



HAL
open science

Some Asymptotic Regimes for Quantile Estimation

Marvin K Nakayama, Bruno Tuffin

► **To cite this version:**

Marvin K Nakayama, Bruno Tuffin. Some Asymptotic Regimes for Quantile Estimation. 2024 Winter Simulation Conference, Dec 2024, Orlando (FL), United States. hal-04551483

HAL Id: hal-04551483

<https://inria.hal.science/hal-04551483v1>

Submitted on 18 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Some Asymptotic Regimes for Quantile Estimation

Marvin K. Nakayama
Computer Science Department
New Jersey Institute of Technology
Newark, NJ 07102, USA

Bruno Tuffin
Inria, Univ Rennes, CNRS, IRISA
Campus de Beaulieu, 263 Avenue Général Leclerc
35042 Rennes, FRANCE

April 18, 2024

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. [6] review estimating ξ_p via Monte Carlo (MC) methods, including *simple random sampling* (SRS) and *importance sampling* (IS) [1, Chapter V.1].

We consider here families of stochastic models indexed by a parameter r and examine the behavior of 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 $r \rightarrow \infty$ or $r \rightarrow 0$. Or r could be the quantile level p , and we let $p \rightarrow 0$

or 1. The paper examines the *relative errors* (REs) of quantile estimators under different asymptotic regimes as $r \rightarrow r_0$, and provides examples where the RE is vanishing, bounded, or unbounded.

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 $p \rightarrow 1$ or $p \rightarrow 0$ 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 independent and identically distributed (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 n using \mathfrak{M} , construct an estimator $\hat{\varphi}_{\mathfrak{M},n} \equiv \hat{\varphi}_{\mathfrak{M},n,r}$ of φ_r , and for each fixed r , we assume a *central limit theorem* (CLT) holds: $\sqrt{n}[\hat{\varphi}_{\mathfrak{M},n} - \varphi] \Rightarrow \mathcal{N}(0, \zeta_{\mathfrak{M}}^2)$ as $n \rightarrow \infty$, where \Rightarrow denotes weak convergence [18, Section 1.2.4], $\mathcal{N}(a, b^2)$ is a normal random variable with mean a and variance b^2 , and $\zeta_{\mathfrak{M}}^2 \equiv \zeta_{\mathfrak{M},r}^2 \in (0, \infty)$ is the CLT's *asymptotic variance*. When $\varphi \neq 0$, the *relative error* of the \mathfrak{M} estimator of φ is

$$\text{RE}_{\mathfrak{M},r}[\varphi] = \zeta_{\mathfrak{M}}/|\varphi| \equiv \zeta_{\mathfrak{M},r}/|\varphi_r|; \quad (1)$$

e.g., see [13] and Chapter VI of [1]. To motivate our interest in RE, let $\hat{\zeta}_{\mathfrak{M},n}$ be a consistent estimator of $\zeta_{\mathfrak{M}}$ in the sense that $\hat{\zeta}_{\mathfrak{M},n} \Rightarrow \zeta_{\mathfrak{M}}$ as $n \rightarrow \infty$. For a given confidence level $\gamma \in (0, 1)$, e.g., $\gamma = 0.95$, we then get $(\hat{\varphi}_{\mathfrak{M},n} \pm z_{1-(1-\gamma)/2} \hat{\zeta}_{\mathfrak{M},n} / \sqrt{n})$ as an approximate *confidence interval* (CI) for φ for large n from the CLT and Slutsky's theorem [18, 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 $(\hat{\varphi}_{\mathfrak{M},n} \pm \varepsilon |\hat{\varphi}_{\mathfrak{M},n}|)$) entails a sample size n roughly proportional to the squared $\text{RE}_{\mathfrak{M},r}[\varphi]$, showing the relevance of RE.

We want to study the behavior of $\text{RE}_{\mathfrak{M},r}[\varphi]$ as $r \rightarrow r_0$ for some limiting value $r_0 \in \overline{\mathfrak{R}} \equiv [-\infty, \infty]$, with each $r \in \mathfrak{R} \equiv (-\infty, \infty)$. If $\text{RE}_{\mathfrak{M},r}[\varphi]$ remains bounded as $r \rightarrow r_0$, we say that the estimator $\hat{\varphi}_{\mathfrak{M},n}$ has *bounded relative error* (BRE). If $\text{RE}_{\mathfrak{M},r}[\varphi] \rightarrow 0$ as $r \rightarrow r_0$, then the estimator $\hat{\varphi}_{\mathfrak{M},n}$ has *vanishing relative error* (VRE). If $\text{RE}_{\mathfrak{M},r}[\varphi] \rightarrow \infty$ as $r \rightarrow r_0$, then the estimator $\hat{\varphi}_{\mathfrak{M},n}$ has *unbounded relative error* (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 $\mu = \mathbb{E}[Y] = \int y dF(y)$, 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, \dots, Y_n from F . Compute the *empirical distribution* $\widehat{F}_{\text{SRS},n}$ as an estimator of F , with

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

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 $\widehat{\xi}_{\text{SRS},n} = \widehat{F}_{\text{SRS},n}^{-1}(p)$. For any fixed $p \in (0, 1)$ such that $f(\xi_p) > 0$, which will be assumed throughout, $\widehat{\xi}_{\text{SRS},n}$ obeys a CLT as $n \rightarrow \infty$ [18, Section 2.3.3]:

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

$$\kappa_{\text{SRS}}^2 \equiv \kappa_{\text{SRS},p}^2 = \psi_{\text{SRS}}^2 / f^2(\xi_p), \quad \text{with} \quad \psi_{\text{SRS}}^2 \equiv \psi_{\text{SRS},p}^2 = p(1-p). \quad (4)$$

