ξ as the promised completion date. Hence, according to the model, the project will complete by time ξ with probability $p = 0.95$. □

We assume that the complexity of the stochastic model prevents F from being computed, but we can simulate the model using Monte Carlo to produce an output $Y \sim F$, where the notation $\sim$ means "is distributed as." Thus, our goal is to use Monte Carlo simulation to develop an estimator of ξ and also to provide a confidence interval for ξ as a measure of the estimator's statistical error..

A special case of our framework arises when the random variable Y has the form

$$Y = c_Y(U_1,U_2,\ldots,U_d) \sim F \tag{5}$$

for a given function $c_Y : [0,1)^d \to \Re$, and $U_1, U_2, \ldots, U_d$ are independent and identically distributed (i.i.d.) unif$[0,1)$, where unif$[0,1)$ denotes a (continuous) uniform distribution on the interval $[0,1)$. We can think of c_Y as a computer code that takes $\mathbf{U} \equiv (U_1, U_2, \ldots, U_d)$ as input, transforms it into a random vector having a specified joint CDF with some (stochastic) dependence structure (independence being a special case), performs computations using the random vector, and then finally outputs Y. When Y satisfies (5), we can express its CDF F in (3) as

$$F(y) = \mathsf{P}(c_Y(\mathbf{U}) \le y) = \mathsf{E}[I(c_Y(\mathbf{U}) \le y)] = \int_{\mathbf{u} \in [0,1)^d} I(c_Y(\mathbf{u}) \le y)\, \mathrm{d}\mathbf{u}$$

for any constant $y \in \Re$, which we will later exploit in Section 5.3 when considering a VRT known as Latin hypercube sampling. For smooth integrands, computing a d-dimensional integral when d is small (say no more than 4 or 5) can be more efficiently handled through numerical quadrature techniques [14] rather than Monte Carlo simulation. But when d is large or the integrand is not smooth, Monte Carlo may be more attractive.

As we will later see in Section 5.4 when considering a VRT known as importance sampling, it is sometimes more convenient to instead consider Y having the form

$$Y = c'_Y(X_1, X_2, \ldots, X_{d'}) \sim F \tag{6}$$

for a given function $c'_Y : \Re^{d'} \to \Re$, and $\mathbf{X} = (X_1, X_2, \ldots, X_{d'})$ is a random vector with known joint CDF G from which we can generate observations. The joint CDF G specifies a dependence structure (independence being a special case) for $\mathbf{X}$, and the marginal distributions of the components of $\mathbf{X}$ may differ. We can see that (5) is a special case of (6) by taking $d' = d$, and assuming that $X_1, X_2, \ldots, X_{d'}$ are i.i.d. unif$[0,1)$. When Y has the form in (6), the CDF F in (3) satisfies

$$F(y) = \mathsf{P}(c'_Y(\mathbf{X}) \le y) = \mathsf{E}[I(c'_Y(\mathbf{x}) \le y)] = \int_{\mathbf{x} \in \Re^{d'}} I(c'_Y(\mathbf{x}) \le y)\, \mathrm{d}G(\mathbf{x}). \tag{7}$$

Let G_j be the marginal CDF of X_j. In the special case when $X_1, X_2, \ldots, X_{d'}$ are mutually independent under G and each G_j has a density g_j, we have that $\mathrm{d}G(\mathbf{x}) = \prod_{j=1}^{d'} g_j(x_j)\, \mathrm{d}x_j$ for $\mathbf{x} = (x_1, x_2, \ldots, x_{d'})$.

Example 1 (continued). For our SAN model in Figure 1 with Y in (2),

$$c'_Y(X_1, X_2, \ldots, X_{d'}) = \max(X_1 + X_2,\ X_4 + X_5,\ X_1 + X_3 + X_5)$$

is the function c'_Y in (6), where $d' = d = 5$. To define the function c_Y in (5) for this model, let $U_1, U_2, \ldots, U_d$ be $d = 5$ i.i.d. unif$[0,1)$ random variables. For each activity $j = 1, 2, \ldots, d$, we can use the inverse transform method (e.g., Section II.2a of [4] or Section 2.2.1 of [22]) to convert $U_j \sim$ unif$[0,1)$ into $X_j \sim G_j$ by letting $X_j = G_j^{-1}(U_j) = -\ln(1 - U_j)$. Hence, for $(u_1, u_2, \ldots, u_d) \in [0,1)^d$,

$$c_Y(u_1,u_2,\ldots,u_d) = \max(G_1^{-1}(u_1)+G_2^{-1}(u_2),\ G_4^{-1}(u_4)+G_5^{-1}(u_5),\\ G_1^{-1}(u_1)+G_3^{-1}(u_3)+G_5^{-1}(u_5)) \tag{8}$$

specifies the function c_Y in (5) to generate $Y \sim F$. □

3 Quantile Point Estimation via Monte Carlo

As seen in (4), the p-quantile ξ is the (generalized) inverse of the true CDF F evaluated at p. Thus, a common (but not the only) approach for devising a point estimator for ξ follows a generic recipe.

Step 1. Use a Monte Carlo method to construct $\hat{F}_n$ as an estimator of F, where n denotes the computational budget, typically the number of times the simulation model (e.g., a computer code as in (5)) is run.

Step 2. Compute $\hat{\xi}_n = \hat{F}_n^{-1}(p)$ as an estimator of ξ.

How we accomplish Step 1 depends on the particular Monte Carlo method being applied. Different methods will yield different CDF estimators, which in turn will produce different quantile estimators in Step 2.

3.1 Naive Monte Carlo

We next illustrate how to accomplish the two steps when applying naive Monte Carlo (NMC). Alternatively called crude Monte Carlo, standard simulation, and simple random sampling, NMC simply employs Monte Carlo without applying any variance-reduction technique. Note that (3) suggests estimating $F(y)$ by averaging i.i.d. copies of $I(Y \le y)$. To do this, generate $Y_1, Y_2, \ldots, Y_n$ as n i.i.d. copies of $Y \sim F$. We then compute the NMC estimator $\hat{F}_{\mathrm{NMC},n}$ of the CDF F as

$$\hat{F}_{\mathrm{NMC},n}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \le y), \tag{9}$$

completing Step 1. For each y, $\hat{F}_{\mathrm{NMC},n}(y)$ is an unbiased estimator of $F(y)$ because

$$\mathsf{E}[\hat{F}_{\mathrm{NMC},n}(y)] = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}[I(Y_i \le y)] = \frac{1}{n}\sum_{i=1}^{n} \mathsf{P}(Y_i \le y) = F(y) \tag{10}$$

as each $Y_i \sim F$. Then applying Step 2 yields the NMC quantile estimator

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

We can compute $\hat{\xi}_{\mathrm{NMC},n}$ in (11) via *order statistics*. Let $Y_{1:n} \le Y_{2:n} \le \cdots \le Y_{n:n}$ be the sorted values of the sample $Y_1, Y_2, \ldots, Y_n$, so $Y_{i:n}$ is the ith smallest value. Then we have that $\hat{\xi}_{\mathrm{NMC},n} = Y_{\lceil np \rceil:n}$, where $\lceil \cdot \rceil$ is the ceiling (or round-up) function. Although (10) shows that the CDF estimator is unbiased, the p-quantile estimator typically has bias; e.g., see Proposition 2 of [6].

In the special case of (5), we can obtain n i.i.d. copies of $Y \sim F$ by generating $n \times d$ i.i.d. unif$[0,1)$ random numbers $U_{i,j}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, d$, which we arrange in an $n \times d$ grid:

$$\begin{array}{cccc} U_{1,1} & U_{1,2} & \ldots & U_{1,d} \\ U_{2,1} & U_{2,2} & \ldots & U_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ U_{n,1} & U_{n,2} & \ldots & U_{n,d} \end{array} . \tag{12}$$

Now apply the function c_Y in (5) to each row to get

$$\begin{array}{l} Y_1 = c_Y(U_{1,1}, U_{1,2}, \ldots, U_{1,d}), \\ Y_2 = c_Y(U_{2,1}, U_{2,2}, \ldots, U_{2,d}), \\ \quad \vdots \quad \vdots \qquad \vdots \qquad \vdots \quad \ddots \quad \vdots \\ Y_n = c_Y(U_{n,1}, U_{n,2}, \ldots, U_{n,d}). \end{array} \tag{13}$$

Because each row i of (12) has d independent unif$[0,1)$ random numbers, we see that $Y_i \sim F$ by (5). Moreover, the independence of the rows of (12) ensures that $Y_1, Y_2, \ldots, Y_n$ are also independent.

Example 1 (continued). To apply NMC to our SAN model, we employ c_Y from (8) with $d = 5$ in (13) to obtain $Y_1, Y_2, \ldots, Y_n$, which are used to compute the NMC CDF estimator $\hat{F}_{\mathrm{NMC},n}$ in (9) and the NMC p-quantile estimator $\hat{\xi}_{\mathrm{NMC},n}$ in (11). □

We have considered the NMC p-quantile estimator $\hat{\xi}_{\mathrm{NMC},n}$ in (11) obtained by inverting the CDF estimator $\hat{F}_{\mathrm{NMC},n}$ in (9), but other NMC quantile estimators have also been developed. For example, we may replace the step function $\hat{F}_{\mathrm{NMC},n}$ with a linearly interpolated version, and [30] examines several such variants. Although these alternative quantile estimators may behave differently when the sample size n is small, they typically share the same large-sample properties (to be discussed in Section 4) as (11).

3.2 A General Approach to Construct a CDF Estimator

In addition to NMC, there are other ways of accomplishing Steps 1 and 2 of Section 3 to obtain CDF and quantile estimators. Constructing another CDF estimator often entails deriving and exploiting an alternative representation for F. To do this, we may perform Step 1 through the following:

Step 1a. Identify a random variable $J(y)$, whose value depends on y, such that

$$\mathsf{E}[J(y)] = F(y) \text{ for each } y \in \Re. \tag{14}$$

Step 1b. Construct the estimator $\hat{F}_n(y)$ of $F(y)$ as the sample average of n identically distributed copies of $J(y)$, possibly with some adjustments.

Note that we built the NMC CDF estimator $\hat{F}_{\mathrm{NMC},n}$ in (9) by following Steps 1a and 1b, with $J(y) = I(Y \le y)$, which satisfies (14) by (3).

Section 5 will review other Monte Carlo methods for performing Steps 1 and 2 of Section 3. Many (but not all) of the approaches handle Step 1 via Steps 1a and 1b. The n copies of $J(y)$ in Step 1b are often generated independently, but certain Monte Carlo methods sample them in a dependent manner; e.g., see Section 5.3.

