CORRELATION-INDUCTION TECHNIQUES FOR ESTIMATING QUANTILES IN SIMULATION EXPERIMENTS

ATHANASSIOS N. AVRAMIDIS
SABRE Decision Technologies, Paris, France

JAMES R. WILSON
North Carolina State University, Raleigh, North Carolina

(November 7, 1996)

A simulation-based quantile estimator measures the level of system performance that can be delivered with a prespecified probability. To estimate selected quantiles of the response of a finite-horizon simulation, we develop procedures based on correlation-induction techniques for variance reduction, with emphasis on antithetic variates and Latin hypercube sampling. These procedures achieve improved precision by controlling the simulation's random-number inputs as an integral part of the experimental design. The proposed multiple-sample quantile estimator is the average of negatively correlated quantile estimators computed from disjoint samples of the simulation response, where negative correlation is induced between corresponding responses in different samples while mutual independence of responses is maintained within each sample. The proposed single-sample quantile estimator is computed from negatively correlated simulation responses within one all-inclusive sample. The single-sample estimator based on Latin hypercube sampling is shown to be asymptotically normal and unbiased with smaller variance than the comparable direct-simulation estimator based on independent replications. Similar asymptotic comparisons of the multiple-sample and direct-simulation estimators focus on bias and mean square error. Monte Carlo results suggest that the proposed procedures can yield significant reductions in bias, variance, and mean square error when estimating quantiles of the completion time of a stochastic activity network.

Subject classifications: Simulation, efficiency: variance reduction techniques. Simulation, design of experiments: antithetic variates, Latin hypercube sampling. Simulation, statistical analysis: single- and multiple-sample quantile estimators.

Area of review: SIMULATION.

In this paper we formulate and analyze procedures for estimating selected quantiles of the response Y of a finite-horizon stochastic simulation experiment based on the variance reduction technique of correlation induction. Let $F(\cdot)$ denote the (unknown) cumulative distribution function (c.d.f.) of Y. For any r with $0 < r < 1$, the rth quantile ξ_r of the random variable Y is the smallest value t such that $F(t) \equiv \Pr\{Y \leq t\} \geq r$. Most of the literature on simulation output analysis is concerned with estimating the mean of the response Y or the mean of some known function of Y. Unfortunately estimation of a quantile is fundamentally different from estimation of a mean in the sense that an arbitrary quantile of Y cannot generally be expressed as the mean of a known function of Y. Quantiles provide additional information about the distribution of Y, and in certain cases they may be of more interest than the mean. For example, to meet the scheduled completion date δ of a large engineering project with a specified degree of confidence (say, 95%), the project manager may use a simulation model of the project to obtain an estimator $\widehat{\xi}_{0.95}$ of the 95th percentile (quantile) $\xi_{0.95}$ of the project duration Y; and then the required project starting time is estimated by $\delta - \widehat{\xi}_{0.95}$.

As detailed in Subsection 1.1 below, the direct-simulation method for estimating the rth quantile ξ_r of the response Y is based on the order statistics of a sample of independent identically distributed (i.i.d.) observations of Y. Variance reduction techniques seek to restructure the simulation experiment to improve the efficiency of the estimation procedure—that is, to reduce the estimation error for a fixed computing budget. The problem of variance reduction for quantile estimation has received relatively little attention in the simulation literature. To address this problem, Ressler and Lewis (1990) extended the method of control variates to apply to a nonlinear transformation of an auxiliary simulation response that has known quantiles. Specifically they proposed using as a control variate a transformation of the direct-simulation estimator of the rth quantile of the auxiliary response, where the transformation is chosen to improve the correlation between the target response and the control variate. Hsu and Nelson (1990) also used a control variate with known quantiles that results from inverting a classical linear control-variate estimator for probabilities. Hesterberg and Nelson (1995) exploited control variates with known quantiles to estimate either the target c.d.f. at selected cutoff values or selected quantiles of the target distribution. In a Monte Carlo performance evaluation involving inventory, project-planning, and queueing examples, Hesterberg and Nelson obtained reductions in mean square error ranging from 10% to 80% when estimating quantiles of order 0.90, 0.95, and 0.99. In practice the main drawback of all of these quantile-estimation methods seems to be the difficulty of identifying control variates with known quantiles (as opposed to identifying control variates with known means) that are strongly correlated with the response variable.

The objective of this work is to develop practical, effective variance reduction techniques for estimating selected quantiles of the response in large-scale, finite-horizon simulation experiments. The rest of this paper is organized as follows. In Section 1 we begin by discussing quantile estimation via direct simulation; and we establish some basic results on correlation-induction techniques for variance reduction, with emphasis on the methods of antithetic variates and Latin hypercube sampling.

In Section 2 we formulate and analyze multiple-sample quantile estimators wherein negative correlation is induced between the corresponding simulation responses in disjoint samples while mutual independence of the simulation responses is maintained within each sample. Section 3 treats quantile estimators resulting from correlation induction within a single sample of simulation responses. In Section 4 we summarize the results of a Monte Carlo study designed to gauge the reductions in bias, variance, and mean square error that are achieved by the proposed quantile-estimation techniques in the context of simulating stochastic activity networks; moreover, we validate all the assumptions underlying our main theoretical results in a broad class of activity-network simulations that includes the two networks used in the Monte Carlo study. Finally in Section 5 we recapitulate our main findings, and we make recommendations for follow-up work.

1 BACKGROUND

1.1 Quantile Estimation via Direct Simulation

We consider finite-horizon simulation experiments in which the response has the form $Y \equiv y(\mathbf{U})$, where $\mathbf{U} \equiv (U_1, \ldots, U_d)$ is composed of d independent *random numbers*—i.e., random variables that are uniformly distributed on the unit interval (0, 1). The dimension d of the random-number input vector $\mathbf{U}$ is a finite constant. In terms of the (unknown) *inverse c.d.f.* of Y,

$$F^{-1}(u) \equiv \min\{t : F(t) \geq u\} \quad \text{for all} \quad u \in (0,\, 1)\,,$$

the rth quantile of the distribution of Y is

$$\xi \equiv F^{-1}(r) \quad \text{for} \quad 0 < r < 1.$$

Throughout the rest of this paper, we assume that a single value of r is specified; and we suppress the dependence of ξ on r for notational simplicity.

In a direct-simulation experiment, we perform n independent replications that yield i.i.d. observations $\{Y_i : i = 1, \ldots, n\}$ of the target response. The *direct-simulation* estimator of ξ based on n independent replications is denoted by

$$\widehat{\xi}_{\mathrm{DS}}(\psi, n) \equiv \psi(Y_1, Y_2, \ldots, Y_n), \tag{1}$$

where we will consider several choices for the function $\psi(\cdot)$. The most natural approach to estimating ξ is to use $\widehat{\xi}_{\mathrm{DS}} = \min\{t : F_n(t) \geq r\}$, where $F_n(\cdot)$ is the empirical c.d.f. based on the sample $\{Y_i : i = 1, \ldots, n\}$. In terms of the *order statistics* $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ obtained by sorting the observations $\{Y_i : i = 1, \ldots, n\}$ in ascending order, the usual definition of $F_n(\cdot)$ is

$$F_n(t) \equiv \begin{cases} 0, & \text{if } \; t < Y_{(1)}\,, \\ i/n, & \text{if } \; Y_{(i)} \leq t < Y_{(i+1)} \;\text{ and }\; 1 \leq i \leq n-1\,, \\ 1, & \text{if } \; Y_{(n)} \leq t\,; \end{cases} \tag{2}$$

and this choice for $F_n(\cdot)$ corresponds to taking $\psi(\cdot) = \psi_1(\cdot)$ in the general definition (1) of the direct-simulation quantile estimator, where

$$\psi_1(Y_1, Y_2, \ldots, Y_n) \equiv F_n^{-1}(r) = Y_{(\lceil nr \rceil)} \tag{3}$$

and $\lceil x \rceil$ denotes the smallest integer that is greater than or equal to x. See David (1981) for properties of $\widehat{\xi}_{\mathrm{DS}}(\psi_1, n)$.

A second quantile estimator that is used, for example, in the S statistical package (Becker and Chambers 1984) results from taking a piecewise linear version $\widetilde{F}_n(\cdot)$ of the empirical c.d.f. such that $\widetilde{F}_n[Y_{(i)}] \equiv (i-0.5)/n$ for $i = 1, \ldots, n-1$ and $\lim_{\varepsilon \to 0+} \widetilde{F}_n[Y_{(n)} - \varepsilon] \equiv (n-0.5)/n$, with $\widetilde{F}_n(t) \equiv 0$ for $t < Y_{(1)}$ and $\widetilde{F}_n(t) \equiv 1$ for $t \geq Y_{(n)}$. This choice for the empirical c.d.f. corresponds to taking $\psi(\cdot) = \psi_2(\cdot)$ in the general definition (1), where

$$\psi_2(Y_1, Y_2, \ldots, Y_n) \equiv \widetilde{F}_n^{-1}(r) = \tag{4}$$
$$\begin{cases} Y_{(1)}, & \text{if } \ r \leq 0.5/n\,, \\ \vartheta_n Y_{(\lceil nr+0.5 \rceil - 1)} + (1-\vartheta_n) Y_{(\lceil nr+0.5 \rceil)}, & \text{if } \ 0.5/n < r < (n-0.5)/n\,, \\ Y_{(n)}, & \text{if } \ (n-0.5)/n \leq r, \end{cases}$$

and

$$\vartheta_n \equiv \lceil nr + 0.5 \rceil - (nr + 0.5) \quad \text{for} \quad n = 1, 2, \ldots\,. \tag{5}$$

A more general quantile estimator can be based on a linear combination of the order statistics; and in this case we take $\psi(\cdot) = \psi_3(\cdot)$ in the general definition (1), where

$$\psi_3(Y_1, Y_2, \ldots, Y_n) = \sum_{i=1}^{n} \lambda_{i,n} Y_{(i)}\,, \tag{6}$$

and $\{\lambda_{i,n} : i = 1, \ldots, n\}$ are constants. For example, the quantile estimator proposed by Yang (1985) has the form (6). The statistic proposed by Kappenman (1987) appears to be the only quantile estimator in the literature that is a function of the $\{Y_i : i = 1, \ldots, n\}$ alone but yet is not a linear combination of the corresponding order statistics.

1.2 A General Scheme for Correlation Induction

To provide a general framework for correlation induction, we introduce the notion of negative quadrant dependence, which was defined by Lehmann (1966).

Definition 1 *The bivariate random vector $(A_1, A_2)^{\mathrm{T}}$ is negatively quadrant dependent (n.q.d.) if*

$$\Pr\{A_1 \leq a_1,\, A_2 \leq a_2\} \leq \Pr\{A_1 \leq a_1\} \cdot \Pr\{A_2 \leq a_2\} \quad \textit{for all } a_1, a_2.$$

Equivalently, we will say that the distribution of $(A_1, A_2)^{\mathrm{T}}$ is n.q.d. We will exploit this concept in Result 2 below to provide the desired sufficient condition for negatively correlated simulation responses. Moreover, we use the concept of negative quadrant dependence to define a special class $\mathcal{G}$ of distributions for the random-number inputs. Every distribution $G \in \mathcal{G}$ must have the following correlation-induction properties:

CI_1 For some $k \geq 2$, G is a k-variate distribution with univariate marginals that are uniform on the unit interval (0, 1).

CI_2 Each bivariate marginal of G is n.q.d.

When it is desirable to indicate explicitly that a distribution in $\mathcal{G}$ is k-variate, we will write $G^{(k)} \in \mathcal{G}$ rather than $G \in \mathcal{G}$.

Next we discuss how k-dimensional vectors of (uniform) random numbers sampled according to $G^{(k)} \in \mathcal{G}$ are used to generate k negatively correlated observations of the simulation response Y, where each Y-value is assigned to a different sample of simulation responses; and if this procedure is independently replicated m times, then we obtain k random samples of size m from the target distribution $F(\cdot)$ such that the quantile estimators of the form (1) computed from each of these k samples are also negatively correlated. Using a k-variate distribution $G^{(k)}$ selected from the special class $\mathcal{G}$ of distributions, we induce negative quadrant dependence between k replications of Y as follows. We perform k dependent replications yielding outputs

$$Y^{(i)} = y\left[U_1^{(i)}, \ldots, U_d^{(i)}\right] \quad \text{for} \quad i = 1, \ldots, k \tag{7}$$

by sampling the column vectors of input random numbers,

$$\mathcal{U}_j \equiv \left[U_j^{(1)}, \ldots, U_j^{(k)}\right]^{\mathrm{T}} \quad \text{for} \quad j = 1, \ldots, d, \tag{8}$$

according to a scheme satisfying the following conditions:

SC_1 For each j $(1 \leq j \leq d)$, the random vector $\mathcal{U}_j$ has distribution $G^{(k)}$.

SC_2 The column vectors $\mathcal{U}_1, \ldots, \mathcal{U}_d$ are mutually independent.

Sampling condition SC_1 specifies that for each $j \in \{1, \ldots, d\}$, we induce dependence between the outputs $\left\{Y^{(i)} : i = 1, \ldots, k\right\}$ by arranging a negative quadrant dependence between the jth random numbers sampled on each pair of replications. Sampling condition SC_2 requires mutual independence of the random numbers used within the ith replication to generate the output $Y^{(i)}$; and together with property CI_1, this guarantees that each $Y^{(i)}$ has the target distribution $F(\cdot)$.

Definition 2 *The sample* $\left\{Y^{(i)} : i = 1, \ldots, k\right\}$ *is called a* $G^{(k)}$*-sample of* Y *if it is generated as in* (7) *and* (8) *subject to conditions* SC_1 *and* SC_2.

The next two results provide the justification for using correlation-induction techniques to reduce the variance of simulation-generated statistics.

Result 1 *If $G^{(k)} \in \mathcal{G}$, if $\left\{Y^{(i)} : i = 1, \ldots, k\right\}$ is a $G^{(k)}$-sample of Y, and if $y(\cdot)$ is a monotone function of each argument individually, then $\left[Y^{(i)}, Y^{(\ell)}\right]^{\mathrm{T}}$ is n.q.d. for $i \neq \ell$.*

Result 1 is essentially Theorem 1(ii) of Lehmann (1966).

Result 2 *If the bivariate random vector $(A_1, A_2)^{\mathrm{T}}$ is n.q.d., then $\mathrm{Cov}(A_1, A_2) \leq 0$, with equality holding if and only if A_1 and A_2 are independent.*

Result 2 is Lemma 3 of Lehmann (1966).

For an elaboration of the general framework for correlation induction presented in this section, see Avramidis and Wilson (1996). In the next subsection we give examples of correlation-induction techniques that are special cases of the general scheme described above, and in each case we prove that the relevant distribution G belongs to the class $\mathcal{G}$. Avramidis and Wilson (1995) provide additional examples illustrating the correlation-induction scheme detailed in this section.