Note that $\widehat{F}_{\text{SRS},n}(\xi_p)$ has variance $\psi_{\text{SRS},p}^2/n$, and [6] 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

$$\text{RE}_{\text{SRS}}[\xi_p] = \kappa_{\text{SRS}} / |\xi_p|, \quad (5)$$

as in (1). Let $\widehat{\kappa}_{\text{SRS},n}$ be a consistent estimator of κ_{SRS} , e.g., from Corollary 2.5.2 of [18]. 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

$$U_n = \widehat{\xi}_{\text{SRS},n} + z_\gamma \widehat{\kappa}_{\text{SRS},n} / \sqrt{n}. \quad (6)$$

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 $\widehat{\mu}_{\text{SRS},n} = (1/n) \sum_{i=1}^n Y_i$. 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 $\sqrt{n}[\widehat{\mu}_{\text{SRS},n} - \mu] \Rightarrow \mathcal{N}(0, \sigma^2)$ as $n \rightarrow \infty$ [18, Section 1.9.1], so as in (1) when $\mu \neq 0$, the RE of the SRS estimator of the mean is given by

$$\text{RE}_{\text{SRS}}[\mu] = \sigma / |\mu|. \quad (7)$$

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

$$\text{RE}_{\text{IS}}[\mu] = \sigma_{\text{IS}}/|\mu|. \quad (8)$$

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

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

e.g., see [7] and [3]. Note that $\hat{F}_{\text{IS},n}(\xi_p)$ has variance $\psi_{\text{IS},p}^2/n$, which depends on \tilde{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

$$\text{RE}_{\text{IS}}[\xi_p] = \kappa_{\text{IS}}/|\xi_p|. \quad (10)$$

We will examine the RE of estimators of ξ_p and μ under various asymptotic regimes when $r \rightarrow r_0$. 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 $r \rightarrow r_0$. 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 $r \rightarrow r_0$ 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 $r \rightarrow r_0$ 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 $r \rightarrow r_0$ if $\lim_{r \rightarrow r_0} h_1(r)/h_2(r) = 0$ (resp., $\lim_{r \rightarrow r_0} h_2(r)/h_1(r) = 0$).

3 Location-Scale Family of Distributions

Suppose that $Y = a + bX$ for a scalar random variable X , where $X \sim G$, and $a, b \in \mathfrak{R}$. For example, if X is a mean-1 exponential with CDF $G(x) = [1 - e^{-x}]I(x \geq 0)$, then 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 $f(\xi_p) = \frac{1}{b}g\left(\frac{\xi_p - a}{b}\right) = \frac{1}{b}g(\eta_p)$ [17], so

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

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 $\text{RE}_{\text{SRS}}[\xi_p] = \frac{\sqrt{p(1-p)}}{g(\eta_{1-p})|\frac{a}{b} + \eta_{1-p}|}$. 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 $\text{RE}_{\text{SRS}}[\xi_p]$ remains bounded as $b \rightarrow \infty$ 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 $\text{RE}_{\text{SRS}}[\mu] = b\tau/a \rightarrow \infty$ as $b \rightarrow \infty$. Thus, in this case for SRS, the mean becomes more difficult to estimate (URE) as $b \rightarrow \infty$, but the p -quantile has BRE for each fixed $p \in (0, 1)$.

If instead $\nu \neq 0$, then $|\mu| = |a + b\nu| \rightarrow \infty$ as $b \rightarrow \infty$. In this case, $\text{RE}_{\text{SRS}}[\mu] = b\tau/|a + b\nu| \rightarrow \tau/|\nu|$ as $b \rightarrow \infty$, 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 $b \rightarrow 0$. By (11) and (7), both $\text{RE}_{\text{SRS}}[\xi_p]$ and $\text{RE}_{\text{SRS}}[\mu]$ shrink to 0 (VRE) as $b \rightarrow 0$.

4 SRS Estimation of p -quantile as $p \rightarrow 1$ or $p \rightarrow 0$ for Specific Distributions

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

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

$$\text{RE}_{\text{SRS}}[\xi_p] = \frac{-1}{k \ln(1-p)} \sqrt{\frac{p}{1-p}} \rightarrow \infty \quad (\text{URE}) \quad \text{as } p \rightarrow 1 \text{ or } p \rightarrow 0. \quad (12)$$

2. (Generalized) Pareto with $(\vartheta > 0, a_0 \in \mathfrak{R}, b_0 > 0) = (\text{shape, location, scale})$ 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 $b_0 > 2$. The p -quantile is $\xi_p = a_0 + b_0[(1-p)^{-1/\vartheta} - 1]$, whose SRS estimator has asymptotic variance $\kappa_{\text{SRS}}^2 = b_0^2 p / [\vartheta^2(1-p)^{(\vartheta+2)/\vartheta}]$. If $a_0 = 0$,

$$\text{RE}_{\text{SRS}}[\xi_p] = \frac{1}{\vartheta[1-(1-p)^{1/\vartheta}]} \sqrt{\frac{p}{1-p}} \rightarrow \infty \quad (\text{URE}) \quad \text{as } p \rightarrow 1 \text{ or } p \rightarrow 0. \quad (13)$$

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 $\kappa_{\text{SRS}}^2 = p(1-p) / [\phi(\eta_p)/\sigma]^2$, and it can be shown that (5) results in

$$\text{RE}_{\text{SRS}}[\xi_p] = \frac{\sigma \sqrt{p(1-p)}}{\phi(\eta_p)(\mu + \sigma \eta_p)} \rightarrow \infty \quad (\text{URE}) \quad \text{as } p \rightarrow 1 \text{ or } p \rightarrow 0.$$

