Control Variates for Probability and Quantile Estimation

Timothy C. Hesterberg • Barry L. Nelson
MathSoft, 1700 Westlake Avenue N., Suite 500, Seattle, Washington 98109
Department of Industrial Engineering and Management Science, 2145 Sheridan Road,
Northwestern University, Evanston, Illinois 62208-3119

In stochastic systems, quantiles indicate the level of system performance that can be delivered with a specified probability, while probabilities indicate the likelihood that a specified level of system performance can be achieved. We present new estimators for use in simulation experiments designed to estimate such quantiles or probabilities of system performance. All of the estimators exploit control variates to increase their precision, which is especially important when extreme quantiles (in the tails of the distribution of system performance) or extreme probabilities (near zero or one) are of interest. Control variates are auxiliary random variables with known properties—in this case, known quantiles—and a strong stochastic association with the performance measure of interest. Since transforming a control variate can increase its effectiveness, we propose both continuous and discrete approximations to the optimal (variance-minimizing) transformation for estimating probabilities, and then invert the probability estimators to obtain corresponding quantile estimators. We also propose a direct control-variate quantile estimator that is not based on inverting a probability estimator. An empirical study using queueing, inventory and project-planning examples shows that substantial reductions in mean squared error can be obtained when estimating the 0.9, 0.95, and 0.99 quantiles.

(*Simulation*; *Variance Reduction*; *Control Variates*; *Statistics*)

1. Introduction

Variance-reduction research in the discrete-event simulation literature has focussed on estimating expected values—especially means or first moments, which includes probabilities. There are, however, many practical problems in which a quantile is a more relevant performance measure.

As a prototype example, consider a stochastic activity network that represents the time to complete a large project. In order to bid the project correctly, the planners might wish to know an upper bound on the completion time that will hold with high probability, perhaps 0.85. Therefore, the value they desire is the 0.85 quantile, also called the 85th percentile, of project completion time.

As a second example, consider the design of a commercial on-line database system in which response time to customer queries is important. When evaluating possible configurations of hardware and software, analysts may be interested in determining a response time that will only rarely be exceeded. Again, the desired performance measure is a quantile.

The need for quantile estimation is more familiar to statisticians. The critical values for test statistics, confidence intervals and sequential-sampling procedures are quantiles, traditionally the 0.9, 0.95, and 0.99 quantiles. For statistics with complicated sampling distributions, simulation could be required to estimate these quantiles. When the critical values are computed in real time by a software package, then fast, precise quantile estimation is crucial. For example, the ADJUST=SIMULATE setting under the LSMEANS option of PROC GLM and PROC MIXED of *SAS* Version 6.11 estimates critical quantiles as needed for multiple comparison procedures, and the

0025-1909/98/4409/1295$05.00

bootstrap confidence intervals in *S-PLUS* Version 4.0 use estimated critical quantiles.

In this paper, we form improved probability and quantile estimators for ''terminating'' (finite-horizon) simulations using the method of control variates. That is, we exploit known information about certain other random variables in the simulation to more precisely estimate probabilities and quantiles of the random variable of interest. Previous attempts to use control variates in this way are reviewed and unified in §2. The review is important because there are several competitors, and our ideas pull together and extend many of them. For a different approach to improving quantile estimation using correlation-induction techniques, see Avramidis (1992) and Avramidis and Wilson (1998).

The remainder of the paper is organized as follows. In §3 we show how control-variate estimators can be viewed as weighted averages, and how we exploit this perspective to directly form probability estimators and indirectly form quantile estimators. This observation is critical because it is much easier to derive effective control-variate probability estimators than it is to derive effective control-variate quantile estimators. Since the effectiveness of a given control variate can be enhanced via transformation, §4 provides continuous and discrete *approximations* to the optimal (variance-minimizing) transformation for estimating probabilities; the optimal transformation is typically unattainable. In §5 we solve a longstanding technical problem that allows us to form control-variate quantile estimators directly, without inverting a probability estimator. Section 6 presents an empirical evaluation of the competitors, which reveals that a simple discrete approximation to the optimal control variate is the best choice for both probability and quantile estimation. Some summary conclusions are offered in §7.

2. Background

Let Y be a random variable with absolutely continuous cumulative distribution function (cdf) F_Y. For $0 < q < 1$, let y_q denote the unique value such that $F_Y(y_q) = \Pr\{Y \le y_q\} = q$. In other words, y_q is the q quantile of Y. Notice that since F_Y is continuous, $y_q = F_Y^{-1}(q)$ also defines y_q, and this observation suggests that estimators of y_q may be obtained by inverting some form of the empirical cdf of Y.

For example, suppose that we obtain independent and identically distributed (i.i.d.) observations $Y_1, Y_2, \ldots, Y_n$ of Y, and form the empirical cdf

$$\hat{F}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{I}(Y_i \le y), \tag{1}$$

where $\mathcal{I}$ is the indicator function. This leads to the *standard estimator*

$$\hat{y}_q = \inf\{y : \hat{F}_Y(y) \ge q\} = Y_{(\lceil nq \rceil)}, \tag{2}$$

where $Y_{(k)}$ is the kth order statistic. This estimator can (and often is) refined by smoothing, interpolating, etc. (see Dielman et al. 1994 for a number of possibilities). In the empirical study Y will represent the time to complete a stochastic activity network, the delay in queue of a customer and the average cost of an inventory policy.

Clearly quantile estimation and probability estimation are intimately connected, since a better estimator of the cdf implies a better quantile estimator. One way to classify control-variate variance-reduction techniques is whether they attempt to directly improve upon the quantile estimator $Y_{(\lceil nq \rceil)}$ itself, or indirectly improve the quantile estimator by directly improving upon the empirical cdf (1). We review both approaches below, with an eye toward enhancing their effectiveness.

2.1. Direct Methods

Let X be another random variable that is observed along with Y and whose distribution F_X, or at least some aspect of it such as its mean or certain quantiles, is known. Depending on the estimator, we might have no specific requirements for F_X (as in this section) or rather strict requirements; we therefore introduce the requirements as they are needed throughout the paper.

If Y and X are dependent, then X may be used as a *control variate* to aid in estimating probabilities or quantiles of F_Y. In the empirical study, X will represent the time to complete the path with the longest expected length in the stochastic activity network, the sum of the service times of the preceding customers in the queueing system and the sum of the demands in the inventory example. Although we focus on exploiting a single control variate, if there are s control variates then we denote them by $X^{(1)}, X^{(2)}, \ldots, X^{(s)}$.

Suppose we simulate i.i.d. pairs (X_i, Y_i), for $i = 1, 2, \ldots, n$. A direct application of control variates to quantile estimation is to form the estimator

$$\hat{y}_q^{\text{dir}} = \hat{y}_q - \beta(\hat{x}_q - x_q), \tag{3}$$

where x_q is the q quantile of X and $\hat{x}_q$ is the standard estimator of it. The performance of this estimator is very sensitive to the strength of the correlation between $\hat{y}_q$ and $\hat{x}_q$, and the value chosen for the multiplier β. The optimal (variance-minimizing) value of β is $\beta^* = \text{Cov}[\hat{x}_q, \hat{y}_q] / \text{Var}[\hat{x}_q]$, which is typically unknown.

