Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

SOME ASYMPTOTIC REGIMES FOR QUANTILE ESTIMATION

Marvin K. Nakayama[1] and Bruno Tuffin[2]

[1]Dept. of Computer Science, New Jersey Institute of Technology, Newark, NJ, USA
[2]Inria, Univ Rennes, CNRS, IRISA Campus de Beaulieu, 35042 Rennes, FRANCE

ABSTRACT

The paper examines the relative errors (REs) of quantile estimators of various stochastic models under different asymptotic regimes. Depending on the particular limit considered and the Monte Carlo method applied, the RE may be vanishing, bounded, or unbounded. We provide examples of these possibilities.

1 INTRODUCTION

Analysts employ *quantiles* to assess risk in application areas such as finance, manufacturing, and nuclear engineering. Also called a *value-at-risk* or *percentile*, the p-quantile, $p \in (0,1)$, of a continuous random variable is a constant ξ_p for which exactly p of its distribution's mass lies below ξ_p, so the median is the 0.5-quantile. For example, a manufacturer specifying a widget's warranty length as the 0.1-quantile of its time to failure would lead to about 10% of sales resulting in warranty claims. Dong and Nakayama (2019) review estimating ξ_p via Monte Carlo (MC) methods, including *simple random sampling* (SRS) and *importance sampling* (IS) (Asmussen and Glynn 2007, Chapter V.1).

We focus here on the efficiency of quantile estimators. For an estimator obeying a central limit theorem (CLT) as the sample size grows large, the estimator's *relative error* (RE) is the ratio of the square root of the CLT's asymptotic variance over the estimand; e.g., see L'Ecuyer et al. (2010). We consider families of stochastic models indexed by a parameter r and examine the REs of SRS and IS estimators of $\xi_p \equiv \xi_{p,r}$ as r approaches some limiting value r_0. For example, r may represent the mean or variance of a random variable, and we let $r \to \infty$ or $r \to 0$. Or r could be the quantile level p, and we let $p \to 0$ or 1. The paper provides examples under various asymptotic regimes as $r \to r_0$ of vanishing, bounded, or unbounded RE.

Many previous papers have studied quantile estimation, but there does not appear to be much analyzing the estimators' theoretical efficiency under asymptotic regimes. The asymptotic variance of a quantile estimator often has a ratio form, with the numerator as the asymptotic variance of an estimator of the probability of exceeding ξ_p, and the denominator the squared density at ξ_p. Prior results analyzing the IS asymptotics as $p \to 1$ of just the variance's numerator (ignoring the denominator) include Glynn (1996) for a sum of independent and identically distributed (i.i.d.) random variables, and Deo and Murthy (2021) for a black-box model. Kohler and Krzyżak (2019) examine the rate of convergence of quantile estimators based on surrogate models with sample size n for a quantile level $p = p_n \to 1$ as $n \to \infty$. These works do not consider RE, but Li et al. (2024) does for an i.i.d. sum as $p \to 1$, as described in Section 6.2.

The rest of the paper unfolds as follows. Section 2 develops the mathematical framework. Section 3 considers location-scale families of distributions. We examine SRS estimators of ξ_p as $p \to 1$ or $p \to 0$ for some specific parametric families of distributions in Section 4. Section 5 analyzes IS for the exponential distribution. In Section 6 we consider averages and sums of i.i.d. random variables. Section 7 reviews results for a quantile of the hitting time to a rarely visited set of states for a regenerative process. We compare hypothesis tests for quantiles and tail probabilities in Section 8.

2 MATHEMATICAL FRAMEWORK

Consider a family of stochastic models indexed by a parameter r, with $\varphi \equiv \varphi_r$ an estimand (e.g., a quantile or mean) to be estimated via a MC method $\mathfrak{M}$ (e.g., SRS or IS). From a simulation with sample size

979-8-3315-3420-2/24/$31.00 ©2024

n using $\mathfrak{M}$, construct an estimator $\widehat{\varphi}_{\mathfrak{M},n} \equiv \widehat{\varphi}_{\mathfrak{M},n,r}$ of φ_r, and for each fixed r, we assume a CLT holds: $\sqrt{n}[\widehat{\varphi}_{\mathfrak{M},n} - \varphi] \Rightarrow \mathcal{N}(0, \varsigma^2_{\mathfrak{M}})$ as $n \to \infty$, where $\Rightarrow$ denotes weak convergence (Serfling 1980, Section 1.2.4), $\mathcal{N}(a,b^2)$ is a normal random variable with mean a and variance b^2, and $\varsigma^2_{\mathfrak{M}} \equiv \varsigma^2_{\mathfrak{M},r} \in (0,\infty)$ is the CLT's *asymptotic variance*. When $\varphi \neq 0$, the *relative error* of the $\mathfrak{M}$ estimator of φ is

$$\mathrm{RE}_{\mathfrak{M},r}[\varphi] = \varsigma_{\mathfrak{M}}/|\varphi| \equiv \varsigma_{\mathfrak{M},r}/|\varphi_r|; \tag{1}$$

e.g., see L'Ecuyer et al. (2010) and Chapter VI of Asmussen and Glynn (2007). To motivate our interest in RE, let $\widehat{\varsigma}_{\mathfrak{M},n}$ be a consistent estimator of $\varsigma_{\mathfrak{M}}$ in the sense that $\widehat{\varsigma}_{\mathfrak{M},n} \Rightarrow \varsigma_{\mathfrak{M}}$ as $n \to \infty$. For a given confidence level $\gamma \in (0,1)$, e.g., $\gamma = 0.95$, we then get $(\widehat{\varphi}_{\mathfrak{M},n} \pm z_{1-(1-\gamma)/2}\widehat{\varsigma}_{\mathfrak{M},n}/\sqrt{n})$ as an approximate *confidence interval* (CI) for φ for large n from the CLT and Slutsky's theorem (Serfling 1980, Theorem 1.5.4), where $\Phi(z_q) = q$ for $q \in (0,1)$ and Φ is the *cumulative distribution function* (CDF) of $\mathcal{N}(0,1)$. Obtaining a CI with a pre-specified relative half-width $\varepsilon > 0$ (i.e., a CI of the form $(\widehat{\varphi}_{\mathfrak{M},n} \pm \varepsilon|\widehat{\varphi}_{\mathfrak{M},n}|)$) entails a sample size n roughly proportional to the squared $\mathrm{RE}_{\mathfrak{M},r}[\varphi]$, showing the relevance of RE.

We want to study the behavior of $\mathrm{RE}_{\mathfrak{M},r}[\varphi]$ as $r \to r_0$ for some limiting value $r_0 \in \overline{\mathfrak{R}} \equiv [-\infty,\infty]$, with each $r \in \mathfrak{R} \equiv (-\infty,\infty)$. If $\mathrm{RE}_{\mathfrak{M},r}[\varphi]$ vanishes (resp., is bounded or unbounded) as $r \to r_0$, we say that the estimator $\widehat{\varphi}_{\mathfrak{M},n}$ has *vanishing* (resp., *bounded* or *unbounded*) *relative error* (VRE) (resp., BRE or URE).

We often consider estimands φ in (1) related to a random variable $Y \equiv Y_r$ with CDF $F \equiv F_r$, which we denote by $Y \sim F$, and let $f \equiv f_r$ be the derivative (when it exists) of F. For example, φ can be the mean $\mu = \mathbb{E}[Y] = \int y \, \mathrm{d}F(y)$, but our main focus will be on estimating a quantile of $Y \sim F$. For each $p \in (0,1)$, the p-quantile $\xi \equiv \xi_p$ of F (or equivalently of Y) is $\xi = F^{-1}(p) = \inf\{y : F(y) \geq p\}$.

We may estimate the p-quantile of $Y \sim F$ using SRS as follows. First generate a sample of n i.i.d. observations $Y_1, Y_2, \ldots, Y_n$ from F. Compute the *empirical distribution* $\widehat{F}_{\mathrm{SRS},n}$ as an estimator of F, with

$$\widehat{F}_{\mathrm{SRS},n}(y) = (1/n)\textstyle\sum_{i=1}^n I(Y_i \leq y), \tag{2}$$

where $I(\cdot)$ denotes the indicator function, which equals 1 (resp., 0) when its argument is true (resp., false). Then the SRS estimator of $\xi = F^{-1}(p)$ is $\widehat{\xi}_{\mathrm{SRS},n} = \widehat{F}^{-1}_{\mathrm{SRS},n}(p)$. For any fixed $p \in (0,1)$ such that $f(\xi_p) > 0$, which will be assumed throughout, $\widehat{\xi}_{\mathrm{SRS},n}$ obeys a CLT as $n \to \infty$ (Serfling 1980, Section 2.3.3):

$$\sqrt{n}[\widehat{\xi}_{\mathrm{SRS},n} - \xi] \Rightarrow \mathcal{N}(0, \kappa^2_{\mathrm{SRS}}), \quad \text{where} \tag{3}$$

$$\kappa^2_{\mathrm{SRS}} \equiv \kappa^2_{\mathrm{SRS},p} = \psi^2_{\mathrm{SRS}}/f^2(\xi_p), \ \text{ with } \ \psi^2_{\mathrm{SRS}} \equiv \psi^2_{\mathrm{SRS},p} = p(1-p). \tag{4}$$

Note that $\widehat{F}_{\mathrm{SRS},n}(\xi_p)$ has variance $\psi^2_{\mathrm{SRS},p}/n$, and Dong and Nakayama (2019) explain similar connections between estimating ξ_p and $F(\xi_p)$ for many MC methods. When $\xi_p \neq 0$, the SRS quantile estimator has

$$\mathrm{RE}_{\mathrm{SRS}}[\xi_p] = \kappa_{\mathrm{SRS}}/|\xi_p|, \tag{5}$$

as in (1). Let $\widehat{\kappa}_{\mathrm{SRS},n}$ be a consistent estimator of κ_{SRS}, e.g., from Corollary 2.5.2 of Serfling (1980). For $\gamma \in (0,1)$, an approximate γ-level *upper confidence bound* (UCB) U_n for ξ_p based on a sample size of n is

$$U_n = \widehat{\xi}_{\mathrm{SRS},n} + z_\gamma \widehat{\kappa}_{\mathrm{SRS},n}/\sqrt{n}. \tag{6}$$

We can similarly build an approximate γ-level *lower confidence bound* (LCB) or two-sided CI for ξ_p.