$1-p$	$\mathcal{N}(2,1)$	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

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

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

For any continuous F , examine now the SRS estimator $\widehat{F}_{\text{SRS},n}(y)$ in (2) of $F(y)$ for y in the tails. For $u \in \{0, 1\}$, let $y_u = \lim_{p \rightarrow u} F^{-1}(p) \in \overline{\mathfrak{R}}$. As $\mathbb{V}[\widehat{F}_{\text{SRS},n}(y)] = F(y)[1-F(y)]/n = \mathbb{V}[1-\widehat{F}_{\text{SRS},n}(y)]$ for all $y \in \mathfrak{R}$, we get $\text{RE}_{\text{SRS}}[\widehat{F}_{\text{SRS},n}(y)] = \sqrt{[1-F(y)]/F(y)} \rightarrow \infty$ as $y \downarrow y_0$, and $\text{RE}_{\text{SRS}}[1-\widehat{F}_{\text{SRS},n}(y)] = \sqrt{F(y)/[1-F(y)]} \rightarrow \infty$ 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 $\lambda=1/v$ 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 [14] to determine the value of $\widetilde{\lambda} = \widetilde{\lambda}_p$ that minimizes $\kappa_{\text{IS}}^2 = \kappa_{\text{IS},p}^2$ in (9) and $\text{RE}_{\text{IS}}[\xi_p]$ in (10). For $\psi_{\text{IS},p}^2$ in (9), its first term (i.e., second moment under

IS) satisfies

$$\begin{aligned}
\mathbb{E}_{\tilde{G}}[I(X > \xi_p)R^2(X)] &= \int_{\xi_p}^{\infty} \frac{\lambda^2 e^{-2\lambda x}}{\tilde{\lambda} e^{-\tilde{\lambda} x}} dx \\
&= \frac{\lambda^2}{\tilde{\lambda}} \int_{\xi_p}^{\infty} e^{-(2\lambda - \tilde{\lambda})x} dx \\
&= \begin{cases} \infty & \text{if } \tilde{\lambda} \geq 2\lambda, \\ \frac{\lambda^2}{\tilde{\lambda}(2\lambda - \tilde{\lambda})} e^{-(2\lambda - \tilde{\lambda})\xi_p} & \text{otherwise,} \end{cases} \quad (14)
\end{aligned}$$

so we assume that $\tilde{\lambda} \in (0, 2\lambda)$. Then (14) is minimized at $\tilde{\lambda} = \lambda + \frac{1}{\xi_p} - \left(\lambda^2 + \frac{1}{\xi_p^2}\right)^{1/2}$, which also minimizes $\kappa_{\text{IS},p}^2$ in (9) as $f(\xi_p) = \lambda e^{-\lambda \xi_p} = \lambda(1-p)$ does not depend on $\tilde{\lambda}$. Using $\xi_p = -\ln(1-p)/\lambda$ leads to

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

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

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

so $\text{RE}_{\text{IS}}[\xi_p]$ 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: By item 1 of Section 3, the SRS quantile estimator has BRE when $b \equiv 1/\lambda \rightarrow \infty$. For the IS estimator, (16) implies $\kappa_{\text{IS},p}^2 \rightarrow \infty$ as $\lambda \rightarrow 0$ (since q does not depend on λ), and λ does not appear in $\text{RE}_{\text{IS}}[\xi_p]$. 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 $p \rightarrow 1$,

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

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

$$\begin{aligned}
\text{RE}_{\text{IS}}[\xi_p] &= \left(\frac{(1-p)^{[1+o(1)]/\ln(1-p)}}{[-(1+o(1))\ln(1-p)][2+(1+o(1))/\ln(1-p)]} - \frac{1}{[-\ln(1-p)]^2} \right)^{1/2} \\
&= \left(\frac{e^{1+o(1)}}{[2+o(1)][-\ln(1-p)]} - o(1) \right)^{1/2}, \quad (18)
\end{aligned}$$