When estimating the mean $E[Y]$ using a control variate X (as described in Section 3), there is a natural estimator for β^*, since $\text{Cov}[\bar{X}, \bar{Y}] / \text{Var}[\bar{X}] = \text{Cov}[X, Y] / \text{Var}[X]$, which depends on *individual* values of Y and X. But for quantiles estimated as in (2), the *entire* sample is required to obtain a *single* estimator $\hat{y}_q$ and $\hat{x}_q$, leaving no degrees of freedom to estimate their variances or covariance. To solve this problem, Ressler and Lewis (1990) suggest partitioning the size n sample into subsamples, calculating estimates of x_q and y_q from each subsample, and then estimating β^* from the subsample quantile estimates. Unfortunately, this substantially increases the bias of $\hat{y}_q$ when q is extreme. *In* §5 *we introduce a method that estimates β^* implicitly, without resorting to subsamples.*

A second problem is that the correlation between $\hat{y}_q$ and $\hat{x}_q$ might not be large, especially when q is extreme. Ressler and Lewis (1990) attempt to alleviate this problem by transforming $\hat{x}_q$ to increase its correlation with $\hat{y}_q$. Their estimator can be represented as

$$\hat{y}_q^{\text{rl}} = \hat{y}_q - \beta\{g(\hat{x}_q, \boldsymbol{\alpha}) - E[g(\hat{x}_q, \boldsymbol{\alpha})]\}, \tag{4}$$

where g is a parametric function of a vector of unknown parameters $\boldsymbol{\alpha}$. They use ACE (Breiman and Friedman 1985) to suggest a functional form for g, and a nonlinear regression using subsamples to estimate $\boldsymbol{\alpha}$ and β. A difficulty that they encounter is determining the $E[g(\hat{x}_q, \boldsymbol{\alpha})]$, so they ultimately restrict attention to certain classes of strictly monotone g. *In* §4 *we propose much simpler approximations to the optimal transformation.*

2.2. Indirect Methods

Since a quantile estimator can be obtained by inverting an estimator of the cdf, control variates can also be used to indirectly improve quantile estimators by improving the cdf estimators on which they are based.

The usual control variate estimator of $F_Y(y)$ is

$$\hat{F}_Y^{\text{cv}}(y) = \hat{F}_Y(y) - \hat{\beta}(\hat{F}_X(x) - F_X(x)), \tag{5}$$

where $\hat{F}_X(x)$ is the empirical cdf of X and $\hat{\beta}$ is the slope from regressing $\mathcal{I}(Y_i \le y)$ on $\mathcal{I}(X_i \le x)$; notice that there is no need to form subsamples. *In* §4 *we refine this idea by incorporating more effective control variate*(*s*).

Another common estimate is obtained by using a fixed constant in place of $\hat{\beta}$. Using a fixed constant may lead to a variance inflation, however, while $\hat{\beta}$ is asymptotically optimal and turns out to be particularly convenient for quantile estimation, as we discuss in §3.

Both probability and quantile estimators have been based on the following analysis using maximum likelihood methods: Let

$$p_{00}(y) = \Pr\{X \le x_q,\ Y \le y\},$$
$$p_{01}(y) = \Pr\{X \le x_q,\ Y > y\},$$
$$p_{10}(y) = \Pr\{X > x_q,\ Y \le y\},$$
$$p_{11}(y) = \Pr\{X > x_q,\ Y > y\},$$

and let

$$N_{00}(y) = \#\{i\colon X_i \le x_q,\ Y_i \le y\},$$
$$N_{01}(y) = \#\{i\colon X_i \le x_q,\ Y_i > y\},$$
$$N_{10}(y) = \#\{i\colon X_i > x_q,\ Y_i \le y\},$$
$$N_{11}(y) = \#\{i\colon X_i > x_q,\ Y_i > y\},$$
$$N_0 = \#\{i\colon X_i \le x_q\},$$
$$N_1 = \#\{i\colon X_i > x_q\}.$$

We assume here that $\Pr\{X \le x_q\} = q$, which is true if F_X is continuous; the more general approach in §4.3 allows X to have an arbitrary distribution.

Notice that

$$\begin{aligned}
&\Pr\{N_{00}(y) = n_{00},\ N_{01}(y) = n_{01},\\
&\quad N_{10}(y) = n_{10},\ N_{11}(y) = n_{11}\}\\
&\quad = \frac{n!}{n_{00}!n_{01}!n_{10}!n_{11}!}[p_{00}(y)]^{n_{00}}[p_{01}(y)]^{n_{01}}\\
&\qquad \cdot[p_{10}(y)]^{n_{10}}[p_{11}(y)]^{n_{11}}\\
&\quad = \frac{n!}{n_{00}!n_{01}!n_{10}!n_{11}!}[q - p_{01}(y)]^{n_{00}}[p_{01}(y)]^{n_{01}}\\
&\qquad \cdot[p_{10}(y)]^{n_{10}}[1 - q - p_{10}(y)]^{n_{11}},
\end{aligned} \tag{6}$$

a multinomial distribution. Fieller and Hartley (1954), Davidson and MacKinnon (1981), and Rothery (1982) showed that the maximum likelihood estimators of the probabilities implied by (6), conditional on $N_0 > 0$ and $N_1 > 0$, are

$$\hat{p}_{0\ell}(y) = \frac{qN_{0\ell}(y)}{N_0},$$

$$\hat{p}_{1\ell}(y) = \frac{(1-q)N_{1\ell}(y)}{N_1}, \tag{7}$$

for $\ell = 0, 1$. But since $F_Y(y) = p_{00}(y) + p_{10}(y)$, this leads to the maximum likelihood estimator

$$\hat{F}_Y^{\text{ml}}(y) = \hat{p}_{00}(y) + \hat{p}_{10}(y). \tag{8}$$

This estimator is easy to compute, is unbiased when $N_0 > 0$ and $N_1 > 0$, and is effective. *We show in the next section that* (8) *is equivalent to* (5). Also, the cdf estimator (8) can be inverted to obtain a quantile estimator. That is, we can search for a value $\hat{y}_q^{\text{ml}}$ such that $\hat{F}_Y^{\text{ml}}(\hat{y}_q^{\text{ml}}) \doteq q$, or equivalently $\hat{p}_{01}(\hat{y}_q^{\text{ml}}) \doteq \hat{p}_{10}(\hat{y}_q^{\text{ml}})$. This is essentially Hsu and Nelson's (1990) ''ILRT'' quantile estimator; *it is a special case of the discrete approximation to the optimal control variate estimator that we introduce in* §4.3.