For comparison, we sometimes also consider estimating the mean μ of $Y \sim F$ by its SRS estimator $\widehat{\mu}_{\mathrm{SRS},n} = (1/n)\sum_{i=1}^n Y_i$. Assuming that the variance $\mathbb{V}[Y] = \mathbb{E}[(Y-\mu)^2]$ of $Y \sim F$ is $\sigma^2 \in (0,\infty)$, the SRS estimator of μ obeys a CLT $\sqrt{n}[\widehat{\mu}_{\mathrm{SRS},n} - \mu] \Rightarrow \mathcal{N}(0,\sigma^2)$ as $n \to \infty$ (Serfling 1980, Section 1.9.1), so as in (1) when $\mu \neq 0$, the RE of the SRS estimator of the mean is given by

$$\mathrm{RE}_{\mathrm{SRS}}[\mu] = \sigma/|\mu|. \tag{7}$$

We can also estimate ξ_p and μ using some variance-reduction technique (Asmussen and Glynn 2007, Chapter V), such as IS, and the resulting estimators will also obey CLTs under certain assumptions. To develop IS, assume that $Y = v(\mathbf{X})$, where $\mathbf{X} = (X_1, X_2, \ldots, X_d) \sim G$ is an $\Re^d$-valued input random vector that is fed into a function $v : \Re^d \to \Re$ to produce $Y \sim F$. Let $\widetilde{G}$ be another joint CDF on $\Re^d$ whose measure is absolutely continuous with respect to that of G. We can then apply a *change of measure* to write $\mu = \mathbb{E}[Y] = \mathbb{E}_G[v(\mathbf{X})] = \int_{\Re^d} v(\mathbf{x})\,\mathrm{d}G(\mathbf{x}) = \int_{\Re^d} v(\mathbf{x})\,\frac{\mathrm{d}G(\mathbf{x})}{\mathrm{d}\widetilde{G}(\mathbf{x})}\mathrm{d}\widetilde{G}(\mathbf{x}) = \mathbb{E}_{\widetilde{G}}[v(\mathbf{X})R(\mathbf{X})]$, where $\mathbb{E}_K$ (resp., $\mathbb{V}_K$) denotes the expectation (resp., variance) operator when $\mathbf{X} \sim K$, and $R(\mathbf{x}) = \mathrm{d}G(\mathbf{x})/\mathrm{d}\widetilde{G}(\mathbf{x})$ is the *likelihood ratio* (LR). By sampling i.i.d. copies $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ of $\mathbf{X} \sim \widetilde{G}$, we then obtain an unbiased estimator of μ as $\widehat{\mu}_{\mathrm{IS},n} = (1/n)\sum_{i=1}^n v(\mathbf{X}_i)R(\mathbf{X}_i)$. The IS estimator $\widehat{\mu}_{\mathrm{IS},n}$ of the mean obeys a CLT $\sqrt{n}[\widehat{\mu}_{\mathrm{IS},n} - \mu] \Rightarrow \mathcal{N}(0, \sigma_{\mathrm{IS}}^2)$ as $n \to \infty$ when $\sigma_{\mathrm{IS}}^2 \equiv \mathbb{V}_{\widetilde{G}}[v(\mathbf{X})R(\mathbf{X})] \in (0,\infty)$. If $\mu \neq 0$, the relative error of the IS estimator $\widehat{\mu}_{\mathrm{IS},n}$ is

$$\mathrm{RE}_{\mathrm{IS}}[\mu] = \sigma_{\mathrm{IS}}/|\mu|. \tag{8}$$

Glynn (1996) develops IS for estimating ξ_p. A change of measure yields $1 - F(y) = \mathbb{E}_G[I(v(\mathbf{X}) > y)] = \mathbb{E}_{\widetilde{G}}[I(v(\mathbf{X}) > y)R(\mathbf{X})]$, so an unbiased estimator of $F(y)$ is $\widehat{F}_{\mathrm{IS},n}(y) = 1 - (1/n)\sum_{i=1}^n I(v(\mathbf{X}_i) > y)R(\mathbf{X}_i)$ with each $\mathbf{X}_i \sim \widetilde{G}$. Inverting the IS CDF estimator $\widehat{F}_{\mathrm{IS},n}$ leads to the IS p-quantile estimator as $\widehat{\xi}_{\mathrm{IS},n} = \widehat{F}_{\mathrm{IS},n}^{-1}(p)$. If there exists constants $\varepsilon_0 > 0$ and $\lambda_0 > 0$ such that $\mathbb{E}_{\widetilde{G}}[I(v(\mathbf{X}) > \xi_p - \lambda_0)R^{2+\varepsilon_0}(\mathbf{X})] < \infty$, then the IS quantile estimator obeys a CLT $\sqrt{n}[\widehat{\xi}_{\mathrm{IS},n} - \xi_p] \Rightarrow \mathcal{N}(0, \kappa_{\mathrm{IS}}^2)$ as $n \to \infty$, where

$$\kappa_{\mathrm{IS}}^2 \equiv \kappa_{\mathrm{IS},p}^2 = \psi_{\mathrm{IS},p}^2/f^2(\xi_p), \quad \text{with} \quad \psi_{\mathrm{IS},p}^2 = \mathbb{E}_{\widetilde{G}}[I(v(\mathbf{X}) > \xi_p)R^2(\mathbf{X})] - (1-p)^2; \tag{9}$$

e.g., see Glynn (1996) and Chu and Nakayama (2012). Note that $\widehat{F}_{\mathrm{IS},n}(\xi_p)$ has variance $\psi_{\mathrm{IS},p}^2/n$, which depends on $\widetilde{G}$ but $f(\xi_p)$ in (9) does not. As in (1) when $\xi_p \neq 0$, the RE of the IS p-quantile estimator is

$$\mathrm{RE}_{\mathrm{IS}}[\xi_p] = \kappa_{\mathrm{IS}}/|\xi_p|. \tag{10}$$

Our goal is to examine the RE of estimators of ξ_p and μ under various asymptotic regimes when $r \to r_0$. Quantities that we allow to depend on r include the quantile level p or parameters of the CDF F of Y (or both). For example, we can take $p = r$ and consider $r_0 = 0$ or $r_0 = 1$. Another possibility is that the mean or variance of F is r, and let $r_0 = 0$ or $r_0 = \infty$. Our notation often omits r to simplify expressions.

We use the following asymptotic notation to describe the RE behaviors as $r \to r_0$. For functions $h_1(r)$ and $h_2(r)$, we write $h_1(r) = O(h_2(r))$ (resp., $h_1(r) = \Omega(h_2(r))$) as $r \to r_0$ if there is a constant $c_1 > 0$ such that $|h_1(r)| \leq c_1|h_2(r)|$ (resp., $|h_1(r)| \geq c_1|h_2(r)|$) for all sufficiently large (resp., small) r when $r \uparrow r_0$ (resp., $r \downarrow r_0$). Moreover, $h_1(r) = \Theta(h_2(r))$ as $r \to r_0$ if both $h_1(r) = O(h_2(r))$ and $h_1(r) = \Omega(h_2(r))$, and $h_1(r) = o(h_2(r))$ (resp., $h_1(r) = \omega(h_2(r))$) as $r \to r_0$ if $\lim_{r \to r_0} h_1(r)/h_2(r) = 0$ (resp., $\lim_{r \to r_0} h_2(r)/h_1(r) = 0$).

3 LOCATION-SCALE FAMILY OF DISTRIBUTIONS

Suppose that $Y = a + bX$ for a scalar random variable X, where $X \sim G$, and $a, b \in \Re$. For example, if X is a mean-1 exponential with CDF $G(x) = [1 - e^{-x}]I(x \geq 0)$, then for $b > 0$, Y is a mean-b exponential shifted by a. Let ν and $\tau^2 > 0$ be the mean and variance of G, and let g denote the derivative (when it exists) of G. Let $\eta_p = G^{-1}(p)$ be the p-quantile of $X \sim G$, and assume that $g(\eta_p) > 0$. For Y, its CDF F and its derivative f then satisfy $F(y) = G\left(\frac{y-a}{b}\right)$ and $f(y) = \frac{1}{b}g\left(\frac{y-a}{b}\right)$. Also, F has mean $\mu = a + b\nu$ and variance $\sigma^2 = b^2\tau^2$.