so $\text{RE}_{\text{IS}}[\xi_p] \rightarrow 0$ as $p \rightarrow 1$, 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 $p = 1 - \varepsilon$: As in (17) and (18), we get as $\varepsilon \rightarrow 0$ that $q = -[1 + o(1)]/\ln \varepsilon$ and $\text{RE}_{\text{IS}}[\xi_p] = \left(\frac{e^{1+o(1)}}{[2+o(1)]^{[-\ln(\varepsilon)]}} - o(1) \right)^{1/2} \rightarrow 0$ (uniformly in c), so the IS estimator of ξ_p has VRE. 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, \dots, 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 $\mathbf{X} = (X_1, X_2, \dots, X_m)$. 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 $m \rightarrow \infty$, 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. Then $Y(m) \sim F_m$ obeys a CLT $Z(m) \equiv \frac{\sqrt{m}}{\sigma_0} [Y(m) - \mu_0] \Rightarrow \mathcal{N}(0, 1)$ as $m \rightarrow \infty$. For K_m as the CDF of $Z(m)$ and fixed $p \in (0, 1)$, the continuity of Φ^{-1} ensures $K_m^{-1}(p) \rightarrow \Phi^{-1}(p)$ as $m \rightarrow \infty$ [22, 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 $\hat{\xi}_{p,n}(m)$ of $\xi_p(m)$ from a sample Y_1, Y_2, \dots, 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 $\hat{\xi}_{p,n}(m) = \hat{\mu}_n(m) + \hat{\sigma}_n(m) \Phi^{-1}(p)$, where $\hat{\mu}_n(m) = (1/n) \sum_{i=1}^n Y_i$ and $\hat{\sigma}_n^2(m) = (1/(n-1)) \sum_{i=1}^n [Y_i - \hat{\mu}_n(m)]^2$. To simplify the discussion, assume now that each $X_j \sim \mathcal{N}(\mu_0, \sigma_0^2)$, so $Y(m) \sim \mathcal{N}(\mu_0, \sigma_0^2/m)$. Then the independence of $\hat{\mu}_n(m)$ and $\hat{\sigma}_n^2(m)$ [2, Theorem 5.3.1] implies that $\mathbb{V}_{F_m}[\hat{\xi}_{p,n}(m)] = \mathbb{V}_{F_m}[\hat{\mu}_n(m)] + [\Phi^{-1}(p)]^2 \mathbb{V}_{F_m}[\hat{\sigma}_n^2(m)] = [\sigma_0^2/(mn)] + 2[\Phi^{-1}(p)]^2 (\sigma_0^2/m)^2 / (n-1) \rightarrow 0$ as $m \rightarrow \infty$ for fixed n . Also, $\xi_p(m) \rightarrow \mu_0$ as $m \rightarrow \infty$, so $\text{RE}_{\text{SRS}}[\xi_p(m)] \rightarrow 0$ 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 $Y \equiv Y(m) = \sum_{j=1}^m X_j$. For $Y \sim F$, 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 [7] in which $r = m \rightarrow \infty$ and simultaneously the quantile level

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

Thus, under (19), the number m of i.i.d. summands grows large with the quantile level p simultaneously approaching 1 exponentially fast. For SRS and IS with an exponential twist, as developed in [7] and further analyzed in [16], we will see that $\text{RE}_{\text{SRS}}[\xi_p] \rightarrow \infty$, $\text{RE}_{\text{IS}}[\xi_p] \rightarrow 0$, $\text{RE}_{\text{SRS}}[\mu] \rightarrow 0$, and $\text{RE}_{\text{IS}}[\mu] \rightarrow \infty$ as $m \rightarrow \infty$, where the REs are as given in (5), (10), (7), and (8).

To develop the IS, let $M_0(\theta) = \int e^{\theta x} dG_0(x)$, $\theta \in \mathfrak{X}$, be the *moment generating function* (MGF) of G_0 . Let $\Delta = \{\theta \in \mathfrak{X} : M_0(\theta) < \infty\}$ 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 $Q'_0(\theta) = \frac{d}{d\theta} Q_0(\theta)$ as its first derivative. For $\theta \in \Delta^\circ$, the exponential twist $\tilde{G}_{0,\theta}$ of the marginal CDF G_0 of X_j is given by $d\tilde{G}_{0,\theta}(x) = e^{\theta x} dG_0(x)/M_0(\theta) = e^{\theta x - Q_0(\theta)} dG_0(x)$, $x \in \mathfrak{X}$ [1, Section V.1b]. We then define the method IS(θ) with $\theta \in \Delta^\circ$ so that X_1, X_2, \dots, X_m are i.i.d., each with marginal distribution $\tilde{G}_{0,\theta}$; let \tilde{G}_θ denote the resulting joint CDF of (X_1, X_2, \dots, X_m) . When $\mu_0 \neq 0$, Theorem 6 of [16] shows that as $m \rightarrow \infty$,

$$\text{RE}_{\text{SRS}}[\mu] = \Theta(1/\sqrt{m}) \rightarrow 0 \quad \text{and} \quad \text{RE}_{\text{IS}(\theta)}[\mu] = \Omega([\alpha(\theta)]^{m/2}/\sqrt{m}) \rightarrow \infty$$

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 IS(θ) with any $\theta \neq 0$ leads to exponentially increasing RE.

To estimate the p -quantile for p in (19) and $m \rightarrow \infty$, [7] applies IS(θ_*) for $\theta_* > 0$ satisfying

$$-\theta_* Q'_0(\theta_*) + Q_0(\theta_*) = -\beta_0, \tag{20}$$

assuming $\theta_* \in \Delta^\circ$ exists; a sufficient condition ensuring the existence of $\theta_* > 0$ satisfying (20) is that Q_0 is “steep” [5, p. 44], which holds, e.g., when G_0 is normal or gamma. The choice of twisting parameter θ_* for IS is motivated heuristically as follows. When m is large and $x \geq m(\mu_0 + \varepsilon)$ for any constant $\varepsilon > 0$, where $\mu_0 = \mathbb{E}[X_j]$ is the mean of each i.i.d. summand under G_0 , large-deviations theory [5, Section 2.2] leads to approximating the tail probability $\mathbb{P}_G(Y > x) \approx \exp[-m(\theta_x Q'_0(\theta_x) - Q_0(\theta_x))]$ for θ_x as the root of the equation $mQ'_0(\theta_x) = x$, where \mathbb{P}_H denotes the probability when $\mathbf{X} \sim H$. Under IS(θ) for $\theta \in \Delta^\circ$, each i.i.d. summand $X_j \sim \tilde{G}_{0,\theta}$ has mean $\mathbb{E}_{\tilde{G}_{0,\theta}}[X_j] = Q'_0(\theta)$ [5, Lemma 2.2.5(c)], so their sum Y has mean $\mathbb{E}_{\tilde{G}_\theta}[Y] = mQ'_0(\theta)$. The p -quantile ξ_p satisfies $\mathbb{P}_G(Y > \xi_p) = 1 - p = e^{-\beta_0 m}$ by (19), so equating the two tail probabilities leads to θ_* in (20) resulting in Y under IS(θ_*) having roughly mean ξ_p . Hence, getting observations of Y near ξ_p is not rare under IS(θ_*), leading to efficient estimation of ξ_p .