4 Large-Sample Properties of Quantile Estimators

Although $\hat{\xi}_n$ is often *not* a sample average, it still typically obeys a *central limit theorem* (CLT) as the sample size n grows large. To establish this, let f be the derivative (when it exists) of F. Throughout the rest of the paper, whenever examining large-sample properties, we assume that $f(\xi) > 0$, which ensures that $F(\xi) = p$ and that $y = \xi$ is the unique root of the equation $F(y) = p$. Under various conditions that depend on the Monte Carlo method used to construct the CDF estimator $\hat{F}_n$ in Step 1 of Section 3, the corresponding p-quantile estimator $\hat{\xi}_n = \hat{F}_n^{-1}(p)$ satisfies a CLT

$$\sqrt{n}[\hat{\xi}_n - \xi] \Rightarrow N(0, \tau^2), \quad \text{as } n \to \infty, \tag{15}$$

where $\Rightarrow$ denotes convergence in distribution (e.g., see Section 25 of [9]), $N(a, b^2)$ represents a normal random variable with mean a and variance b^2, and the *asymptotic variance* τ^2 has the form

$$\tau^2 = \frac{\psi^2}{f^2(\xi)}. \tag{16}$$

The numerator ψ^2 on the right side of (16) is the asymptotic variance in the CLT for the CDF estimator at ξ:

$$\sqrt{n}[\hat{F}_n(\xi) - p] \Rightarrow N(0, \psi^2), \quad \text{as } n \to \infty, \tag{17}$$

where $p = F(\xi)$ because $f(\xi) > 0$. The CLT (17) typically holds (under appropriate conditions) because $\hat{F}_n(\xi)$ is often a sample average, e.g., as in (9) for NMC.

There are various ways to prove the CLT (15); e.g., see Sections 2.3.3 and 2.5 of [46] for NMC. A particularly insightful approach exploits a *Bahadur representation* [8], which shows for large n, a quantile estimator $\hat{\xi}_n = \hat{F}_n^{-1}(p)$ is well approximated by a linear transformation of its corresponding CDF estimator $\hat{F}_n$ at ξ:

$$\hat{\xi}_n \approx \xi + \frac{p - \hat{F}_n(\xi)}{f(\xi)}. \tag{18}$$

To heuristically justify this, note that $\hat{F}_n(y) \approx F(y)$ for each y when n is large, so

$$\hat{F}_n(\hat{\xi}_n) - \hat{F}_n(\xi) \approx F(\hat{\xi}_n) - F(\xi) \approx f(\xi)[\hat{\xi}_n - \xi] \tag{19}$$

by a first-order Taylor approximation. Also, $\hat{\xi}_n = \hat{F}_n^{-1}(p)$ implies $\hat{F}_n(\hat{\xi}_n) \approx p$, which we put into (19) and rearrange to finally get (18).

Bahadur [8] makes rigorous the heuristic argument for the NMC setting of Section 3.1. Specifically, if F is twice differentiable at ξ (with $f(\xi) > 0$), then

$$\hat{\xi}_n = \xi + \frac{p - \hat{F}_n(\xi)}{f(\xi)} + R_n \tag{20}$$

for $\hat{\xi}_n = \hat{\xi}_{\mathrm{NMC},n}$ from (11), such that with probability 1,

$$R_n = O(n^{-3/4}(\log\log n)^{3/4}), \quad \text{as } n \to \infty. \tag{21}$$

(The statement that "with probability 1, $A_n = O(h(n))$ as $n \to \infty$" for some function $h(n)$ means that there exists an event Ω_0 such that $\mathsf{P}(\Omega_0) = 1$ and for each outcome $\omega \in \Omega_0$, there exists a constant $K(\omega)$ such that $|A_n(\omega)| \le K(\omega)h(n)$ for all n sufficiently large.) (The almost-sure rate at which R_n vanishes in (21) is sharper than what [8] originally proved; see Section 2.5 of [46] for details.) Assuming only $f(\xi) > 0$, [21] proves a weaker result,

$$\sqrt{n}R_n \Rightarrow 0 \quad \text{as } n \to \infty, \tag{22}$$

which is sufficient for most applications. Note that (21) implies (22), and we call (20) combined with (21) (resp., (22)) a *strong* (resp., *weak*) Bahadur representation. The paper [13] provides a general framework for establishing a weak Bahadur representation, which may be verified for different variance-reduction techniques.

A (strong or weak) Bahadur representation ensures that $\hat{\xi}_n$ obeys a CLT. To see why, rearrange (20) and scale by $\sqrt{n}$ to get

$$\sqrt{n}\left[\hat{\xi}_n - \xi\right] = \frac{\sqrt{n}}{f(\xi)}\left[p - \hat{F}_n(\xi)\right] + \sqrt{n}R_n. \tag{23}$$

As $\hat{F}_n(\xi)$ is typically a sample average (e.g., (9)), it satisfies the CLT in (17). The second term on the right side of (23) vanishes weakly (resp., strongly) by (22) (resp., (21)), so Slutsky's theorem (e.g., Theorem 1.5.4 of [46]) verifies the CLT in (15).

When we apply Steps 1 and 2 in Section 3 to obtain $\hat{\xi}_n = \hat{F}_n^{-1}(p)$, (23) clarifies the reason the asymptotic variance τ^2 in the CLT (15) for $\hat{\xi}_n$ has the ratio form $\psi^2/f^2(\xi)$ in (16). The numerator ψ^2 arises from the CLT in (17) for the CDF estimator $\hat{F}_n$ at ξ, so ψ^2 is determined by the particular Monte Carlo method used to construct $\hat{F}_n$. For NMC, the CLT (17) uses $\hat{F}_{\mathrm{NMC},n}(\xi)$ from (9), which averages i.i.d. copies of $I(Y \le \xi)$, and the numerator in (16) is then

$$\psi^2_{\mathrm{NMC}} = \mathsf{Var}[I(Y \le \xi)] = p(1-p), \tag{24}$$

with Var the variance operator. But the denominator $f^2(\xi)$ in (16) is the *same* for each method.

5 Variance-Reduction Techniques for Quantile Estimation

Section 3.1 showed how to construct a quantile estimator when employing NMC. We next illustrate how the general approach of quantile estimation described in Section 3 and Subsection 3.2 can be applied for other Monte Carlo methods using VRTs.

5.1 Control Variates

Suppose that along with the response Y, the simulation model also outputs another random variable V whose mean $\mu_V = \mathsf{E}[V]$ is known. The method of control variates (CV) exploits this additional information to produce an estimator with typically reduced variance compared to its NMC counterpart. Section V.2 of [4] and Section 4.1 of [22] review CV for estimating a mean, and [29, 27, 13] apply this approach to estimate the CDF F, which is inverted to obtain an estimator of the p-quantile ξ.

When Y has the form in (5), we assume that the control variate V is generated by

$$V = c_V(U_1, U_2, \ldots, U_d) \tag{25}$$

for some function $c_V : [0,1)^d \to \Re$, where again we require that $\mu_V = \mathsf{E}[V]$ is known. Because the inputs $U_1, U_2, \ldots, U_d$ are the same in (25) and (5), V and Y are typically dependent. As will be later seen in (35), the CV method works best when V is strongly (positively or negatively) correlated with $I(Y \le \xi)$.

Example 1 (continued). Figure 1 has $r = 3$ paths from source to sink in (1). Of those, the length $T_3 = X_1 + X_3 + X_5$ of path $\mathscr{P}_3$ has the largest mean. We then choose the CV as $V = I(T_3 \le \zeta)$, where ζ is the p-quantile of the CDF $\widetilde{G}_3$ of T_3. As X_1, X_3, X_5 are i.i.d. exponential with mean 1, the CDF $\widetilde{G}_3$ is an Erlang with shape parameter 3 and scale parameter 1; i.e., $\widetilde{G}_3(x) = 1 - (1 + x + x^2)e^{-x}$ for $x \ge 0$. We can then compute $\zeta = \widetilde{G}_3^{-1}(p)$, and $\mu_V = p$. Hence,

$$c_V(U_1, U_2, \ldots, U_d) = I(G_1^{-1}(U_1) + G_3^{-1}(U_3) + G_5^{-1}(U_5) \le \zeta)$$

is the function c_V in (25). □

To design a CDF estimator when applying CV, we can follow the approach described in Section 3. For any constant $\beta \in \Re$, note that

$$F(y) = \mathsf{E}[I(Y \le y) + \beta(V - \mu_V)] \tag{26}$$

as $\mathsf{E}[V] = \mu_V$. Thus, take $J(y) = I(Y \le y) + \beta(V - \mu_V)$ in Step 1a of Section 3.2, and Step 1b suggests estimating $F(y)$ by averaging copies of $I(Y \le y) + \beta(V - \mu_V)$. Specifically, let (Y_i, V_i), $i = 1, 2, \ldots, n$, be i.i.d. copies of (Y, V), and define

$$\hat{F}'_{\mathrm{CV},\beta,n}(y) = \frac{1}{n}\sum_{i=1}^{n}[I(Y_i \le y) - \beta(V_i - \mu_V)] \tag{27}$$

$$= \hat{F}_{\mathrm{NMC},n}(y) - \beta(\hat{\mu}_{V,n} - \mu_V), \tag{28}$$

where $\hat{F}_{\mathrm{NMC},n}(y)$ is the NMC CDF estimator in (9), and $\hat{\mu}_{V,n} = (1/n)\sum_{i=1}^{n} V_i$. For each y and β, $\hat{F}'_{\mathrm{CV},\beta,n}(y)$ is an unbiased estimator of $F(y)$ by (26) and (27).