If $b > 0$, the p-quantile of $Y \sim F$ is $\xi_p = a + b\eta_p$, and $f(\xi_p) = \frac{1}{b}g\left(\frac{\xi_p - a}{b}\right) = \frac{1}{b}g(\eta_p)$ (Parzen 2004), so

$$\mathrm{RE}_{\mathrm{SRS}}[\xi_p] = \frac{\sqrt{p(1-p)}}{(1/b)g(\eta_p)|a+b\eta_p|} = \frac{\sqrt{p(1-p)}}{g(\eta_p)|\frac{a}{b}+\eta_p|} \tag{11}$$

by (5). (If instead $b < 0$ and $g(\eta_{1-p}) > 0$ for $\eta_{1-p} = G^{-1}(1-p)$, then $\xi_p = a + b\eta_{1-p}$, and $\mathrm{RE}_{\mathrm{SRS}}[\xi_p] = \frac{\sqrt{p(1-p)}}{g(\eta_{1-p})|\frac{a}{b}+\eta_{1-p}|}$. For simplicity, we will only consider the case when $b > 0$ from now on.)

We consider different possible choices for the parameters (r, r_0) in the asymptotics.

1. Let $(r, r_0) = (b, \infty)$, with a and p fixed. For estimand $\xi_p \equiv \xi_{p,b}$, (11) implies that $\mathrm{RE}_{\mathrm{SRS}}[\xi_p]$ remains bounded as $b \to \infty$ when $\eta_p \neq 0$, so the SRS estimator of ξ_p has BRE in this asymptotic regime. In contrast, suppose the estimand is the mean $\mu \equiv \mu_b$ of F when $\nu = 0$ and $a \neq 0$, so $\mu = a \neq 0$. The variance $\sigma^2 \equiv \sigma_b^2$ of F grows as $b^2\tau^2$. Hence, the RE from (7) behaves as $\mathrm{RE}_{\mathrm{SRS}}[\mu] = b\tau/a \to \infty$ as $b \to \infty$. Thus, in this case for SRS, the mean becomes more difficult to estimate (URE) as $b \to \infty$, but the p-quantile has BRE for each fixed $p \in (0,1)$.
   If instead $\nu \neq 0$, then $|\mu| = |a + b\nu| \to \infty$ as $b \to \infty$. In this case, $\mathrm{RE}_{\mathrm{SRS}}[\mu] = b\tau/|a+b\nu| \to \tau/|\nu|$ as $b \to \infty$, so the SRS estimator of μ has BRE, just like the SRS quantile estimator.

2. Let $(r, r_0) = (b, 0)$, with $a \neq 0$ and $p \in (0,1)$ fixed. Thus, the variance $\sigma^2 \equiv \sigma_b^2$ of F shrinks as $b^2\tau^2$ as $b \to 0$. By (11) and (7), both $\mathrm{RE}_{\mathrm{SRS}}[\xi_p]$ and $\mathrm{RE}_{\mathrm{SRS}}[\mu]$ shrink to 0 (VRE) as $b \to 0$.

4 SRS ESTIMATION OF p-QUANTILE AS $p \to 1$ OR $p \to 0$ FOR SPECIFIC DISTRIBUTIONS

For Y from some specific parametric families of distributions F, we now examine the RE in (5) of the SRS estimator of the p-quantile as $p \to 1$ or $p \to 0$. The mean μ and variance σ^2 of $Y \sim F$ do not vary with p, so $\mathrm{RE}_{\mathrm{SRS}}[\mu]$ is then fixed and finite when $\mu \neq 0$ and $\sigma^2 < \infty$.

1. Weibull with scale parameter $\upsilon > 0$ and shape parameter $k > 0$: The CDF is $F(y) = [1 - e^{-(y/\upsilon)^k}]I(y \geq 0)$ with density $f(y) = (k/\upsilon)(y/\upsilon)^{k-1}e^{-(y/\upsilon)^k}I(y \geq 0)$, so $k = 1$ results in an exponential distribution. For fixed $\upsilon > 0$, the right tail of the density gets heavier as k shrinks, always with finite variance. The p-quantile is $\xi_p = \upsilon[-\ln(1-p)]^{1/k}$. From (4), the asymptotic variance of the SRS estimator of ξ_p is $\kappa_{\mathrm{SRS}}^2 = \upsilon^2 p[-\ln(1-p)]^{(2/k)-2}/[k^2(1-p)]$, so (5) leads to
$$\mathrm{RE}_{\mathrm{SRS}}[\xi_p] = \tfrac{-1}{k\ln(1-p)}\sqrt{\tfrac{p}{1-p}} \to \infty \quad \text{(URE)} \quad \text{as } p \to 1 \text{ or } p \to 0. \tag{12}$$

2. (Generalized) Pareto with $(\vartheta > 0, a_0 \in \Re, b_0 > 0) = (\text{shape}, \text{location}, \text{scale})$ parameters: The CDF is $F(y) = 1 - [1 + ((y - a_0)/b_0)]^{-\vartheta}I(y \geq a_0)$ with density $f(y) = (\vartheta/b_0)[1 + ((y-a_0)/b_0)]^{-\vartheta-1}I(y \geq a_0)$, and $\sigma^2 < \infty$ if $\vartheta > 2$. The p-quantile is $\xi_p = a_0 + b_0[(1-p)^{-1/\vartheta} - 1]$, whose SRS estimator has asymptotic variance $\kappa_{\mathrm{SRS}}^2 = {b_0}^2 p/[\vartheta^2(1-p)^{(\vartheta+2)/\vartheta}]$. If $a_0 = 0$,
$$\mathrm{RE}_{\mathrm{SRS}}[\xi_p] = \tfrac{1}{\vartheta[1-(1-p)^{1/\vartheta}]}\sqrt{\tfrac{p}{1-p}} \to \infty \quad \text{(URE)} \quad \text{as } p \to 1 \text{ or } p \to 0. \tag{13}$$

3. Normal: Suppose that F is normal with mean μ and variance $\sigma^2 > 0$, so the density is $f(y) = \frac{1}{\sigma}\phi((y-\mu)/\sigma)$, with $\phi(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}$. The CDF is $F(y) = \Phi((y-\mu)/\sigma)$, with $\Phi(z) = \int_{-\infty}^{z}\phi(y)\,dy$. Let $\eta_p = \Phi^{-1}(p)$, so the p-quantile of F is $\xi_p = \mu + \sigma\eta_p$. From (4), the asymptotic variance of the SRS estimator of ξ_p is $\kappa_{\mathrm{SRS}}^2 = p(1-p)/[\phi(\eta_p)/\sigma]^2$, and it can be shown that (5) results in
$$\mathrm{RE}_{\mathrm{SRS}}[\xi_p] = \tfrac{\sigma\sqrt{p(1-p)}}{\phi(\eta_p)(\mu+\sigma\eta_p)} \to \infty \quad \text{(URE)} \quad \text{as } p \to 1 \text{ or } p \to 0.$$

Table 1 gives values of $\mathrm{RE}_{\mathrm{SRS}}[\xi_p]$ for p near 1, where each distribution's parameters are chosen so that the median is 2, and $a_0 = 0$ for the Pareto. The right tail of the Weibull (resp., Pareto) gets heavier as k (resp., ϑ) shrinks, and the RE grows by (12) (resp., (13)). Also, the Pareto has heavier tails and larger REs than the Weibull; e.g., for fixed k and ϑ, the ratio of (13) to (12) satisfies $-k\ln(1-p)/(\vartheta[1-(1-p)^{1/\vartheta}]) \to \infty$

Table 1: REs of SRS estimator of p-quantile ξ_p for normal, Weibull, and Pareto, each with median 2.

$1-p$	$\mathscr{N}(2,1)$	Weib$(k=2)$	Weib$(k=1)$	Weib$(k=1/2)$	Par$(\vartheta=3)$	Par$(\vartheta=2)$	Par$(\vartheta=1)$
10^{-1}	5.2E-01	6.5E-01	1.3E+00	2.6E+00	1.9E+00	2.2E+00	3.3E+00
10^{-2}	8.6E-01	1.1E+00	2.2E+00	4.3E+00	4.2E+00	5.5E+00	1.0E+01
10^{-3}	1.8E+00	2.3E+00	4.6E+00	9.2E+00	1.2E+01	1.6E+01	3.2E+01
10^{-4}	4.4E+00	5.4E+00	1.1E+01	2.2E+01	3.5E+01	5.1E+01	1.0E+02
10^{-5}	1.1E+01	1.4E+01	2.7E+01	5.5E+01	1.1E+02	1.6E+02	3.2E+02

as $p \to 1$. But not all F lead to URE for extreme quantiles; e.g., a uniform distribution on $[1,3]$ has $\mathrm{RE}_{\mathrm{SRS}}[\xi_p] = 2\sqrt{p(1-p)}/[1+2p] \to 0$ as $p \to \{0,1\}$, so VRE. (Section 8 gives examples of F where the SRS estimators of ξ_p have URE and VRE in other limiting regimes as $r \to r_0$ for fixed $0 \ll p \ll 1$.)

For any continuous F, examine now the SRS estimator $\widehat{F}_{\mathrm{SRS},n}(y)$ in (2) of $F(y)$ for y in the tails. For $u \in \{0,1\}$, let $y_u = \lim_{p\to u} F^{-1}(p) \in \overline{\Re}$. As $\mathbb{V}[\widehat{F}_{\mathrm{SRS},n}(y)] = F(y)[1-F(y)]/n = \mathbb{V}[1-\widehat{F}_{\mathrm{SRS},n}(y)]$ for all $y \in \Re$, we get $\mathrm{RE}_{\mathrm{SRS}}[F(y)] = \sqrt{[1-F(y)]/F(y)} \to \infty$ as $y \downarrow y_0$, and $\mathrm{RE}_{\mathrm{SRS}}[1-F(y)] = \sqrt{F(y)/[1-F(y)]} \to \infty$ as $y \uparrow y_1$, so SRS estimators of tail probabilities always have URE, in contrast to extreme quantiles.