For the SRS and IS(θ_*) estimators of the p -quantile ξ_p , [7] analyzes the numerators (but not the denominator $f(\xi_p)$) of their asymptotic variances in (4) and

(9), showing that $\text{IS}(\theta_*)$ can substantially reduce variance compared to SRS. Theorem 7 of [16] provides a fuller picture by further handling the denominator $f(\xi_p)$ in (4) and (9) through a “saddlepoint approximation” [11, Chapter 2], establishing that as $m \rightarrow \infty$,

$$\text{RE}_{\text{SRS}}[\xi_p] = \omega(e^{(\beta_0/2)m - \sqrt{m}}/\sqrt{m}) \rightarrow \infty \quad \text{and} \quad \text{RE}_{\text{IS}(\theta_*)}[\xi_p] = O(1/\sqrt{m}) \rightarrow 0$$

when the characteristic function $C_0(\theta) = M_0(\theta\sqrt{-1})$, defined for $\theta \in \mathfrak{R}$, of G_0 satisfies $\int_{\mathfrak{R}} |C_0(\theta)|^{q_0} d\theta < \infty$ for some $q_0 \geq 1$. Thus, the SRS estimator of ξ_p has exponentially increasing RE, but the $\text{IS}(\theta_*)$ 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. [9] consider another limiting setting using a different weak-convergence result for exponential rarity in geometric sums [12]. Consider a (non-delayed) *regenerative process* $X = [X(t) : t \geq 0]$ evolving on a state space $\mathcal{S} \subseteq \mathfrak{R}^d$ [1, Section IV.6b], so the process “probabilistically restarts” at a sequence of regeneration times $0 = \Gamma_0 < \Gamma_1 < \Gamma_2 < \dots$ 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 i th (regenerative) *cycle*, which has length $\tau_i = \Gamma_i - \Gamma_{i-1}$. The cycles are i.i.d. by the regenerative property.

For $\mathcal{A} \subset \mathcal{S}$, let $T = \inf\{t \geq 0 : X(t) \in \mathcal{A}\}$ be the *hitting time* to \mathcal{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 [10]. 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 $T_i = \inf\{t \geq 0 : X(\Gamma_{i-1} + t) \in \mathcal{A}\}$ be the time elapsing after Γ_{i-1} until the next hit to \mathcal{A} , and $M = \min\{i \geq 1 : T_i < \tau_i\} - 1$ is the number of cycles completed before the one in which \mathcal{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 [10] and [8]).

Consider a parameterization of the model by r such that $\zeta \equiv \zeta_r \rightarrow 0$ as $r \rightarrow r_0$; e.g., in a stable GI/G/1 queue-length process, we can take r as a large queue threshold with $T \equiv T_r$ being the hitting time of r , and $r_0 = \infty$. Thus, hitting \mathcal{A} in a cycle becomes rarer as $r \rightarrow r_0$, and the scaled hitting time T/μ converges weakly to an exponential: for $\mathbb{P} \equiv \mathbb{P}_r$ as the probability measure for parameter r ,

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

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 $\hat{\mu}_n$ be an estimator of μ based on a sample of n cycles, so an estimator of ξ_p is

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

Several efficient estimators have been developed for μ [9]. For any fixed quantile level $p \in (0, 1)$, $\hat{\xi}_{p,n}$ has BRE, VRE or URE as $r \rightarrow r_0$ if and only if $\hat{\mu}_n$ has the same corresponding property. The proof is immediate since $\hat{\xi}_{p,n}$ is simply a linear transformation of $\hat{\mu}_n$.

While (22) is based on the limiting result (21), a real system has fixed r , leading to $\hat{\xi}_{p,n}$ having bias that does not vanish as $n \rightarrow \infty$. 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 Hypothesis Testing Via Quantile or CDF

Hypotheses about a quantile can often also be equivalently expressed instead in terms of the CDF, and the associated statistical tests can have quite different efficiencies. To see why, consider the asymptotic variance κ_{SRS}^2 in (4) of the SRS estimator of the p -quantile ξ_p . The numerator ψ_{SRS}^2 is the variance of $\sqrt{n}\hat{F}_{\text{SRS},n}(\xi_p)$ from (2). But the denominator $f^2(\xi_p)$ of κ_{SRS}^2 in (4) can be large or small, depending on F , even for a non-extreme quantile level $0 \ll p \ll 1$. We will demonstrate 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 $\mathbb{P}(Y \leq t_0) > p_0$ for a specified (large) 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 $Y \sim F$:

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

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 [20] considers (23) with the load Y as the peak cladding temperature (PCT) having a capacity of $t_0 = 2200^\circ$ Fahrenheit and $p_0 = 0.95$, and this must be established with at least confidence level $\gamma = 1 - \alpha = 0.95$ [19, Section 3.2]. This is known

as a “ $\gamma|p_0$ criterion” [21, 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

$$L_n > p_0. \quad (24)$$

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

$$\sqrt{n} \left[\widehat{F}_{\text{SRS},n}(t_0) - F(t_0) \right] \Rightarrow \mathcal{N}(0, \sigma_0^2) \quad (25)$$

as $n \rightarrow \infty$, where $\sigma_0^2 = F(t_0)[1 - F(t_0)]$, leading to an approximate γ -level LCB for $F(t_0)$ as

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

with $\widehat{\sigma}_{0,n}$ a consistent estimator of σ_0 , e.g., using $\widehat{\sigma}_{0,n}^2 = \widehat{F}_{\text{SRS},n}(t_0)[1 - \widehat{F}_{\text{SRS},n}(t_0)]$.