1.3 Special Cases of Correlation Induction

1.3.1 Antithetic Variates (AV)

To generate $k = 2$ correlated replications of the simulation response by the method of antithetic variates, we sample the random numbers $\{U_j^\star : j = 1, \ldots, d\}$ independently and compute the column vectors of (8) according to the relation

$$\mathcal{U}_j = \left(U_j^\star, 1 - U_j^\star\right)^{\mathrm{T}} \quad \text{for} \quad j = 1, 2, \ldots, d\,.$$

(Throughout the rest of this paper, we reserve the notation $U^\star$ to denote a random number that is sampled independently.) We let $G_{\mathrm{AV}}^{(2)}$ denote the distribution of $\mathcal{U}_j$. It is straightforward to check that $G_{\mathrm{AV}}^{(2)}$ satisfies conditions CI_1 and CI_2 so that $G_{\mathrm{AV}}^{(2)} \in \mathcal{G}$. The method of antithetic variates is clearly a special case of the general correlation-induction scheme described by (7) and (8).

1.3.2 Latin Hypercube Sampling (LHS)

To generate k correlated replications of the simulation response via Latin Hypercube Sampling (LHS) for $k \geq 2$, we compute the input random numbers according to the relation

$$U_j^{(i)} = \frac{\pi_j(i) - U_{ij}^\star}{k} \quad \text{for} \quad i = 1, \ldots, k \quad \text{and} \quad j = 1, \ldots, d, \tag{9}$$

where

a. $\pi_1(\cdot), \ldots, \pi_d(\cdot)$ are permutations of the integers $\{1, \ldots, k\}$ that are randomly sampled with replacement from the set of $k!$ such permutations, with $\pi_j(i)$ denoting the ith element in the jth randomly sampled permutation; and

b. $\{U^{\star}_{ij} : j = 1, \ldots, d,\ i = 1, \ldots, k\}$ are random numbers sampled independently of each other and of the permutations $\pi_1(\cdot), \ldots, \pi_d(\cdot)$.

We let $G^{(k)}_{\mathrm{LH}}$ denote the distribution of each k-dimensional column vector of input random numbers generated in this way so that

$$\mathcal{U}_j \sim G^{(k)}_{\mathrm{LH}} \iff \mathcal{U}_j = \left[U^{(1)}_j, \ldots, U^{(k)}_j\right]^{\mathrm{T}} \text{ is generated according to (9).} \tag{10}$$

The key property of LHS is that for each j ($j = 1, \ldots, d$), the components of the column vector $\mathcal{U}_j$ form a stratified sample of size k from the uniform distribution on the unit interval (0, 1) such that there is a single observation in each stratum and the observations within the sample are negatively quadrant dependent; moreover, different stratified samples of size k are independent. Since $\pi_j(\cdot)$ is a random permutation of the integers $\{1, \ldots, k\}$, each element $\pi_j(i)$ for $i = 1, \ldots, k$ has the discrete uniform distribution on the set $\{1, \ldots, k\}$; and thus in the definition (9), the variate $\pi_j(i)$ randomly indexes a subinterval (stratum) of the form $\big((\ell - 1)/k,\ \ell/k\big]$ for some $\ell \in \{1, \ldots, k\}$. Since $U^{\star}_{ij}$ is a random number sampled independently of $\pi_j(i)$, we see that $U^{(i)}_j$ is uniformly distributed in the subinterval indexed by $\pi_j(i)$; and it follows that $U^{(i)}_j$ is uniformly distributed on the unit interval (0, 1). Moreover, since $\pi_j(\cdot)$ is a permutation of $\{1, \ldots, k\}$, every subinterval (stratum) of the form $\big((\ell - 1)/k,\ \ell/k\big]$ for $\ell = 1, \ldots, k$ contains exactly one of the negatively quadrant dependent random numbers $\left\{U^{(i)}_j : i = 1, \ldots, k\right\}$ so that the components of $\mathcal{U}_j$ constitute a stratified sample of the uniform distribution on (0, 1). Finally, we notice that the column vectors $\mathcal{U}_1, \ldots, \mathcal{U}_d$ are independent since the random permutations $\{\pi_j(\cdot) : j = 1, \ldots, d\}$ and the random numbers $\{U^{\star}_{ij} : i = 1, \ldots, k; j = 1, \ldots, d\}$ are all generated independently. This discussion is formalized in the following result.

Proposition 1 *For any $k \geq 2$, the distribution $G^{(k)}_{\mathrm{LH}}$ defined in* (10) *is in the class $\mathcal{G}$.*

Proof. Choose $j \in \{1, \ldots, d\}$ arbitrarily. For the random permutation $\pi_j(\cdot)$ of the integers $\{1, \ldots, k\}$, it is straightforward to check that the random vector $[\pi_j(1), \pi_j(2)]^{\mathrm{T}}$ is n.q.d.—this is done, for example, in the proof of the theorem in McKay, Beckman, and Conover (1979, p. 245). If $U^{\star}_{1j}$ and $U^{\star}_{2j}$ are random numbers sampled independently of each other and of $\pi_j(\cdot)$, then

$$\left[U^{(1)}_j, U^{(2)}_j\right]^{\mathrm{T}} = \left[\frac{\pi_j(1) - U^{\star}_{1j}}{k}, \frac{\pi_j(2) - U^{\star}_{2j}}{k}\right]^{\mathrm{T}} \quad \text{is n.q.d.}$$

by Theorem 1(iii) of Lehmann (1966); and it is obvious that the distribution of $\left[U^{(1)}_j, U^{(2)}_j\right]^{\mathrm{T}}$ coincides with all bivariate marginals of $G^{(k)}_{\mathrm{LH}}$. Clearly $U^{(1)}_j$ has a uniform distribution on the interval $(0, 1)$ and all univariate marginals of $G^{(k)}_{\mathrm{LH}}$ are equal, so conditions CI_1 and CI_2 are satisfied by $G^{(k)}_{\mathrm{LH}}$ for any $k \geq 2$. This completes the proof of Proposition 1.

In view of Proposition 1, we can take $G^{(k)} = G^{(k)}_{\mathrm{LH}}$ in (7) and (8); and thus we see that LHS is a special case of correlation induction. First devised by McKay, Beckman, and Conover (1979), LHS was subsequently studied by Stein (1987) and Owen (1992a, b).

2 CORRELATION INDUCTION ACROSS SAMPLES

Motivated by the need to estimate the variance of a quantile estimator, Schafer (1974) suggested computing the sample mean and the corresponding sample standard error from k independent quantile estimators that are respectively based on k disjoint samples, where each sample consists of $m = \lfloor n/k \rfloor$ independent observations of the response Y. To simplify the exposition, we assume throughout this paper that n is an integral multiple of k so that $n = km$. Letting $\widehat{\xi}^{(i)}_{\mathrm{DS}}(\psi, m)$ denote the direct-simulation estimator of ξ computed by applying the function $\psi(\cdot)$ to the ith random sample of size m for $i = 1, \ldots, k$, we define the *direct simulation–multiple sample* estimator of ξ,

$$\widehat{\xi}_{\mathrm{DS-MS}}(\psi, k, n) \equiv k^{-1} \sum_{i=1}^{k} \widehat{\xi}^{(i)}_{\mathrm{DS}}(\psi, n/k), \tag{11}$$

where we have substituted n/k for m on the right-hand side of (11) to emphasize the exact dependence of $\widehat{\xi}_{\mathrm{DS-MS}}(\psi, k, n)$ on the function $\psi(\cdot)$, the parameter k, and the total sample size n. Although the direct simulation–multiple sample estimator does not use any variance reduction techniques, we introduce it because it will simplify the statement of some of our results.

If we forgo having a variance estimator associated with our estimator of ξ, then we can improve upon (11) by inducing negative correlation between the k direct-simulation quantile estimators that are averaged to obtain (11). Let $G^{(k)}$ be a k-variate distribution selected from $\mathcal{G}$. We generate m column samples of the simulation response with the following properties:

CI–MS$_1$ The jth column sample $\left[Y^{(1)}_j, \ldots, Y^{(k)}_j\right]^{\mathrm{T}}$ is a $G^{(k)}$-sample of Y for $j = 1, \ldots, m$.

CI–MS$_2$ The column samples $\left\{\left[Y^{(1)}_j, \ldots, Y^{(k)}_j\right]^{\mathrm{T}} : j = 1, \ldots, m\right\}$ are mutually independent.

The total set of Y-observations can also be arranged in k row samples,

$$\left\{\left[Y^{(i)}_1, \ldots, Y^{(i)}_m\right] : i = 1, \ldots, k\right\}. \tag{12}$$

Condition CI–MS$_2$ guarantees that each row sample consists of m independent observations of Y, and condition CI–MS$_1$ suggests that we have induced dependence between the row vectors enumerated in (12). From this sampling scheme we can compute $\widehat{\xi}_{\mathrm{CI-MS}}\left(\psi, G^{(k)}, n\right)$, the *correlation induction–multiple sample* estimator of ξ based on the k-variate distribution $G^{(k)}$. Specifically, $\widehat{\xi}_{\mathrm{CI-MS}}\left(\psi, G^{(k)}, n\right)$ is obtained by applying the function $\psi(\cdot)$ to the ith row sample,

$$\widehat{\xi}^{(i)}_{\mathrm{DS}}(\psi, m) = \psi\left[Y^{(i)}_1, \ldots, Y^{(i)}_m\right] \quad \text{for} \quad i = 1, \ldots, k,$$

and then averaging the resulting correlated quantile estimators to obtain

$$\widehat{\xi}_{\mathrm{CI-MS}}\left(\psi, G^{(k)}, n\right) \equiv k^{-1} \sum_{i=1}^{k} \widehat{\xi}^{(i)}_{\mathrm{DS}}(\psi, n/k). \tag{13}$$

We substituted n/k for m in the right-hand side of (13) to emphasize the exact dependence of the estimator $\widehat{\xi}_{\text{CI–MS}}\left(\psi, G^{(k)}, n\right)$ on the function $\psi(\cdot)$, the distribution $G^{(k)}$, the parameter k, and the total sample size n. We will occasionally suppress the dependence of $\widehat{\xi}_{\text{CI–MS}}$ on some or all of its three arguments when no confusion can result from this usage.

2.1 Mean Square Error of Multiple-Sample Quantile Estimators

With respect to the performance measure of mean square error (MSE), we compare the correlation induction–multiple sample estimator $\widehat{\xi}_{\text{CI–MS}}\left(\psi, G^{(k)}, n\right)$ with the direct simulation–multiple sample estimator $\widehat{\xi}_{\text{DS–MS}}(\psi, k, n)$.

Theorem 1 *If $y(\cdot)$ and $\psi(\cdot)$ are monotone functions of each of their arguments individually, then*

$$\mathrm{MSE}\left[\widehat{\xi}_{\text{CI–MS}}\left(\psi, G^{(k)}, n\right)\right] \le \mathrm{MSE}\left[\widehat{\xi}_{\text{DS–MS}}(\psi, k, n)\right] \tag{14}$$

for any k-variate distribution $G^{(k)} \in \mathcal{G}$ and any sample size n.

Proof. Consider the scheme for generating the column samples of the simulation response that are required to compute (13). To index specific elements $Y_j^{(i)}$ and $Y_j^{(\ell)}$ of the jth column sample, choose $i, \ell \in \{1, \ldots, k\}$ arbitrarily so that $i \neq \ell$. Since $y(U_1, \ldots, U_d)$ is a monotone function of each of its arguments individually, it follows from (7), (8), and Result 1 in Subsection 1.2 that the random vector $\left[Y_j^{(i)},\, Y_j^{(\ell)}\right]^{\mathrm{T}}$ is n.q.d. for each $j = 1, \ldots, m$. Moreover, by property CI–MS$_2$, the pairs $\left\{\left[Y_j^{(i)},\, Y_j^{(\ell)}\right]^{\mathrm{T}} : j = 1, \ldots, m\right\}$ are mutually independent. Since $\psi(Y_1, \ldots, Y_m)$ is a monotone function of each of its arguments individually, it also follows from Theorem 1(ii) of Lehmann (1966) that $\left[\widehat{\xi}_{\text{DS}}^{(i)},\, \widehat{\xi}_{\text{DS}}^{(\ell)}\right]^{\mathrm{T}}$ is n.q.d.; and by Result 2 in Subsection 1.2, we see that $\mathrm{Cov}\left[\widehat{\xi}_{\text{DS}}^{(i)},\, \widehat{\xi}_{\text{DS}}^{(\ell)}\right] \le 0$. Thus we have

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\xi}_{\text{CI–MS}}\left(\psi, G^{(k)}, n\right)\right] &= k^{-2}\left\{\sum_{i=1}^{k} \mathrm{Var}\left[\widehat{\xi}_{\text{DS}}^{(i)}\right] + 2\sum_{i=1}^{k-1}\sum_{\ell=i+1}^{k} \mathrm{Cov}\left[\widehat{\xi}_{\text{DS}}^{(i)},\, \widehat{\xi}_{\text{DS}}^{(\ell)}\right]\right\} \\
&\le k^{-1}\mathrm{Var}\left[\widehat{\xi}_{\text{DS}}^{(1)}\right] \\
&= \mathrm{Var}\left[\widehat{\xi}_{\text{DS–MS}}(\psi, k, n)\right],
\end{aligned}$$

since $\widehat{\xi}_{\text{DS–MS}}(\psi, k, n)$ is the average of k independent replications of $\widehat{\xi}_{\text{DS}}(\psi, n/k)$. Clearly

$$\mathrm{E}\left[\widehat{\xi}_{\text{CI–MS}}\left(\psi, G^{(k)}, n\right)\right] = \mathrm{E}\left[\widehat{\xi}_{\text{DS}}^{(1)}\right] = \mathrm{E}\left[\widehat{\xi}_{\text{DS–MS}}(\psi, k, n)\right],$$

so

$$\mathrm{Bias}\left[\widehat{\xi}_{\text{CI–MS}}\left(\psi, G^{(k)}, n\right)\right] = \mathrm{Bias}\left[\widehat{\xi}_{\text{DS–MS}}(\psi, k, n)\right] = \mathrm{Bias}\left[\widehat{\xi}_{\text{DS}}(\psi, n/k)\right], \tag{15}$$

and the desired result (14) follows from the basic relation $\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Var}$.

Remark 1 Typically the function $\psi(\cdot)$ satisfies the monotonicity requirement in Theorem 1. Both $\psi_1(\cdot)$ and $\psi_2(\cdot)$ satisfy this requirement, and so does $\psi_3(\cdot)$ in (6) if $\lambda_{i,n} \geq 0$ for all i and n.