Another indirect estimator is Hsu and Nelson's ''MED UNB'' estimator. Let $K_0 = \#\{i\colon Y_i \le y_q\}$. While $N_0 = \#\{i\colon X_i \le x_q\}$ is observable, K_0 is not; however, N_0 and K_0 are clearly dependent since $N_0 = N_{00}(y_q) + N_{01}(y_q)$ and $K_0 = N_{00}(y_q) + N_{10}(y_q)$. MED UNB exploits this dependence in the following way: Since $\hat{F}_X(x_q) = N_0/n$, $\hat{F}_Y(y_q) = K_0/n$ and $E[N_0/n] = E[K_0/n] = q$, MED UNB inverts $\hat{F}_Y(y) = N_0/n$ to estimate y_q. The resulting estimator is $Y_{(N_0)}$. *In* §5 *we refine this idea by accounting for the strength of the dependence between Y and X.*

Based on an extensive empirical study, Hsu and Nelson concluded that ILRT is the best of the three quantile estimators (ILRT, MED UNB, and ''NPMLE,'' which is not discussed here), so it is the one against which we compete. Notice, however, that ILRT and MED UNB are both difficult to generalize to multiple control variates, while our new estimators are not.

In the following sections we introduce estimators that refine and extend ideas presented in this section. We primarily derive probability estimators that can be inverted to obtain quantile estimators, because these estimators are more versatile and because their performance characteristics can be more easily established. However, in §5 we also derive a direct control-variate quantile estimator that implicitly estimates β^*.

3. Control Variates as Weighted Averages

This section reformulates the standard linear control-variate estimator for a mean in a way that is particularly advantageous for quantile estimation, and demonstrates the potential benefit of transforming a control variate to increase its effectiveness in reducing variance.

Suppose that we observe i.i.d. pairs (C_i, Z_i), for $i = 1, 2, \ldots, n$, and our goal is to estimate $\mu_Z = E[Z]$ when $\mu_C = E[C]$ is known. In the sequel we let $Z = \mathcal{I}(Y \le y)$ and $C = g(X)$ for some transformation g, but for the moment we leave the presentation general. Let $\sigma_Z^2 = \text{Var}[Z]$ and $R_{ZC} = \text{Cov}[Z, C]/(\text{Var}[C]\,\text{Var}[Z])^{1/2}$, the correlation coefficient.

The usual linear control-variate estimator of μ_Z is

$$\hat{\mu}_Z^{\text{cv}} = \bar{Z} - \hat{\beta}(\bar{C} - \mu_C), \tag{9}$$

where $\bar{Z}$ and $\bar{C}$ are sample means, and $\hat{\beta}$ is the slope estimator obtained from a least-squares regression of Z_i on C_i. Two important properties of this estimator are that $\hat{\mu}_Z^{\text{cv}} \xrightarrow{P} \mu_Z$ and $\sqrt{n}(\hat{\mu}_Z^{\text{cv}} - \mu_Z) \Rightarrow N(0, \sigma^2)$ as $n \to \infty$, where $\xrightarrow{P}$ denotes convergence in probability, ⇒ denotes convergence in distribution, and $\sigma^2 = (1 - R_{ZC}^2)\sigma_Z^2$ (Nelson 1990). Thus, the linear control-variate estimator is consistent and has asymptotically smaller variance than $\bar{Z}$ as long as $R_{ZC}^2 > 0$.

Hesterberg (1993) and others noticed that (9) can be rewritten as

$$\hat{\mu}_Z^{\text{cv}} = \sum_{i=1}^{n}\left(\frac{1}{n} + \frac{(\bar{C} - \mu_C)(\bar{C} - C_i)}{\sum_{j=1}^{n}(C_j - \bar{C})^2}\right)Z_i = \sum_{i=1}^{n} W_i Z_i. \tag{10}$$

Since the $\sum_{i=1}^{n} W_i = 1$ (Nelson 1990, Appendix A), the linear control-variate estimator can be viewed as a *weighted average* of the Z_i values.

There are several advantages to this representation when the goal is probability or quantile estimation. Let $Z = \mathcal{I}(Y \le y)$ and $C = g(X)$ for some function g. Then the control-variate estimator of $F_Y(y)$ is

$$\hat{F}_Y^{\text{cv}}(y) = \sum_{i=1}^{n} \mathcal{I}(Y_i \le y) W_i = \sum_{\{i:Y_i \le y\}} W_i. \tag{11}$$

This estimator is consistent and can be used to estimate $F_Y(y)$ for *any* value of y without recomputing W_i. In addition, quantile estimators are obtained by inverting the weighted cdf (11). Specifically, the control-variate estimator of y_q is $\hat{y}_q^{\text{cv}} = Y_{(L)}$ where

$$L = \min\left\{ j \colon \sum_{i=1}^{j} W_{[i]} \geq q \right\},$$

and $W_{[i]}$ is the weight associated with $Y_{(i)}$. In practice we interpolate between the $W_{[i]}$ to smooth this estimator.

The weighted-average representation also generalizes directly to multiple control variates via a linear regression on multiple controls: Let $\mathbf{C}_i = (C_i^{(1)}, C_i^{(2)}, \ldots, C_i^{(s)})'$ be the $s \times 1$ vector of control variates from replication i with expected value $\boldsymbol{\mu}_{\boldsymbol{C}}$. Further, let $\bar{C}^{(j)} = n^{-1} \Sigma_{i=1}^{n} C_i^{(j)}$ denote the sample mean of the jth control variate across all n replications, and let $\bar{\mathbf{C}} = (\bar{C}^{(1)}, \bar{C}^{(2)}, \ldots, \bar{C}^{(s)})'$. Finally, let $\mathbf{M}$ be the $s \times s$ matrix with (j, k)th element $M_{jk} = \Sigma_{i=1}^{n} (C_i^{(j)} - \bar{C}^{(j)})(C_i^{(k)} - \bar{C}^{(k)})$. Then the control-variate weights when using all s control variates are

$$W_i = \frac{1}{n} + (\bar{\mathbf{C}} - \boldsymbol{\mu}_{\mathbf{C}})'\mathbf{M}^{-1}(\bar{\mathbf{C}} - \mathbf{C}_i), \qquad (12)$$

for $i = 1, 2, \ldots, n$. The asymptotic variance of the control-variate estimator of $E[Z]$ is $(1 - R_{Z\mathbf{C}}^2)\sigma_Z^2$, where $R_{Z\mathbf{C}}$ is the multiple correlation between Z and $\mathbf{C}$; this reduces to $(1 - R_{Z\mathbf{C}}^2)q(1 - q)$ when $Z = \mathcal{I}(Y \leq y_q)$. The asymptotically guaranteed variance reduction for the probability estimator leads us to expect a variance reduction by the same factor for the corresponding quantile estimator. This can be seen as follows.