5 IMPORTANCE SAMPLING FOR THE EXPONENTIAL

The previous sections discuss SRS estimation of the p-quantile ξ_p. For the special case of an exponential random variable $Y = v(X) = X \sim F = G$ with rate $\lambda > 0$ (i.e., Weibull with $k=1$ and $\lambda = 1/\upsilon$ in Section 4), we now consider estimating $\xi_p = -\ln(1-p)/\lambda$ via IS in which X is sampled from $\widetilde{G}$ that simply changes the exponential's rate to $\widetilde{\lambda} > 0$, leading to the LR $R(X) = \lambda e^{-\lambda x}/\widetilde{\lambda} e^{-\widetilde{\lambda} x}$ in (9). We will argue as in L'Ecuyer et al. (2009) to determine the value of $\widetilde{\lambda} = \widetilde{\lambda}_p$ that minimizes $\kappa^2_{\mathrm{IS}} = \kappa^2_{\mathrm{IS},p}$ in (9) and $\mathrm{RE}_{\mathrm{IS}}[\xi_p]$ in (10). For $\psi^2_{\mathrm{IS},p}$ in (9), its first term (i.e., second moment under IS), which is finite if and only if $\widetilde{\lambda} \in (0, 2\lambda)$, satisfies

$$\mathbb{E}_{\widetilde{G}}[I(X > \xi_p) R^2(X)] = \int_{\xi_p}^{\infty} \tfrac{\lambda^2 e^{-2\lambda x}}{\widetilde{\lambda} e^{-\widetilde{\lambda} x}}\, \mathrm{d}x = \tfrac{\lambda^2}{\widetilde{\lambda}} \int_{\xi_p}^{\infty} e^{-(2\lambda - \widetilde{\lambda})x}\, \mathrm{d}x = \tfrac{\lambda^2}{\widetilde{\lambda}(2\lambda - \widetilde{\lambda})} e^{-(2\lambda-\widetilde{\lambda})\xi_p}. \tag{14}$$

Then (14) is minimized at $\widetilde{\lambda} = \lambda + \frac{1}{\xi_p} - \left(\lambda^2 + \frac{1}{\xi^2}\right)^{1/2}$, which also minimizes $\kappa^2_{\mathrm{IS},p}$ in (9) as $f(\xi_p) = \lambda e^{-\lambda \xi_p} = \lambda(1-p)$ does not depend on $\widetilde{\lambda}$. Using $\xi_p = -\ln(1-p)/\lambda$ leads to

$$\widetilde{\lambda} = \lambda q \equiv \lambda \left[1 - \tfrac{1}{\ln(1-p)} - \left(1 + \tfrac{1}{[\ln(1-p)]^2}\right)^{1/2}\right], \tag{15}$$

and as a consequence, (9) and (14) imply $\psi^2_{\mathrm{IS},p} = \frac{(1-p)^{2-q}}{q(2-q)} - (1-p)^2$,

$$\kappa^2_{\mathrm{IS},p} = \tfrac{1}{\lambda^2}\left[\tfrac{(1-p)^{-q}}{q(2-q)} - 1\right], \quad \text{and} \quad \mathrm{RE}_{\mathrm{IS}}[\xi_p] = \left(\tfrac{(1-p)^{-q}}{q(2-q)[-\ln(1-p)]^2} - \tfrac{1}{[-\ln(1-p)]^2}\right)^{1/2}, \tag{16}$$

so $\mathrm{RE}_{\mathrm{IS}}[\xi_p]$ does not depend on λ. We consider three possibilities for the (parameter, limit) pair (r, r_0):

1. $(r, r_0) = (\lambda, 0)$ with p fixed (corresponding, e.g., to a highly reliable Markovian system (HRMS), where component failure rates go to 0; Goyal et al. 1992): By item 1 of Section 3, the SRS quantile estimator has BRE when $b \equiv 1/\lambda \to \infty$. For the IS estimator, (16) implies $\kappa^2_{\mathrm{IS},p} \to \infty$ as $\lambda \to 0$ (since q does not depend on λ), and λ does not appear in $\mathrm{RE}_{\mathrm{IS}}[\xi_p]$. Thus, as with SRS, the IS estimator of ξ_p also has BRE.
2. $(r, r_0) = (p, 1)$ with λ fixed: For $q \equiv q_p$ in (15), we have that as $p \to 1$,

$$q = 1 - \tfrac{1}{\ln(1-p)} - \left[1 + \tfrac{1}{[\ln(1-p)]^2}\right]^{1/2} = 1 - \tfrac{1}{\ln(1-p)} - \left[1 + \tfrac{1+o(1)}{2[\ln(1-p)]^2}\right] = -\tfrac{1+o(1)}{\ln(1-p)} \tag{17}$$

since $\ln(1-p) \to -\infty$. Putting (17) into (16) yields

$$\mathrm{RE}_{\mathrm{IS}}[\xi_p] = \Big(\tfrac{(1-p)^{[1+o(1)]/\ln(1-p)}}{[-(1+o(1))\ln(1-p)][2+(1+o(1))/\ln(1-p)]} - \tfrac{1}{[-\ln(1-p)]^2}\Big)^{1/2} = \Big(\tfrac{e^{1+o(1)}}{[2+o(1)][-\ln(1-p)]} - o(1)\Big)^{1/2}, \tag{18}$$

so $\mathrm{RE}_{\mathrm{IS}}[\xi_p] \to 0$ as $p \to 1$, i.e., the IS (resp., SRS) estimator of ξ_p has VRE (resp., URE by (12)).

3. $(r, r_0) = (\varepsilon, 0)$ with $\lambda = \varepsilon^c$ for some constant $c > 0$ and $p = 1 - \varepsilon$ (arising, e.g., when considering extreme quantiles for an HRMS): As in (17) and (18), we get as $\varepsilon \to 0$ that $q = -[1+o(1)]/\ln \varepsilon$ and $\mathrm{RE}_{\mathrm{IS}}[\xi_p] = (\frac{e^{1+o(1)}}{[2+o(1)][-\ln(\varepsilon)]} - o(1))^{1/2} \to 0$ (uniformly in c), so the IS estimator of ξ_p has VRE. The uniform convergence in c ensures that how extreme the two parameters for the rate and quantile become is not a critical issue. The SRS RE in (12) does not depend on λ, so SRS yields URE.

6 AVERAGES AND SUMS OF I.I.D. RANDOM VARIABLES

In this section, let $X_1, X_2, \ldots, X_m$ be m i.i.d. random variables with each $X_j \sim G_0$, where G_0 does not depend on m, $\mu_0 \equiv \mathbb{E}[X_j]$ and $\sigma_0^2 \equiv \mathbb{V}[X_j] \in (0, \infty)$. Let G denote the joint CDF of $\mathbf{X} = (X_1, X_2, \ldots, X_m)$. Below we will consider $Y \equiv Y(m)$ for $Y = (1/m)\sum_{j=1}^m X_j$ and $Y = \sum_{j=1}^m X_j$ as $m \to \infty$, so $(r, r_0) = (m, \infty)$.

6.1 Y is an Average of I.I.D. Summands

Consider $Y \equiv Y(m) = (1/m)\sum_{j=1}^m X_j$ as a sample average under SRS, assuming here that our simulation code is a black box that outputs only Y but not the individual summands X_j. Then $Y(m) \sim F_m$ obeys a CLT $Z(m) \equiv \frac{\sqrt{m}}{\sigma_0}[Y(m) - \mu_0] \Rightarrow \mathscr{N}(0,1)$ as $m \to \infty$. For K_m as the CDF of $Z(m)$ and fixed $p \in (0,1)$, the continuity of Φ^{-1} ensures $K_m^{-1}(p) \to \Phi^{-1}(p)$ as $m \to \infty$ (Van Der Vaart 1998, Lemma 21.2). As in Section 3, the p-quantile of $Y(m)$ satisfies $\xi_p(m) = \mu_0 + \frac{\sigma_0}{\sqrt{m}} K_m^{-1}(p) \approx \mu_0 + \frac{\sigma_0}{\sqrt{m}}\Phi^{-1}(p)$ for large m. We then construct an estimator $\widehat{\xi}_{p,n}(m)$ of $\xi_p(m)$ from a sample $Y_1, Y_2, \ldots, Y_n$ of n i.i.d. copies of $Y = Y(m)$ as follows. Note that $\mu(m) \equiv \mathbb{E}[Y(m)] = \mu_0$ and $\sigma^2(m) \equiv \mathbb{V}_{F_m}[Y(m)] = \sigma_0^2/m$, so we define $\widehat{\xi}_{p,n}(m) = \widehat{\mu}_n(m) + \widehat{\sigma}_n(m)\Phi^{-1}(p)$, where $\widehat{\mu}_n(m) = (1/n)\sum_{i=1}^n Y_i$ and $\widehat{\sigma}_n^2(m) = (1/(n-1))\sum_{i=1}^n [Y_i - \widehat{\mu}_n(m)]^2$. To simplify the discussion, assume now that each $X_j \sim \mathscr{N}(\mu_0, \sigma_0^2)$, so $Y(m) \sim \mathscr{N}(\mu_0, \sigma_0^2/m)$. Then the independence of $\widehat{\mu}_n(m)$ and $\widehat{\sigma}_n^2(m)$ (Casella and Berger 2002, Theorem 5.3.1) implies that $\mathbb{V}_{F_m}[\widehat{\xi}_{p,n}(m)] = \mathbb{V}_{F_m}[\widehat{\mu}_n(m)] + [\Phi^{-1}(p)]^2 \mathbb{V}_{F_m}[\widehat{\sigma}_n^2(m)] = [\sigma_0^2/(mn)] + 2[\Phi^{-1}(p)]^2(\sigma_0^2/m)^2/(n-1) \to 0$ as $m \to \infty$ for fixed n. Also, $\xi_p(m) \to \mu_0$ as $m \to \infty$, so $\mathrm{RE}_{\mathrm{SRS}}[\xi_p(m)] \to 0$ when $\mu_0 \neq 0$ for any fixed $p \in (0,1)$, i.e., SRS yields VRE.