Next we wish to compare the correlation induction–multiple sample estimator $\widehat{\xi}_{\text{CI–MS}}\Big(\psi, G^{(k)}, n\Big)$ with the direct-simulation estimator $\widehat{\xi}_{\text{DS}}(\psi, n)$ based on a single overall sample of n independent simulation responses. Without any assumptions about the distribution of the simulation response, it is difficult to compare the bias, variance, and MSE of these two estimators for finite n because there are no closed-form expressions for the bias and variance of the relevant order statistics. To characterize adequately the asymptotic behavior of the moments of the relevant order statistics together with the resulting bias and variance of the estimators $\widehat{\xi}_{\text{DS}}(\psi, n)$ and $\widehat{\xi}_{\text{CI–MS}}\Big(\psi, G^{(k)}, n\Big)$, we introduce the following regularity conditions for an arbitrary inverse c.d.f. $Q(\cdot)$ and its derivatives.

RC_1 There exist nonnegative integers a and b such that $Q(u)u^a(1-u)^b$ is bounded for $u \in (0,\, 1)$.

RC_2 There is an open subinterval $\mathcal{S} \subset (0,\, 1)$ containing r such that the second derivative $Q''(\cdot)$ is continuous on $\mathcal{S}$.

RC_3 The first derivative $Q'(r) > 0$.

RC_4 The third derivative $Q'''(\cdot)$ is bounded on the subinterval $\mathcal{S}$ of condition RC_2.

The following lemma describes the bias and variance of the direct-simulation quantile estimator $\widehat{\xi}_{\text{DS}}(\psi_c, n)$ for $c = 1$ and 2, respectively, as these quantities depend on the sample size n. This result is proved in the Appendix.

Lemma 1 *If conditions* RC_1–RC_3 *hold for* $Q(\cdot) = F^{-1}(\cdot)$, *then*

$$\left.\begin{aligned} \text{Bias}\Big[\widehat{\xi}_{\text{DS}}(\psi_c, n)\Big] &= o\Big(n^{-1/2}\Big) \\ \text{Var}\Big[\widehat{\xi}_{\text{DS}}(\psi_c, n)\Big] &= \frac{r(1-r)}{n[F'(\xi)]^2} + o\Big(n^{-1}\Big) \end{aligned}\right\} \quad \textit{for} \quad c = 1, 2. \tag{16}$$

Now we are able to make an asymptotic comparison of the MSEs of the single- and multiple-sample direct-simulation quantile estimators that are based on $\psi_1(\cdot)$ or $\psi_2(\cdot)$. Let k be fixed. In view of (16), we have

$$\begin{aligned} \lim_{n\to\infty} n\text{MSE}\Big[\widehat{\xi}_{\text{DS–MS}}(\psi_c, k, n)\Big] &= \lim_{n\to\infty} n\text{Bias}^2\Big[\widehat{\xi}_{\text{DS–MS}}(\psi_c, k, n)\Big] + \lim_{n\to\infty} n\text{Var}\Big[\widehat{\xi}_{\text{DS–MS}}(\psi_c, k, n)\Big] \\ &= \lim_{n\to\infty} n\text{Bias}^2\Big[\widehat{\xi}_{\text{DS}}(\psi_c, n/k)\Big] + \lim_{n\to\infty} \frac{n}{k}\text{Var}\Big[\widehat{\xi}_{\text{DS}}(\psi_c, n/k)\Big] \\ &= 0 + \frac{r(1-r)}{[F'(\xi)]^2} \\ &= \lim_{n\to\infty} n\text{MSE}\Big[\widehat{\xi}_{\text{DS}}(\psi_c, n)\Big] \quad \text{for} \quad c = 1, 2. \end{aligned} \tag{17}$$

Finally we compare, in an asymptotic MSE sense, the correlation induction–multiple sample estimator $\widehat{\xi}_{\text{CI–MS}}\Big(\psi_c, G^{(k)}, n\Big)$ with the direct-simulation estimator $\widehat{\xi}_{\text{DS}}(\psi_c, n)$ based on a single overall sample of n independent simulation responses, where $c = 1$ and 2, respectively.

Theorem 2 *If $y(\cdot)$ is a monotone function of each of its arguments individually and if conditions* RC_1–RC_3 *hold for $Q(\cdot) = F^{-1}(\cdot)$, then*

$$\limsup_{n\to\infty} n\mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{CI\text{–}MS}}\Big(\psi_c, G^{(k)}, n\Big)\Big] \le \lim_{n\to\infty} n\mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_c, n)\Big] = \frac{r(1-r)}{[F'(\xi)]^2} \tag{18}$$

for any distribution $G^{(k)} \in \mathcal{G}$ and for $c = 1, 2$.

Proof. For $c = 1$ and 2 respectively, the result follows by taking $\psi(\cdot) = \psi_c(\cdot)$ in (14), multiplying both sides of (14) by n, taking the limit superior of each side as $n \to \infty$, and applying (17).

2.2 Bias of Direct-Simulation Quantile Estimators

For each quantile estimator $\widehat{\xi}$ discussed in the previous section, we saw that as the total sample size n becomes large, the ratio $\mathrm{Bias}^2\big[\,\widehat{\xi}\,\big] \big/ \mathrm{Var}\big[\,\widehat{\xi}\,\big]$ goes to zero, i.e., the component of MSE due to bias becomes insignificant relative to the component of MSE due to variance. However, the bias component can be significant or can even dominate the variance component when the sample size is small, especially when we estimate extreme quantiles. This consideration is particularly relevant for correlation induction–multiple sample estimators because, as seen in (15), a multiple-sample estimator based on k samples with total sample size n has the same bias as the direct-simulation estimator with total sample size n/k. Thus we would like to identify a function $\psi(\cdot)$ with good bias properties, i.e. with small bias even for small sample sizes. Based on the following result, we will argue that the function $\psi_2(\cdot)$ is preferable to $\psi_1(\cdot)$ for quantile estimation. The proof of this result is given in the Appendix.

Proposition 2 *If conditions* RC_1–RC_4 *hold for $Q(\cdot) = F^{-1}(\cdot)$, then*

$$\mathrm{Bias}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_1, n)\Big] = \frac{1}{n}\left\{\frac{\lceil nr\rceil - nr - r}{F'(\xi)} - \frac{r(1-r)F''(\xi)}{2[F'(\xi)]^3}\right\} + o\Big(n^{-1}\Big) \tag{19}$$

and

$$\mathrm{Bias}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_2, n)\Big] = \frac{1}{n}\left\{\frac{0.5 - r}{F'(\xi)} - \frac{r(1-r)F''(\xi)}{2[F'(\xi)]^3}\right\} + o\Big(n^{-1}\Big)\,. \tag{20}$$

The bias expansions in Proposition 2 explain the well-known fact that bias is severe when we estimate extreme quantiles, since typically these quantiles are associated with very small values of $F'(\cdot)$. Suppose the assumptions of Proposition 2 are valid and suppose that $F'(\cdot)$ is unimodal with mode t_{mo}, so that $F''(\xi) > 0$ if $\xi < t_{\mathrm{mo}}$ and $F''(\xi) < 0$ if $\xi > t_{\mathrm{mo}}$. Equation (20) then shows that the two leading terms in the bias expansion for $\widehat{\xi}_{\mathrm{DS}}(\psi_2, n)$ are of opposite sign for all n if

$$r < \min\{F(t_{\mathrm{mo}}), 0.5\} \quad \text{or} \quad \max\{F(t_{\mathrm{mo}}), 0.5\} < r. \tag{21}$$

Moreover, (19) shows that $\widehat{\xi}_{\mathrm{DS}}(\psi_1, n)$ does not have this property. Specifically, for all $r \in (0, 1)$, we have the following behavior. The sign of the first term within the large curly braces on the right-hand side of (19) alternates infinitely often as $n \to \infty$, except for $r = 1/2$, in which case the first term vanishes for all odd values of n. On the other hand, the second term within this set of curly braces has the same sign for all n. These observations suggest that $\widehat{\xi}_{\mathrm{DS}}(\psi_2, n)$ should be preferred over $\widehat{\xi}_{\mathrm{DS}}(\psi_1, n)$ when (21) holds. Of course, since $F(t_{\mathrm{mo}})$ is unknown to the simulation practitioner, it is generally impossible to verify that (21) holds. However, as discussed above, typically bias is only important for small values of $F'(\xi)$. In the case of a unimodal distribution, small values of $F'(\xi)$ occur for r near 0 or 1; and this is precisely the situation described by (21). Based on these considerations, we recommend using the function $\psi_2(\cdot)$ rather than $\psi_1(\cdot)$ in all the quantile estimators discussed so far—especially when bias is expected to contribute significantly to MSE. Further evidence of the effectiveness of $\psi_2(\cdot)$ in reducing bias is given by the Monte Carlo results in Section 5.

3 CORRELATION INDUCTION WITHIN A SAMPLE

As discussed in the first paragraph of Subsection 2.2, multiple-sample quantile estimators are more prone to suffer from bias than single-sample estimators. If the bias component of MSE is expected to be dominant (due to a small sample size n, a value of r near 0 or 1, or both of these conditions), then using a multiple-sample quantile estimator might actually increase MSE by increasing bias as well as variance. This is the motivation for considering correlation induction within a sample—we use a single-sample estimator based on an all-inclusive set of dependent observations of the simulation response. In Section 3.1 we discuss a general quantile estimator based on correlation induction within a sample, and in Section 3.2 we study a special case of this estimator based on Latin hypercube sampling.

3.1 Correlation Induction–Single Sample Estimators

We compute $\widehat{\xi}_{\mathrm{CI\text{–}SS}}\left(\psi, G^{(n)}\right)$, the *correlation induction–single sample* estimator of ξ based on the function $\psi(\cdot)$ and the n-variate distribution $G^{(n)} \in \mathcal{G}$, by generating a $G^{(n)}$-sample of Y and applying $\psi(\cdot)$ to this single comprehensive sample of n correlated responses:

$$\widehat{\xi}_{\mathrm{CI\text{–}SS}}\left(\psi, G^{(n)}\right) \equiv \psi\left(Y^{(1)}, \ldots, Y^{(n)}\right), \text{ where } \left\{Y^{(1)}, \ldots, Y^{(n)}\right\} \text{ is a } G^{(n)}\text{-sample of } Y.$$

The dependence of $\widehat{\xi}_{\mathrm{CI\text{–}SS}}$ on the sample size n is implicit in the distribution $G^{(n)}$.

We emphasize the requirement that the distribution $G^{(n)}$ used for inducing dependence must have dimension equal to the sample size n. Thus in order for $\widehat{\xi}_{\mathrm{CI\text{–}SS}}$ to be well defined for all sample sizes, we must use distributions in $\mathcal{G}$ that are defined for any given dimension. Now strictly speaking, the distribution $G_{\mathrm{AV}}^{(2)}$ is only defined as a two-dimensional distribution; and there is no clear-cut extension of $G_{\mathrm{AV}}^{(2)}$ to higher dimensions such that negative correlation is achieved between each pair of coordinates in the column vector of input random numbers (8). On the other hand, for each n the distribution $G_{\mathrm{LH}}^{(n)}$ is readily defined and sampled, achieving a uniform negative correlation between each pair of coordinates in the column vector of input random numbers (8); and thus $G_{\mathrm{LH}}^{(n)}$ is particularly appropriate for use with $\widehat{\xi}_{\mathrm{CI\text{–}SS}}$. See Avramidis and Wilson (1995) for additional examples of distribution families that are readily used with $\widehat{\xi}_{\mathrm{CI\text{–}SS}}$.

The estimator $\widehat{\xi}_{\mathrm{CI\text{–}SS}}$ is fundamentally different from the estimators discussed previously—it is computed by applying the function $\psi(\cdot)$ to a single sample of *dependent* observations of the simulation response Y, whereas the estimators $\widehat{\xi}_{\mathrm{DS\text{–}MS}}$ and $\widehat{\xi}_{\mathrm{CI\text{–}MS}}$ of Section 2 are computed by repeatedly applying $\psi(\cdot)$ to samples of *independent* Y-observations. To motivate the estimator $\widehat{\xi}_{\mathrm{CI\text{–}SS}}$, we show that for each cutoff value t, the estimator $F_n(t)$ based on (2) has smaller variance if we induce negative quadrant dependence between each pair of observations in the sample $\left\{Y^{(i)} : i = 1, \ldots, n\right\}$. If we let $\mathbf{1}\{\cdot\}$ denote the indicator function for an event so that

$$\mathbf{1}\left\{Y^{(i)} \le t\right\} \equiv \begin{cases} 1, & \text{if } \; Y^{(i)} \le t, \\ 0, & \text{otherwise,} \end{cases}$$

then we have

$$\begin{aligned} \operatorname{Var}[F_n(t)] &= \operatorname{Var}\left[n^{-1}\sum_{i=1}^{n}\mathbf{1}\left\{Y^{(i)} \le t\right\}\right] \\ &= n^{-1}F(t)[1-F(t)] + 2n^{-2}\sum_{i=1}^{n-1}\sum_{\ell=i+1}^{n}\operatorname{Cov}\left[\mathbf{1}\left\{Y^{(i)} \le t\right\}, \mathbf{1}\left\{Y^{(\ell)} \le t\right\}\right]. \end{aligned} \tag{22}$$

Notice that $\mathbf{1}\left\{Y^{(i)} \le t\right\}$ is a nonincreasing function of $Y^{(i)}$ for each fixed t and for $i = 1, \ldots, n$. Thus if each pair of Y-observations is n.q.d., then each covariance on the right-hand side of (22) is nonpositive by Theorem 1(ii) of Lehmann (1966) and Result 2 of Subsection 1.2. On the other hand, if the Y-observations are i.i.d., then the variance of $F_n(t)$ is given by the first term on the right-hand side of (22). It follows that inducing a negative quadrant dependence between each pair of Y-observations in a single sample will yield an empirical c.d.f. $F_n(\cdot)$ that is everywhere a more accurate estimator of the underlying theoretical c.d.f. $F(\cdot)$ than could be obtained with random sampling. Since all the proposed quantile estimators that use the function $\psi_1(\cdot)$ are based on the inverse of an empirical c.d.f. having the form of $F_n(\cdot)$, it is plausible that inducing negative correlation between the Y-observations in a single sample will yield a more precise quantile estimator based on $\psi_1(\cdot)$ than a comparable single- or multiple-sample estimator based on applying $\psi_1(\cdot)$ to independent Y-observations. Although a similar argument for the quantile estimators based on $\psi_2(\cdot)$ is not obvious, in Theorem 3 below we show that for a special case of $\widehat{\xi}_{\mathrm{CI\text{–}SS}}$ based on Latin hypercube sampling, the same asymptotic performance is achieved with $\psi_1(\cdot)$ and $\psi_2(\cdot)$.