If Y has a nonzero, continuous density f_Y at y_q, then it is well known that

$$\sqrt{n}(\hat{y}_q - y_q) \Rightarrow N\left(0, \frac{q(1-q)}{f_Y^2(y_q)}\right)$$

(David 1981). An analogous derivation shows that

$$\sqrt{n}(\hat{y}_q^{\text{cv}} - y_q) \Rightarrow N\left(0, (1 - R_{Z\mathbf{C}}^2)\frac{q(1-q)}{f_Y^2(y_q)}\right),$$

revealing an asymptotic variance reduction of $1 - R_{Z\mathbf{C}}^2$ relative to $\hat{y}_q$. This large-sample relationship is confirmed in the small-sample empirical results presented in §6.

We note that the control variate weights W_i can be negative. In practice this occurs only with multiple control variates having long-tailed distributions in very small sample sizes when the observed sample means of the control variates differ substantially from the expected values. Negative weights cannot occur in any of the discrete control variates discussed below, and they are highly unlikely to occur in the estimated optimal control variate described in §4.1. In general, the probability of any weight being negative is $o(n^{-p/2})$ when $E[C^p] < \infty$, and it decreases faster than exponentially if all of the control variates are bounded.

An open question is, what transformation of X should be used to obtain the most benefit as a control variate? When $Z = \mathcal{I}(Y \leq y_q)$, a natural choice is $C = \mathcal{I}(X \leq x_q)$, the indicator function for the corresponding quantile of the control variate. In this case the weights have a particularly simple form

$$W_i = \begin{cases} \dfrac{q}{N_0}, & X_i \leq x_q, \\[2ex] \dfrac{1-q}{N_1}, & X_i > x_q \end{cases} \qquad (13)$$

(Hesterberg 1993), and the point estimator becomes

$$\begin{aligned} \hat{F}_Y^{\text{cv}}(y) &= \sum_{i=1}^{n} \mathcal{I}(Y_i \leq y)W_i \\ &= N_{00}(y)\frac{q}{N_0} + N_{10}(y)\frac{(1-q)}{N_1} \\ &= \hat{p}_{00}(y) + \hat{p}_{10}(y). \end{aligned}$$

That is, when $C = \mathcal{I}(X \leq x_q)$, the control-variate estimator (11) is equivalent to the maximum likelihood estimator (8) for probabilities, or the ILRT for quantiles; it is therefore unbiased for probabilities, conditional on $N_0 > 0$ and $N_1 > 0$. Davidson and MacKinnon (1992) also noted the equivalence of (11) and (8) when $C = \mathcal{I}(X \leq x_q)$. Furthermore, for this choice of C, (11) is the same as poststratified sampling on two strata (Hesterberg 1993), an insight we exploit later.

Although $C = \mathcal{I}(X \leq x_q)$ has nice properties as a control variate, the variance reduction that is achieved depends on the squared correlation between $\mathcal{I}(Y \leq y)$ and the control. The optimal transformation—the one that maximizes the squared correlation—is

$$C^* = g^*(X) = E[\mathcal{I}(Y \leq y) \mid X] = \Pr\{Y \leq y \mid X\} \qquad (14)$$

(Rao 1973, p. 264–265). Of course, if g^* were known then we might avoid simulation altogether by integrating $\int g^*(x)dF_X(x)$ to obtain $\Pr\{Y \leq y\}$, and numerically solving $\Pr\{Y \leq y\} = q$ to obtain y_q with no sampling error. In practice, the transformation g^* is typically not known, but an *approximation* of it can be used as a control variate to obtain the benefits of increased correlation. We illustrate this by the following example.

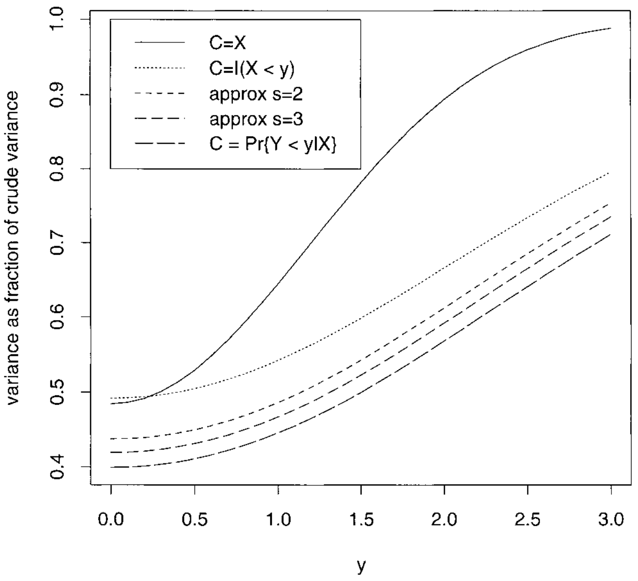
Suppose that the joint distribution of (X, Y) is $N(0, 0, 1, 1, \rho)$; that is, they are bivariate standard normal with correlation ρ. For this case we can numerically evaluate the asymptotic variance of the following estimators of $F_Y(y)$:

1. The linear control-variate estimator with $g(X) = X$; this is the usual mean-based control.
2. The linear control-variate estimator with $g(X) = \mathcal{I}(X \leq y)$; this is the natural control for estimating a probability.
3. The linear control-variate estimator with $g(X) = \Pr\{Y \leq y \mid X\}$; this is the optimal control based on X.
4. Two approximations to the optimal linear control-variate estimator that we introduce in §4.3.

For the case $\rho = 0.9$, Figure 1 shows the ratio of the asymptotic variance of each estimator to that of the crude estimator for $0 \leq y \leq 3\sigma_Y$. The crude estimator is $\hat{F}_Y(y)$, which has asymptotic variance $F_Y(y)(1 - F_Y(y))$. The figure illustrates the potential benefit to be obtained from an approximation to C^*, relative to using either X or $\mathcal{I}(X \leq y)$ as controls. The mean-based control X has almost no effect for extreme probabilities. The natural control $\mathcal{I}(X \leq y)$ is much better, but the optimal control is better still. And our approximations to the optimal control are very nearly optimal. The figure also illustrates the diminishing correlation (and therefore variance reduction) between the response and control at more extreme values of y.

In the following sections we propose techniques to approximate the optimal control (14). We begin with continuous approximations in §§4.1–4.2. These techniques are conceptually similar, but simpler and easier to implement, than those of Ressler and Lewis. Nevertheless, they do require nonlinear least squares to fix the transformation, and may require numerical integration to calculate the expected value of the control. *We use one of these estimated optimal control variates as a standard against which to compare approximations that are no more difficult to implement than the linear control-variate estimator*; *these estimators are introduced in* §§4.3 *and* 5.