Although the choice of β does not affect the mean of $\hat{F}'_{\mathrm{CV},\beta,n}(y)$ by (26), it does have an impact on its variance, which by (27) equals

$$\begin{aligned}\mathsf{Var}[\hat{F}'_{\mathrm{CV},\beta,n}(y)] &= \frac{1}{n}\mathsf{Var}\big[I(Y \le y) - \beta(V - \mu_V)\big] \\ &= \frac{1}{n}\Big(F(y)[1 - F(y)] + \beta^2\mathsf{Var}[V] - 2\beta\mathsf{Cov}[I(Y \le y), V]\Big),\end{aligned} \tag{29}$$

where Cov denotes the covariance operator. As (29) is a quadratic function in β, we can easily find the value $\beta = \beta^*_y$ minimizing (29) as

$$\beta^*_y = \frac{\mathsf{Cov}[I(Y \le y), V]}{\mathsf{Var}[V]} = \frac{\mathsf{E}[I(Y \le y)V] - \mathsf{E}[I(Y \le y)]\,\mathsf{E}[V]}{\mathsf{E}[(V - \mu_V)^2]}. \tag{30}$$

The values of $\mathsf{Var}[V]$ and $\mathsf{Cov}[I(Y \le y), V]$ may be unknown, so we estimate them from our data (Y_i, V_i), $i = 1, 2, \ldots, n$. We then arrive at an estimator for β^*_y in (30) as

$$\hat{\beta}^*_{y,n} = \frac{[(1/n)\sum_{i=1}^{n} I(Y_i \le y)V_i] - \hat{F}_{\mathrm{NMC},n}(y)\hat{\mu}_{V,n}}{(1/n)\sum_{i=1}^{n}(V_i - \hat{\mu}_{V,n})^2}, \tag{31}$$

which uses $\hat{F}_{\mathrm{NMC},n}(y)$ to estimate $\mathsf{E}[I(Y \le y)] = F(y)$. Replacing β in (28) with the estimator $\hat{\beta}^*_{y,n}$ of its optimal value leads to the CV estimator of $F(y)$ as

$$\hat{F}_{\mathrm{CV},n}(y) = \hat{F}_{\mathrm{NMC},n}(y) - \hat{\beta}^*_{y,n}(\hat{\mu}_{V,n} - \mu_V). \tag{32}$$

For any constant β, (26) ensures that $\hat{F}'_{\mathrm{CV},\beta,n}(y)$ in (27) is an unbiased estimator of $F(y)$ for each $y \in \Re$, but the estimator $\hat{F}_{\mathrm{CV},n}(y)$ typically no longer enjoys this property as $\hat{\beta}^*_{y,n}$ and $\hat{\mu}_{V,n}$ are dependent. We finally obtain the CV p-quantile estimator

$$\hat{\xi}_{\mathrm{CV},n} = \hat{F}^{-1}_{\mathrm{CV},n}(p). \tag{33}$$

Computing the inverse in (33) appears to be complicated by the fact that the estimator $\hat{\beta}^*_{y,n}$ in (31) of the optimal β^*_y depends on y. However, [27] derives an algebraically equivalent representation for $\hat{F}_{\mathrm{CV},n}(y)$ that avoids this complication. It turns out that we can rewrite the CV CDF estimator in (32) as

$$\hat{F}_{\mathrm{CV},n}(y) = \sum_{i=1}^{n} W_i I(Y_i \le y) \quad \text{with} \quad W_i = \frac{1}{n} + \frac{(\hat{\mu}_{V,n} - V_i)(\hat{\mu}_{V,n} - \mu_V)}{\sum_{\ell=1}^{n}(V_\ell - \hat{\mu}_{V,n})^2}, \tag{34}$$

which satisfies $\sum_{i=1}^{n} W_i = 1$. While it is possible for $W_i < 0$, [27] notes it is unlikely. Because of (34), we can view $\hat{F}_{\mathrm{CV},n}(y)$ as a weighted average of the $I(Y_i \le y)$.

The weights W_i reduce to a simple form when the control $V = I(\widetilde{V} \le \zeta)$, where $\widetilde{V}$ is an auxiliary random variable, and ζ is the (known) p-quantile of the CDF of $\widetilde{V}$. (This is the setting of Example 1, in which $\widetilde{V} = T_3$.) Let $(Y_i, \widetilde{V}_i)$, $i = 1, 2, \ldots, n$, be i.i.d. copies of $(Y, \widetilde{V})$, and define $V_i = I(\widetilde{V}_i \le \zeta)$. Also, let $M = \sum_{i=1}^{n} V_i$. Then each weight becomes $W_i = p/M$ if $V_i = 1$, and $W_i = (1-p)/(n-M)$ if $V_i = 0$.

The key point of the representation in (34) is that each W_i does not depend on the argument y at which the CDF estimator $\hat{F}_{\mathrm{CV},n}$ is evaluated, simplifying the computation of its inverse. Specifically, let $Y_{i:n}$ be the ith smallest value among $Y_1, Y_2, \ldots, Y_n$, and let $W_{i:n}$ correspond to $Y_{i:n}$. Then the CV p-quantile estimator in (33) satisfies $\hat{\xi}_{\mathrm{CV},n} = Y_{i_p:n}$, where $i_p = \min\{k : \sum_{i=1}^{k} W_{i:n} \ge p\}$.

When $0 < \mathsf{Var}[V] < \infty$, the CV p-quantile estimator $\hat{\xi}_{\mathrm{CV},n}$ in (33) satisfies the CLT in (15), where ψ^2 in (16) is given by

$$\psi^2_{\mathrm{CV}} = p(1-p) - \frac{(\mathsf{Cov}[I(Y \le \xi), V])^2}{\mathsf{Var}[V]} = (1-\rho^2)p(1-p), \tag{35}$$

and $\rho = \mathsf{Cov}[I(Y \le \xi), V] / \sqrt{\mathsf{Var}[I(Y \le \xi)]\mathsf{Var}[V]}$ is the (Pearson) correlation coefficient of $I(Y \le \xi)$ and V; see [27, 13]. Thus, (35) shows that the more strongly (negatively or positively) correlated the CV V and $I(Y \le \xi)$ are, the smaller the asymptotic variance of the CV p-quantile estimator is, by (16). Also, [13] establishes that $\hat{\xi}_{\mathrm{CV},n}$ satisfies a weak Bahadur representation, as in (20) and (22).

We have developed the CV method when there is a single control V. But the idea extends to multiple controls $V^{(1)}, V^{(2)}, \ldots, V^{(m)}$, in which case the CDF estimator corresponds to a linear-regression estimator on the multiple CVs; see [27] for details. Also, rather than following the framework in Section 3 of constructing a p-quantile estimator as $\hat{\xi}_n = \hat{F}_n^{-1}(p)$, [44, 29, 27] consider an alternative CV estimator $\hat{\xi}'_{\mathrm{CV},n} \equiv \hat{\xi}_{\mathrm{NMC},n} - \beta(\hat{\zeta}_{\mathrm{NMC},n} - \zeta)$, where $\hat{\zeta}_{\mathrm{NMC},n}$ is the NMC estimator of the p-quantile ζ (assumed known) of the CDF of a random variable $\widetilde{V}$ (e.g., $\widetilde{V} = T_3$ in Example 1).

5.2 Stratified Sampling

Stratified sampling (SS) partitions the sample space into a finite number of subsets, known as *strata*, and allocates a fixed fraction of the overall sample size to sample from each stratum. Section 4.3 of [22] provides an overview of SS to estimate a mean, and [23, 12, 13] apply SS to estimate a quantile.

One way to partition the sample space for SS, as developed in [23], is as follows. Let S be an auxiliary random variable that is generated at the same time as the output Y. When Y has the form in (5), we assume that S is computed as

$$S = c_S(U_1, U_2, \ldots, U_d) \tag{36}$$

for some function $c_S : [0,1)^d \to \Re$, where $U_1, U_2, \ldots, U_d$ are the same uniforms used to generate Y in (5).

We next use S as a *stratification variable* to partition the sample space of (Y,S) by splitting the support of S into $t \geq 1$ disjoint subsets. Let $\mathscr{A}$ be the support of S, so $\mathsf{P}(S \in \mathscr{A}) = 1$. We then partition $\mathscr{A} = \cup_{s=1}^{t} \mathscr{A}_{\langle s \rangle}$ for some user-specified integer $t \geq 1$, where $\mathscr{A}_{\langle s \rangle} \cap \mathscr{A}_{\langle s' \rangle} = \emptyset$ for $s \neq s'$. For each $s = 1, 2, \ldots, t$, let $\lambda_{\langle s \rangle} = \mathsf{P}(S \in \mathscr{A}_{\langle s \rangle})$. The law of total probability implies

$$\begin{aligned} F(y) = \mathsf{P}(Y \leq y) &= \sum_{s=1}^{t} \mathsf{P}(Y \leq y, S \in \mathscr{A}_{\langle s \rangle}) \\ &= \sum_{s=1}^{t} \mathsf{P}(S \in \mathscr{A}_{\langle s \rangle}) \mathsf{P}(Y \leq y \mid S \in \mathscr{A}_{\langle s \rangle}) = \sum_{s=1}^{t} \lambda_{\langle s \rangle} F_{\langle s \rangle}(y), \end{aligned} \tag{37}$$

where $F_{\langle s \rangle}(y) \equiv \mathsf{P}(Y \leq y \mid S \in \mathscr{A}_{\langle s \rangle})$. In (37), $\lambda = (\lambda_{\langle s \rangle} : s = 1, 2, \ldots, t)$ is assumed known, but we need to estimate each $F_{\langle s \rangle}(y)$. We further assume that we have a way of sampling $Y_{\langle s \rangle} \sim F_{\langle s \rangle}$. A simple (but not necessarily the most efficient) way is through rejection sampling: generate (Y,S), and accept (resp., reject) Y as an observation from $F_{\langle s \rangle}$ if $S \in \mathscr{A}_{\langle s \rangle}$ (resp., if $S \notin \mathscr{A}_{\langle s \rangle}$).

To construct our SS estimator of F, we define $\gamma = (\gamma_{\langle s \rangle} : s = 1, 2, \ldots, t)$ as a vector of positive constants satisfying $\sum_{s=1}^{t} \gamma_{\langle s \rangle} = 1$. Then for our overall sample size n, we allocate a portion $n_{\langle s \rangle} \equiv \gamma_{\langle s \rangle} n$ to estimate $F_{\langle s \rangle}$ for stratum index s, where we assume that each $n_{\langle s \rangle}$ is integer-valued, so that $\sum_{s=1}^{t} n_{\langle s \rangle} = n$. For each $s = 1, 2, \ldots, t$, let $Y_{\langle s \rangle, i}$, $i = 1, 2, \ldots, n_{\langle s \rangle}$, be i.i.d. observations from $F_{\langle s \rangle}$, so our estimator of $F_{\langle s \rangle}$ is given by

$$\hat{F}_{\langle s \rangle, \gamma, n}(y) = \frac{1}{n_{\langle s \rangle}} \sum_{i=1}^{n_{\langle s \rangle}} I(Y_{\langle s \rangle, i} \leq y). \tag{38}$$

Replacing each $F_{\langle s \rangle}(y)$ in (37) by its estimator $\hat{F}_{\langle s \rangle, \gamma, n}(y)$ gives

$$\hat{F}_{\mathrm{SS}, \gamma, n}(y) = \sum_{s=1}^{t} \lambda_{\langle s \rangle} \hat{F}_{\langle s \rangle, \gamma, n}(y) \tag{39}$$

as the SS estimator of F. Inverting $\hat{F}_{\mathrm{SS}, \gamma, n}$ leads to the SS p-quantile estimator

$$\hat{\xi}_{\mathrm{SS}, \gamma, n} = \hat{F}^{-1}_{\mathrm{SS}, \gamma, n}(p). \tag{40}$$

While (39) and (40) follow the general approach of Steps 1 and 2 of Section 3, the way we constructed (39) does not exactly fit into the scheme of Steps 1a and 1b of Section 3.2, although the estimator $\hat{F}_{\langle s \rangle, \gamma, n}(y)$ in (38) applies the same idea.