3.2 Latin Hypercube–Single Sample Estimators

We define the *Latin hypercube–single sample* estimator of ξ as

$$\widehat{\xi}_{\mathrm{LH\text{–}SS}}(\psi, n) \equiv \widehat{\xi}_{\mathrm{CI\text{–}SS}}\left(\psi, G_{\mathrm{LH}}^{(n)}\right).$$

Thus $\widehat{\xi}_{\text{LH–SS}}(\psi, n)$ is a function $\psi(\cdot)$ of a $G_{\text{LH}}^{(n)}$-sample of Y; and we will refer to such a sample as a *Latin hypercube sample of size* n. By Proposition 1 in Subsection 1.3.2 and by Results 1 and 2 in Subsection 1.2, any two Y-observations in a Latin hypercube sample of any size are n.q.d. and hence negatively correlated if $y(\cdot)$ is a monotone function of each of its arguments individually; and in such a case, we have the intuitive motivation for inducing correlation between responses that was elaborated in the last paragraph of the previous subsection. However, we will see that monotonicity of $y(\cdot)$ is not necessary to guarantee improved precision in our quantile estimators, at least for the special case of Latin hypercube sampling. We will derive the asymptotic distribution of $\widehat{\xi}_{\text{LH–SS}}(\psi_c, n)$ for $c = 1$ and 2 under appropriate conditions on the response Y; and as a by-product, we will see that the Latin hypercube–single sample estimators $\widehat{\xi}_{\text{LH–SS}}(\psi_c, n)$ for $c = 1, 2$ are asymptotically more precise than the direct-simulation estimators $\widehat{\xi}_{\text{DS}}(\psi_c, n)$ for $c = 1, 2$.

Some additional nomenclature is required to proceed. Let $\mathbf{U} \equiv (U_1, \ldots, U_d)$ denote the vector of random-number inputs to the simulation and let $\mathbf{u} \equiv (u_1, \ldots, u_d)$ denote a realization of $\mathbf{U}$. Given an arbitrary real-valued, square-integrable function $\varphi(\cdot)$ defined on the d-dimensional unit cube $[0, 1]^d$, we decompose $\varphi(\cdot)$ as in Stein (1987). We define the following functionals of $\varphi(\cdot)$: a) the *mean* of $\varphi(\cdot)$,

$$\mu_\varphi \equiv \mathrm{E}[\varphi(\mathbf{U})] = \int_{[0,\,1]^d} \varphi(\mathbf{u})\, d\mathbf{u}\,;$$

b) the jth *main effect* of $\varphi(\cdot)$,

$$\begin{aligned}\varphi_j(u_j) &\equiv \mathrm{E}[\varphi(\mathbf{U})|U_j = u_j]\\ &= \int_{[0,\,1]^{d-1}} \varphi(u_1, \ldots, u_j, \ldots, u_d) \prod_{\substack{\alpha = 1\\ \alpha \neq j}}^{d} du_\alpha \quad \text{for} \quad u_j \in [0,\,1] \quad \text{and} \quad j = 1, \ldots, d;\end{aligned}$$

c) the *additive part* of $\varphi(\cdot)$,

$$\varphi_{\text{add}}(\mathbf{u}) \equiv \sum_{j=1}^{d} \varphi_j(u_j) - (d-1)\mu_\varphi \quad \text{for} \quad \mathbf{u} \in [0,\,1]^d;$$

and d) the *residual from additivity* of $\varphi(\cdot)$,

$$\varphi_{\text{res}}(\mathbf{u}) \equiv \varphi(\mathbf{u}) - \varphi_{\text{add}}(\mathbf{u}) \quad \text{for} \quad \mathbf{u} \in [0,\,1]^d.$$

We observe that $\mathrm{E}[\varphi_j(U_j)] = \mathrm{E}[\varphi_{\text{add}}(\mathbf{U})] = \mu_\varphi$ for each j, and $\mathrm{E}[\varphi_{\text{res}}(\mathbf{U})] = 0$. Moreover,

$$\mathrm{E}\Big[\varphi^2_{\text{res}}(\mathbf{U})\Big] = \mathrm{Var}[\varphi_{\text{res}}(\mathbf{U})] = \mathrm{Var}[\varphi(\mathbf{U})] - \sum_{j=1}^{d} \mathrm{Var}[\varphi_j(U_j)]\,, \tag{23}$$

where the last equality follows by observing that $\mathrm{Cov}[\varphi(\mathbf{U}),\, \varphi_j(U_j)] = \mathrm{Var}[\varphi_j(U_j)]$ for each j.

Recalling the representation of the simulation response $Y = y(\mathbf{U})$ as a function of the input random vector $\mathbf{U}$, we define $\chi(\mathbf{u}) \equiv \mathbf{1}\{y(\mathbf{u}) \leq \xi\}$; and we let $\chi_j(\cdot)$, $\chi_{\text{add}}(\cdot)$, and $\chi_{\text{res}}(\cdot)$ respectively denote the jth main effect, the additive part, and the residual from additivity of $\chi(\cdot)$. The asymptotic distribution of $\widehat{\xi}_{\text{LH–SS}}(\psi_c, n)$ for $c = 1, 2$ is given by Theorem 3 below in which $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution (Billingsley 1986, pp. 338–339) and $N(\mu, \sigma^2)$ denotes a normal random variable with mean μ and variance σ^2.

Detailed proofs of the next two key results (namely, Theorem 3 and Proposition 3 below) are given in Avramidis and Wilson (1995); and only the main steps in the proof Theorem 3 are outlined here. We believe that the assumptions underlying both of these results are reasonable in practice, and in Subsection 4.2 below we explicitly validate these assumptions in our simulations of stochastic activity networks.

Theorem 3 *Suppose that the following continuity conditions hold:*

CC_1 *The function $y(\cdot)$ has a finite set of discontinuities $\mathcal{D}$.*

CC_2 *There exists a neighborhood $\mathcal{N}(\xi)$ of ξ such that for each $x \in \mathcal{N}(\xi)$ and for each $j = 1, \ldots, d$, there exists a finite set $\mathcal{Q}_j(x)$ such that*

$$\Pr\{y(\mathbf{U}) = x | \, U_j = u_j\} = 0 \quad \textit{for every } u_j \in (0,\, 1) - \mathcal{Q}_j(x). \tag{24}$$

If $F(\cdot)$ has a bounded second derivative in a neighborhood of ξ, and if $F'(\xi) \neq 0$, then

$$n^{1/2}\Big[\widehat{\xi}_{\text{LH–SS}}(\psi_c, n) - \xi\Big] \xrightarrow{\mathcal{D}} N\Big(0,\, \sigma^2_{\text{LH–SS}}\Big) \quad \textit{as } n \to \infty \textit{ for } c = 1, 2, \tag{25}$$

where

$$\sigma^2_{\text{LH–SS}} = \frac{\text{Var}[\chi_{\text{res}}(\mathbf{U})]}{[F'(\xi)]^2}\,.$$

Proof. First we prove the result for $\psi_1(\cdot)$ and then easily extend the proof for $\psi_2(\cdot)$. Let $F_n^{(\text{LH})}(\cdot)$ denote the empirical c.d.f., defined as in (2), based on a Latin hypercube sample of size n. Fix a real t. Then

$$\begin{aligned}
\gamma_n(t) &\equiv \Pr\Big\{n^{1/2}\Big[\widehat{\xi}_{\text{LH–SS}}(\psi_1, n) - \xi\Big] \leq t\Big\} \\
&= \Pr\Big\{\widehat{\xi}_{\text{LH–SS}}(\psi_1, n) \leq \xi + n^{-1/2}t\Big\} \\
&= \Pr\Big\{r \leq F_n^{(\text{LH})}(\xi + n^{-1/2}t)\Big\} \\
&= \Pr\Big\{r \leq F(\xi + n^{-1/2}t) + F_n^{(\text{LH})}(\xi + n^{-1/2}t) - F(\xi + n^{-1/2}t)\Big\}\,.
\end{aligned} \tag{26}$$

Taking a second-order Taylor expansion of $F(\cdot)$ centered at ξ and using the assumed properties that $F'(\xi) \neq 0$ and $F''(\cdot)$ is bounded in a neighborhood of ξ, we have

$$F(\xi + n^{-1/2}t) = r + n^{-1/2}tF'(\xi) + O(n^{-1}) \quad \text{as } n \to \infty. \tag{27}$$

Defining

$$\Delta_n^{(\mathrm{LH})} \equiv F_n^{(\mathrm{LH})}(\xi + n^{-1/2}t) - F(\xi + n^{-1/2}t)\,, \tag{28}$$

we complete the proof of Theorem 3 by exploiting the following key result,

$$n^{1/2}\Delta_n^{(\mathrm{LH})} \xrightarrow{\mathcal{D}} N\{0,\, \mathrm{Var}[\chi_{\mathrm{res}}(\mathbf{U})]\} \quad \text{as} \quad n \to \infty\,, \tag{29}$$

which requires the continuity conditions CC_1 and CC_2. The Appendix contains a sketch of the main steps in the proof of (29), and a complete justification of this result is provided in Lemma 6 of Avramidis and Wilson (1995). Using (27), we can rewrite (26) as

$$\gamma_n(t) = \Pr\left\{ \frac{n^{1/2}\Delta_n^{(\mathrm{LH})}}{F'(\xi)} + O(n^{-1/2}) \geq -t \right\}. \tag{30}$$

From (29), (30), and Slutsky's Theorem (Serfling 1980, p. 19), it follows that

$$\lim_{n\to\infty} \gamma_n(t) = 1 - \Phi\left\{ -tF'(\xi) \Big/ \sqrt{\mathrm{Var}[\chi_{\mathrm{res}}(\mathbf{U})]} \right\} = \Phi\left\{ tF'(\xi) \Big/ \sqrt{\mathrm{Var}[\chi_{\mathrm{res}}(\mathbf{U})]} \right\}, \tag{31}$$

where $\Phi(\cdot)$ is the standard normal c.d.f. This completes the proof of (25) for $\psi_1(\cdot)$.

To prove (25) for $\psi_2(\cdot)$, we let $\widetilde{F}_n^{(\mathrm{LH})}(\cdot)$ denote the piecewise linear version of the empirical c.d.f. that corresponds to the definition (4) of $\psi_2(\cdot)$ and that is based on the same Latin hypercube sample of size n as for $F_n^{(\mathrm{LH})}(\cdot)$. Fix a real t. We proceed along the lines of the proof of (25) for $\psi_1(\cdot)$ to obtain

$$\begin{aligned}
\widetilde{\gamma}_n(t) &\equiv \Pr\Big\{ n^{1/2}\Big[\widehat{\xi}_{\mathrm{LH-SS}}(\psi_2, n) - \xi \Big] \leq t \Big\} \\
&= \Pr\Big\{ \widehat{\xi}_{\mathrm{LH-SS}}(\psi_2, n) \leq \xi + n^{-1/2}t \Big\} \\
&= \Pr\Big\{ r \leq F(\xi + n^{-1/2}t) + \Delta_n^{(\mathrm{LH})} + \Omega_n^{(\mathrm{LH})} \Big\}\,,
\end{aligned}$$

where $\Delta_n^{(\mathrm{LH})}$ is defined in (28) and

$$\Omega_n^{(\mathrm{LH})} \equiv \widetilde{F}_n^{(\mathrm{LH})}(\xi + n^{-1/2}t) - F_n^{(\mathrm{LH})}(\xi + n^{-1/2}t)\,.$$

¿From the definitions of $F_n^{(\mathrm{LH})}(\cdot)$ and $\widetilde{F}_n^{(\mathrm{LH})}(\cdot)$, it follows immediately that

$$\Big| \widetilde{F}_n^{(\mathrm{LH})}(z) - F_n^{(\mathrm{LH})}(z) \Big| \leq 1/(2n) \quad \text{for all} \quad z;$$

and thus

$$n^{1/2}\Omega_n^{(\mathrm{LH})} \xrightarrow{\mathcal{P}} 0 \quad \text{as} \quad n \to \infty, \tag{32}$$

where $\xrightarrow{\mathcal{P}}$ denotes convergence in probability (Serfling 1980, p. 6). In this situation the analogue of (30) is

$$\widetilde{\gamma}_n(t) = \Pr\left\{ \frac{n^{1/2}\Delta_n^{(\mathrm{LH})}}{F'(\xi)} + \frac{n^{1/2}\Omega_n^{(\mathrm{LH})}}{F'(\xi)} + O(n^{-1/2}) \geq -t \right\}. \tag{33}$$

From (29), (32), (33), and Slutsky's Theorem, it follows that $\lim_{n\to\infty} \widetilde{\gamma}_n(t)$ is also equal to the right-hand side of (31). This completes the proof of (25) for $\psi_2(\cdot)$.

To compare the asymptotic performance of the Latin hypercube–single sample quantile estimators $\widehat{\xi}_{\text{LH–SS}}(\psi_c, n)$ for $c = 1, 2$ with the asymptotic performance of the direct-simulation estimators $\widehat{\xi}_{\text{DS}}(\psi_c, n)$ for $c = 1, 2$, we establish results analogous to Theorem 3 for the direct-simulation estimators. If $F(\cdot)$ is differentiable at ξ and $F'(\xi) \neq 0$, then $\widehat{\xi}_{\text{DS}}(\psi_1, n)$ is asymptotically normal:

$$n^{1/2}\Big[\widehat{\xi}_{\text{DS}}(\psi_1, n) - \xi\Big] \xrightarrow{\mathcal{D}} N\Big(0,\, \sigma^2_{\text{DS}}\Big) \quad \text{as } n \to \infty, \tag{34}$$

where

$$\sigma^2_{\text{DS}} = \frac{r(1-r)}{[F'(\xi)]^2} = \frac{\text{Var}[\chi(\mathbf{U})]}{[F'(\xi)]^2}$$

(Corollary 2.3.3.A of Serfling 1980). The following proposition establishes that $\widehat{\xi}_{\text{DS}}(\psi_2, n)$ has the same asymptotic distribution as $\widehat{\xi}_{\text{DS}}(\psi_1, n)$.

Proposition 3 *If $F(\cdot)$ is differentiable at ξ and $F'(\xi) \neq 0$, then*

$$n^{1/2}\Big[\widehat{\xi}_{\text{DS}}(\psi_2, n) - \xi\Big] \xrightarrow{\mathcal{D}} N\Big(0,\, \sigma^2_{\text{DS}}\Big) \quad \text{as} \quad n \to \infty\,. \tag{35}$$

The proof of Proposition 3, which is similar to that of Theorem 3, is given in Avramidis and Wilson (1995).

Applying (23) to the function $\chi(\cdot)$, we see that

$$\sigma^2_{\text{LH–SS}} = \text{Var}[\chi(\mathbf{U})] - \sum_{j=1}^{d} \text{Var}[\chi_j(U_j)] \le \text{Var}[\chi(\mathbf{U})] = \sigma^2_{\text{DS}}.$$

Hence Theorem 3, result (34), and Proposition 3 ensure that quantile estimators of the form $\widehat{\xi}_{\text{LH–SS}}(\psi_c, n)$ for $c = 1, 2$ are asymptotically more accurate than quantile estimators of the form $\widehat{\xi}_{\text{DS}}(\psi_c, n)$ for $c = 1, 2$. This result and Proposition 2 strongly suggest that $\widehat{\xi}_{\text{LH–SS}}(\psi_2, n)$ is superior to all of the other quantile estimators considered in this paper that are based on direct simulation or LHS. In the next section we quantify the improvements in accuracy that are achievable with the various single- and multiple-sample quantile estimators based on Latin hypercube sampling and antithetic variates in some common simulation applications.