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, $F(t_0) \leq p_0$ if and only if $\xi_{p_0} \geq t_0$, so we can equivalently express (23) as hypotheses

$$H'_0 : \xi_{p_0} \geq t_0 \quad \text{vs.} \quad H'_A : \xi_{p_0} < t_0, \quad (27)$$

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

$$U_n < t_0. \quad (28)$$

[23] provides an asymptotically equivalent approach; see Section 2.6.1 of [18].

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 (28), or vice versa. We will illustrate these two possibilities through specific examples of normal mixtures.

Let F be the CDF of a mixture of two normals: for mixing weights $w_1, w_2 \geq 0$, $w_1 + w_2 = 1$,

$$F(y) = w_1 \Phi_{\mu_1, \sigma_1}(y) + w_2 \Phi_{\mu_2, \sigma_2}(y), \quad (29)$$

where $\Phi_{a,b}$ 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, leading to vastly different behaviors for the RE of the SRS estimator $\widehat{\xi}_{\text{SRS},n} = \widehat{F}_{\text{SRS},n}^{-1}(p_0)$ of ξ_{p_0} , with fixed $p_0 \in (0, 1)$, for $\widehat{F}_{\text{SRS},n}$ in (2).

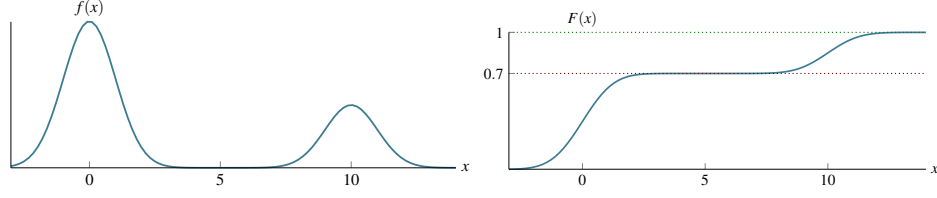


Figure 1: The density function $f(x)$ (left) and CDF $F(x)$ (right) are plotted for a normal mixture in (29), where $w_1 = 0.7$, $w_2 = 0.3$, $\mu_1 = 0$, $\mu_2 = 10$, and $\sigma_1^2 = \sigma_2^2 = 1$. For $p_0 = w_1 = 0.7$, the p_0 -quantile is $\xi_{p_0} \doteq 5$, where $f(\xi_{p_0}) \doteq 1.49\text{E}-06$.

8.1 Example where SRS estimator of ξ_{p_0} for fixed p_0 has URE

First assume that the component means in (29) 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 $r \rightarrow r_0 = \infty$, 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 bumps, 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 bumps. As r grows, the two bumps 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 $\xi_{p_0,r} \rightarrow c_0$ as $r \rightarrow \infty$. 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 $\xi_{p_0,r} \doteq c_0 = 5$, and we then get $f_r(\xi_{p_0,r}) \doteq 1.49\text{E}-06$ and $\text{RE}_{\text{SRS}}[\xi_{p_0,r}] \doteq 6.16\text{E}+04$ by (5). The asymptotic variance of the SRS estimator (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 $f_r(\xi_{p_0,r}) = w_1 \phi(r/\sigma_1)/\sigma_1 + w_2 \phi(-r/\sigma_2)/\sigma_2 \rightarrow 0$ exponentially fast as roughly $e^{-r^2/2}$ as $r \rightarrow \infty$ when $\sigma_1 = \sigma_2 = 1$. Thus, as $r \rightarrow \infty$, 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 $r \rightarrow \infty$, even though $0 \ll p_0 \ll 1$, so not just extreme quantiles (as in Section 4) may be difficult to estimate.

For the normal mixture in Figure 1, we now compare the efficiencies of using (24) with (26) and (28) with $p = p_0$ in (6) to test the hypotheses in (23) and (27), respectively, when the threshold is $t_0 = \mu_{2,r}$. We will evaluate the tests through the sample sizes needed to reject the null hypotheses with a given probability (power) β . The two means in the mixture components are very far apart, with the p_0 -

quantile ξ_{p_0} in the middle of the two modes. Note that $F(\xi_{p_0}) = p_0$, and the CDF F of Y is nearly flat in a large neighborhood of ξ_{p_0} because the density is very small there; see the right plot of Figure 1. Specifically, for any small $\delta > 0$, there will be a large $d > 0$ such that for all $y \in (\xi_{p_0} - d, \xi_{p_0} + d)$, we will have that $|F(y) - F(\xi_{p_0})| \leq \delta$. Hence, while the true value of the p_0 -quantile is very sensitive to small perturbations in p_0 , the CDF $F(y)$ changes little from large shifts in y around ξ_{p_0} . Thus, in this case, it may not be appropriate to solely focus on the p_0 -quantile ξ_{p_0} and the difficulty in estimating it because of the large RE. Instead, the SRS estimator of $F(y)$ will have small RE for y in a large neighborhood of ξ_{p_0} .