Figure 1 **Ratio of the Asymptotic Variance of Each Improved Estimator to the Variance of the Crude Estimator of $F_Y(y)$**



4. Approximations of the Optimal Control Variate

This section presents the central results of the paper. We provide approximations to the optimal control variate based on X, namely $C^* = g^*(X) = \Pr\{Y \leq y \mid X\}$, for estimating $\Pr\{Y \leq y\}$. We first introduce a continuous approximation that is highly effective, but difficult to implement. Taking the continuous approximation as the standard, we derive discrete approximations that are nearly as effective and much easier to use. In all cases quantile estimators are formed by inverting the probability estimator.

4.1. Continuous Approximation from a Binary-Response Regression

We first approximate the optimal control variate C^* given in (14) by using the result of a nonlinear regression of $Z = \mathcal{I}(Y \leq y)$ on X. We then use the *estimated optimal control variate*, denoted $\hat{C}^*$, in the control variate estimator $\hat{F}_Y^{\mathrm{cv}}(y)$.

There are a number of existing procedures for regression when the response variable is dichotomous, as $Z = \mathcal{I}(Y \leq y)$ is. These include logistic regression and certain generalized linear models (Dobson 1990) or generalized additive models (Hastie and Tibshirani 1990). Depending on the joint distribution of X and Z, and on the procedure used, the approximation $\hat{C}^*$ may be consistent for C^*, in the sense that the function $\hat{g}^*$ is pointwise consistent for g^*. In particular, some of the nonparametric procedures described in Hastie and Tibshirani (1990) and implemented in *S-PLUS* (Chambers and Hastie 1992) give consistent estimators under mild conditions on the joint distribution.

In practice, there is no requirement that $\hat{C}^*$ be consistent for C^*, only that the correlation between $\hat{C}^*$ and C^* be reasonably high. A nonconsistent estimator may trade some loss of statistical efficiency for easier implementation. The greatest implementation hurdle in our context is the need to evaluate of $E[\hat{C}^*] = \int \hat{g}^*(x) f_X(x) dx$, after the transformation $\hat{g}^*(x)$ is fixed, so that we can use $\hat{C}^*$ as a control variate. This may be done by numerical integration (assuming that the distribution of X is known). We hope to avoid a difficult numerical integration by restricting the class of curves that may be fit, in particular to curves of the form

$$C = g(X) = \Phi(a_0 + a_1 X), \tag{15}$$

where Φ is the standard normal cdf. Then

$$\begin{aligned} E[C] &= \int_{-\infty}^{\infty} \Phi(a_0 + a_1 x) f_X(x) dx \\ &= \Pr\{W < a_0 + a_1 X\} = \Pr\{W - a_1 X < a_0\}, \end{aligned} \tag{16}$$

where W is a standard normal random variable that is independent of X. In the special case when X is normally distributed this reduces to

$$E[C] = \Phi\left(\frac{a_0 + a_1 \mu_X}{\sqrt{1 + a_1^2 \sigma_X^2}}\right). \tag{17}$$

In addition to numerical integration, Edgeworth approximation or the saddlepoint formula of Lugannani and Rice (see Daniels 1987) may be used to approximate $\Pr\{W - a_1 X < a_0\}$.

Our continuous approximation is implemented as follows.

1. Given an i.i.d. sample, (X_i, Z_i), $i = 1, 2, \ldots, n$, use nonlinear least squares to fit the model

$$Z_i = \Phi(a_0 + a_1 X_i) + \epsilon_i,$$

yielding estimators $\hat{a}_0$ and $\hat{a}_1$. We used the Gauss-Newton method implemented in *S-PLUS* as function `nls`.

2. Form the control variates $\hat{C}_i^* = \Phi(\hat{a}_0 + \hat{a}_1 X_i)$, $i = 1, 2, \ldots, n$.

3. Approximate $E[\hat{C}^*] = \Pr\{W - \hat{a}_1 X < \hat{a}_0\}$, treating $\hat{a}_0$ and $\hat{a}_1$ as known constants.

4. Form the control-variate estimator $\hat{F}_Y^{\text{cv}}$ using the weighted-average interpretation (11), and invert this estimator to estimate quantiles.

Notice that while this procedure yields a cdf estimate which may be inverted to obtain quantile estimates for any quantile, it is most effective for quantiles very near the value of y used in the definition of the binary response variable $Z = \mathcal{I}(Y \le y)$. Typically y would be a preliminary estimate of a particular desired quantile, such as the standard estimator (2).

If the conditional distribution of Y given X is normal with constant variance and a mean that is linear in X, then $\hat{C}^*$ defined by (15) is consistent for the optimal control variate C^*. But, again, the key is not that the model is generally consistent, but that it can serve as an approximation that captures much of the potential increase in correlation attainable with the optimal transformation. Because we are using $\hat{C}^*$ as a control, our only interest is in enhancing the correlation, not in precisely determining the transformation.

This is a benchmark estimator; it is usable in practice, but is complex and computationally expensive. We will attain nearly the same performance with simpler approaches.

4.2. Other Continuous Approximations

Another continuous approximation to C^* is obtained by starting with a "scatterplot smooth" of Y_i against X_i; many procedures are suitable for this purpose, including smoothing splines, regression splines, kernel smoothing and lowess. A number of these are discussed in Hastie and Tibshirani (1990).

Let $\hat{d}_0$ $\hat{d}_1$ be the intercept and slope of the line tangent to the scatterplot smooth at x_q, and let $\hat{\sigma}_\epsilon$ be the estimated residual standard deviation at x_q (estimated using residuals in a neighborhood of x_q). Approximating the local relationship between Y and X as linear with normal residuals leads to the control variate:

$$\begin{aligned} \hat{C}_i^* &= \widehat{\Pr}\{Y \le y_q \mid X = X_i\} \\ &= \Phi[(y_q - (\hat{d}_0 + \hat{d}_1 X_i)) / \hat{\sigma}_\epsilon]. \end{aligned}$$

This procedure has the advantage that scatterplot smooths are somewhat easier to perform than regression with a binary response, and that the same smooth may be used to define multiple control variates, each tailored for estimating a different quantile y_q. The asymptotic efficiency of this procedure is (slightly) less than the procedure in §4.1, unless the residuals are in fact normally distributed and the relationship between Y and X is linear. If the residuals are noticeably nonnormal then a robust smooth, rather than one that uses least-squares, can be used to reduce the effect of outliers.

Such a continuous approximation is used in the same way as the continuous approximation from a binary-response regression in the previous section; in particular, once the functional form has been determined, steps 2–4 are the same in both cases. We further note that the functional form may be estimated using only a subset of the observations. This is particularly advantageous in simulation experiments with very large n. We do not pursue or compare other continuous approximations, since our goal is to develop simple discrete approximations.

4.3. Discrete Approximations

In this section we consider the use of piecewise-constant approximations to C^*. This method has two advantages: the approximation can be estimated by linear (rather than nonlinear) least squares, and the expected value of C depends on the cdf of X only at the points of discontinuity. Furthermore, the method is equivalent to post-stratified sampling, so the piecewise constant approximation need not even be explicitly computed.