4 APPLICATION TO STOCHASTIC ACTIVITY NETWORKS

We illustrate the application of our quantile-estimation techniques to the simulation of stochastic activity networks (SANs). In Section 4.1 we describe the simulation experiments that were performed. In Section 4.2 we validate the assumptions required to apply the main theoretical results of this paper in our activity-network simulations. In Section 4.3 we summarize the results of our Monte Carlo experiments.

4.1 Description of the Simulation Experiments

The Monte Carlo study is designed to estimate the reductions in bias, variance, and mean square error that are achieved by the proposed multiple- and single-sample quantile estimators in the context of simulating SANs. Specifically, we estimate the 5th, 25th, 50th, 75th, and 95th percentiles (quantiles) of the network completion time, i.e., the longest directed path from the source node to the sink node; and we use two SANs for the experimental performance evaluation. To facilitate our description of the simulation experiments as well as our validation of the assumptions underlying each quantile-estimation technique, we define some general notation for specifying and simulating an arbitrary SAN.

The graph-theoretic structure of a stochastic activity network is described by the pair $(\mathcal{W}, \mathcal{A})$, where the nodes in the network constitute the set $\mathcal{W} \equiv \{1, \ldots, \nu\}$ and the activities in the network constitute the set

$$\mathcal{A} \equiv \Big\{(\beta_j, \gamma_j) \;:\; \text{activity } j \text{ has start node } \beta_j \in \mathcal{W} \text{ and end node } \gamma_j \in \mathcal{W},\ j = 1, \ldots, d\Big\}.$$

The network is assumed to be acyclic with a source node and a sink node in $\mathcal{W}$. Each activity j has a random duration V_j with c.d.f. $H_j(\cdot)$, and the individual activity durations are independently sampled by inversion so that

$$V_j = H_j^{-1}(U_j) \quad \text{for} \quad j = 1, \ldots, d\,. \tag{36}$$

The objective of simulating the network is to estimate the rth quantile of the time to realize the sink node for $r = 0.05$, 0.25, 0.50, 0.75, and 0.95. Let τ denote the number of directed paths from source to sink, and let

$$\mathcal{A}(\omega) \equiv \{j \;:\; \text{activity } j \text{ is on } \omega\text{th source-to-sink path}\} \quad \text{for} \quad \omega = 1, \ldots, \tau. \tag{37}$$

The duration of the ωth path is the random variable

$$P_\omega \equiv \sum_{j \in \mathcal{A}(\omega)} V_j \qquad \text{for} \qquad \omega = 1, \ldots, \tau; \tag{38}$$

and thus the basic simulation response is the *network completion time*

$$Y \equiv \max_{1 \le \omega \le \tau} \{P_\omega\} = \max_{1 \le \omega \le \tau} \left\{ \sum_{j \in \mathcal{A}(\omega)} H_j^{-1}(U_j) \right\} \equiv y(U_1, \ldots, U_d)\,. \tag{39}$$

For the duration V_j of the jth activity in a given network, the associated distribution is taken to be either a) a normal distribution with a nominal mean μ_j and standard deviation $\sigma_j = \mu_j/4$ whose probability mass below the origin has been relocated to the origin; or b) an exponential distribution with a specified mean μ_j. Network 1 is taken from Elmaghraby (1977, p. 275), and it is depicted in Figure 1. For network 1 the set of activities with "adjusted" normal durations as in a) is taken to be

$\{(1,2),\ (1,3),\ (2,4),\ (6,9),\ (7,8)\}$; all other activities are taken to be exponentially distributed as in b). Network 2 is taken from Antill and Woodhead (1990, Figure 8.5(b), p. 180), and it is depicted in Figure 2. For network 2 the set of activities with "adjusted" normal durations is taken to be $\{(0,2),\ (1,5),\ (1,3),\ (7,10),\ (9,12),\ (11,17),\ (15,16),\ (16,20),\ (16,22),\ (16,18),\ (17,18),\ (22,23)\}$. In Figures 1 and 2, the mean duration is shown next to each nontrivial activity.

The following quantile estimators are evaluated for $c = 1$ and 2, respectively: the conventional direct-simulation estimator $\widehat{\xi}_{\mathrm{DS}}(\psi_c, n)$; the antithetic variates–multiple sample estimator $\widehat{\xi}_{\mathrm{CI–MS}}\left(\psi_c, G_{\mathrm{AV}}^{(2)}, n\right)$; the antithetic variates–single sample estimator $\widehat{\xi}_{\mathrm{AV–SS}}(\psi_c,\ n)$ that results from applying $\psi_c(\cdot)$ to a sample of size n consisting of $n/2$ independent pairs of antithetic responses; the Latin hypercube–multiple sample estimator $\widehat{\xi}_{\mathrm{CI–MS}}\left(\psi_c, G_{\mathrm{LH}}^{(k)}, n\right)$ for a set of selected values of k; and the Latin hypercube–single sample estimator $\widehat{\xi}_{\mathrm{LH–SS}}\left(\psi_c, G_{\mathrm{LH}}^{(n)}\right)$.

4.2 Validation of the Quantile-Estimation Procedures

The key assumptions in the development of Sections 1–3 are the monotonicity of $y(\cdot)$ together with requirements RC_1–RC_4, CC_1, and CC_2; and in the context of simulating networks 1 and 2, we validate all these assumptions along with the standard requirement for estimating the quantile ξ that $F'(\xi) \neq 0$ and $F''(\cdot)$ is bounded in a neighborhood of ξ. Since the network completion time (39) is a nondecreasing function of the arc durations $\{V_j\}$ and each V_j is a nondecreasing function (36) of the corresponding random number U_j, it follows that $y(U_1, \ldots, U_d)$ is a monotone function of each of its arguments individually as required by Theorems 1 and 2. For the sake of clarity and simplicity, we limit most of the following discussion on validating requirements RC_1–RC_4, CC_1, and CC_2 to consideration of network 1; but for each requirement to be checked, the corresponding analysis also applies to network 2.

We check the regularity requirements RC_2–RC_4 as follows. In each network, we can choose a *uniformly directed cutset* $\mathcal{L}$—that is, a set of activities such that each directed path from source to sink contains exactly one activity in $\mathcal{L}$ (Sigal, Pritsker, and Solberg 1980); and then we obtain an integral expression for $F(\cdot)$ by applying the law of total probability and conditioning on all activity durations *except* for the activities in $\mathcal{L}$. To exploit this representation of $F(\cdot)$, we require the following additional properties of the network:

i) For each activity $g \in \mathcal{L}$, the first three derivatives $H_g'(v)$, $H_g''(v)$, and $H_g'''(v)$ of the activity-time c.d.f. $H_g(v)$ are bounded and continuous for all $v > 0$.

ii) For each activity j $(1 \leq j \leq d)$, the first derivative $H_j'(v) > 0$ for all $v > 0$.

iii) For each activity $j \notin \mathcal{L}$, the activity-time c.d.f. $H_j(\cdot)$ has at most a single discontinuity at zero.

In network 1, we take $\mathcal{L} = \{(3,6),\ (2,6),\ (4,5),\ (4,7)\}$; and in network 2, we take $\mathcal{L} = \{(2,9),\ (4,7),\ (5,7),\ (1,6),\ (3,6),\ (3,8)\}$. Since all nontrivial activity durations have "adjusted" normal or exponential distributions, clearly networks 1 and 2 possess properties i) through iii).

For every activity j in network 1 ($1 \le j \le d$), we define $\mathcal{Z}(j) \equiv \{\omega : 1 \le \omega \le \tau \text{ and } j \in \mathcal{A}(\omega)\}$ to be the collection of indexes of source-to-sink paths that contain activity j. Let

$$\{j_c : c = 1, \ldots, \alpha\} \equiv \{1, \ldots, d\} - \mathcal{L}$$

be an enumeration of all activities that are not in the cutset $\mathcal{L}$. Using this notation, we can express the target c.d.f. $F(x)$ for each $x > 0$ as

$$\begin{aligned} F(x) &= \underset{\substack{-\infty < v_{j_c} < \infty \\ c = 1, \ldots, \alpha}}{\int \cdots \int} \Pr\Big\{ Y \le x \,\Big|\, V_{j_c} = v_{j_c} \text{ for } c = 1, \ldots, \alpha \Big\} \prod_{c=1}^{\alpha} dH_{j_c}(v_{j_c}) \\ &= \int_{v_{j_1}=0}^{\lambda_1(x)} \cdots \int_{v_{j_\alpha}=0}^{\lambda_\alpha(x)} \prod_{g \in \mathcal{L}} H_g\left[x - \max_{\omega \in \mathcal{Z}(g)} \left\{ \sum_{j_c \,\in\, \mathcal{A}(\omega)} v_{j_c} \right\} \right] dH_{j_\alpha}(v_{j_\alpha}) \cdots dH_{j_1}(v_{j_1}), \end{aligned} \tag{40}$$

where for $c = 1, \ldots, \alpha$, the upper limit $\lambda_c(x)$ of the cth outermost iterated integral on the right-hand side of (40) is defined as follows:

$$\lambda_1(x) \equiv x \quad \text{and} \quad \lambda_c(x) \equiv x - \max_{\omega \in \mathcal{Z}(j_c)} \left\{ \sum_{\substack{j_s \,\in\, \mathcal{A}(\omega) \\ 1 \le s \le c-1}} v_{j_s} \right\} \quad \text{for} \quad c = 2, \ldots, \alpha. \tag{41}$$

Property i) of network 1 ensures that both the integrand (40) and the limits of integration in (41) have first-, second-, and third-order derivatives with respect to x that are bounded and continuous for all $x > 0$. We can therefore apply the chain rule for partial differentiation to the right-hand side of (40) three times to conclude that the first, second, and third derivatives (that is, $F'(x)$, $F''(x)$, and $F'''(x)$, respectively) of the target c.d.f. are bounded, continuous functions of x for all $x > 0$. Applying property ii) of network 1 to the integral expression that results from taking the first derivative with respect to x of the right-hand side of (40), we see that $F'(x) > 0$ for all $x > 0$. We have assumed throughout the paper that $r > 0$, which implies $\xi > 0$. Since $Q'(r) = 1/F'(\xi)$, condition RC_3 is satisfied; moreover it follows from the existence of $F''(\cdot)$ and $F'''(\cdot)$ in a neighborhood of ξ together with the chain rule for ordinary differentiation that $Q''(\cdot)$ and $Q'''(\cdot)$ are bounded and continuous in a neighborhood of r so that conditions RC_2 and RC_4 are also satisfied. Notice that as a by-product of this argument, we also validated the requirement that $F'(\xi) \ne 0$ and $F''(\cdot)$ is bounded in a neighborhood of ξ.

The following auxiliary result provides a technique for validating regularity requirement RC_1 that should be useful in applications of the proposed quantile-estimation techniques to more general stochastic simulations. The proof of this lemma is given in Avramidis and Wilson (1995).

Lemma 2 *If the random variable X has c.d.f. $F_X(\cdot)$ and inverse c.d.f. $Q_X(\cdot)$ each with a finite number of discontinuities and if* $\mathrm{E}[|X|] < \infty$, *then $u(1-u)Q_X(u)$ is bounded for all $u \in (0, 1)$ so that* RC_1 *is satisfied for the inverse c.d.f. $Q_X(\cdot)$ with $a = b = 1$.*

We show that in network 1, the completion time Y satisfies the conditions of Lemma 2. From the discussion following (41), we have already seen that $F'(x) > 0$ for all $x > 0$; and it follows that $F(\cdot)$ has at most a single discontinuity at zero while $Q(\cdot)$ is continuous on (0, 1). Moreover, the response Y has a finite mean since $\mathrm{E}[Y] \leq \sum_{j=1}^{d} \mu_j < \infty$ so that Lemma 2 applies to Y.

Finally we consider the applicability to network 1 of the continuity requirements CC_1 and CC_2. Since the network completion time (39) is a continuous function of the arc durations $\{V_j\}$ and each V_j is a continuous function (36) of the corresponding random number U_j, it follows that $y(U_1, \ldots, U_d)$ is a continuous function of $(U_1, \ldots, U_d)$ and the requirement CC_1 is satisfied with $\mathcal{D} = \emptyset$.

To check the continuity requirement CC_2 in network 1, we select an arbitrary positive cutoff value x. Given the jth input random number $U_j = u_j$ $(1 \leq j \leq d)$, we see by properties i) and iii) of network 1 and by Lemma 1 in Section V.4 of Feller (1971) that the conditional c.d.f. of the ωth path duration P_ω $(1 \leq \omega \leq \tau)$ has at most the following single discontinuity:

a. at zero, if $j \notin \mathcal{A}(\omega)$ and $H_c(\cdot)$ has a discontinuity at zero for each $c \in \mathcal{A}(\omega)$; or

b. at $H_j^{-1}(u_j)$, if $j \in \mathcal{A}(\omega)$ and $H_c(\cdot)$ has a discontinuity at zero for each $c \in \mathcal{A}(\omega) - \{j\}$.

It follows that if $x > 0$ and we take $\mathcal{Q}_j(x)$ to be the one-element set $\{H_j(x)\}$ in condition CC_2, then for the ωth path in network 1 $(1 \leq \omega \leq \tau)$, the conditional c.d.f. $F_{P_\omega|U_j}(\cdot \,|\, u_j)$ of the path duration P_ω given $U_j = u_j \in [0, 1] - \mathcal{Q}_j(x)$ is continuous at x. Moreover, we have

$$\begin{aligned} \Pr\{y(\mathbf{U}) = x |\, U_j = u_j\} &= \lim_{z \to x-} \Pr\Big\{ z < Y \leq x \,\Big|\, U_j = u_j \Big\} \\ &\leq \lim_{z \to x-} \sum_{\omega=1}^{\tau} \Pr\Big\{ z < P_\omega \leq x \,\Big|\, U_j = u_j \Big\} \\ &= \sum_{\omega=1}^{\tau} \Big\{ \lim_{z \to x-} \Big[F_{P_\omega|U_j}(x \,|\, u_j) - F_{P_\omega|U_j}(z \,|\, u_j) \Big] \Big\} = 0, \end{aligned} \tag{42}$$

which establishes the continuity requirement CC_2 with $\mathcal{N}(\xi) = (0, \infty)$ and $\mathcal{Q}_j(x) = \{H_j(x)\}$.

To summarize, we validated all the assumptions underlying Propositions 2–3 and Theorems 1–3 for networks 1 and 2; and it is clear that the main theoretical results of this paper can be applied to a large class of SANs that includes our networks. Development of general techniques for validating these assumptions in a broader class of stochastic simulations is the subject of ongoing work.