We now demonstrate this on the specific normal mixture in Figure 1 with $p_0 = w_1 = 0.7$ and $r = 5$. Because $t_0 = \mu_{2,r} = 10$, the p_0 -quantile $\xi_{p_0} \doteq 5$ lies below t_0 , and

$$\begin{aligned} F(t_0) &= w_1 \Phi_{\mu_{1,r}, \sigma_1}(\mu_{2,r}) + w_2 \Phi_{\mu_{2,r}, \sigma_2}(\mu_{2,r}) \\ &= w_1 \Phi((\mu_{2,r} - \mu_{1,r})/\sigma_1) + w_2/2 \doteq 0.85 > p_0 = 0.7. \end{aligned} \quad (30)$$

Thus, H'_A holds in (27), and H_A holds in (23). First consider trying to establish this 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 $\hat{\xi}_{\text{SRS},n} = \hat{F}_{\text{SRS},n}^{-1}(p_0)$. Using ideas related to Pitman efficiency [18, Section 10.2], we want to determine the sample size n so that (28) holds with a given probability (power) $\beta \in (1 - \gamma, 1)$. The probability of event (28) is

$$\begin{aligned} \mathbb{P}(U_n < t_0) &= \mathbb{P}\left(\frac{\sqrt{n}}{\hat{\kappa}_{\text{SRS},n}} \left[\hat{\xi}_{\text{SRS},n} - \xi_{p_0} \right] < -z_\gamma + \frac{\sqrt{n}}{\hat{\kappa}_{\text{SRS},n}} [t_0 - \xi_{p_0}]\right) \\ &\approx \Phi\left(-z_\gamma + \frac{\sqrt{n}}{\kappa_{\text{SRS}}} [t_0 - \xi_{p_0}]\right) \end{aligned} \quad (31)$$

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

$$n \equiv n_{r,\gamma,\beta} = \frac{(z_\gamma + z_\beta)^2 \kappa_{\text{SRS}}^2}{(t_0 - \xi_{p_0})^2}. \quad (32)$$

For large r , while $(t_0 - \xi_{p_0})^2 \approx (\mu_{2,r} - c_0)^2 = r^2$ in the denominator of (32), κ_{SRS}^2 in the numerator grows as roughly e^{r^2} because of $f_r^2(\xi_{p_0,r})$ in the bottom of κ_{SRS}^2 (see (4)), so $n \equiv n_{r,\gamma,\beta}$ grows exponentially in r^2 . For example, the specified parameter values and $\beta = 0.95$ result in $n \approx 4.11\text{E}+10$.

By (30), H_A holds in (23) (i.e., $F(t_0) > p_0$ for $p_0 = 0.7$), and consider trying to establish this through (24) using a LCB L_n in (26) at confidence level $\gamma = 0.95$.

We want to determine the sample size $n = n'$ so that (24) holds with probability $\beta \in (1 - \gamma, 1)$. The probability of (24) occurring is

$$\begin{aligned} \mathbb{P}(L_n > p_0) &= \mathbb{P}\left(\frac{\sqrt{n}}{\widehat{\sigma}_{0,n}} \left[\widehat{F}_{\text{SRS},n}(t_0) - F(t_0) \right] > z_\gamma + \frac{\sqrt{n}}{\widehat{\sigma}_{0,n}} [p_0 - F(t_0)]\right) \\ &\approx \overline{\Phi}\left(z_\gamma + \frac{\sqrt{n}}{\sigma_0} [p_0 - F(t_0)]\right) \end{aligned} \quad (33)$$

when n is large by the CLT (25) and the consistency of $\widehat{\sigma}_{0,n}$, where $\overline{\Phi} = 1 - \Phi$. Equating the right side of (33) to β yields $z_\gamma + \frac{\sqrt{n}}{\sigma_0} [p_0 - F(t_0)] = \Phi^{-1}(1 - \beta) = z_{1-\beta} = -z_\beta$, and using $\sigma_0^2 = F(t_0)[1 - F(t_0)]$ gives

$$n' \equiv n'_{r,\gamma,\beta} = \frac{(z_\gamma + z_\beta)^2 F_r(t_0) [1 - F_r(t_0)]}{[p_0 - F_r(t_0)]^2} \quad (34)$$

as the required sample size. Because $t_0 = \mu_{2,r}$ in our hypotheses in (23), (30) implies that $F_r(t_0) \rightarrow w_1 + w_2/2 = 0.85$ as $r \rightarrow \infty$, so n' is bounded in r . In particular, we have that $F_r(t_0) \doteq 0.85$ and $p_0 = 0.7$ by (30), so (32) gives $n' \doteq 61$ for our specific parameter values and $\beta = 0.95$. 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 (29) with different parameters for which the SRS estimator of the p_0 -quantile has VRE. The parameters in (29) are now $w_1 = w_2 = 0.5$, $\mu_1 = 0$, $\mu_2 = \Phi^{-1}(0.9) \doteq 1.2816$, $\sigma_1 = 1$, and $\sigma_2 = 1/r$ 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 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 $\text{RE}_{\text{SRS}}[\xi_{p_0,r}] \doteq 0.0018$ by (4) and (5). As $r \rightarrow \infty$ with $p_0 = 0.7$ fixed, the p_0 -quantile remains $\xi_{p_0,r} = \mu_2$, but $f_r(\xi_{p_0,r}) \rightarrow \infty$, so $\kappa_{\text{SRS},p_0}^2 \equiv \kappa_{\text{SRS},p_0,r}^2 \rightarrow 0$ in (4) and $\text{RE}_{\text{SRS}}[\xi_{p_0,r}] \rightarrow 0$. Thus, the SRS estimator of $\xi_{p_0,r}$ has VRE as $r \rightarrow \infty$.