We can compute $\hat{\xi}_{\mathrm{SS},\gamma,n}$ in (40) as follows. Let $D_k = Y_{\langle s\rangle,i}$ for $k = \sum_{\ell=1}^{s-1} n_{\langle\ell\rangle} + i$, and let $W'_k = \lambda_{\langle s\rangle}/n_{\langle s\rangle}$, which satisfies $\sum_{k=1}^{n} W'_k = 1$. Next define $D_{1:n} \le D_{2:n} \le \cdots \le D_{n:n}$ as the order statistics of $D_1, D_2, \ldots, D_n$, and let $W'_{i::n}$ be the W'_k associated with $D_{i:n}$. Then we have that $\hat{\xi}_{\mathrm{SS},\gamma,n} = D_{i'_p:n}$ for $i'_p = \min\{\ell : \sum_{i=1}^{\ell} W'_{i::n} \ge p\}$.

Example 1 (continued). Let the stratification variable in (36) be

$$S = X_1 + X_3 + X_5 = G_1^{-1}(U_1) + G_3^{-1}(U_3) + G_5^{-1}(U_5) \equiv c_S(U_1, U_2, \ldots, U_5), \quad (41)$$

the (random) length of the path $\mathscr{P}_3$ in (1), which has largest expectation among all paths in (1). As in Section 5.1, the CDF $\widetilde{G}_S$ of S is then an Erlang with shape parameter 3 and scale parameter 1. One way of partitioning the support $\mathscr{A}$ of S into $t \ge 1$ intervals takes $\mathscr{A}_{\langle s\rangle} = [\widetilde{G}_S^{-1}((s-1)/t), \widetilde{G}_S^{-1}(s/t))$ for each $s = 1, 2, \ldots, t$.

As in [23] we can use a "bin tossing" approach to sample the $Y_{\langle s\rangle,i}$, $s = 1, 2, \ldots, t$, $i = 1, 2, \ldots, n_{\langle s\rangle}$. In one run, generate $U_1, U_2, \ldots, U_5$ as i.i.d. unif$[0,1)$ random numbers, and compute $Y = c_Y(U_1, U_2, \ldots, U_5)$ for c_Y in (8) and $S = c_S(U_1, U_2, \ldots, U_5)$ for c_S in (41). If $S \in \mathscr{A}_{\langle s\rangle}$, then use Y as an observation from the stratum with index s. Keep independently sampling $(U_1, U_2, \ldots, U_5)$ and computing (Y, S) until each stratum index s has $n_{\langle s\rangle}$ observations, discarding any extras in a stratum. □

The SS p-quantile estimator $\hat{\xi}_{\mathrm{SS},\gamma,n}$ in (40) satisfies the CLT in (15) with

$$\psi^2_{\mathrm{SS},\gamma} = \sum_{s=1}^{t} \frac{\lambda^2_{\langle s\rangle}}{\gamma_{\langle s\rangle}} F_{\langle s\rangle}(\xi)[1 - F_{\langle s\rangle}(\xi)] \quad (42)$$

in (16); see [23, 12, 13]. Also, [13] shows that $\hat{\xi}_{\mathrm{CV},n}$ satisfies a weak Bahadur representation, as in (20) and (22). The value of $\psi^2_{\mathrm{SS},\gamma}$ depends on how the user specifies the sampling-allocation parameter γ. Setting $\gamma = \lambda$, known as the *proportional allocation*, ensures that $\psi^2_{\mathrm{SS},\lambda} \le \psi^2_{\mathrm{NMC}}$, so the proportional allocation guarantees no greater asymptotic variance than NMC. The optimal value of γ to minimize $\psi^2_{\mathrm{SS},\gamma}$ is $\gamma^* = (\gamma^*_{\langle s\rangle} : s = 1, 2, \ldots, t)$ with $\gamma^*_{\langle s\rangle} = \kappa_{\langle s\rangle}/(\sum_{s'=1}^{t} \kappa_{\langle s'\rangle})$, where $\kappa_{\langle s\rangle} = \lambda_{\langle s\rangle}(F_{\langle s\rangle}(\xi)[1 - F_{\langle s\rangle}(\xi)])^{1/2}$; e.g., see p. 217 of [23] and [12]. Although the $\kappa_{\langle s\rangle}$ are unknown, [12] employs pilot runs to estimate them, which are then used to estimate γ^*, and then performs additional runs with the estimated γ^*.

5.3 Latin Hypercube Sampling

Latin hypercube sampling (LHS) can be thought of as an efficient way of implementing SS in high dimensions. Section 5.4 of [22] provides an overview of LHS to estimate a mean, and [6, 31, 15, 16, 25, 38] develop LHS for quantile estimation.

To motivate how we apply LHS to estimate ξ, recall that for NMC, (10) shows that $\hat{F}_{\mathrm{NMC},n}(y)$ in (9) is an unbiased estimator of $F(y)$ for each y. While NMC uses a sample $Y_1, Y_2, \ldots, Y_n$ that are i.i.d. with CDF F, (10) still holds if we replace the sample with $Y_1^*, Y_2^*, \ldots, Y_n^*$ that are *dependent*, with each $Y_i^* \sim F$. Moreover, as

$$\mathsf{Var}\Big[\sum_{i=1}^{n} I(Y_i^* \le y)\Big] = \sum_{i=1}^{n} \mathsf{Var}[I(Y_i^* \le y)] + 2 \sum_{1 \le i < j \le n} \mathsf{Cov}[I(Y_i^* \le y), I(Y_j^* \le y)],$$

if $I(Y_i^* \le y)$ and $I(Y_j^* \le y)$ are negatively correlated for each $i \ne j$, then the average of the $I(Y_i^* \le y)$ will have lower variance than the average of the $I(Y_i \le y)$. We next show for the setting of (5) how LHS samples the $Y_i^* \sim F$ in a dependent manner.

Recall that d is the number of uniform inputs to c_Y in (5). For each $j = 1, 2, \ldots, d$, let $\pi_j = (\pi_j(1), \pi_j(2), \ldots, \pi_j(n))$ be a uniform random permutation of $(1, 2, \ldots, n)$, where $\pi_j(i)$ denotes the number in $\{1, 2, \ldots, n\}$ to which i maps. Thus, π_j equals one of the particular $n!$ permutations with probability $1/n!$. Let $\pi_1, \pi_2, \ldots, \pi_d$, be d mutually independent permutations, and also independent of the $n \times d$ grid of i.i.d. unif$[0,1)$ random numbers $U_{i,j}$ in (12). Then define

$$U_{i,j}^* = \frac{U_{i,j} + \pi_j(i) - 1}{n}, \quad \text{for } i = 1, 2, \ldots, n, \;\; j = 1, 2, \ldots, d. \tag{43}$$

It is easy to show that each $U_{i,j}^* \sim \text{unif}[0,1)$. Next arrange the $U_{i,j}^*$ in an $n \times d$ grid:

$$\begin{array}{cccc} U_{1,1}^* & U_{1,2}^* & \cdots & U_{1,d}^* \\ U_{2,1}^* & U_{2,2}^* & \cdots & U_{2,d}^* \\ \vdots & \vdots & \ddots & \vdots \\ U_{n,1}^* & U_{n,2}^* & \cdots & U_{n,d}^* \end{array} . \tag{44}$$

Each column j in (44) depends on π_j but not on any other permutation, making the d columns independent because $\pi_1, \pi_2, \ldots, \pi_d$ are. But the rows in (44) are *dependent* because for each column j, its entries $U_{i,j}^*$, $i = 1, 2, \ldots, n$, share the same permutation π_j. Now apply the function c_Y in (5) to each row of (44) to get

$$\begin{array}{cccccc} Y_1^* & = & c_Y(U_{1,1}^*, & U_{1,2}^*, & \ldots, & U_{1,d}^*), \\ Y_2^* & = & c_Y(U_{2,1}^*, & U_{2,2}^*, & \ldots, & U_{2,d}^*), \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ Y_n^* & = & c_Y(U_{n,1}^*, & U_{n,2}^*, & \ldots, & U_{n,d}^*). \end{array} \tag{45}$$

Because each row i of (44) has d i.i.d. unif$[0,1)$ random numbers, we see that $Y_i^* \sim F$ by (5). But $Y_1^*, Y_2^*, \ldots, Y_n^*$ are dependent because (44) has dependent rows.

Consider any column $j = 1, 2, \ldots, d$, in (44), and an interval $I_{k,n} = [(k-1)/n, k/n)$ for any $k = 1, 2, \ldots, n$. By (43), exactly one $U_{i,j}^*$ from column j lies in $I_{k,n}$. Thus, each column j forms a stratified sample of size n of unif$[0,1)$ random numbers, so LHS simultaneously stratifies each input coordinate $j = 1, 2, \ldots, d$.

We form the LHS estimator of the CDF F as

$$\hat{F}_{\text{LHS},n}(y) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i^* \le y) \tag{46}$$

and the LHS p-quantile estimator as

$$\hat{\xi}_{\mathrm{LHS},n} = \hat{F}^{-1}_{\mathrm{LHS},n}(p). \tag{47}$$

We can compute (47) by $\hat{\xi}_{\mathrm{LHS},n} = Y^*_{\lceil np \rceil:n}$, where $Y^*_{i:n}$ is the ith smallest value among $Y^*_1, Y^*_2, \ldots, Y^*_n$ in (45). Note that (46) and (47) fit into the framework of Section 3, where Step 1 is implemented through Steps 1a and 1b of Section 3.2, with $J(y) = I(Y \le y)$. But in contrast to the other methods considered, Step 1b generates n *dependent* copies of $I(Y \le y)$ as $I(Y^*_i \le y)$, $i = 1, 2, \ldots, n$, where each $Y^*_i \sim F$.

Example 1 (continued). To apply LHS to our SAN model, we employ c_Y from (8) in (45) to obtain $Y^*_1, Y^*_2, \ldots, Y^*_n$, which are then used in (46) and (47) to compute the LHS CDF estimator $\hat{F}_{\mathrm{LHS},n}$ and the LHS p-quantile estimator $\hat{\xi}_{\mathrm{LHS},n}$. □

Under regularity conditions, [6] proves that the LHS p-quantile estimator $\hat{\xi}_{\mathrm{LHS},n}$ in (47) obeys the CLT (15), and gives the specific form of $\psi^2 = \psi^2_{\mathrm{LHS}}$ in (16). Also, [17] shows that $\hat{\xi}_{\mathrm{LHS},n}$ satisfies a weak Bahadur representation, as in (20) and (22).

5.4 Importance Sampling

Importance sampling (IS) is a variance-reduction technique that can be particularly effective when studying rare events. The basic idea is to change the distributions driving the stochastic model to cause the rare event of interest to occur more frequently, and then unbias the outputs by multiplying by a correction factor. Section V.1 and Chapter VI of [4] and Section 4.6 of [22] provide overviews of IS to estimate a mean or tail probability.