4.3 Experimental Results

We begin by presenting some experimental evidence to complement the analysis given after Proposition 2 for the claim that $\widehat{\xi}_{\mathrm{DS}}(\psi_2, n)$ has better bias properties than $\widehat{\xi}_{\mathrm{DS}}(\psi_1, n)$. For networks 1 and 2, the true values of the selected quantiles were estimated by direct simulation using 2,048 independent *macroreplications* of a simulation experiment involving each network; and each simulation experiment consisted of $n = 131,072$ independent replications of the associated network. Thus we took for the "true" value of a selected quantile the grand average of 2,048 independent replications of the corresponding direct-simulation estimator $\widehat{\xi}_{\mathrm{DS}}(\psi_2,\, 131{,}072)$. Figure 3 depicts plots of the estimated bias squared, variance, and MSE of the two direct-simulation quantile estimators when the 95th percentile of the network completion time is to be estimated so that $r = 0.95$. (Each entry used for generating these plots was estimated with a relative error whose magnitude did not exceed 5%.) The erratic behavior of $\widehat{\xi}_{\mathrm{DS}}(\psi_1, n)$ stands in sharp contrast to the smooth behavior of $\widehat{\xi}_{\mathrm{DS}}(\psi_2, n)$. Avramidis and Wilson (1995) contains similar plots of the estimated bias squared, variance, and MSE of the two direct-simulation quantile estimators for $r = 0.05$ and $r = 0.50$.

Based on the analytical results presented in Subsection 2.2, the Monte Carlo results presented here, and additional computational experience not reported here, we conclude that $\widehat{\xi}_{\mathrm{DS}}(\psi_2, n)$ should be preferred over $\widehat{\xi}_{\mathrm{DS}}(\psi_1, n)$—particularly when squared bias is expected to contribute significantly to MSE. As discussed earlier, the significance of bias becomes more important as n becomes smaller or as r approaches 0 or 1. Consequently, we used the function $\psi_2(\cdot)$ for all estimators discussed in the rest of this section.

For the simulation experiments involving network 1, Tables I and II display estimated ratios of the form $\mathrm{MSE}\big[\,\widehat{\xi}_{\mathrm{DS}}\,\big]\big/\mathrm{MSE}\big[\,\widehat{\xi}_{\mathrm{CI}}\,\big]$ between the MSE of the direct-simulation estimator $\widehat{\xi}_{\mathrm{DS}}$ and the MSE of the correlation-induction estimator $\widehat{\xi}_{\mathrm{CI}}$ for several choices of $\widehat{\xi}_{\mathrm{CI}}$ and for the sample sizes $n = 2{,}048$ and $n = 8{,}192$, respectively. For the simulation experiments involving network 2, Tables III and IV display the corresponding results for the sample sizes $n = 2{,}048$ and $n = 8{,}192$, respectively. Immediately below each estimated MSE ratio is its estimated standard error.

As expected from the results of Section 2, the multiple-sample estimators based on the methods of antithetic variates and Latin hypercube sampling achieve reductions in mean square error when compared to the direct-simulation estimators. In almost all cases, the Latin hypercube–multiple sample estimator $\widehat{\xi}_{\mathrm{CI\text{–}MS}}\Big(\psi_2, G_{\mathrm{LH}}^{(k)}, n\Big)$ with $k > 2$ performed significantly better than the single- and multiple-sample antithetic-variates estimators. This behavior was also observed in several other experiments (not reported here); thus we recommend the Latin hypercube estimator $\widehat{\xi}_{\mathrm{CI\text{–}MS}}\Big(\psi_2, G_{\mathrm{LH}}^{(k)}, n\Big)$ with $k > 2$ rather than any of the quantile estimators based on the method of antithetic variates. To use a Latin hypercube–multiple sample estimator, a practitioner would probably have to choose k given the total number of replications n. For fixed n, variance is typically decreasing in k (because of the more complete stratification), while bias is typically increasing in k (because of the smaller subsample size $m = n/k$). The net effect is that $\mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{CI\text{–}MS}}\Big(\psi_2, G_{\mathrm{LH}}^{(k)}, n\Big)\Big]$ is typically decreasing for $k \le k_0$ and increasing for $k \ge k_0$, with the critical value k_0 being an increasing function of n. For example, in Table IV we see that $k_0 \approx 64$ for $r = 0.50$.

The Latin hypercube–single sample quantile estimator $\widehat{\xi}_{\mathrm{LH\text{–}SS}}(\psi_2, n)$ achieved substantial MSE reductions relative to the direct-simulation estimator $\widehat{\xi}_{\mathrm{DS}}(\psi_2, n)$. In particular, $\widehat{\xi}_{\mathrm{LH\text{–}SS}}(\psi_2, n)$ achieved the largest MSE reductions of all estimators considered here for the case $r = 0.95$. Although $\widehat{\xi}_{\mathrm{LH\text{–}SS}}(\psi_2, n)$ was outperformed by the Latin hypercube–multiple sample estimators or the antithetic-variates estimators in some cases for which $r \le 0.75$, the overall performance of

Table I

Estimated Ratio $\mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_2,\, 2048)\Big] \Big/ \mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{CI}}(\psi_2,\, \cdot\,,\, 2{,}048)\Big]$ ± Its Estimated Standard Error for Various Order-r Quantile Estimators $\widehat{\xi}_{\mathrm{CI}}(\psi_2,\, \cdot\,,\, 2{,}048)$ in Network 1

Estimator	Order r				
	0.05	0.25	0.50	0.75	0.95
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(2)}_{\mathrm{AV}},\, 2{,}048)$	1.096 ±0.018	1.261 ±0.020	1.211 ±0.019	1.082 ±0.016	1.013 ±0.015
$\widehat{\xi}_{\mathrm{AV-SS}}(\psi_2,\, 2{,}048)$	1.099 ±0.018	1.265 ±0.020	1.212 ±0.018	1.075 ±0.016	1.009 ±0.015
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(2)}_{\mathrm{LH}},\, 2{,}048)$	1.068 ±0.016	1.263 ±0.023	1.238 ±0.018	1.123 0.019	1.010 0.014
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(4)}_{\mathrm{LH}},\, 2{,}048)$	1.178 ±0.019	1.523 ±0.026	1.665 ±0.024	1.311 ±0.023	1.033 ±0.016
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(8)}_{\mathrm{LH}},\, 2{,}048)$	1.231 ±0.021	1.704 ±0.025	2.091 ±0.033	1.849 ±0.031	1.146 ±0.018
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(16)}_{\mathrm{LH}},\, 2{,}048)$	1.203 ±0.018	1.810 ±0.030	2.446 ±0.033	2.498 ±0.040	1.332 ±0.021
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(32)}_{\mathrm{LH}},\, 2{,}048)$	1.060 ±0.018	1.752 ±0.028	2.517 ±0.035	3.093 ±0.050	1.830 ±0.026
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(64)}_{\mathrm{LH}},\, 2{,}048)$	0.533 ±0.008	1.367 ±0.021	2.256 ±0.034	2.974 ±0.049	2.261 ±0.035
$\widehat{\xi}_{\mathrm{LH-SS}}(\psi_2,\, 2{,}048)$	1.162 ±0.018	1.848 ±0.029	2.664 ±0.038	3.440 ±0.057	3.827 ±0.057

Table II

Estimated Ratio $\mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_2,\, 8{,}192)\Big] \Big/ \mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{CI}}(\psi_2,\, \cdot\,,\, 8{,}192)\Big]$ ± Its Estimated Standard Error for Various Order-r Quantile Estimators $\widehat{\xi}_{\mathrm{CI}}(\psi_2,\, \cdot\,,\, 8{,}192)$ in Network 1

Estimator	Order r				
	0.05	0.25	0.50	0.75	0.95
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(2)}_{\mathrm{AV}},\, 8{,}192)$	1.085 ±0.029	1.290 ±0.044	1.162 ±0.033	1.090 ±0.034	1.044 ±0.033
$\widehat{\xi}_{\mathrm{AV-SS}}(\psi_2,\, 8{,}192)$	1.074 ±0.029	1.293 ±0.043	1.170 ±0.032	1.088 ±0.034	1.054 ±0.035
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(2)}_{\mathrm{LH}},\, 8{,}192)$	1.054 ±0.031	1.280 ±0.042	1.227 ±0.039	1.143 ±0.033	1.011 ±0.030
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(4)}_{\mathrm{LH}},\, 8{,}192)$	1.183 ±0.037	1.514 ±0.055	1.694 ±0.058	1.377 ±0.044	1.052 ±0.036
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(8)}_{\mathrm{LH}},\, 8{,}192)$	1.218 ±0.043	1.727 ±0.057	2.036 ±0.066	1.826 ±0.061	1.100 ±0.031
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(16)}_{\mathrm{LH}},\, 8{,}192)$	1.202 ±0.033	1.807 ±0.060	2.435 ±0.075	2.472 ±0.079	1.318 ±0.040
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(32)}_{\mathrm{LH}},\, 8{,}192)$	1.174 ±0.044	1.926 ±0.068	2.506 ±0.084	3.075 ±0.094	1.784 ±0.060
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(64)}_{\mathrm{LH}},\, 8{,}192)$	0.960 ±0.029	1.896 ±0.062	2.698 ±0.090	3.145 ±0.088	2.304 ±0.084
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(128)}_{\mathrm{LH}},\, 8{,}192)$	0.601 ±0.017	1.504 ±0.051	2.413 ±0.077	3.340 ±0.109	2.777 ±0.093
$\widehat{\xi}_{\mathrm{LH-SS}}(\psi_2,\, 8{,}192)$	0.966 ±0.027	1.725 ±0.051	2.578 ±0.080	3.470 ±0.129	3.873 ±0.132

Table III

Estimated Ratio $\mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_2,\, 2{,}048)\Big] \Big/ \mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{CI}}(\psi_2,\, \cdot,\, 2{,}048)\Big]$ ± Its Estimated Standard Error for Various Order-r Quantile Estimators $\widehat{\xi}_{\mathrm{CI}}(\psi_2,\, \cdot,\, 2{,}048)$ in Network 2

Estimator	Order r				
	0.05	0.25	0.50	0.75	0.95
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(2)}_{\mathrm{AV}},\, 2{,}048)$	1.065 ±0.017	1.199 ±0.019	1.145 ±0.017	1.090 ±0.017	1.036 ±0.016
$\widehat{\xi}_{\mathrm{AV-SS}}(\psi_2,\, 2{,}048)$	1.055 ±0.016	1.198 ±0.019	1.145 ±0.017	1.088 ±0.017	1.031 ±0.016
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(2)}_{\mathrm{LH}},\, 2{,}048)$	1.071 ±0.016	1.205 ±0.018	1.162 ±0.018	1.086 ±0.018	1.014 ±0.017
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(4)}_{\mathrm{LH}},\, 2{,}048)$	1.147 ±0.016	1.405 ±0.019	1.431 ±0.022	1.270 ±0.021	1.095 ±0.018
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(8)}_{\mathrm{LH}},\, 2{,}048)$	1.202 ±0.021	1.604 ±0.025	1.809 ±0.026	1.546 ±0.023	1.173 ±0.018
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(16)}_{\mathrm{LH}},\, 2{,}048)$	1.187 ±0.018	1.637 ±0.025	2.058 ±0.031	2.021 ±0.034	1.340 ±0.022
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(32)}_{\mathrm{LH}},\, 2{,}048)$	1.168 ±0.019	1.683 ±0.028	2.207 ±0.034	2.374 ±0.041	1.637 ±0.025
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(64)}_{\mathrm{LH}},\, 2{,}048)$	0.690 ±0.009	1.397 ±0.020	2.035 ±0.030	2.383 ±0.039	1.844 ±0.032
$\widehat{\xi}_{\mathrm{LH-SS}}(\psi_2,\, 2{,}048)$	1.172 ±0.016	1.766 ±0.028	2.253 ±0.029	2.584 ±0.040	2.346 ±0.029

Table IV

Estimated Ratio $\mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_2,\, 8{,}192)\Big] \Big/ \mathrm{MSE}\Big[\widehat{\xi}_{\mathrm{CI}}(\psi_2,\, \cdot,\, 8{,}192)\Big]$ ± Its Estimated Standard Error for Various Order-r Quantile Estimators $\widehat{\xi}_{\mathrm{CI}}(\psi_2,\, \cdot,\, 8{,}192)$ in Network 2

Estimator	Order r				
	0.05	0.25	0.50	0.75	0.95
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(2)}_{\mathrm{AV}},\, 8{,}192)$	1.092 ±0.039	1.258 ±0.044	1.127 ±0.032	1.114 ±0.035	1.111 ±0.046
$\widehat{\xi}_{\mathrm{AV-SS}}(\psi_2,\, 8{,}192)$	1.099 ±0.039	1.250 ±0.043	1.117 ±0.030	1.114 ±0.034	1.090 ±0.043
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(2)}_{\mathrm{LH}},\, 8{,}192)$	1.098 ±0.034	1.295 ±0.042	1.231 ±0.049	1.092 ±0.033	1.074 ±0.038
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(4)}_{\mathrm{LH}},\, 8{,}192)$	1.165 ±0.041	1.446 ±0.046	1.414 ±0.043	1.258 ±0.040	1.133 ±0.035
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(8)}_{\mathrm{LH}},\, 8{,}192)$	1.223 ±0.042	1.717 ±0.051	1.846 ±0.059	1.572 ±0.047	1.225 ±0.044
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(16)}_{\mathrm{LH}},\, 8{,}192)$	1.209 ±0.037	1.825 ±0.053	2.107 ±0.063	2.048 ±0.060	1.342 ±0.041
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(32)}_{\mathrm{LH}},\, 8{,}192)$	1.263 ±0.046	1.838 ±0.060	2.177 ±0.068	2.300 ±0.089	1.620 ±0.055
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(64)}_{\mathrm{LH}},\, 8{,}192)$	1.058 ±0.041	1.794 ±0.050	2.307 ±0.062	2.526 ±0.082	1.808 ±0.056
$\widehat{\xi}_{\mathrm{CI-MS}}(\psi_2,\, G^{(128)}_{\mathrm{LH}},\, 8{,}192)$	0.818 ±0.028	1.542 ±0.046	2.270 ±0.083	2.596 ±0.082	1.876 ±0.060
$\widehat{\xi}_{\mathrm{LH-SS}}(\psi_2,\, 8{,}192)$	1.239 ±0.050	1.760 ±0.054	2.161 ±0.076	2.514 ±0.084	2.468 ±0.085

$\widehat{\xi}_{\text{LH–SS}}(\psi_2, n)$ was superior. Moreover, the asymptotic distribution of $\widehat{\xi}_{\text{LH–SS}}(\psi_2, n)$ is given by Theorem 3, and we derived an exact expression for the corresponding asymptotic variance parameter $\sigma^2_{\text{LH–SS}}$ that clearly reveals the mechanism for achieving efficiency increases via Latin hypercube sampling. In contrast, the asymptotic distribution of correlation induction–multiple sample estimators is unknown—even for the special case of the Latin hypercube–multiple sample estimator. Based on these considerations, we strongly recommend $\widehat{\xi}_{\text{LH–SS}}(\psi_2, n)$ of all the estimators considered in this paper.