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

$$F_r(t_0) = w_1 \Phi_{\mu_1, \sigma_1}(t_0) + w_2 \Phi_{\mu_2, \sigma_2}(t_0) \approx w_1 \Phi(\Phi^{-1}(0.91)) + w_2 = 0.955 > 0.7 = p_0$$

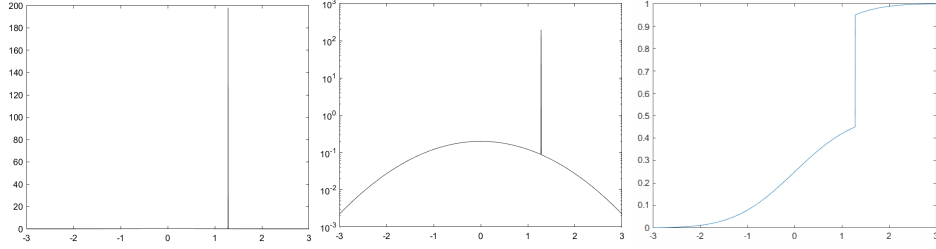


Figure 2: The density function $f(x)$ and CDF $F(x)$ are plotted for a normal mixture $w_1\Phi_1 + w_2\Phi_2$, where $w_1 = w_2 = 1/2$, and for $k = 1, 2$, Φ_k is the CDF of a $\mathcal{N}(\mu_k, \sigma_k^2)$ distribution, with $\mu_1 = 0$, $\mu_2 = \Phi^{-1}(0.9) \doteq 1.2816$, $\sigma_1^2 = 1$, and $\sigma_2^2 = 10^{-3}$. 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.

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 $n = n_{r, \gamma, \beta} \rightarrow 0$ as $r \rightarrow \infty$ because $\kappa_{\text{SRS}, p_0, r}^2 \rightarrow 0$. But the necessary sample size $n' = n'_{r, \gamma, \beta}$ from (34) so that (24) roughly holds with probability β remains bounded as $r \rightarrow \infty$ because then $\sigma_0^2 = \sigma_{0, r}^2 = F_r(t_0)[1 - F_r(t_0)]$ is bounded away from 0 since $\sigma_{0, r}^2 \rightarrow (0.955)(0.045) = 0.042975$ as $r \rightarrow \infty$. Our specific parameter values lead to $n_{r, \gamma, \beta} = 0.016$ from (32) and $n'_{r, \gamma, \beta} \doteq 7.15$ from (34), although both sample sizes should be taken much larger for the CLT approximations in (31) and (33) to roughly hold.

Acknowledgements

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

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

- [2] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury, Pacific Grove, Calif., 2nd edition, 2002.
- [3] F. Chu and M. K. Nakayama. Confidence intervals for quantiles when applying variance-reduction techniques. *ACM Transactions On Modeling and Computer Simulation*, 22(2):10:1–10:25, 2012.
- [4] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, C, 3rd edition, 2022.
- [5] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, second edition, 1998.
- [6] H. Dong and M. K. Nakayama. A tutorial on quantile estimation via Monte Carlo. In P. L'Ecuyer and B. Tuffin, editors, *Monte Carlo & Quasi-Monte Carlo Methods:MCQMC 2018*, volume 324 of *Springer Proceedings in Mathematics & Statistics*, pages 3–30. Springer, Cham, Switzerland, 2019.
- [7] P. W. Glynn. Importance sampling for Monte Carlo estimation of quantiles. In S. M. Ermakov and V. B. Melas, editors, *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation*, pages 180–185, St. Petersburg, Russia, 1996. Publishing House of St. Petersburg Univ.
- [8] P. W. Glynn, M. K. Nakayama, and B. Tuffin. On the estimation of the mean time to failure by simulation. In *2017 Winter Simulation Conference (WSC)*, pages 1844–1855, 2017.
- [9] P. W. Glynn, M. K. Nakayama, and B. Tuffin. Using simulation to calibrate exponential approximations to tail-distribution measures of hitting times to rarely visited sets. In *2018 Winter Simulation Conference (WSC)*, pages 1802–1813, 2018.
- [10] A. Goyal, P. Shahabuddin, P. Heidelberger, V. Nicola, and P. W. Glynn. A unified framework for simulating Markovian models of highly dependable systems. *IEEE Transactions on Computers*, C-41(1):36–51, 1992.
- [11] J. L. Jensen. *Saddlepoint Approximations*. Oxford University Press, New York, 1995.
- [12] V. Kalashnikov. *Geometric Sums: Bounds for Rare Events with Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

- [13] P. L'Ecuyer, J. H. Blanchet, B. Tuffin, and P. W. Glynn. Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation*, 20(1):Article 6, 2010.
- [14] P. L'Ecuyer, M. Mandjes, and B. Tuffin. Importance sampling and rare event simulation. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation using Monte Carlo Methods*, chapter 2, pages 17–38. Wiley, Chichester, UK, 2009.
- [15] E. L. Lehmann. *Elements of Large-Sample Theory*. Springer, New York, 1999.
- [16] Y. Li, Z. T. Kaplan, and M. K. Nakayama. Monte Carlo methods for economic capital. *INFORMS Journal on Computing*, 2024. To appear.
- [17] E. Parzen. Quantile probability and statistical data modeling. *Statistical Science*, 19(4):652–662, 2004.
- [18] R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons, New York, 1980.
- [19] 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). Technical report, U.S. Nuclear Regulatory Commission, Washington, DC, 2005.
- [20] 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.
- [21] 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.
- [22] A. W. Van Der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK, 1998.
- [23] S. S. Wilks. Determination of sample sizes for setting tolerance limits. *Annals of Mathematical Statistics*, 12:91–96, 1941.