For IS quantile estimation [24, 23, 48, 13], it is more natural to consider Y having the form in (6) rather than (5), i.e., $Y = c'_Y(\mathbf{X})$ for random vector $\mathbf{X} \in \Re^{d'}$ with joint CDF G. Let H be another joint CDF on $\Re^{d'}$ such that G is *absolutely continuous* with respect to H. For example, if G (resp., H) has a joint density function g (resp., h), then G is absolutely continuous with respect to H if $g(\mathbf{x}) > 0$ implies $h(\mathbf{x}) > 0$. In general, let P_G and E_G (resp., P_H and E_H) be the probability measure and expectation operator when $\mathbf{X} \sim G$ (resp., $\mathbf{X} \sim H$). The absolute continuity permits us to apply a *change of measure* to express the tail distribution corresponding to (7) as

$$\begin{aligned} 1 - F(y) = \mathsf{P}_G(Y > y) = \mathsf{E}_G[I(c'_Y(\mathbf{X}) > y)] &= \int_{\mathbf{x}\in\Re^{d'}} I(c'_Y(\mathbf{x}) > y)\,\mathrm{d}G(\mathbf{x}) \\ &= \int_{\mathbf{x}\in\Re^{d'}} I(c'_Y(\mathbf{x}) > y)\frac{\mathrm{d}G(\mathbf{x})}{\mathrm{d}H(\mathbf{x})}\,\mathrm{d}H(\mathbf{x}) = \int_{\mathbf{x}\in\Re^{d'}} I(c'_Y(\mathbf{x}) > y)L(\mathbf{x})\,\mathrm{d}H(\mathbf{x}) \\ &= \mathsf{E}_H[I(c'_Y(\mathbf{X}) > y)L(\mathbf{X})], \end{aligned} \tag{48}$$

where $L(\mathbf{x}) = \mathrm{d}G(\mathbf{x})/\mathrm{d}H(\mathbf{x})$ is the *likelihood ratio* or Radon-Nikodym derivative of G with respect to H; see Section 32 of [9]. In the special case when $\mathbf{X} = (X_1, X_2, \ldots, X_{d'})$ has mutually independent components under G (resp., H) with each marginal CDF G_j (resp., H_j) of X_j having a density function g_j (resp., h_j),

the likelihood ratio becomes $L(\mathbf{x}) = \prod_{j=1}^{d'} g_j(x_j)/h_j(x_j)$. By (48), we can obtain an unbiased estimator of $1-F(y)$ by averaging i.i.d. copies of $I(c'_Y(\mathbf{X}) > y)L(\mathbf{X})$, with $\mathbf{X} \sim H$. Specifically, let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be i.i.d., with each $\mathbf{X}_i \sim H$. Then we get an IS estimator of F as

$$\hat{F}_{\mathrm{IS},n}(y) = 1 - \frac{1}{n}\sum_{i=1}^{n} I(c'_Y(\mathbf{X}_i) > y)L(\mathbf{X}_i). \tag{49}$$

An IS p-quantile estimator is then

$$\hat{\xi}_{\mathrm{IS},n} = \hat{F}^{-1}_{\mathrm{IS},n}(p). \tag{50}$$

Note that (49) and (50) follow the general approach of Steps 1 and 2 of Section 3, where (49) is obtained through Steps 1a and 1b of Section 3.2, with $J(y) = 1 - I(c'_Y(\mathbf{X}) > y)L(\mathbf{X})$, which satisfies (14) by (48).

As shown in [24], we can compute $\hat{\xi}_{\mathrm{IS},n}$ in (50) as follows. Let $Y_i = c'_Y(\mathbf{X}_i)$, and define $Y_{i:n}$ as the ith smallest value among $Y_1, Y_2, \ldots, Y_n$. Also, let $L_{i::n} = L(\mathbf{X}_j)$ for $\mathbf{X}_j$ corresponding to $Y_{i:n}$. Then $\hat{\xi}_{\mathrm{IS},n} = Y_{i'_p:n}$ for $i'_p = \max\{k : \sum_{i=k}^{n} L_{i::n} \le (1-p)n\}$.

The key to effective application of IS is choosing an appropriate IS distribution H for $\mathbf{X}$ so that the quantile estimator $\hat{\xi}_{\mathrm{IS},n}$ has small variance. As seen in Section 4, the asymptotic variance τ^2 of the IS p-quantile estimator $\hat{\xi}_{\mathrm{IS},n}$ is closely related to the asymptotic variance $\psi^2 = \psi^2_{\mathrm{IS}}$ in the CLT (17) for $\hat{F}_{\mathrm{IS},n}(\xi)$. Let Var_H denote the variance operator when $\mathbf{X} \sim H$, and as $\mathbf{X}_i$, $i = 1, 2, \ldots, n$, are i.i.d., we have that

$$\mathsf{Var}_H[\hat{F}_{\mathrm{IS},n}(\xi)] = \frac{1}{n}\mathsf{Var}_H[I(c'_Y(\mathbf{X}) > \xi)L(\mathbf{X})] \equiv \frac{1}{n}\psi^2_{\mathrm{IS}}. \tag{51}$$

A "good" choice for H is problem specific, and a poorly designed H can actually increase the variance (or even produce infinite variance). The papers [24, 23, 13] discuss particular ways of selecting H in various problem settings.

Example 1 (continued). For a SAN as in Figure 1, [13] estimates the p-quantile when $p \approx 1$ via an IS scheme that combines ideas from [34] and [24]. Recall that Figure 1 has $r = 3$ paths from source to sink, which are given in (1). When estimating the SAN tail probability $\mathsf{P}_G(Y > y)$ for large y, we want to choose H so that the event $\{Y > y\}$ occurs more frequently. To do this, [34] specifies H as a mixture of r CDFs $H^{(1)}, H^{(2)}, \ldots, H^{(r)}$; i.e., $H(\mathbf{x}) = \sum_{k=1}^{r} \alpha^{(k)} H^{(k)}(\mathbf{x})$, where each $\alpha^{(k)}$ is a nonnegative constant such that $\sum_{k=1}^{r} \alpha^{(k)} = 1$. Each $H^{(k)}$ keeps all activity durations as independent exponentials but increases the mean of X_j for edges $j \in \mathscr{P}_k$, making $\{Y > y\}$ more likely. (More generally, one could choose $H^{(k)}$ to not only have different means for activities $j \in \mathscr{P}_k$ but further to have entirely different distributions.) Also, $H^{(k)}$ leaves unaltered the CDF of $X_{j'}$ for each $j' \notin \mathscr{P}_k$. Changing the mean of X_j corresponds to *exponentially twisting* its original CDF G_j; see Example 4.6.2 of [22] and Section V.1b of [4] for details on exponential twisting. The exponential twist requires specifying a twisting parameter $\theta \in \Re$, and [13] employs an approach in [24] to choose a value for $\theta = \theta^{(k)}$ for each $H^{(k)}$ in the mixture. Also, by adapting

a heuristic from [34] for estimating a tail probability to instead handle a quantile, [13] determines the mixing weights $\alpha^{(k)}$, $k = 1,2,\ldots,r$, by first obtaining an approximate upper bound for the second moment $\mathsf{E}_H[(I(c'_Y(\mathbf{X}) > \xi)L(\mathbf{X}))^2]$ in terms of the $\alpha^{(k)}$, and then choosing the $\alpha^{(k)}$ to minimize the approximate upper bound. Note that the mixture H used for IS does *not* satisfy the special case mentioned after (48), so the likelihood ratio $L(\mathbf{X}) = \mathrm{d}G(\mathbf{X})/\mathrm{d}H(\mathbf{X})$ is *not* simply the product $\prod_{j=1}^{d'} g_j(X_j)/h_j(X_j)$; see equation (33) of [13] for details. □

Glynn [24] develops other estimators of the CDF F using IS, leading to different IS quantile estimators. Through a simple example, he shows that of the IS p-quantile estimators he considers, $\hat{\xi}_{\mathrm{IS},n}$ in (50) can be the most effective in reducing variance when $p \approx 1$, but another of his IS p-quantile estimators can be better when $p \approx 0$.

Under a variety of different sets of assumptions (see [24, 1, 13]), the IS p-quantile estimator $\hat{\xi}_{\mathrm{IS},n}$ in (50) satisfies the CLT in (15), where ψ^2 in (16) equals ψ^2_{IS} in (51). Also, [13] shows that $\hat{\xi}_{\mathrm{CV},n}$ satisfies a weak Bahadur representation, as in (20) and (22). Moreover, [48] shows another IS p-quantile estimator from [24] obeys a strong Bahadur representation.

5.5 Conditional Monte Carlo

Conditional Monte Carlo (CMC) reduces variance by analytically integrating out some of the variability; see Section V.4 of [4] for an overview of CMC to estimate a mean. We next explain how to employ CMC for estimating a quantile, as developed in [40, 16, 3, 18], which fits into the general framework given in Section 3.

Let $\mathbf{Z}$ be an $\Re^{\bar{d}}$-valued random vector that is generated along with the output Y. In the special case when Y has the form in (5), we assume that

$$\mathbf{Z} = c_{\mathbf{Z}}(U_1, U_2, \ldots, U_d) \tag{52}$$

for a given function $c_{\mathbf{Z}} : [0,1)^d \to \Re^{\bar{d}}$. Because (52) and (5) utilize the same unif$[0,1)$ inputs $U_1, U_2, \ldots, U_d$, we see that $\mathbf{Z}$ and Y are dependent. In general, by using iterated expectations (e.g., p. 448 of [9]), we express the CDF F of Y as

$$F(y) = \mathsf{P}(Y \le y) = \mathsf{E}[\mathsf{P}(Y \le y \mid \mathbf{Z})] = \mathsf{E}[q(\mathbf{Z}, y)], \tag{53}$$

where the function $q : \Re^{\bar{d}+1} \to \Re$ is defined for each $\mathbf{z} \in \Re^{\bar{d}}$ as

$$q(\mathbf{z}, y) = \mathsf{P}(Y \le y \mid \mathbf{Z} = \mathbf{z}) = \mathsf{E}[I(Y \le y) \mid \mathbf{Z} = \mathbf{z}]. \tag{54}$$

We assume that $q(\mathbf{z}, y)$ can be computed, analytically or numerically, for each possible $\mathbf{z}$ and $y \in \Re$. By (53), we can obtain an unbiased estimator of $F(y)$ by averaging i.i.d. copies of $q(\mathbf{Z}, y)$. Specifically, let $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_n$ be i.i.d. replicates of the conditioning vector $\mathbf{Z}$. We then define the CMC estimator of the CDF F by

$$\hat{F}_{\mathrm{CMC},n}(y) = \frac{1}{n}\sum_{i=1}^{n} q(\mathbf{Z}_i, y), \tag{55}$$

which uses copies of $\mathbf{Z}$ but not of Y. We finally get the CMC p-quantile estimator

$$\hat{\xi}_{\mathrm{CMC},n} = \hat{F}^{-1}_{\mathrm{CMC},n}(p). \tag{56}$$

Thus, we obtained (55) and (56) by following Steps 1a, 1b, and 2 of Section 3 and Subsection 3.2, where in Step 1a, we take $J(y) = q(\mathbf{Z}, y)$, which satisfies (14) by (53). Computing the inverse in (56) typically requires employing an iterative root-finding method, such as the bisection method or Newton's method (e.g., Chapter 7 of [41]), incurring some computation cost.