6.2 Y is an I.I.D. Sum

Now consider $Y \equiv Y(m) = \sum_{j=1}^m X_j$, assuming here that SRS entails sampling the summands $X_j \sim G_0$ and returning the sum Y, and we compare SRS with IS that changes G_0. For $Y \sim F \equiv F_m$, let $\mu \equiv \mu(m) = \mathbb{E}[Y]$ and $\xi_p \equiv \xi_p(m) = F^{-1}(p)$. We will study an asymptotic regime from Glynn (1996) in which the number of summands $r = m \to \infty$ and simultaneously the quantile level p approaches 1 exponentially fast:

$$p \equiv p_m = 1 - e^{-\beta_0 m} \quad \text{for some constant } \beta_0 > 0. \tag{19}$$

For SRS and IS with an exponential twist, as developed in Glynn (1996) and further analyzed in Li et al. (2024), we will see that $\mathrm{RE}_{\mathrm{SRS}}[\xi_p] \to \infty$, $\mathrm{RE}_{\mathrm{IS}}[\xi_p] \to 0$, $\mathrm{RE}_{\mathrm{SRS}}[\mu] \to 0$, and $\mathrm{RE}_{\mathrm{IS}}[\mu] \to \infty$ as $m \to \infty$, where the REs are as given in (5), (10), (7), and (8).

To develop the IS, let $M_0(\theta) = \int e^{\theta x}\, \mathrm{d}G_0(x)$, $\theta \in \Re$, be the *moment generating function* (MGF) of G_0. Let $\Delta = \{\theta \in \Re : M_0(\theta) < \infty\}$ be the MGF's domain, and assume that its interior Δ° contains 0, which restricts us to light-tailed summands. Let $Q_0(\theta) = \ln M_0(\theta)$ be the *cumulant generating function*, and define $Q_0'(\theta) = \frac{d}{d\theta}Q(\theta)$ as its first derivative. For $\theta \in \Delta^\circ$, the exponential twist $\widetilde{G}_{0,\theta}$ of the marginal CDF

G_0 of X_j is given by $\mathrm{d}\widetilde{G}_{0,\theta}(x) = e^{\theta x}\,\mathrm{d}G_0(x)/M_0(\theta) = e^{\theta x - Q_0(\theta)}\,\mathrm{d}G_0(x)$, $x \in \Re$ (Asmussen and Glynn 2007, Section V.1b). We then define the method $\mathrm{IS}(\theta)$ with $\theta \in \Delta^\circ$ so that $X_1, X_2, \ldots, X_m$ are i.i.d., each with marginal distribution $\widetilde{G}_{0,\theta}$; let $\widetilde{G}_\theta$ denote the resulting joint CDF of $(X_1, X_2, \ldots, X_m)$. When $\mu_0 \neq 0$, Theorem 6 of Li et al. (2024) shows that as $m \to \infty$, $\mathrm{RE}_{\mathrm{SRS}}[\mu] = \Theta(1/\sqrt{m}) \to 0$ and $\mathrm{RE}_{\mathrm{IS}(\theta)}[\mu] = \Omega([\alpha(\theta)]^{m/2}/\sqrt{m}) \to \infty$ for *any* $\theta \neq 0$ with $\pm\theta \in \Delta^\circ$, where $\alpha(\theta) \equiv M_0(\theta)M_0(-\theta) \in (1, \infty)$. Thus, when estimating the mean μ of $Y \sim F$, SRS yields VRE, but $\mathrm{IS}(\theta)$ with any $\theta \neq 0$ leads to exponentially increasing RE.

To estimate the p-quantile for p in (19) and $m \to \infty$, Glynn (1996) applies $\mathrm{IS}(\theta_\star)$ for $\theta_\star > 0$ satisfying

$$-\theta_\star Q_0'(\theta_\star) + Q_0(\theta_\star) = -\beta_0, \tag{20}$$

assuming $\theta_\star \in \Delta^\circ$ exists. Motivated by large-deviations theory, (20) has a root $\theta_\star > 0$ if Q_0 is "steep" (Dembo and Zeitouni 1998, p. 44), which holds, e.g., when G_0 is normal or gamma.

For the SRS and $\mathrm{IS}(\theta_\star)$ estimators of the p-quantile ξ_p, Glynn (1996) analyzes the numerators (but not the denominator $f^2(\xi_p)$) of their asymptotic variances in (4) and (9), showing that $\mathrm{IS}(\theta_\star)$ can substantially reduce variance compared to SRS. Theorem 7 of Li et al. (2024) provides a fuller picture by further handling the denominator $f^2(\xi_p)$ in (4) and (9) through a "saddlepoint approximation" (Jensen 1995, Chapter 2), establishing that as $m \to \infty$, $\mathrm{RE}_{\mathrm{SRS}}[\xi_p] = \omega(e^{(\beta_0/2)m - \sqrt{m}}/\sqrt{m}) \to \infty$ and $\mathrm{RE}_{\mathrm{IS}(\theta_\star)}[\xi_p] = O(1/\sqrt{m}) \to 0$ when the characteristic function $C_0(\theta) = M_0(\theta\sqrt{-1})$, $\theta \in \Re$, of G_0 satisfies $\int_{\Re} |C_0(\theta)|^{q_0}\,\mathrm{d}\theta < \infty$ for some $q_0 \geq 1$. Thus, the SRS estimator of ξ_p has exponentially increasing RE, but the $\mathrm{IS}(\theta_\star)$ estimator has VRE.

7 HITTING TIME TO A RARELY VISITED SET OF REGENERATIVE SYSTEM

When Y is the average of m i.i.d. random variables (Section 6.1), an asymptotic regime may be based on a CLT. Glynn et al. (2018) consider another limiting setting using a different weak-convergence result for exponential rarity in geometric sums (Kalashnikov 1997). Consider a (non-delayed) *regenerative process* $X = [X(t) : t \geq 0]$ evolving on a state space $\mathscr{S} \subseteq \Re^d$ (Asmussen and Glynn 2007, Section IV.6b), so the process "probabilistically restarts" at a sequence of regeneration times $0 = \Gamma_0 < \Gamma_1 < \Gamma_2 < \cdots$ of X. Examples of regenerative times include the starts of busy periods of a stable GI/G/1 queue, or successive entrance times to a fixed state in a positive-recurrent Markov chain. We call $[X(\Gamma_{i-1} + s) : 0 \leq s < \tau_i]$ the ith (regenerative) *cycle*, which has length $\tau_i = \Gamma_i - \Gamma_{i-1}$. The cycles are i.i.d. by the regenerative property.

For $\mathscr{A} \subset \mathscr{S}$, let $T = \inf\{t \geq 0 : X(t) \in \mathscr{A}\}$ be the *hitting time* to $\mathscr{A}$. Typical examples are the hitting time to a large level for a stable GI/G/1 queue-length process, or the first time to failure for highly reliable Markovian systems (Goyal et al. 1992). For F as the CDF of T, our aim is to determine a p-quantile $\xi_p = F^{-1}(p)$ for fixed $p \in (0,1)$. For $i \geq 1$, let $T_i = \inf\{t \geq 0 : X(\Gamma_{i-1} + t) \in \mathscr{A}\}$ be the time elapsing after Γ_{i-1} until the next hit to $\mathscr{A}$, and $M = \min\{i \geq 1 : T_i < \tau_i\} - 1$ is the number of cycles completed before the one in which $\mathscr{A}$ is first hit. Thus, the hitting time is $T = \sum_{i=1}^{M} \tau_i + T_{M+1}$. The regenerative structure allows writing the expected hitting time $\mu = \mathbb{E}[T]$ as a ratio $\mu = \mathbb{E}[\min(T, \tau)]/\zeta$ with $\tau = \tau_1$ and $\zeta = \mathbb{P}(T < \tau)$ (e.g., see Goyal et al. 1992 and Glynn et al. 2017).

Now parameterize the model by r so that $\zeta \equiv \zeta_r \to 0$ as $r \to r_0$; e.g., in a stable GI/G/1 queue-length process, let r be a large queue threshold with $T \equiv T_r$ as the hitting time of r, and $r_0 = \infty$. Thus, hitting $\mathscr{A}$ in a cycle becomes rarer as $r \to r_0$, and the scaled hitting time T/μ converges weakly to an exponential under various assumptions (Kalashnikov 1997, Chapter 3): for $\mathbb{P} \equiv \mathbb{P}_r$ as the probability measure,

$$\mathbb{P}(T/\mu \leq x) \to 1 - e^{-x} \quad \text{as } r \to r_0 \tag{21}$$

for each $x \geq 0$. The advantage of this asymptotic property is that determining quantiles when ζ is small can be reduced to computing the mean $\mu \equiv \mu_r$ since (21) implies $\xi_p \approx -\mu \ln(1-p)$ for fixed $p \in (0,1)$. Specifically, let $\widehat{\mu}_n$ be an estimator of μ based on a sample of n cycles, so an estimator of ξ_p is

$$\widehat{\xi}_{p,n} = -\widehat{\mu}_n \ln(1-p). \tag{22}$$

Several efficient estimators have been developed for μ (Glynn et al. 2018). For any fixed quantile level $p \in (0,1)$, $\widehat{\xi}_{p,n}$ has BRE, VRE or URE as $r \to r_0$ if and only if $\widehat{\mu}_n$ has the same corresponding property. The proof is immediate since $\widehat{\xi}_{p,n}$ is simply a linear transformation of $\widehat{\mu}_n$.