5 CONCLUSIONS AND RECOMMENDATIONS

Both the theoretical and experimental results presented in this paper provide substantial evidence that some of the proposed correlation-induction techniques for estimating quantiles can yield worthwhile improvements in estimator accuracy relative to direct simulation. In particular, the Latin hypercube–single sample estimator appears to be effective for estimating the upper extreme quantiles of the network completion time of a stochastic activity network.

Although several issues require follow-up investigation, perhaps the most urgent need is for a more extensive experimental evaluation of the proposed quantile estimators. In particular, it is unclear whether the efficiency improvements observed for the Latin hypercube–single sample quantile estimator are typical of the gains that can be anticipated in practice. In the spirit of Avramidis, Bauer, and Wilson (1991) for example, a comprehensive experimental evaluation is required for the correlation-induction quantile estimators developed in this paper. Special emphasis should be given to a study of the robustness of the proposed quantile estimators against violations of the assumptions underlying the main theoretical results of this work (namely, Theorems 1, 2, and 3 together with Propositions 2 and 3).

Follow-up methodological work is required in the following areas:

1. Extension of the theoretical development to cover a larger class of simulation experiments, including formulation and justification of assumptions that yield the main results but are simpler and less restrictive than the current assumptions (namely, RC_1–RC_4, CC_1, and CC_2); and

2. Development of convenient methods for checking (validating) the assumptions underlying the main results.

Although our development is limited to simulations for which the dimension d of the vector of random-number inputs is fixed, we believe that much of this development can ultimately be extended to simulations where d is random. Such a complication naturally arises in the following situations: a) a finite-horizon simulation involving, for example, the acceptance-rejection method for generating random variates; and b) an infinite-horizon simulation potentially involving the generation of an unlimited number of random variates. Moreover, we believe that all of our results can be extended to multiresponse simulations.

In light of the demonstrated effectiveness of Latin hypercube sampling (LHS), we believe that emphasis should be given to this technique in future research. The asymptotic distribution should be established for the multiple-sample estimators $\widehat{\xi}_{\text{CI–MS}}\Big(\psi_c, G_{\text{LH}}^{(k)}, n\Big)$ with $c = 1, 2$ as $n, k \to \infty$. It would also be highly desirable to have an analogue of LHS for infinite-horizon simulations. Another direction along which LHS can be generalized is to stratify the marginal distributions of subvectors of the vector of input random numbers, where the dimension of the subvectors is higher than one (Owen 1992b). Finally, practical methods should be developed for constructing asymptotically valid confidence regions for a vector of selected quantiles under Latin hypercube sampling.

APPENDIX

Proof of Lemma 1

The proof of Lemma 1 requires the following auxiliary result.

Lemma 3 *Let $\{A_{(1)}, \ldots, A_{(n)}\}$ denote the order statistics of a random sample of size n from a distribution having inverse c.d.f. $Q(\cdot)$. Let $\{i_n\}$ denote a sequence of positive integers such that $i_n/n = r + O(1/n)$, where $r \in (0,\, 1)$. If conditions* RC_1 *and* RC_2 *hold for the function $Q(\cdot)$, then*

$$\mathrm{E}\Big[A_{(i_n)}\Big] = Q\Big(\frac{i_n}{n+1}\Big) + Q''\Big(\frac{i_n}{n+1}\Big)\frac{\mu_2(i_n, n)}{2} + o(n^{-1}), \tag{43}$$

and for every even positive integer w,

$$\mathrm{E}\Big[\Big\{A_{(i_n)} - \mathrm{E}\Big[A_{(i_n)}\Big]\Big\}^w\Big] = \Big[Q'\Big(\frac{i_n}{n+1}\Big)\Big]^w \mu_w(i_n, n) + o\Big(n^{-(w+1)/2}\Big)\ , \tag{44}$$

where

$$\mu_\alpha(\ell, n) \equiv \int_0^1 \Big(u - \frac{\ell}{n+1}\Big)^\alpha h_{\ell,n}(u)\, du \quad \textit{and} \quad h_{\ell,n}(u) \equiv \frac{n!}{(\ell-1)!(n-\ell)!} u^{\ell-1}(1-u)^{n-\ell}$$

respectively denote the αth central moment $(\alpha = 2, 3, \ldots)$ and the probability density function (p.d.f.) of the beta distribution with shape parameters ℓ and $n - \ell + 1$.

Proof. If condition RC_1 holds and if condition RC_2 holds with $\mathcal{S} = (0,\, 1)$, then (43) and (44) are the conclusions of Lemmas 3.2.2 and 3.2.3 of van Zwet (1964), respectively. To handle situations in which condition RC_2 holds with $\mathcal{S} \neq (0,\, 1)$, we modify van Zwet's arguments; and since the modifications are very similar for the two results, we provide full details for (43) while merely outlining the main steps in the proof of (44). Let

$$\varepsilon_n \equiv \max\left\{\left|\frac{i_n}{n+1} - r\right|,\ 2\left|\frac{i_n - a}{n-a-b+1} - r\right|,\ \frac{1}{\log(n+1)}\right\} \quad \text{for} \quad n \geq a + b, \tag{45}$$

where a and b are the nonnegative integer constants of condition RC_1. Clearly $\lim_{n\to\infty} \varepsilon_n = 0$; and since $r \in \mathcal{S}$, we can find a sufficiently large integer N such that

$$[r - \varepsilon_n,\, r + \varepsilon_n] \subset \mathcal{S}, \quad i_n > a, \quad \text{and} \quad n - i_n > b \quad \text{for all} \quad n \geq N, \tag{46}$$

where $\mathcal{S}$ is the subinterval containing r in condition RC_2.

We establish first that the expectation on the left-hand side of (43) is finite for all $n \geq N$. From Stirling's formula (Billingsley 1986, p. 246), we see that there is a finite constant M_1 such that

$$\frac{n!}{(i_n - 1)!(n - i_n)!} \cdot \frac{(i_n - a - 1)!(n - i_n - b)!}{(n - a - b)!} \leq M_1 \quad \text{for all} \quad n \geq N. \tag{47}$$

Condition RC_1 and relation (47) imply there is a finite constant M_2 such that

$$\begin{aligned} &|Q(u)h_{i_n,n}(u)| \\ &\quad = \left[\frac{n!}{(i_n - 1)!(n - i_n)!} \cdot \frac{(i_n - a - 1)!(n - i_n - b)!}{(n - a - b)!}\right] \cdot \left|Q(u)u^a(1-u)^b\right| \cdot h_{i_n - a, n-a-b}(u) \\ &\quad \leq M_2 h_{i_n-a,n-a-b}(u) \quad \text{for all} \quad u \in [0,\, 1] \quad \text{and} \quad n \geq N. \end{aligned} \tag{48}$$

Representing the order statistic $A_{(i_n)}$ according to the relation $A_{(i_n)} = Q\Big[U_{(i_n)}\Big]$, where $U_{(i_n)}$ is the corresponding order statistic for a sample of n uniform random numbers so that $U_{(i_n)}$ has p.d.f. $h_{i_n,n}(\cdot)$, we see that

$$\mathrm{E}\Big[A_{(i_n)}\Big] = \int_0^1 Q(u)h_{i_n,n}(u)\, du; \tag{49}$$

and since the definition (46) implies that $i_n - a > 0$ and $n - i_n - b > 0$ for all $n \geq N$, display (48) ensures that $\mathrm{E}\Big[A_{(i_n)}\Big]$ exists for all $n \geq N$.

Starting from (49) and taking

$$\Upsilon_n(u) \equiv Q(u) - Q\Big(\frac{i_n}{n+1}\Big) - Q'\Big(\frac{i_n}{n+1}\Big)\Big(u - \frac{i_n}{n+1}\Big) - \frac{1}{2}Q''\Big(\frac{i_n}{n+1}\Big)\Big(u - \frac{i_n}{n+1}\Big)^2 \tag{50}$$

for all $u \in [0,\, 1]$ and $n \geq N$, we see that

$$\int_0^1 \Upsilon_n(u)\, h_{i_n,n}(u)\, du = \mathrm{E}\Big[A_{(i_n)}\Big] - Q\Big(\frac{i_n}{n+1}\Big) - Q''\Big(\frac{i_n}{n+1}\Big)\frac{\mu_2(i_n, n)}{2}\,. \tag{51}$$

To establish (43), we will show that the left-hand side of (51) is $o(n^{-1})$. This is accomplished by decomposing the left-hand side of (51) into two integrals over the subregions $\{u : 0 \leq u \leq 1 \text{ and } |u - r| \leq \varepsilon_n\}$ and $\{u : 0 \leq u \leq 1 \text{ and } |u - r| > \varepsilon_n\}$, respectively, for $n \geq N$, where the constraint $0 \leq u \leq 1$ is implicit in all subsequent integrals involving beta p.d.f.'s. ¿From the definition (45) of ε_n, (47), and (48), we see that there is a finite constant M_3 such that

$$\left| \int\limits_{|u-r|>\varepsilon_n} \Upsilon_n(u)\, h_{i_n,n}(u)\, du \right| \leq \int\limits_{\left|u - \frac{i_n - a}{n-a-b+1}\right| > \frac{1}{2}\varepsilon_n} M_3 h_{i_n-a,n-a-b}(u)\, du \tag{52}$$

$$\leq \quad M_3 \frac{\mu_4(i_n - a, n - a - b)}{(\varepsilon_n/2)^4} \tag{53}$$

$$= \quad O\left\{\mu_4(i_n - a, n - a - b)\Big[\log(n+1)\Big]^4\right\} \tag{54}$$

$$= \quad O\left\{\frac{1}{n^2}\Big[\log(n+1)\Big]^4\right\} = o(n^{-1}) \text{ for all } n \geq N. \tag{55}$$

The Markov inequality (Billingsley 1986, p. 74) yields (53), and the definition (45) yields (54). Relation (55) follows from Lemma 3.2.1 of van Zwet (1964) or from standard formulas for the variance and fourth central moment of a beta distribution with shape parameters $i_n - a$ and $n - b - i_n + 1$; see Hahn and Shapiro (1967, p. 128).

Next we consider the other component $\int_{|u-r|\leq\varepsilon_n} \Upsilon_n(u)\, h_{i_n,n}(u)\, du$ of the integral on the left-hand side of (51). By (46) and condition RC_2, we see that for each $n \geq N$ and for each $u \in [r - \varepsilon_n,\, r + \varepsilon_n]$, we can take a second-order Taylor expansion of $Q(u)$ centered at $i_n/(n+1)$ with remainder of the form $\frac{1}{2}Q''[z(u,n)][u - i_n/(n+1)]^2$, where $z(u,n)$ is a point between u and $i_n/(n+1)$. Combining the definition (50) of $\Upsilon_n(\cdot)$ with this Taylor expansion of $Q(\cdot)$, we obtain

$$\left| \int_{|u-r|\leq\varepsilon_n} \Upsilon_n(u)\, h_{i_n,n}(u)\, du \right|$$

$$= \left| \int_{|u-r|\leq\varepsilon_n} \frac{1}{2}\left\{Q''[z(u,n)] - Q''\!\left(\frac{i_n}{n+1}\right)\right\}\left(u - \frac{i_n}{n+1}\right)^2 h_{i_n,n}(u)\, du \right| \tag{56}$$

$$\leq \left[\max_{r-\varepsilon_n\leq u_1,u_2\leq r+\varepsilon_n} \Big|Q''(u_1) - Q''(u_2)\Big|\right] \frac{\mu_2(i_n,n)}{2} = o(n^{-1}) \text{ for all } n \geq N \tag{57}$$

since $Q''(\cdot)$ is continuous on $[r - \varepsilon_n,\, r + \varepsilon_n]$ for $n \geq N$. Combining (51), (55), and (57), we obtain (43).

The main steps in the proof of (44) are similar to the main steps in the proof of (43). Given a fixed positive even integer w, we define

$$\varepsilon_n^* \equiv \max\left\{\left|\frac{i_n}{n+1} - r\right|, 2\left|\frac{i_n - wa}{n - wa - wb + 1} - r\right|, \frac{1}{\log(n+1)}\right\} \text{ for } n \geq w(a+b), \tag{58}$$

where a and b are the nonnegative integer constants of condition RC_1. Along the lines of (46), we can find a sufficiently large integer N^* such that

$$[r - \varepsilon_n^*,\, r + \varepsilon_n^*] \subset \mathcal{S}, \quad i_n > wa, \quad \text{and} \quad n - i_n > wb \text{ for all } n \geq N^*, \tag{59}$$

where $\mathcal{S}$ is the subinterval containing r in condition RC_2. Moreover, there is a finite constant M^* for which we have the following counterpart of (48),

$$|Q^w(u) h_{i_n,n}(u)| \leq M^* h_{i_n - wa, n - wa - wb}(u) \text{ for all } u \in [0,\, 1] \text{ and } n \geq N^*; \tag{60}$$

and since $i_n - wa > 0$ and $n - i_n - wb > 0$ for all $n \geq N^*$, it follows that the wth central moment (44) of $A_{(i_n)}$ exists for all $n \geq N^*$.

Using an argument similar to (52)–(55), we can show that

$$\left| \int_{|u-r|>\varepsilon_n} \Big\{ Q(u) - \mathrm{E}\Big[A_{(i_n)}\Big] \Big\}^w \, h_{i_n,n}(u)\, du \right| = o\Big(n^{-(w+1)/2}\Big) \quad \text{for all} \quad n \geq N^*. \tag{61}$$

Finally applying regularity condition RC_2 to obtain a second-order Taylor expansion of $Q(u)$ centered at $i_n/(n+1)$ and applying (43) to obtain a comparable expansion for $\mathrm{E}\Big[A_{(i_n)}\Big]$, we can show that

$$\int_{|u-r|\leq\varepsilon_n} \Big\{ Q(u) - \mathrm{E}\Big[A_{(i_n)}\Big] \Big\}^w \, h_{i_n,n}(u)\, du = \left[Q'\Big(\frac{i_n}{n+1}\Big) \right]^w \mu_w(i_n, n) + o\Big(n^{-(w+1)/2}\Big) \;; \tag{62}$$

and paralleling the justification of (57), the key properties required to establish (62) are that $Q''(\cdot)$ is continuous on $[r - \varepsilon_n^*,\, r + \varepsilon_n^*]$ for $n \geq N^*$ and that $\lim_{n\to\infty} \varepsilon_n^* = 0$. Combining (61) and (62), we finally obtain (44). This completes the proof of Lemma 3.