Example 1 (continued). For a SAN, [47] develops a CMC approach for estimating the CDF F of Y, which we apply as follows. Let the conditioning vector $\mathbf{Z}$ be the (random) durations of the activities on the path $\mathscr{P}_3 = \{1,3,5\}$, so $\mathbf{Z} = (X_1, X_3, X_5) \in \Re^{\bar{d}}$ with $\bar{d} = 3$. Thus, the function $c_{\mathbf{Z}}$ in (52) is given by

$$c_{\mathbf{Z}}(U_1, U_2, \ldots, U_5) = (G_1^{-1}(U_1), G_3^{-1}(U_3), G_5^{-1}(U_5)).$$

Recall that for each $k = 1,2,3$, we defined $T_k = \sum_{j \in \mathscr{P}_k} X_k$, the (random) length of path $\mathscr{P}_k$ in (1). Since $\{Y \le y\} = \{T_1 \le y, T_2 \le y, T_3 \le y\}$ by (2), we can compute the function $q(\mathbf{z}, y)$ in (54) for any constant $\mathbf{z} = (x_1, x_3, x_5) \in \Re^{\bar{d}}$ as

$$\begin{aligned}
&q((x_1,x_3,x_5), y) = \mathsf{P}(Y \le y \mid X_1 = x_1, X_3 = x_3, X_5 = x_5) \\
&= \mathsf{P}(X_1 + X_2 \le y, X_4 + X_5 \le y, X_1 + X_3 + X_5 \le y \mid X_1 = x_1, X_3 = x_3, X_5 = x_5) \\
&= \mathsf{P}(X_2 \le y - x_1, X_4 \le y - x_5, x_1 + x_3 + x_5 \le y \mid X_1 = x_1, X_3 = x_3, X_5 = x_5) \\
&= \mathsf{P}(X_2 \le y - x_1)\, \mathsf{P}(X_4 \le y - x_5)\, \mathsf{P}(x_1 + x_3 + x_5 \le y) \\
&= (1 - e^{-(y - x_1)})\, (1 - e^{-(y - x_5)})\, I(x_1 + x_3 + x_5 \le y)
\end{aligned}$$

because $X_1, X_2, \ldots, X_5$ are i.i.d. exponential with mean 1. □

Applying a variance decomposition (e.g., problem 34.10 of [9]) yields

$$\begin{aligned}
\mathsf{Var}[I(Y \le y)] &= \mathsf{Var}[\mathsf{E}[I(Y \le y) \mid \mathbf{Z}]] + \mathsf{E}[\mathsf{Var}[I(Y \le y) \mid \mathbf{Z}]] \\
&\ge \mathsf{Var}[\mathsf{E}[I(Y \le y) \mid \mathbf{Z}]] = \mathsf{Var}[q(\mathbf{Z}, y)]
\end{aligned}$$

for each y, where the inequality uses the nonnegativity of conditional variance, and the last step holds by (54). Hence, for each y, averaging i.i.d. copies of $q(\mathbf{Z}, y)$, as is done in constructing $\hat{F}_{\mathrm{CMC},n}(y)$ in (55), leads to smaller variance than averaging i.i.d. copies of $I(Y \le y)$, as in the estimator $\hat{F}_{\mathrm{NMC},n}(y)$ in (9). We thus conclude that CMC provides a CDF estimator with lower variance at each point than NMC.

The CMC p-quantile estimator $\hat{\xi}_{\mathrm{CMC},n}$ in (56) obeys the CLT (15) with ψ^2 in (16) as

$$\psi^2_{\mathrm{CMC}} = \mathsf{Var}[q(\mathbf{Z}, \xi)] \le \mathsf{Var}[I(Y \le \xi)] = p(1-p) = \psi^2_{\mathrm{NMC}} \tag{57}$$

by (24), so the CMC p-quantile estimator has no greater asymptotic variance than that of NMC; see [3, 16, 18, 40]. Also, $\hat{\xi}_{\mathrm{CMC},n}$ has a weak Bahadur representation, as in (20), (22).

While we have applied CMC by conditioning on a random vector $\mathbf{Z}$, the method can be more generally applied by instead conditioning on a sigma-field; see [3].

5.6 Other Approaches

LHS in Section 5.3 reduces variance by inducing negative correlation among the outputs, and [6] examines quantile estimation via other correlation-induction schemes, including *antithetic variates* (AV); see also [13]. (Randomized) quasi-Monte Carlo has been applied for quantile estimation [42, 32, 26]. Other simulation-based methods for estimating ξ do not follow the approach in Steps 1 and 2 of Section 3. For example, [45] considers quantile estimation as a root-finding problem, and applies stochastic approximation to solve it.

We can also combine different variance-reduction techniques to estimate a quantile. The integrated methods can sometimes (but not always) behave synergistically, outperforming each approach by itself. Some particularly effective mergers include combined IS+SS [23, 13], CMC+LHS [16], and SS+CMC+LHS [18].

5.7 Numerical Results of Point Estimators for Quantiles

We now provide numerical results comparing some of the methods discussed in Sections 3.1 and 5.1–5.6 applied to the SAN model in Example 1. Using 10^3 independent replications, we estimated the bias, variance, and mean-square error (MSE) of quantile estimators with sample size $n = 640$, where we numerically computed (without simulation) the true values of the p-quantile ξ as approximately $\xi = 3.58049$ for $p = 0.6$ and $\xi = 6.66446$ for $p = 0.95$. For each method x, we computed the MSE improvement factor (IF) of x as the ratio of the MSEs for NMC and x.

Table 1 shows that each VRT reduces the variance and MSE compared to NMC. Each VRT also produces less bias for $p = 0.95$, but not always for $p = 0.6$, especially for IS. The IS approach (Section 5.4) for the SAN is designed to estimate the p-quantile when $p \approx 1$, and it leads to substantial MSE improvement for $p = 0.95$. But for $p = 0.6$, IS only slightly outperforms NMC. Also, observe that the IF of the combination CMC+LHS is larger than the product of the IFs of CMC and LHS, illustrating that their combination can work synergistically together.

Table 1 Bias, variance, and mean-square error of p-quantile estimators for $p = 0.6$ and 0.95, where a method's MSE improvement factor (IF) is the ratio of the MSEs of NMC and the method.

	$p = 0.6$				$p = 0.95$			
Method	Bias ($\times 10^{-3}$)	Variance ($\times 10^{-3}$)	MSE ($\times 10^{-3}$)	MSE IF	Bias ($\times 10^{-2}$)	Variance ($\times 10^{-2}$)	MSE ($\times 10^{-2}$)	MSE IF
NMC	1.32	7.18	7.18	1.00	−3.00	5.15	5.24	1.00
CV	1.45	3.88	3.89	1.85	0.69	2.15	2.15	2.44
LHS	−0.87	2.78	2.78	2.58	−1.74	2.36	2.39	2.19
IS	12.39	6.43	6.58	1.09	1.46	1.01	1.03	5.09
CMC	3.39	5.26	5.27	1.36	0.03	4.01	4.01	1.31
CMC+LHS	0.84	1.32	1.32	5.42	−0.36	1.67	1.67	3.14

6 Confidence Intervals for a Quantile

Example 1 (continued). The contractor understands that her p-quantile estimator $\hat{\xi}_n$ does not exactly equal the true p-quantile ξ due to Monte Carlo's sampling noise. To account for the statistical error, she also desires a 90% confidence interval $\mathscr{C}_n$ for ξ, so she can be highly confident that the true value of ξ lies in $\mathscr{C}_n$. □

We want a confidence interval (CI) $\mathscr{C}_n$ for ξ based on a sample size n satisfying

$$\mathsf{P}(\xi \in \mathscr{C}_n) = 1 - \alpha \tag{58}$$

for a user-specified constant $0 < \alpha < 1$, where $1 - \alpha$ is the desired *confidence level*, e.g., $1 - \alpha = 0.9$ for a 90% CI. In a few limited cases, we can design a CI for which (58) or $\mathsf{P}(\xi \in \mathscr{C}_n) \geq 1 - \alpha$ holds for a fixed n. But for most Monte Carlo methods, we instead have to be satisfied with a large-sample CI $\mathscr{C}_n$ for which

$$\mathsf{P}(\xi \in \mathscr{C}_n) \to 1 - \alpha, \quad \text{as } n \to \infty. \tag{59}$$

6.1 Small-Sample CIs

Consider applying NMC as in Section 3.1 with a fixed sample size n. Let $Y_1, Y_2, \ldots, Y_n$ be an i.i.d. sample from F, which we assume is continuous at ξ, ensuring that $\mathsf{P}(Y_i \leq \xi) = p$. Then $B_{n,p} \equiv n\hat{F}_{\mathrm{NMC},n}(\xi) = \sum_{i=1}^{n} I(Y_i \leq \xi)$ has a binomial(n, p) distribution by (9). Recall that $Y_{i:n}$ is the ith smallest value in the sample, so $\{Y_{i:n} \leq \xi\} = \{B_{n,p} \geq i\}$, which is equivalent to $\{Y_{i:n} > \xi\} = \{B_{n,p} < i\}$, Thus, for any integers $1 \leq i_1 < i_2 \leq n$, we see that

$$\mathsf{P}(Y_{i_1:n} \leq \xi < Y_{i_2:n}) = \mathsf{P}(i_1 \leq B_{n,p} < i_2) = 1 - \mathsf{P}(B_{n,p} < i_1) - \mathsf{P}(B_{n,p} \geq i_2).$$

If we select i_1 and i_2 such that $\mathsf{P}(B_{n,p} < i_1) + \mathsf{P}(B_{n,p} \geq i_2) \leq \alpha$, then

$$\mathscr{C}_{\text{bin},n} \equiv [Y_{i_1:n}, Y_{i_2:n}) \tag{60}$$

is a CI for ξ with confidence level at least $1-\alpha$. For example, we may pick i_1 and i_2 so that $\mathsf{P}(B_{n,p} < i_1) \le \alpha/2$ and $\mathsf{P}(B_{n,p} \ge i_2) \le \alpha/2$. We call (60) the *binomial CI*, also known as a *distribution-free CI*; Section 2.6.1 of [46] provides more details.

This idea unfortunately breaks down when applying a Monte Carlo method other than NMC because $n\hat{F}_n(\xi)$ no longer has a binomial distribution in general. But [29] extends the binomial approach to a multinomial for the alternative CV p-quantile estimator $\hat{\xi}'_{\text{CV},n}$ described in the last paragraph of Section 5.1.