Let $-\infty = b_0 < b_1 < \cdots < b_s < b_{s+1} = \infty$ partition the range of X into $s + 1$ intervals, for some $s \geq 1$. We call $\mathbf{b} = (b_1, b_2, \ldots, b_s)$ *cutpoints*, and we can express $F_Y(y)$ as

$$F_Y(y) = \sum_{\ell=0}^{s} \Pr\{Y \leq y \mid b_\ell < X \leq b_{\ell+1}\} \times \Pr\{b_\ell < X \leq b_{\ell+1}\}. \quad (18)$$

We shall approximate $g^*(x) = \Pr\{Y \leq y \mid X = x\}$ by the piecewise constant function that takes the value $\Pr\{Y \leq y \mid b_\ell < X \leq b_{\ell+1}\}$ on the interval $(b_\ell, b_{\ell+1}]$. The finer the grid of cutpoints $b_1, b_2, \ldots, b_s$, the closer the two representations. However, bivariate normal examples (below) and empirical studies (§6) show that most of the potential variance reduction can be obtained with s as small as 2.

For example, suppose that (X, Y) are $N(0, 0, 1, 1, \rho)$, and consider estimating $F_Y(y_{0.975})$. Figure 2 shows the asymptotic variance of the linear control-variate estimator employing the following control variates: the mean-based control $C = X$; our discrete approximation to the optimal control with $s = 1, 2, 3$ cutpoints; and the optimal control variate C^*. For the fixed quantile $y_{0.975}$, the plot shows the asymptotic variance as a function of the conditional standard deviation of Y given X, namely $\sqrt{1 - \rho^2}$. Clearly the discrete approximation achieves nearly the same variance reduction as the optimal control with two or three cuts.

To define the discrete approximation, which we refer to as the *poststratified sampling estimator*, we extend the notation in §2.2. Let

$$p_{\ell 0}(y, \mathbf{b}) = \Pr\{b_\ell < X \leq b_{\ell+1},\ Y \leq y\},$$

$$p_{\ell 1}(y, \mathbf{b}) = \Pr\{b_\ell < X \leq b_{\ell+1},\ Y > y\},$$

$$p_\ell(\mathbf{b}) = \Pr\{b_\ell < X \leq b_{\ell+1}\},$$

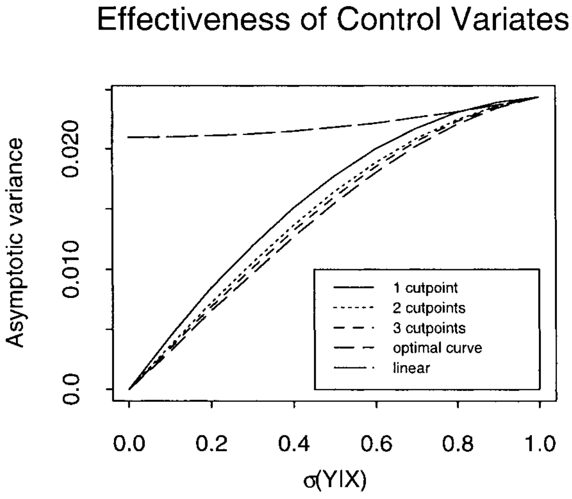
and

$$N_{\ell 0}(y, \mathbf{b}) = \#\{i : b_\ell < X_i \leq b_{\ell+1},\ Y_i \leq y\},$$

$$N_{\ell 1}(y, \mathbf{b}) = \#\{i : b_\ell < X_i \leq b_{\ell+1},\ Y_i > y\},$$

$$N_\ell(\mathbf{b}) = \#\{i : b_\ell < X_i \leq b_{\ell+1}\},$$

Figure 2 **Asymptotic Variance of the Linear Control-Variate Estimator of $F_Y(y_{0.975})$ for Bivariate Normal Data and Various Choices of Control Variate**



for $\ell = 0, 1, \ldots, s$. The poststratified sampling estimator of $F_Y(y)$ is therefore

$$\begin{aligned}\hat{F}_Y^{\mathrm{ps}}(y) &= \sum_{\ell=0}^{s} \widehat{\Pr}\{Y \le y \,|\, b_\ell < X \le b_{\ell+1}\} \\ &\quad \times \Pr\{b_\ell < X \le b_{\ell+1}\} \\ &= \sum_{\ell=0}^{s} \frac{N_{\ell 0}(y, \mathbf{b})}{N_\ell(\mathbf{b})} p_\ell(\mathbf{b}). \end{aligned} \tag{19}$$

Tedious algebra shows that this estimator is algebraically equivalent[1] to the linear control-variate estimator with control variates $C^{(\ell)} = \mathcal{I}(X \le b_\ell)$, for $\ell = 1, 2, \ldots, s$. Therefore, the estimator is a weighted average of the form (11), with weights that are independent of y; specifically,

$$W_i = \frac{p_\ell(\mathbf{b})}{N_\ell(\mathbf{b})}, \tag{20}$$

when X_i falls in stratum ℓ. These weights reduce to (13) when $s = 1$.

This result is especially important for quantile estimation, because it means that the poststratified probability estimator, no matter how many strata are employed, is as easy to invert as the linear control-variate estimator, because it *is* a linear control-variate estimator. Thus, we can capture some of the nonlinear relationship between Y and X without any additional computational or conceptual complexity. Appropriate selection of the cutpoints is addressed in the next section.

4.4. Strata Selection

In this section we consider the choice of the cutpoints $b_1, b_2, \ldots, b_s$ that define the strata for our poststratified estimator. The results are a combination of extensive simulation experience in sample problems and asymptotic analysis. The asymptotic analysis follows two threads, as $n \to \infty$ and as the correlation between X and Y increases to 1. Here is a summary of our results:

- For a single cutpoint we can barely improve upon the simple choice of $b_1 = x_q$, at least for reasonably large sample sizes. In other words, when we use a single cutpoint we use the q-quantile of X to improve the estimator of the q-quantile of Y.
- For multiple cutpoints we recommend using two or three cutpoints of the form

$$b_\ell = x_q + c_\ell \hat{\sigma}_\epsilon / \hat{d}_1, \tag{21}$$

where $c_1 \doteq -0.674$ and $c_2 \doteq 0.674$ when $s = 2$ cutpoints, $c_1 = -1$, $c_2 = 0$ and $c_3 = 1$ when $s = 3$ cutpoints, and $\hat{d}_1$ and $\hat{\sigma}_\epsilon$ are the slope and local residual standard deviation from a scatterplot smooth (i.e., nonlinear regression) of Y on X, as described in §4.2, or from a linear regression of Y on X using the $n^{6/7}$ observations with X_i closest to x_q (which is what we do in our experiments). However, the cutpoints should be adjusted if necessary so that the expected number of observations in any stratum is at least 30.

Readers interested only in applications can skip the remainder of this section, in which we derive the particular recommendations above.