While (22) is based on the limiting result (21), a real system has fixed r, leading to $\widehat{\xi}_{p,n}$ having bias that does not vanish as $n \to \infty$. But the REs in (5) and (10) do not account for such non-vanishing bias, so the asymptotic variance in the RE should be replaced with the mean squared error, which can be decomposed as the sum of the squared biased and variance. This is left to future research.

8 SRS HYPOTHESIS TESTING VIA QUANTILE OR CDF

Hypotheses about a quantile can often be expressed equivalently through the CDF (Serfling 1980, Section 2.3.7), and the associated statistical tests can have quite different efficiencies. A reason stems from the asymptotic variance κ^2_{SRS} in (4) of the SRS estimator of the p-quantile ξ_p. The numerator ψ^2_{SRS} is the variance of $\sqrt{n}[\widehat{F}_{\text{SRS},n}(\xi_p) - p]$ from (2). But the denominator $f^2(\xi_p)$ of κ^2_{SRS} can be large or small, even for a non-extreme quantile level $0 \ll p \ll 1$. We will show this through specific examples.

To motivate the discussion, consider a system with a capacity t_0 to withstand a random load $Y \sim F$, and the system fails (e.g., damage occurs) when $Y > t_0$. A regulator overseeing the system requires strong evidence that $\mathbb{P}(Y \le t_0) > p_0$ for a specified probability $p_0 \in (0,1)$, e.g., $p_0 = 0.95$ or $p_0 = 0.999$. We then define the following null and alternative hypotheses about $Y \sim F$:

$$H_0 : F(t_0) \le p_0 \quad \text{vs.} \quad H_A : F(t_0) > p_0, \tag{23}$$

and a statistical decision is to be made at a given significance level $\alpha \in (0,1)$, say $\alpha = 0.05$. Thus, establishing regulatory compliance entails rejecting H_0 in favor of H_A at level α. For example, for a postulated accident of a nuclear power plant, the U.S. Nuclear Regulatory Commission (2010) considers (23) with the load Y as the peak cladding temperature (PCT) having a capacity of $t_0 = 2200°$ Fahrenheit and $p_0 = 0.95$, and H_A must be established with at least confidence level $\gamma = 1-\alpha = 0.95$ (U.S. Nuclear Regulatory Commission 2005, Section 3.2). This is known as a "$\gamma|p_0$ criterion" (U.S. Nuclear Regulatory Commission 2011, Section 24.9), which can be demonstrated via SRS (in a large-sample setting) through a level $\alpha = 1-\gamma$ hypothesis test as follows. From an i.i.d. sample $Y_1, Y_2, \dots, Y_n$ of size n from F, construct an approximate γ-level lower confidence bound L_n for $F(t_0)$, and reject H_0 in favor of H_A when

$$L_n > p_0. \tag{24}$$

We can obtain such an LCB L_n by noting that the empirical distribution function in (2) obeys a CLT

$$\sqrt{n}\left[\widehat{F}_{\text{SRS},n}(t_0) - F(t_0)\right] \Rightarrow \mathscr{N}(0, \psi_0^2) \tag{25}$$

as $n \to \infty$, where $\psi_0^2 = F(t_0)[1-F(t_0)]$, leading to an approximate (for large n) γ-level LCB for $F(t_0)$ as

$$L_n = \widehat{F}_{\text{SRS},n}(t_0) - z_\gamma \widehat{\psi}_{0,n}/\sqrt{n} \tag{26}$$

with $\widehat{\psi}_{0,n}$ a consistent estimator of ψ_0, e.g., using $\widehat{\psi}^2_{0,n} = \widehat{F}_{\text{SRS},n}(t_0)[1-\widehat{F}_{\text{SRS},n}(t_0)]$.

We can also study (23) through the p_0-quantile ξ_{p_0}, where we assume that $f(\xi_{p_0}) > 0$. In this case, $F(t_0) \le p_0$ if and only if $\xi_{p_0} \ge t_0$, so we can equivalently express (23) as hypotheses

$$H_0' : \xi_{p_0} \ge t_0 \quad \text{vs.} \quad H_A' : \xi_{p_0} < t_0, \tag{27}$$

and use the same significance level α. For an approximate γ-level upper confidence bound U_n for ξ_{p_0}, e.g., as in (6) with $p = p_0$ and sample size n, we make a statistical decision that the $\gamma|p_0$ criterion holds if

$$U_n < t_0. \tag{28}$$

Wilks (1941) provides an asymptotically equivalent approach; see Section 2.6.1 of Serfling (1980).

In certain cases, when H_A in (23) and equivalently H'_A in (27) are actually true, a statistical test may be able to reject H_0 through (24) with a much smaller sample size than would be required to reject H'_0 via a test (28), or vice versa. First consider trying to establish H_A holds in (23) with SRS through (24) using a LCB L_n in (26) at confidence level $\gamma = 0.95$. Applying ideas related to Pitman efficiency (Serfling 1980, Section 10.2), we want to determine the sample size n so that (24) holds with a given probability (power) $\beta \in (1-\gamma, 1)$. For $\overline{\Phi}(\cdot) = 1 - \Phi(\cdot)$, the probability of (24) occurring is

$$\mathbb{P}(L_n > p_0) = \mathbb{P}\left(\tfrac{\sqrt{n}}{\widehat{\psi}_{0,n}}\left[\widehat{F}_{\mathrm{SRS},n}(t_0) - F(t_0)\right] > z_\gamma + \tfrac{\sqrt{n}}{\widehat{\psi}_{0,n}}\left[p_0 - F(t_0)\right]\right) \approx \overline{\Phi}\left(z_\gamma + \tfrac{\sqrt{n}}{\psi_0}\left[p_0 - F(t_0)\right]\right) \tag{29}$$

when n is large by the CLT (25) and the consistency of $\widehat{\psi}_{0,n}$. Equating the right side of (29) to β yields $z_\gamma + \frac{\sqrt{n}}{\psi_0}[p_0 - F(t_0)] = \Phi^{-1}(1-\beta) = z_{1-\beta} = -z_\beta$, so $\psi_0^2 = F(t_0)[1-F(t_0)]$ gives the needed sample size

$$n \equiv n_{\gamma,\beta} = \tfrac{(z_\gamma + z_\beta)^2 F(t_0)[1-F(t_0)]}{[F(t_0) - p_0]^2}. \tag{30}$$

Next instead consider trying to establish H'_A holds in (27) by showing a $\gamma = 0.95$ level UCB U_n from (6) for the p_0-quantile satisfies (28), for $p = p_0$ in (4) and (6) with $\widehat{\xi}_{\mathrm{SRS},n} = \widehat{F}^{-1}_{\mathrm{SRS},n}(p_0)$. We now determine the sample size $n = n'$ so that (28) holds with a given power $\beta \in (1-\gamma, 1)$. Event (28) has probability

$$\mathbb{P}(U_n < t_0) = \mathbb{P}\left(\tfrac{\sqrt{n}}{\widehat{\kappa}_{\mathrm{SRS},n}}\left[\widehat{\xi}_{\mathrm{SRS},n} - \xi_{p_0}\right] < -z_\gamma + \tfrac{\sqrt{n}}{\widehat{\kappa}_{\mathrm{SRS},n}}\left[t_0 - \xi_{p_0}\right]\right) \approx \Phi\left(-z_\gamma + \tfrac{\sqrt{n}}{\kappa_{\mathrm{SRS}}}\left[t_0 - \xi_{p_0}\right]\right) \tag{31}$$

when n is large by the CLT (3) and the consistency of $\widehat{\kappa}_{\mathrm{SRS},n}$. Equating the right side of (31) to β yields $-z_\gamma + \frac{\sqrt{n}}{\kappa_{\mathrm{SRS}}}[t_0 - \xi_{p_0}] = \Phi^{-1}(\beta) = z_\beta$, leading to the required sample size, by (4), as

$$n' \equiv n'_{\gamma,\beta} = \tfrac{(z_\gamma + z_\beta)^2 \kappa^2_{\mathrm{SRS}}}{(t_0 - \xi_{p_0})^2} = \tfrac{(z_\gamma + z_\beta)^2 p_0 (1-p_0)}{(t_0 - \xi_{p_0})^2 f^2(\xi_{p_0})}. \tag{32}$$

Below we will compare n in (30) and n' in (32) for two examples of normal mixtures. We take CDF

$$F(y) = w_1 \Phi_{\mu_1,\sigma_1}(y) + w_2 \Phi_{\mu_2,\sigma_2}(y), \tag{33}$$

where $w_1, w_2 \geq 0$ are mixing weights with $w_1 + w_2 = 1$, and $\Phi_{a,b}$ is a normal CDF with mean a and variance b^2, so $\Phi_{a,b}(x) = \Phi((x-a)/b)$. Let $\phi_{a,b}$ be the density of $\Phi_{a,b}$, so $\phi_{a,b}(x) = (1/b)\phi((x-a)/b)$, where ϕ is the density of Φ, and $f(y) = w_1 \phi_{\mu_1,\sigma_1}(y) + w_2 \phi_{\mu_2,\sigma_2}(y)$ is the density of F. We will consider various choices for the parameters (including r for the asymptotic regime), leading to vastly different behaviors for the RE of the SRS estimator $\widehat{\xi}_{\mathrm{SRS},n} = \widehat{F}^{-1}_{\mathrm{SRS},n}(p_0)$ of ξ_{p_0}, with fixed $p_0 \in (0,1)$, for $\widehat{F}_{\mathrm{SRS},n}$ in (2).