6.2 Consistent Estimation of Asymptotic Variance

We can also build a *large-sample* CI $\mathscr{C}_n$ for ξ satisfying (59) by exploiting the CLT in (15) or the (weak) Bahadur representation in (20) and (22), which both hold for the Monte Carlo methods we considered in Sections 3 and 5. One approach based on the CLT (15) requires a *consistent* estimator $\hat{\tau}_n^2$ of τ^2 from (16); i.e., $\hat{\tau}_n^2 \Rightarrow \tau^2$ as $n \to \infty$. Then we can obtain a CI $\mathscr{C}_n$ for which (59) holds as

$$\mathscr{C}_{\text{con},n,b} = [\hat{\xi}_n \pm z_\alpha \hat{\tau}_n/\sqrt{n}], \tag{61}$$

where $z_\alpha = \Phi^{-1}(1-\alpha/2)$ and Φ is the $N(0,1)$ CDF; e.g., $z_\alpha = 1.645$ for $1-\alpha = 0.9$. A way to construct a consistent estimator $\hat{\tau}_n^2$ of $\tau^2 = \psi^2/f^2(\xi)$ devises a consistent estimator $\hat{\psi}_n^2$ of the numerator ψ^2 and also one for the denominator $f^2(\xi)$.

To handle ψ^2, [13] develops consistent estimators $\hat{\psi}_n^2$ when ψ^2 equals ψ^2_{CV} in (35) for CV, $\psi^2_{\text{SS},\gamma}$ in (42) for SS, and ψ^2_{IS} in (51) for IS, as well as for IS+SS. Also, [40] provides an estimator for ψ^2_{CMC} in (57), and [17] handles LHS. For NMC, (24) shows that $\psi^2_{\text{NMC}} = p(1-p)$, which does not require estimation.

Several techniques have been devised to consistently estimate $f(\xi)$ appearing in the denominator of (16). One approach exploits the fact that

$$\eta \equiv \frac{1}{f(\xi)} = \frac{\mathrm{d}}{\mathrm{d}p}F^{-1}(p) = \lim_{\delta \to 0} \frac{F^{-1}(p+\delta) - F^{-1}(p-\delta)}{2\delta} \tag{62}$$

by the chain rule of differentiation, which suggests estimating η by a *finite difference*

$$\hat{\eta}_n = \frac{\hat{F}_n^{-1}(p+\delta_n) - \hat{F}_n^{-1}(p-\delta_n)}{2\delta_n}, \tag{63}$$

for some user-specified *bandwidth* $\delta_n > 0$. For the case of NMC, [10, 11] establish the consistency of $\hat{\eta}_n$ when $\delta_n \to 0$ and $n\delta_n \to \infty$ as $n \to \infty$, and [13, 17] develop similar results when applying various variance-reduction techniques. Then in (61), we can use $\hat{\tau}_n^2 = \hat{\psi}_n^2 \hat{\eta}_n^2$ to consistently estimate τ^2. Kernel methods [43, 19, 37] have also been employed to estimate $f(\xi)$.

6.3 Batching, Sectioning, and Other Methods

An issue with the finite-difference estimator in (63) and with kernel methods is that for a given sample size n, the user must specify an appropriate bandwidth δ_n, which can be difficult to do in practice. To avoid this complication, we can instead build a CI for ξ via a method that does not try to consistently estimate the asymptotic variance τ^2 in (16).

Batching is such an approach; e.g., see p. 491 of [22]. Rather than computing one p-quantile estimator from a single sample, batching instead generates $b \geq 2$ independent samples, each called a *batch* (or *subsample*), and builds a p-quantile estimator from each batch. We then construct a CI from the sample average and sample variance of the b i.i.d. p-quantile estimators. Specifically, to keep the overall sample size as n, we generate the b independent batches to each have size $m = n/b$. In practice, setting $b = 10$ is often a reasonable choice. For example, for NMC with an overall sample $Y_1, Y_2, \ldots, Y_n$ of size n, batch $\ell = 1, 2, \ldots, b$, comprises observations $Y_{(\ell-1)m+i}$, $i = 1, 2, \ldots, m$. From each batch $\ell = 1, 2, \ldots, b$, we compute a p-quantile estimator $\hat{\xi}_{m,\ell}$, which is roughly normally distributed when the batch size $m = n/b$ is large, by the CLT in (15). As the batches are independent, we have that $\hat{\xi}_{m,\ell}$, $\ell = 1, 2, \ldots, b$, are i.i.d. From their sample average $\bar{\xi}_{n,b} = (1/b)\sum_{\ell=1}^{b} \hat{\xi}_{m,\ell}$ and sample variance $S^2_{n,b} = (1/(b-1))\sum_{\ell=1}^{b} [\hat{\xi}_{m,\ell} - \bar{\xi}_{n,b}]^2$, we obtain the *batching CI* as

$$\mathscr{C}_{\text{bat},n,b} = [\bar{\xi}_{n,b} \pm t_{b-1,\alpha} S_{n,b}/\sqrt{b}], \tag{64}$$

where $t_{b-1,\alpha} = \Gamma_{b-1}^{-1}(1-\alpha/2)$ with Γ_{b-1} as the CDF of a Student-t random variable with $b-1$ degrees of freedom; e.g., $t_{b-1,\alpha} = 1.83$ when $b = 10$ and $1-\alpha = 0.9$. The batching CI $\mathscr{C}_{\text{bat},n,b}$ uses a Student t critical point $t_{b-1,\alpha}$ rather than z_α from a normal, as in (61), because $\mathscr{C}_{\text{bat},n,b}$ has a *fixed* (small) number b of batches, and the quantile estimator $\hat{\xi}_{m,\ell}$ from each batch ℓ is approximately normally distributed. (When applying LHS as in Section 5.3, each batch is an LHS sample, as in (45), but of size m. We then sample the b batches independently; see [17] for details.)

While the batching CI $\mathscr{C}_{\text{bat},n,b}$ in (64) is asymptotically valid in the sense that (59) holds for any fixed $b \geq 2$, it can have poor performance when the overall sample size n is not large. Specifically, for a generic CI $\mathscr{C}_n$ for ξ, define the CI's *coverage* as $\mathsf{P}(\xi \in \mathscr{C}_n)$, which may differ from the nominal confidence level $1-\alpha$ for any fixed n even though (59) holds. The issue with the batching CI stems from quantile estimators being biased in general; e.g., see Proposition 2 of [6] for the case of NMC. While the bias typically vanishes as the sample size $n \to \infty$, the bias can be significant when n is not large. The bias of the batching point estimator $\bar{\xi}_{n,b}$ is determined by the batch size $m = n/b < n$, so $\bar{\xi}_{n,b}$ may be severely contaminated by bias. Hence, the batching CI $\mathscr{C}_{\text{bat},n,b}$ is centered at the wrong point on average, which can lead to poor coverage when n is small.

Sectioning can produce a CI with better coverage than batching. Introduced in Section III.5a of [4] for NMC and extended by [39, 17] to apply when employing different VRTs, sectioning modifies batching to center its CI at the p-quantile estima-

tor $\hat{\xi}_n$ based on the entire sample size n rather than at the batching point estimator $\bar{\xi}_{n,b}$. For example, for NMC, we use $\hat{\xi}_n = \hat{\xi}_{\mathrm{NMC},n}$ from (11). We also replace $S_{n,b}$ in (64) with $S'_{n,b}$, where $S'^2_{n,b} = (1/(b-1))\sum_{\ell=1}^{b}[\hat{\xi}_{m,\ell} - \hat{\xi}_n]^2$. The *sectioning CI* is then

$$\mathscr{C}_{\mathrm{sec},n,b} = [\hat{\xi}_n \pm t_{b-1,\alpha} S'_{n,b}/\sqrt{b}]. \tag{65}$$

Because we center $\mathscr{C}_{\mathrm{sec},n,b}$ at $\hat{\xi}_n$ instead of the typically more-biased $\bar{\xi}_{n,b}$, the sectioning CI $\mathscr{C}_{\mathrm{sec},n,b}$ can have better coverage than the batching CI $\mathscr{C}_{\mathrm{bat},n,b}$ when n is small. By exploiting a weak Bahadur representation, as in (20) and (22), we can rigorously justify replacing the batching point estimator $\bar{\xi}_{n,b}$ in (64) with the overall point estimator $\hat{\xi}_n$ and still maintain the asymptotic validity in (59).

For NMC, bootstrap CIs for ξ have been developed in in [36, 7]. Also, [35] develops bootstrap CI for ξ when applying IS.

6.4 Numerical Results of CIs for Quantiles

Table 2 provides numerical results comparing the methods discussed in Sections 6.1–6.3 to construct nominal 90% CIs for a p-quantile ξ of the longest path Y in the SAN model in Example 1 for different values of p. We built the CIs using NMC with different overall sample sizes n. For the consistent CI in (61), we estimated $\eta = 1/f(\xi)$ in (62) via the finite difference in (63) with bandwidth $\delta_n = 1/\sqrt{n}$. For a given CI $\mathscr{C}_n$ based on an overall sample size n, we estimated its coverage $\mathsf{P}(\xi \in \mathscr{C}_n)$ from 10^4 independent replications. Also, we computed for each method the average relative half width (ARHW), defined as the average half-width of the CI divided by the true p-quantile ξ, computed numerically; Section 5.7 gives the values.

Comparing the results for $p = 0.6$ and $p = 0.95$, we see that the more extreme quantile is harder to estimate, which is typically the case. For example, for the same n, the ARHW for $p = 0.95$ is larger than for $p = 0.6$. To see why, recall that the NMC p-quantile estimator's asymptotic variance is $p(1-p)/f^2(\xi)$ by (16) and (24). Although the numerator shrinks as p approaches 1, the denominator $f^2(\xi)$ decreases much faster. Moreover, while each method's coverage for $p = 0.6$ is close to the nominal 0.9 for each n, the consistent CI and the batching CI from (64) for $p = 0.95$ exhibit coverages that substantially depart from 0.9 when n is small, with overcoverage (resp., undercoverage) for the consistent (resp., batching) CI. When n is large, both methods produce CIs with close to nominal coverage, illustrating their asymptotic validity. As explained in Section 6.3, the batching CI can suffer from poor coverage for small n because the batching point estimator can be significantly biased. In contrast, the binomial CI in (60) and sectioning CI from (65) have coverage close to 0.9 for all n. It is important to remember that the binomial CI does not apply in general when applying VRTs, but sectioning does.

Table 2 Average relative half width (ARHW) and coverage of nominal 90% CIs for the p-quantile for $p = 0.6$ and 0.95 with different sample sizes n when applying NMC. Batching and sectioning use $b = n/10$ batches.