We first justify the recommendation of using a small number of strata, from 2 to 4 (from 1 to 3 cutpoints). Using the standard variance decomposition we can write

$$\begin{aligned}\mathrm{Var}[Z] &= \mathrm{Var}[E(Z \,|\, X)] \\ &\quad + E[\mathrm{Var}(Z \,|\, X)] = \sigma_1^2 + \sigma_2^2,\end{aligned} \tag{22}$$

where $Z = \mathcal{I}(Y \le y)$. Using the optimal control corresponds to eliminating the σ_1^2 term. Cochran (1977, §5A.8) notes that stratification on X reduces the first term but leaves the second unchanged, and that the addition of strata quickly reaches a point of diminishing returns, beyond which the residual term σ_2^2 dominates the $\mathrm{Var}[Z]$.

Further, there is good reason to believe that the first term, σ_1^2, could be disappointingly small when estimating extreme probabilities or quantiles using control variates. For example, in Figure 1 the vertical axis corresponds to the variance-reduction ratio $(1 - R_{ZC}^2)$, which is at best (smallest) about 0.4 for the optimal control variate. Recall that the correlation between Y and X is $\rho = 0.9$ in this figure, implying that the variance-reduction ratio for estimating the mean $E[Y]$ using the control variate $C = X$ is $(1 - \rho^2) = 0.19$. Stated differently, variance is reduced by a factor of roughly 5 when estimating the mean, compared to only a factor of roughly 2 when estimating probabilities or quantiles with the same data.

[1] In fact, *any* set of control variates that form a basis for the piecewise approximation will lead to the same estimator. This is because any two sets of control variates that generate the same column space, when viewed as the independent variables in a least-squares regression, yield the same control-variate point estimator.

This suggests that the benefit of additional strata (closer approximation to the optimal control variate) will drop off quickly.

We also find that the precise placement of the cutpoints is not critical. If the optimal transformation g^* and distribution of the control variate $C^* = g^*(X)$ were known, then we would choose cutpoints on the C scale to minimize the within-strata variance of C^*, then translate to cutpoints on the X scale using the inverse function of g^*. Although g^* is unknown, it is worth noting that under general conditions the shape of g^* is that of a cumulative distribution function flipped vertically, which results in a distribution for C^* which is strongly bimodal, with infinite density at 0 and 1, if X and Y are highly correlated. The left panel of Figure 3 shows the relationship between C^* and X when X and Y are bivariate normal with $\rho = 0.95$; notice how many of the values of C^* are close to either 0 or 1. The right panel shows the bimodal distribution of C^*. Placing three cutpoints in the distribution of C^* gives a within-strata variation of C^* that is much smaller than $\sigma_1^2 = \text{Var}[C^*]$. The variance reduction is relatively insensitive to exactly where the cutpoints are placed, as long as they separate the modes.

We now proceed to derive the recommended placement of the cutpoints, beginning with the single cutpoint recommendation.

The variance of (19), conditional on N_ℓ, $\ell = 0, 1, \ldots, s$, is

$$\text{Var}\{\hat{F}_Y^{\text{ps}}(y)\,|\,N_0, \ldots, N_s\} = \sum_{\ell=0}^{s} \frac{p_{\ell 0}p_{\ell 1}}{N_\ell}, \tag{23}$$

where $p_{\ell k} = p_{\ell k}(y, \mathbf{b})$ depends implicitly (in the remainder of this section) on y and $\mathbf{b}$. A delta method approximation, as in Cochran (1977, equation 5A.41), ignoring the exponentially decreasing probability that $N_\ell = 0$ for some ℓ, yields

$$\text{Var}\{\hat{F}_Y^{\text{ps}}(y)\} = n^{-1} \sum_{\ell=0}^{s} \frac{p_{\ell 0}p_{\ell 1}}{p_\ell} + n^{-2} \sum_{\ell=0}^{s} \frac{p_{\ell 0}p_{\ell 1}(1 - p_\ell)}{p_\ell^2} + O(n^{-3}). \tag{24}$$

Figure 4 shows the leading $O(n^{-1})$ term (multiplied by n) in the variance of $\hat{F}_Y^{\text{ps}}(y_{0.975})$ as a function of a single cutpoint b_1, when X and Y are bivariate standard normal with correlations ρ chosen so that the conditional standard deviation of Y given X takes on values 0, 0.1, 0.2, ..., 1. Notice that 1.96 is the 0.975 quantile of X. The top curve, where X and Y are independent, gives the same asymptotic variance as if no control variate were used. Successively lower curves correspond to successively larger correlations. When the correlation is 1, the optimal choice $b_1 = x_{0.975}$ gives a variance of 0. The optimal cutpoints are shown on each curve with a ○. There appears to be little gain from using the optimal cutpoint relative to the simple choice of $b_1 = x_{0.975}$.

The second panel of Figure 4 includes a subset of the curves shown in the first panel, together with the vari-

Figure 3 **Optimal Transformation $C^* = g(X)$, and Implied Density of C^* when X and Y are Bivariate Normal with High Correlation**

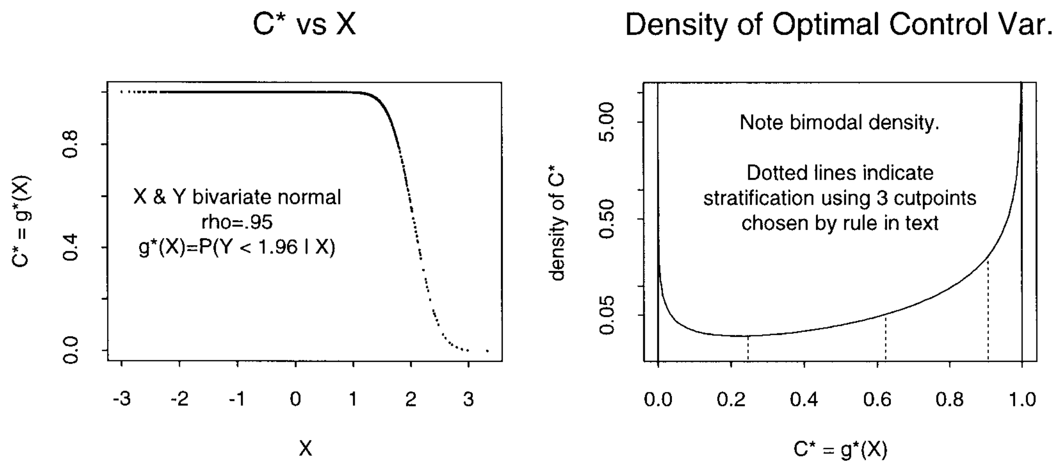
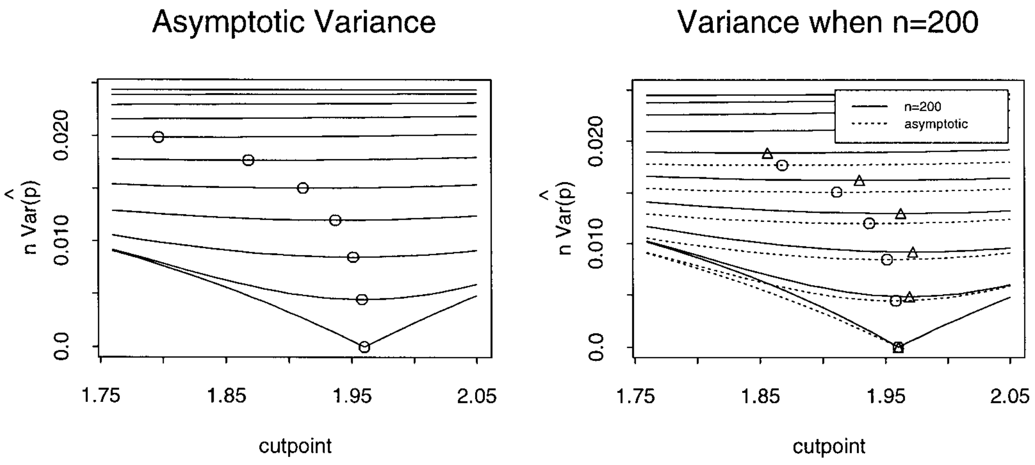