8.1 Example where SRS Estimator of ξ_{p_0} for Fixed p_0 has URE

First assume that the component means in (33) satisfy $\mu_1 \equiv \mu_{1,r} = c_0 - r$ and $\mu_2 \equiv \mu_{2,r} = c_0 + r$ for some fixed constant c_0, and consider the asymptotic regime in which $r \to r_0 = \infty$, where all other parameters are fixed. Under this parameterization, let $F \equiv F_r$, $f \equiv f_r$, and $\xi_{p_0} \equiv \xi_{p_0,r}$ be the CDF, density, and true p_0-quantile, respectively. The density f_r will roughly have two humps, one centered at $\mu_{1,r}$ with approximate mass w_1, and another centered at $\mu_{2,r}$ with roughly mass w_2, with c_0 in the middle between the two humps. As r grows, the two humps move away from each other, and the density f_r at c_0 shrinks to 0. We choose the quantile level $p_0 = w_1$, so the true p_0-quantile $\xi_{p_0,r} \to c_0$ as $r \to \infty$. The left (resp., right) side of Figure 1 shows the density (resp., CDF) of the mixture with $w_1 = 0.7$, $w_2 = 0.3$, $\sigma_1^2 = \sigma_2^2 = 1$, $c_0 = 5$, and $r = 5$, so $\mu_1 = 0$ and $\mu_2 = 10$. In this case, the p_0-quantile is $\xi_{p_0,r} \approx c_0 = 5$, and we then get $f_r(\xi_{p_0,r}) \doteq 1.49\mathrm{E}{-06}$ and $\mathrm{RE}_{\mathrm{SRS}}[\xi_{p_0,r}] \doteq 6.16\mathrm{E}{+04}$ by (5). The asymptotic variance of the SRS estimator

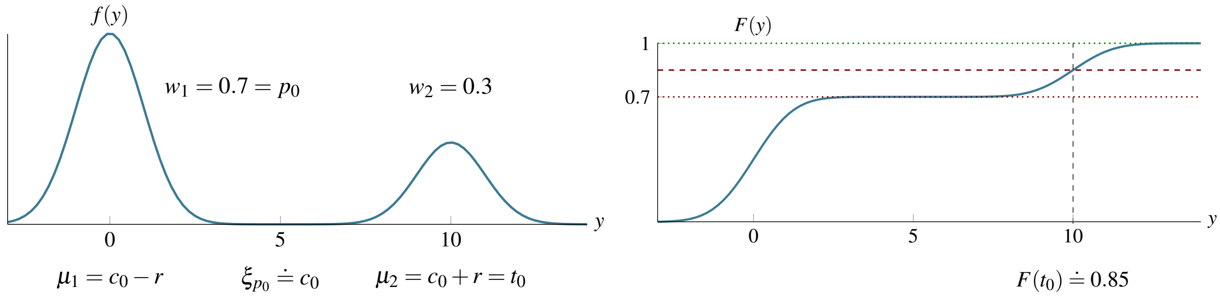

Figure 1: The density function $f(y)$ (left) and CDF $F(y)$ (right) are plotted for a normal mixture in (33), where $w_1 = 0.7$, $w_2 = 0.3$, $\mu_1 = 0$, $\mu_2 = 10$, and $\sigma_1 = \sigma_2 = 1$. For $p_0 = 0.7$, $\xi_{p_0} \doteq 5$ and $f(\xi_{p_0}) \doteq 1.49\text{E}{-06}$.

(and also of many other MC estimators) of the p_0-quantile includes $f_r^2(\xi_{p_0,r})$ in the denominator, as in (4) and (9), and $f_r(\xi_{p_0,r}) \approx w_1\phi(r/\sigma_1)/\sigma_1 + w_2\phi(-r/\sigma_2)/\sigma_2 \to 0$ exponentially fast as roughly $e^{-r^2/2}$ as $r \to \infty$ when $\sigma_1 = \sigma_2 = 1$. Thus, as $r \to \infty$, the asymptotic variance (4) of the SRS estimator of $\xi_{p_0,r}$ blows up, but the true quantile $\xi_{p_0,r}$ tends to c_0. Hence, for $c_0 \neq 0$, the SRS estimator of $\xi_{p_0,r}$ has URE as $r \to \infty$, even though $0 \ll p_0 \ll 1$, so not just extreme quantiles (as in Section 4) may be difficult to estimate.

In contrast we next show that the SRS estimator (2) of $F(y)$ has BRE as $r \to \infty$ for y in a large neighborhood of $\xi_{p_0,r}$ for $p_0 = w_1$. For large r, the two means in the mixture components are far apart, with the p_0-quantile $\xi_{p_0,r}$ in the middle of the two modes. Note that $F(\xi_{p_0,r}) = p_0$, and the CDF F of Y is nearly flat in a large neighborhood of $\xi_{p_0,r}$ as the density is very small there; see the right plot of Figure 1. Specifically, for any (small) $\delta > 0$, there is a large $d > 0$ such that $|F(y) - F(\xi_{p_0,r})| \leq \delta$ for all $y \in (\xi_{p_0,r} - d, \xi_{p_0,r} + d)$. Hence, while the true value of the p_0-quantile is quite sensitive to small perturbations in p_0, the CDF $F(y)$ changes little from large shifts in y around $\xi_{p_0,r}$. Thus, as $r \to \infty$ in this case, even though the SRS estimator of $\xi_{p_0,r}$ has URE, we estimate $F(y)$ with BRE for y in a large neighborhood of $\xi_{p_0,r}$ since $\mathbb{V}[\widehat{F}_{\text{SRS},n}(y)] = F(y)[1-F(y)]/n$ and $F(y)$ is bounded away from 0 and 1.

We now examine the efficiencies of hypothesis tests for (23) and (27) on the normal mixture in Figure 1 with $p_0 = w_1 = 0.7$, $r = 5$, and $t_0 = \mu_{2,r} = 10$. The p_0-quantile $\xi_{p_0} \doteq 5$ lies below t_0, and

$$F(t_0) = w_1\Phi_{\mu_{1,r},\sigma_1}(\mu_{2,r}) + w_2\Phi_{\mu_{2,r},\sigma_2}(\mu_{2,r}) = w_1\Phi((\mu_{2,r} - \mu_{1,r})/\sigma_1) + w_2/2 \doteq 0.85 > p_0 = 0.7. \quad (34)$$

Thus, H'_A holds in (27), and H_A holds in (23). Recall that (32) (resp., (30)) gives the approximate sample size $n' \equiv n'_{r,\gamma,\beta}$ (resp., $n \equiv n_{r,\gamma,\beta}$) needed to reject H'_0 via (28) (resp., H_0 via (24)) with probability β, which we set as $\beta = 0.95$. For large r, while $(t_0 - \xi_{p_0,r})^2 \approx (\mu_{2,r} - c_0)^2 = r^2$ in (32), $f_r^2(\xi_{p_0,r})$ behaves as roughly e^{-r^2}, so $n' \equiv n'_{r,\gamma,\beta}$ grows exponentially in r^2. For example, the specified parameter values result in $n' \doteq 4.11\text{E}{+10}$. In contrast, for n in (30), because $t_0 = \mu_{2,r}$ in our hypotheses in (23), (34) implies that $F_r(t_0) \to w_1 + (w_2/2) = 0.85$ as $r \to \infty$, so n is bounded in r. In particular, we have that $F_r(t_0) \doteq 0.85$ and $p_0 = 0.7$ by (34), so (30) gives $n \doteq 61$ for our specific parameter values. Thus, confirming the $\gamma|p_0$ criterion through $F_r(t_0)$ using (24) requires a much smaller sample size than through ξ_{p_0} using (28).