To prove Lemma 1, we apply (43) and (44) together with the following standard property of the central moments of a beta distribution (see van Zwet's Lemma 3.2.1),

$$\mu_w(i_n, n) = \begin{cases} O\Big(n^{-w/2}\Big), & \text{if } w \text{ is even,} \\[2ex] O\Big(n^{-(w+1)/2}\Big), & \text{if } w \text{ is odd,} \end{cases}$$

to show that

$$\sup_n \mathrm{E}\bigg(\Big\{ n^{1/2} \Big[\widehat{\xi}_{\mathrm{DS}}(\psi_c, n) - \xi \Big] \Big\}^4 \bigg) < \infty \quad \text{for } c = 1, 2. \tag{63}$$

Now condition RC_3 implies that $F'(\xi) = 1/Q'(r) \neq 0$; and thus we can apply (34) and (35) to show that $n^{1/2}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_c, n) - \xi\Big]$ converges weakly to a distribution with mean zero and variance $\sigma^2_{\mathrm{DS}} = r(1-r)/[F'(\xi)]^2$ for $c = 1, 2$. Display (63) and the corollary to Theorem 25.12 of Billingsley (1986) imply that

$$\left. \begin{aligned} &\lim_{n\to\infty} \mathrm{E}\Big\{ n^{1/2}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_c, n) - \xi\Big] \Big\} = 0 \\ &\lim_{n\to\infty} \mathrm{Var}\Big\{ n^{1/2}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_c, n) - \xi\Big] \Big\} = \frac{r(1-r)}{[F'(\xi)]^2} \end{aligned} \right\} \quad \text{for} \quad c = 1, 2, \tag{64}$$

which is the desired conclusion.

Proof of Proposition 2

Using the notation established in Lemma 3 with $i_n \equiv \lceil nr \rceil$ for all n, we detail the proof of (19). Let $\delta_n \equiv \lceil nr \rceil/(n+1) - r = O(n^{-1})$ for all $n \geq N$. Lemma 3 implies that

$$\mathrm{E}\Big[Y_{(\lceil nr \rceil)}\Big] = Q\Big(\frac{\lceil nr \rceil}{n+1}\Big) + Q''\Big(\frac{\lceil nr \rceil}{n+1}\Big)\frac{\mu_2(\lceil nr \rceil, n)}{2} + o(n^{-1})\,. \tag{65}$$

Into (65) we will insert (a) the asymptotic expansion $\mu_2(\lceil nr\rceil, n) = r(1-r)/n + o(n^{-1})$ for all $n \ge N$ (see, for example, van Zwet's equation (3.2.1)); and (b) the following Taylor expansions whose validity is guaranteed by condition RC_4 together with the definitions (45) and (46):

$$\left.\begin{aligned} Q\Big(\frac{\lceil nr\rceil}{n+1}\Big) &= Q(r) + Q'(r)\delta_n + O(n^{-2}) \\ Q''\Big(\frac{\lceil nr\rceil}{n+1}\Big) &= Q''(r) + O(n^{-1}) \end{aligned}\right\} \quad \text{for all} \quad n \ge N.$$

Thus we obtain

$$\mathrm{E}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_1, n)\Big] = \mathrm{E}\Big[Y_{(\lceil nr\rceil)}\Big] = Q(r) + Q'(r)\delta_n + \tfrac{1}{2}Q''(r)\frac{r(1-r)}{n} + o(n^{-1})\,, \tag{66}$$

and (19) follows immediately from the definition of δ_n. To prove (20), we obtain expansions similar to (66) for $\mathrm{E}\Big[Y_{(\lceil nr+0.5\rceil-1)}\Big]$ and $\mathrm{E}\Big[Y_{(\lceil nr+0.5\rceil)}\Big]$; and we combine these expansions in accordance with (4) and (5) to obtain the desired expression for $\mathrm{E}\Big[\widehat{\xi}_{\mathrm{DS}}(\psi_2, n)\Big]$.

Sketch of the Proof of Relation (29)

Some additional notation is required to justify (29). Let $G_{\mathrm{IR}}^{(n)}$ denote the distribution of n independent uniform random numbers. Clearly $G_{\mathrm{IR}}^{(n)}$ satisfies conditions CI_1 and CI_2 so that $G_{\mathrm{IR}}^{(n)} \in \mathcal{G}$. Now the row vectors $\Big\{\mathbf{U}^{(i)} \equiv \Big[U_1^{(i)}, \ldots, U_d^{(i)}\Big] : i = 1, \ldots, n\Big\}$ are generated under Latin hypercube sampling (LHS) if the corresponding column vectors $\Big\{\mathcal{U}_j \equiv \Big[U_j^{(1)}, \ldots, U_j^{(n)}\Big]^{\mathrm{T}} : j = 1, \ldots, d\Big\}$ are randomly sampled from the distribution $G_{\mathrm{LH}}^{(n)}$. Similarly, the row vectors $\Big\{\mathbf{U}^{(i)} : i = 1, \ldots, n\Big\}$ are generated under i.i.d. sampling if the corresponding column vectors $\{\mathcal{U}_j : j = 1, \ldots, d\}$ are randomly sampled from the distribution $G_{\mathrm{IR}}^{(n)}$. Expectations, variances, and covariances under LHS (respectively, i.i.d. sampling) will be denoted by the subscript LH (respectively, IR) when these moments are potentially different under LHS and i.i.d. sampling. For a fixed real t and for any positive integer n, we take

$$\chi^{(n)}(\mathbf{u}) \equiv \mathbf{1}\Big\{y(\mathbf{u}) \le \xi + n^{-1/2}t\Big\} \quad \text{for all} \quad \mathbf{u} \in [0,\, 1]^d,$$

where we suppress the dependence of $\chi^{(n)}(\cdot)$ on t; and we let $\chi_j^{(n)}(\cdot)$, $\chi_{\mathrm{add}}^{(n)}(\cdot)$, and $\chi_{\mathrm{res}}^{(n)}(\cdot)$ respectively denote the jth main effect, the additive part, and the residual from additivity of $\chi^{(n)}(\cdot)$.

The proof of relation (29) is based on arguments given by Stein (1987) and Owen (1992a) to establish respectively the asymptotic variance and the asymptotic distribution of the *mean* of a Latin hypercube sample. The adaptation of these arguments depends on the basic property that $\lim_{n\to\infty}\chi^{(n)}(\mathbf{u}) = \chi(\mathbf{u})$ for all $\mathbf{u} \in [0,1]^d$. By the definition of the sample c.d.f. in (2), $F_n^{(\mathrm{LH})}\left(\xi + n^{-1/2}t\right) = n^{-1}\sum_{i=1}^n \chi^{(n)}\left(\mathbf{U}^{(i)}\right)$, where the $\left\{\mathbf{U}^{(i)} : i = 1,\ldots,n\right\}$ are sampled under LHS. By decomposing $\chi^{(n)}(\cdot)$ into its additive part $\chi_{\mathrm{add}}^{(n)}(\cdot)$ and its residual from additivity $\chi_{\mathrm{res}}^{(n)}(\cdot)$, we have $n^{1/2}\Delta_n^{(\mathrm{LH})} = A_n + B_n$, where

$$A_n \equiv n^{-1/2}\sum_{i=1}^n \left[\chi_{\mathrm{add}}^{(n)}\left(\mathbf{U}^{(i)}\right) - F\left(\xi + n^{-1/2}t\right)\right] \quad\text{and}\quad B_n \equiv n^{-1/2}\sum_{i=1}^n \chi_{\mathrm{res}}^{(n)}\left(\mathbf{U}^{(i)}\right). \tag{67}$$

The first step in the proof of (29) is to show that under LHS,

$$A_n \xrightarrow{\mathcal{P}} 0 \quad\text{as}\quad n\to\infty. \tag{68}$$

Following closely the proof of Theorem 1 in Stein (1987) and using the function $\chi_{\mathrm{add}}^{(n)}(\cdot)$ in place of Stein's function $h(\cdot)$, we obtain the following analogue of Stein's Theorem 1 by applying Lemma 4 of Avramidis and Wilson (1995):

$$\mathrm{Cov}_{\mathrm{LH}}\left[\chi_{\mathrm{add}}^{(n)}\left(\mathbf{U}^{(1)}\right),\, \chi_{\mathrm{add}}^{(n)}\left(\mathbf{U}^{(2)}\right)\right] = dn^{-1}\mathrm{E}^2\left[\chi_{\mathrm{add}}^{(n)}(\mathbf{U})\right] - n^{-1}\sum_{j=1}^d \mathrm{E}\left[\chi_j^2(U_j)\right] + o(n^{-1}). \tag{69}$$

Combining (67) and (69) and applying the bounded convergence theorem (Billingsley 1986, p. 214), we have

$$\begin{aligned}\mathrm{Var}_{\mathrm{LH}}[A_n] &= \mathrm{Var}\left[\chi_{\mathrm{add}}^{(n)}(\mathbf{U})\right] + (n-1)\mathrm{Cov}_{\mathrm{LH}}\left[\chi_{\mathrm{add}}^{(n)}\left(\mathbf{U}^{(1)}\right),\, \chi_{\mathrm{add}}^{(n)}\left(\mathbf{U}^{(2)}\right)\right] \\ &= \mathrm{Var}[\chi_{\mathrm{add}}(\mathbf{U})] - \sum_{j=1}^d \mathrm{Var}[\chi_j(U_j)] + o(1) = o(1) \quad\text{as}\quad n\to\infty. \end{aligned}\tag{70}$$

Since $\mathrm{E}[A_n] = 0$, (68) follows from (70) and Chebyshev's inequality (Billingsley 1986, p. 75).

The second step in the proof of (29) is to show that under LHS,

$$B_n \xrightarrow{\mathcal{D}} N\{0,\, \mathrm{Var}[\chi_{\mathrm{res}}(\mathbf{U})]\}. \tag{71}$$

Adapting the proof of Theorem 1 of Owen (1992a) as detailed in Lemmas 5 and 6 of Avramidis and Wilson (1995), we have that for any integer $\nu \ge 1$,

$$\mathrm{E}_{\mathrm{LH}}[B_n^\nu] = \mathrm{E}_{\mathrm{IR}}[B_n^\nu] + o(1) \quad\text{as}\quad n\to\infty. \tag{72}$$

¿From the analysis given in displays (30.4)–(30.9) of Billingsley (1986), it follows immediately that under i.i.d. sampling each moment of the standardized sum $S_n \equiv B_n \big/ \mathrm{E}^{1/2}\left\{\left[\chi_{\mathrm{res}}^{(n)}(\mathbf{U})\right]^2\right\}$ converges to the corresponding moment of a standard normal random variable so that $\lim_{n\to\infty}\mathrm{E}_{\mathrm{IR}}[S_n^\nu] = \mathrm{E}\{\,[N(0,1)]^\nu\,\}$ for $\nu = 1,2,\ldots$. This last result and the bounded convergence theorem imply that

$$\lim_{n\to\infty}\mathrm{E}_{\mathrm{IR}}[B_n^\nu] = \mathrm{E}\left[\,\left(N\left\{0,\, \mathrm{Var}[\chi_{\mathrm{res}}(\mathbf{U})]\,\right\}\right)^\nu\,\right] \quad\text{for}\quad \nu = 1,2,\ldots . \tag{73}$$

Combining (72) and (73), we see that under LHS each moment of B_n converges to the corresponding moment of a normal distribution with mean zero and variance $\mathrm{Var}[\chi_{\mathrm{res}}(\mathbf{U})]$ as $n \to \infty$. Thus (71) follows from the method-of-moments theorem (Billingsley 1986, Theorem 30.2) and the fact that the normal distribution is determined by its moments (Billingsley 1986, Example 30.1). Finally (29) follows from (68), (71), and Slutsky's Theorem.

ACKNOWLEDGMENTS

This work was partially supported by NSF Grant No. DMS-8717799 and by a David Ross Grant from the Purdue Research Foundation. The authors thank Art Owen for providing an indication of the method of proof of Theorem 1 in Owen (1992a) before that paper appeared in the literature. Thanks also go to Bruce Schmeiser for many enlightening discussions on this paper and to Barry Nelson, the Simulation Area Editor, and two anonymous referees for numerous suggestions that substantially improved the readability of this paper.

REFERENCES

Antill, J. M., and R. W. Woodhead. 1990. *Critical Path Methods in Construction Practice*, 4th edition. John Wiley, New York.

Avramidis, A. N., K. W. Bauer, Jr. and J. R. Wilson. 1991. Simulation of Stochastic Activity Networks Using Path Control Variates. *Naval Research Logistics* **38,** 183–201.

Avramidis, A. N., and J. R. Wilson. 1995. Correlation-Induction Techniques for Estimating Quantiles in Simulation Experiments. Technical Report 95-5, Department of Industrial Engineering, North Carolina State University, Raleigh, N.C.

Avramidis, A. N., and J. R. Wilson. 1996. Integrated Variance Reduction Strategies for Simulation. *Operations Research* **44**, 327–346.

Billingsley, P. 1986. *Probability and Measure*, 2nd ed. John Wiley, New York.

Becker, R. A., and J. M. Chambers. 1984. *S: An Interactive Environment for Data Analysis and Graphics*. Wadsworth, Belmont, Calif.

David, H. A. 1981. *Order Statistics*, 2nd ed. John Wiley, New York.

Elmaghraby, S. E. 1977. *Activity Networks: Project Planning and Control by Network Models*. John Wiley, New York.

Feller, W. 1971. *An Introduction to Probability Theory and Its Applications, Volume II*, 2nd ed. John Wiley, New York.

Hahn, G. J., and S. S. Shapiro. 1967. *Statistical Models in Engineering*. John Wiley, New York.

Hesterberg, T., and B. L. Nelson. 1995. Control Variates for Probability and Quantile Estimation. Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Ill.

Hsu, J. C., and B. L. Nelson. 1990. Control Variates for Quantile Estimation. *Management Science* **36,** 835–851.

Kappenman, R. F. 1987. Improved Distribution Quantile Estimation. *Communications in Statistics—Simulation and Computation* **B16,** 307–320.

Lehmann, E. L. 1966. Some Concepts of Dependence. *Annals of Mathematical Statistics* **37,** 1137–1153.

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

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

Owen, A. B. 1992b. Orthogonal Arrays for Computer Integration and Visualization. *Statistica Sinica* **2,** 2.

Ressler, R. L., and P. A. W. Lewis. 1990. Variance Reduction for Quantile Estimates in Simulations via Nonlinear Controls. *Communications in Statistics—Simulation and Computation* **B19**, 1045–1077.

Schafer, R. E. 1974. On Assessing the Precision of Simulations. *Journal of Statistical Computation and Simulation* **3,** 67–69.

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

Sigal, C. E., A. A. B. Pritsker and J. J. Solberg. 1980. The Stochastic Shortest Route Problem. *Operations Research* **28**, 1122–1129.

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

van Zwet, W. R. 1964. *Convex Transformations of Random Variables.* Mathematical Centre Tracts 7, Mathematisch Centrum, Amsterdam.

Yang, S.-S. 1985. A Smooth Nonparametric Estimator of a Quantile Function. *Journal of the American Statistical Association* **80,** 1004–1011.

Figure 1. Network 1 with mean duration shown next to each activity.

Figure 2. Network 2 with mean duration shown next to each activity.

Figure 3. Bias squared, variance, and MSE of the estimators $\widehat{\xi}_{\mathrm{DS}}(\psi_1, n)$ and $\widehat{\xi}_{\mathrm{DS}}(\psi_2, n)$ as functions of the sample size n in network 1 with $r = 0.95$.