		$p = 0.6$		$p = 0.95$	
n	Method	ARHW	Coverage	ARHW	Coverage
400	Binomial	0.053	0.921	0.082	0.932
400	Consistent	0.051	0.893	0.094	0.952
400	Batching	0.054	0.869	0.069	0.666
400	Sectioning	0.055	0.907	0.075	0.888
1600	Binomial	0.026	0.910	0.038	0.914
1600	Consistent	0.025	0.896	0.039	0.916
1600	Batching	0.027	0.893	0.037	0.838
1600	Sectioning	0.028	0.904	0.038	0.904
6400	Binomial	0.013	0.904	0.018	0.905
6400	Consistent	0.013	0.897	0.018	0.899
6400	Batching	0.014	0.898	0.019	0.885
6400	Sectioning	0.014	0.900	0.019	0.903

7 Summary and Concluding Remarks

This tutorial reviewed various Monte Carlo methods for estimating a p-quantile ξ of the CDF F of a random variable Y. Because $\xi = F^{-1}(p)$, a common approach for estimating ξ first obtains an estimator $\hat{F}_n$ of F, and then inverts $\hat{F}_n$ to obtain a p-quantile estimator $\hat{\xi}_n = \hat{F}_n^{-1}(p)$. Sections 3 and 5 applied this approach to construct quantile estimators based on different Monte Carlo methods. We also discussed techniques for constructing confidence intervals for ξ. In addition to our paper, [28] further surveys simulation procedures for estimating ξ, along with another risk measure $\mathsf{E}[Y \mid Y > \xi]$, which is known as the conditional value-at-risk, expected shortfall, or conditional tail expectation, and often used in finance.

We focused on quantile estimation for the setting in which the outputs are i.i.d., but there has also been work covering the situation when outputs form a dependent sequence, as in a time series or stochastic process. For example, see [52, 2] and references therein.

Acknowledgements This work has been supported in part by the National Science Foundation under Grant No. CMMI-1537322. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Ahn, J.Y., Shyamalkumar, N.D.: Large sample behavior of the CTE and VaR estimators under importance sampling. The North American Actuarial Journal **15**, 393–416 (2011)
2. Alexopoulos, C., Goldsman, D., Mokashi, A.C., Wilson, J.R.: Automated estimation of extreme steady-state quantiles via the maximum transformation. ACM Transactions on Modeling and Computer Simulation **27**(4), 22:1–22:29 (2017)
3. Asmussen, S.: Conditional Monte Carlo for sums, with applications to insurance and finance. Annals of Actuarial Science **12**(2), 455–478 (2018)
4. Asmussen, S., Glynn, P.: Stochastic Simulation: Algorithms and Analysis. Springer, New York (2007)
5. Austin, P., Schull, M.: Quantile regression: A tool for out-of-hospital research. Academic Emergency Medicine **10**(7), 789–797 (2003)
6. Avramidis, A.N., Wilson, J.R.: Correlation-induction techniques for estimating quantiles in simulation. Operations Research **46**, 574–591 (1998)
7. Babu, G.J.: A note on bootstrapping the variance of sample quantile. Annals of the Institute of Statistical Mathematics **38**(3), 439–443 (1986)
8. Bahadur, R.R.: A note on quantiles in large samples. Annals of Mathematical Statistics **37**(3), 577–580 (1966)
9. Billingsley, P.: Probability and Measure, 3rd edn. John Wiley and Sons, New York (1995)
10. Bloch, D.A., Gastwirth, J.L.: On a simple estimate of the reciprocal of the density function. Annals of Mathematical Statistics **39**, 1083–1085 (1968)
11. Bofinger, E.: Estimation of a density function using order statistics. Australian Journal of Statistics **17**, 1–7 (1975)
12. Cannamela, C., Garnier, J., Iooss, B.: Controlled stratification for quantile estimation. Annals of Applied Statistics **2**(4), 1554–1580 (2008)
13. Chu, F., Nakayama, M.K.: Confidence intervals for quantiles when applying variance-reduction techniques. ACM Transactions On Modeling and Computer Simulation **22**(2), 10:1–10:25 (2012)
14. Davis, P.J., Rabinowitz, P.: Methods of Numerical Integration, 2nd edn. Academic Press, San Diego (1984)
15. Dong, H., Nakayama, M.K.: Constructing confidence intervals for a quantile using batching and sectioning when applying Latin hypercube sampling. In: A. Tolk, S.D. Diallo, I.O. Ryzhov, L. Yilmaz, S. Buckley, J.A. Miller (eds.) Proceedings of the 2014 Winter Simulation Conference, pp. 640–651. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey (2014)
16. Dong, H., Nakayama, M.K.: Quantile estimation using conditional Monte Carlo and Latin hypercube sampling. In: W.K.V. Chan, A.D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, E. Page (eds.) Proceedings of the 2017 Winter Simulation Conference, pp. 1986–1997. Institute of Electrical and Electronics Engineers, Piscataway, NJ (2017)
17. Dong, H., Nakayama, M.K.: Quantile estimation with Latin hypercube sampling. Operations Research **65**(6), 1678–1695 (2017)
18. Dong, H., Nakayama, M.K.: Quantile estimation using stratified sampling, conditional Monte Carlo, and Latin hypercube sampling (2018). In preparation
19. Falk, M.: On the estimation of the quantile density function. Statistics and Probability Letters **4**, 69–73 (1986)
20. Garnett, O., Mandelbaum, A., Reiman, M.: Designing a call center with impatient customers. Manufacturing and Service Operations Management **4**(3), 208–227 (2002)
21. Ghosh, J.K.: A new proof of the Bahadur representation of quantiles and an application. Annals of Mathematical Statistics **42**, 1957–1961 (1971)
22. Glasserman, P.: Monte Carlo Methods in Financial Engineering. Springer, New York (2004)
23. Glasserman, P., Heidelberger, P., Shahabuddin, P.: Variance reduction techniques for estimating value-at-risk. Management Science **46**, 1349–1364 (2000)

24. Glynn, P.W.: Importance sampling for Monte Carlo estimation of quantiles. In: S.M. Ermakov, V.B. Melas (eds.) Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation, pp. 180–185. Publishing House of St. Petersburg Univ., St. Petersburg, Russia (1996)
25. Grabaskas, D., Nakayama, M.K., Denning, R., Aldemir, T.: Advantages of variance reduction techniques in establishing confidence intervals for quantiles. Reliability Engineering and System Safety **149**, 187–203 (2016)
26. He, Z., Wang, X.: Convergence of randomized quasi-Monte Carlo sampling for value-at-risk and conditional value-at-risk (2017). ArXiv:1706.00540
27. Hesterberg, T.C., Nelson, B.L.: Control variates for probability and quantile estimation. Management Science **44**, 1295–1312 (1998)
28. Hong, L.J., Hu, Z., Liu, G.: Monte Carlo methods for value-at-risk and conditional value-at-risk: A review. ACM Transactions on Modeling and Computer Simulataion **24**(4), 22:1–22:37 (2014)
29. Hsu, J.C., Nelson, B.L.: Control variates for quantile estimation. Management Science **36**, 835–851 (1990)
30. Hyndman, R.J., Fan, Y.: Sample quantiles in statistical packages. American Statistician **50**(4), 361–365 (1996)
31. Jin, X., Fu, M.C., Xiong, X.: Probabilistic error bounds for simulation quantile estimation. Management Science **49**, 230–246 (2003)
32. Jin, X., Zhang, A.X.: Reclaiming quasi-Monte Carlo efficiency in portfolio value-at-risk simulation through Fourier transform. Management Science **52**(6), 925–938 (2006)
33. Jorion, P.: Value at Risk: The New Benchmark for Managing Financial Risk, 3rd edn. McGraw-Hill, New York (2007)
34. Juneja, S., Karandikar, R., Shahabuddin, P.: 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 (2007)
35. Liu, J., Yang, X.: The convergence rate and asymptotic distribution of bootstrap quantile variance estimator for importance sampling. Advances in Applied Probability **44**, 815–841 (2012)
36. Maritz, J.S., Jarrett, R.G.: A note on estimating the variance of the sample median. Journal of the American Statistical Association **73**, 194–196 (1978)
37. Nakayama, M.K.: Asymptotic properties of kernel density estimators when applying importance sampling. In: S. Jain, R. Creasey, J. Himmelspach, K. White, M. Fu (eds.) Proceedings of the 2011 Winter Simulation Conference, pp. 556–568. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey (2011)
38. Nakayama, M.K.: 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 (2011)
39. Nakayama, M.K.: Confidence intervals using sectioning for quantiles when applying variance-reduction techniques. ACM Transactions on Modeling and Computer Simulation **24**(4), 19:1–19:21 (2014)
40. Nakayama, M.K.: Quantile estimation when applying conditional Monte Carlo. In: SIMULTECH 2014 Proceedings, pp. 280–285 (2014)
41. Ortega, J.M., Rheinboldt, W.C.: terative Solution of Nonlinear Equations in Several Variables. SIAM, Philadelphia (2000)
42. Papageorgiou, A., Paskov, S.H.: Deterministic simulation for risk management. Journal of Portfolio Management **25**(5), 122–127 (1999)
43. Parzen, E.: Nonparametric statistical data modeling. Journal of the American Statistical Association **74**(365), 105–121 (1979)
44. Ressler, R.L., Lewis, P.A.W.: Variance reduction for quantile estimates in simulations via nonlinear controls. Communications in Statistics — Simulation and Computation **B19**(3), 1045–1077 (1990)
45. Robinson, D.W.: Non-parametric quantile estimation through stochastic approximation. Phd thesis, Naval Postgraduate School, Monterey, Calif. (1975)

46. Serfling, R.J.: Approximation Theorems of Mathematical Statistics. John Wiley and Sons, New York (1980)
47. Sigal, C.E., Pritsker, A.A.B., Solberg, J.J.: The use of cutsets in Monte Carlo analysis of stochastic networks. Mathematics and Computers in Simulation **21**(4), 376–384 (1979)
48. Sun, L., Hong, L.J.: Asymptotic representations for importance-sampling estimators of value-at-risk and conditional value-at-risk. Operations Research Letters **38**(4), 246–251 (2010)
49. U.S. Nuclear Regulatory Commission: Final safety evaluation for WCAP-16009-P, revision 0, "realistic large break LOCA evaluation methodology using automated statistical treatment of uncertainty method (ASTRUM)" (TAC no. MB9483). Tech. rep., U.S. Nuclear Regulatory Commission, Washington, DC (2005). https://www.nrc.gov/docs/ML0509/ML050910159.pdf
50. U.S. Nuclear Regulatory Commission: Acceptance criteria for emergency core cooling systems for light-water nuclear power reactors. Title 10, Code of Federal Regulations §50.46, NRC, Washington, DC (2010)
51. U.S. Nuclear Regulatory Commission: Applying statistics. U.S. Nuclear Regulatory Commission Report NUREG-1475, Rev 1, U.S. Nuclear Regulatory Commission, Washington, DC (2011)
52. Wu, W.B.: On the Bahadur representation of sample quantiles for dependent sequences. Annals of Statistics **33**(4), 1934–1963 (2005)