8.2 Example where SRS Estimator of ξ_{p_0} for Fixed p_0 has VRE

We next consider another normal mixture (33) with different parameters for which the SRS estimator of the p_0-quantile has VRE. The parameters in (33) are now $w_1 = w_2 = 0.5$, $\mu_1 = 0$, $\mu_2 = \Phi^{-1}(0.9) \doteq 1.2816$, $\sigma_1 = 1$, and $\sigma_2 = r$ for $r = 10^{-3}$. For these values, the left and center panels of Figure 2 display the density function of the mixture as a standard normal density with a spike near $x = \mu_2$, where the left (resp., center) panel shows the vertical axis on a linear (resp., log) scale. The first component in the mixture is a standard normal, which is clearer when the vertical axis has log scale, so then the standard normal density appears as a quadratic. The right panel of Figure 2 plots the CDF F, showing a sharp increase near $x = \mu_2$ corresponding to the spike

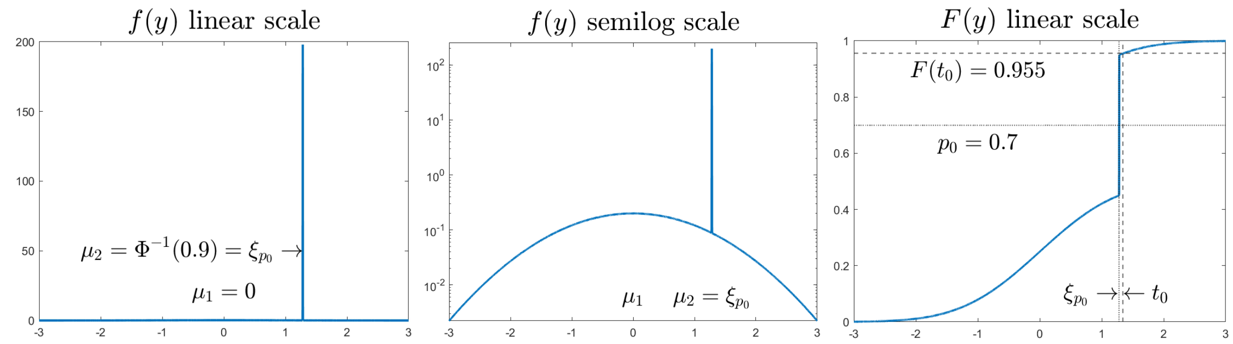

Figure 2: The density function $f(y)$ and CDF $F(y)$ are plotted for the normal mixture in (33), where $w_1 = w_2 = 1/2$, $\mu_1 = 0$, $\mu_2 = \Phi^{-1}(0.9) \doteq 1.2816$, $\sigma_1 = 1$, and $\sigma_2 = 10^{-3}$. For $p_0 = 0.7$, the p_0-quantile $\xi_{p_0} = \mu_2$, where $f(\xi_{p_0}) \doteq 199.56$. The left (resp., center) panel shows the density with the vertical axis in linear (resp., log) scale, and the right panel displays the CDF.

in the mixture density. Note that $F_r(\mu_2) = w_1\Phi_{\mu_1,\sigma_1}(\mu_2) + w_2\Phi_{\mu_2,\sigma_2}(\mu_2) = w_1\Phi(\Phi^{-1}(0.9)) + w_2/2 = 0.7$, so for $p_0 = 0.7$, the p_0-quantile $\xi_{p_0,r} = \mu_2$, and $f_r(\xi_{p_0,r}) \doteq 199.56$, which lead to $\mathrm{RE}_{\mathrm{SRS}}[\xi_{p_0,r}] \doteq 0.0018$ by (4) and (5). As $r \to 0$ with $p_0 = 0.7$ fixed, the p_0-quantile remains $\xi_{p_0,r} = \mu_2$, but $f_r(\xi_{p_0,r}) \to \infty$, so $\kappa^2_{\mathrm{SRS},p_0} \equiv \kappa^2_{\mathrm{SRS},p_0,r} \to 0$ in (4) and $\mathrm{RE}_{\mathrm{SRS}}[\xi_{p_0,r}] \to 0$. Thus, the SRS estimator of $\xi_{p_0,r}$ has VRE as $r \to 0$.

Now consider the hypotheses (23) and (27) with $t_0 = \Phi^{-1}(0.91) \doteq 1.3408$, and $\gamma = 1 - \alpha = 0.95 = \beta$. Then $F_r(t_0) = w_1\Phi_{\mu_1,\sigma_1}(t_0) + w_2\Phi_{\mu_2,\sigma_2}(t_0) \approx w_1\Phi(\Phi^{-1}(0.91)) + w_2 = 0.955 > 0.7 = p_0$ and $\xi_{p_0,r} = \mu_2 \doteq 1.2816 < t_0$, so H_A in (23) and H'_A in (27) are true. Consider trying to establish this using the tests (24) with (26) and (28) with (6). From (32), the sample size n' needed so that (28) roughly holds with probability $\beta = 0.95$ satisfies $n' = n'_{r,\gamma,\beta} \to 0$ as $r \to 0$ because $\kappa^2_{\mathrm{SRS},p_0,r} \to 0$. But the necessary sample size $n = n_{r,\gamma,\beta}$ from (30) so that (24) roughly holds with probability β remains bounded as $r \to 0$ because then $\psi_0^2 \equiv \psi^2_{0,r} = F_r(t_0)[1 - F_r(t_0)]$ is bounded away from 0 since $\psi^2_{0,r} \to (0.955)(0.045) = 0.042975$ as $r \to 0$. Our specific parameter values lead to $n'_{r,\gamma,\beta} \doteq 0.016$ from (32) and $n_{r,\gamma,\beta} \doteq 7.15$ from (30), although both sample sizes should be taken much larger for the CLT approximations in (31) and (29) to roughly hold.

ACKNOWLEDGMENTS

This work has been supported in part by the National Science Foundation under Grant No. CMMI-2345330. 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

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

Casella, G. and R. L. Berger. 2002. *Statistical Inference*. 2nd ed. Pacific Grove, Calif.: Duxbury.

Chu, F. and M. K. Nakayama. 2012. “Confidence Intervals for Quantiles When Applying Variance-Reduction Techniques”. *ACM Transactions On Modeling and Computer Simulation* 22(2):10:1–10:25.

Dembo, A. and O. Zeitouni. 1998. *Large Deviations Techniques and Applications*. second ed. New York: Springer.

Deo, A. and K. Murthy. 2021. “Efficient Black-Box Importance Sampling for VaR and CVaR Estimation”. In *2021 Winter Simulation Conference (WSC)*, 1–12 https://doi.org/https://doi.org/10.1109/WSC52266.2021.9715385.

Dong, H. and M. K. Nakayama. 2019. “A Tutorial on Quantile Estimation via Monte Carlo”. In *Monte Carlo & Quasi-Monte Carlo Methods:MCQMC 2018*, edited by P. L’Ecuyer and B. Tuffin, Volume 324 of *Springer Proceedings in Mathematics & Statistics*, 3–30. Cham, Switzerland: Springer.

Glynn, P. W. 1996. “Importance Sampling for Monte Carlo Estimation of Quantiles”. In *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation*, edited by S. M. Ermakov and V. B. Melas, 180–185. St. Petersburg, Russia: Publishing House of St. Petersburg Univ.

Glynn, P. W., M. K. Nakayama, and B. Tuffin. 2017. “On the Estimation of the Mean Time to Failure by Simulation”. In *2017 Winter Simulation Conference (WSC)*, 1844–1855 https://doi.org/https://doi.org/10.1109/WSC.2017.8247921.

Glynn, P. W., M. K. Nakayama, and B. Tuffin. 2018. “Using Simulation to Calibrate Exponential Approximations to Tail-Distribution Measures of Hitting Times to Rarely Visited Sets”. In *2018 Winter Simulation Conference (WSC)*, 1802–1813 https://doi.org/https://doi.org/10.1109/WSC.2018.8632477.

Goyal, A., P. Shahabuddin, P. Heidelberger, V. Nicola and P. W. Glynn. 1992. “A Unified Framework for Simulating Markovian Models of Highly Dependable Systems”. *IEEE Transactions on Computers* C-41(1):36–51.

Jensen, J. L. 1995. *Saddlepoint Approximations*. New York: Oxford University Press.

Kalashnikov, V. 1997. *Geometric Sums: Bounds for Rare Events with Applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Kohler, M. and A. Krzyżak. 2019. “Estimation of Extreme Quantiles in a Simulation Model”. *Journal of Nonparametric Statistics* 31(2):393–419.

L’Ecuyer, P., J. H. Blanchet, B. Tuffin, and P. W. Glynn. 2010. “Asymptotic Robustness of Estimators in Rare-Event Simulation”. *ACM Transactions on Modeling and Computer Simulation* 20(1):Article 6.

L’Ecuyer, P., M. Mandjes, and B. Tuffin. 2009. “Importance Sampling and Rare Event Simulation”. In *Rare Event Simulation using Monte Carlo Methods*, edited by G. Rubino and B. Tuffin, Chapter 2, 17–38. Chichester, UK: Wiley.

Li, Y., Z. T. Kaplan, and M. K. Nakayama. 2024. “Monte Carlo Methods for Economic Capital”. *INFORMS Journal on Computing*. To appear.

Parzen, E. 2004. “Quantile Probability and Statistical Data Modeling”. *Statistical Science* 19(4):652–662.

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

U.S. Nuclear Regulatory Commission 2005. “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)”. Technical report, U.S. Nuclear Regulatory Commission, Washington, DC.

U.S. Nuclear Regulatory Commission 2010. “Acceptance criteria for emergency core cooling systems for light-water nuclear power reactors”. Title 10, Code of Federal Regulations §50.46, NRC, Washington, DC.

U.S. Nuclear Regulatory Commission 2011. “Applying Statistics”. U.S. Nuclear Regulatory Commission Report NUREG-1475, Rev 1, U.S. Nuclear Regulatory Commission, Washington, DC.

Van Der Vaart, A. W. 1998. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Wilks, S. S. 1941. “Determination of Sample Sizes for Setting Tolerance Limits”. *Annals of Mathematical Statistics* 12:91–96.

AUTHOR BIOGRAPHIES

MARVIN K. NAKAYAMA is a professor in the Department of Computer Science at the New Jersey Institute of Technology. He received an M.S. and a Ph.D. in operations research from Stanford University. His research interests include simulation, modeling, statistics, and risk analysis. His email: marvin@njit.edu.

BRUNO TUFFIN received his PhD degree in applied mathematics from the University of Rennes 1, France, in 1997. Since then, he has been with Inria in Rennes. His research interests include developing Monte Carlo and quasi-Monte Carlo simulation techniques for performance evaluation. His email: bruno.tuffin@inria.fr.