Figure 4 $nO(n^{-1})$ **and** $n(O(n^{-1}) + O(n^{-2}))$ **Terms in Variance of** $\hat{F}_Y^{ps}(y_{0.975})$ **as a Function of Cutpoint (Each curve corresponds to a different correlation between bivariate normal** X **and** Y**.)**



ance curves obtained by including the $O(n^{-2})$ term of (24) in the variance, for $n = 200$. Although the optimal cutpoints (denoted by $\triangle$) are affected by the second-order term, the curves are flat enough for samples sizes of 200 or larger that there is little to be gained by choosing the optimal cutpoint, relative to the easy choice $b_1 = x_q$.

In the appendix we derive the recommendation for multiple cutpoints. The analysis—which is based on several simplifying approximations—was employed to suggest a method for specifying cutpoints, a method which was shown to work well in empirical studies, a portion of which is presented in §6.

When n is not large, the multiple-cutpoint recommendation should be modified so that all expected stratum sizes are at least 30. This is to avoid the increased variability that results when few observations fall in a small stratum. With $E[N_\ell] \geq 30$ the probability that N_ℓ is less than 15 is small. The (second-order) asymptotics indicate that smaller expected sizes would be adequate, but we have encountered inflated variances with smaller sample sizes in simulation trials.

Finally, the recommendation to estimate the slope and residual standard deviation from a regression with $n^{6/7}$ neighbors of x_q is based on a tradeoff between the asymptotic variance of the regression slope, the squared bias of the regression slope, and the accuracy of $\hat{\sigma}_\epsilon^2$.

5. A Direct Quantile Estimator

Recall that the primary difficulty we encounter when deriving a direct control-variate quantile estimator of the form

$$\hat{y}_q^{\text{dir}} = \hat{y}_q - \beta(\hat{x}_q - x_q),$$

is in estimating the optimal multiplier $\beta^* = \text{Cov}[\hat{x}_p, \hat{y}_q] / \text{Var}[\hat{x}_p]$. In this section we propose a new estimator that may be viewed as a way to closely approximate β^* without subsamples. *Specifically, we estimate the optimal multiplier for the asymptotic joint distribution of* $\hat{y}_q$ *and* $\hat{x}_p$, where p need not equal q. We call this asymptotically optimal multiplier β^*_∞, and our estimator of it $\hat{\beta}^*_\infty$.

This direct estimator may also be viewed as a refinement of Hsu and Nelson's (1990) MED UNB estimator $Y_{(N_0)}$, where $N_0 = \#\{i : X_i \leq x_q\}$. Notice that MED UNB may be written as

$$Y_{(N_0)} = Y_{(nq+\hat{\gamma}(N_0-nq))}. \tag{25}$$

Thus, MED UNB is the standard estimator $Y_{(nq)}$, adjusted for the difference between N_0 and its mean. We propose to shrink the adjustment by a factor $\hat{\gamma}$ that depends on the correlation between $N_0 = \#\{i : X_i \leq x_p\}$ and $K_0 = \#\{i : Y_i \leq y_q\}$, and to allow the possibility of using a cutpoint x_p with $p \neq q$. Specifically, the new estimator takes the form

$$\hat{y}_q^{\text{dqe}} = Y_{(nq+\hat{\gamma}(N_0-np))}, \tag{26}$$

where $\hat{\gamma}$ will be defined below, just after (29). Surprisingly, $\hat{y}_q^{\mathrm{dqe}}$ is approximately equal to

$$\hat{y}_q^{\infty} = \hat{y}_q - \hat{\beta}_{\infty}^{*}(\hat{x}_p - x_p), \tag{27}$$

implying that in practice $\hat{\beta}_{\infty}^{*}$ (defined below) need not be computed explicitly if we use (26).

Let (X_i, Y_i), for $i = 1, 2, \ldots, n$, be an i.i.d. sample from an absolutely continuous joint distribution F_{XY} with marginal distributions that satisfy $f_X(x_p) > 0$ and $f_Y(y_q) > 0$, and let $\hat{y}_q = Y_{(nq)}$ and $\hat{x}_p = X_{(np)}$ be the crude estimators of y_q and x_p, respectively; the single cutpoint $b_1 = x_p$ is known. We do not require that $p = q$, and we assume that both nq and np are integer, for simplicity. Then as $n \to \infty$,

$$\begin{pmatrix} \sqrt{n}(\hat{x}_p - x_p) \\ \sqrt{n}(\hat{y}_q - y_q) \end{pmatrix},$$

converges in distribution to a bivariate normal random variable with zero mean vector and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \dfrac{p(1-p)}{f_X^2(x_p)} & \dfrac{F_{XY}(x_p, y_q) - pq}{f_X(x_p) f_Y(y_q)} \\ \dfrac{F_{XY}(x_p, y_q) - pq}{f_X(x_p) f_Y(y_q)} & \dfrac{q(1-q)}{f_Y^2(y_q)} \end{pmatrix}$$

(Weiss 1964). For random variables (W_X, W_Y) with this joint distribution, the optimal control-variate multiplier for estimating the mean of W_Y with W_X as the control is

$$\beta_{\infty}^{*} = \frac{\mathrm{Cov}[W_Y, W_X]}{\mathrm{Var}[W_X]}$$

$$= R\sqrt{\frac{q(1-q)}{p(1-p)}}\,\frac{f_X(x_p)}{f_Y(y_q)} = \gamma\,\frac{f_X(x_p)}{f_Y(y_q)}, \tag{28}$$

where $R = (F_{XY}(x_p, y_q) - pq)/\sqrt{p(1-p)q(1-q)}$ is the correlation between $\mathcal{I}(X \le x_p)$ and $\mathcal{I}(Y \le y_q)$, and $\gamma = R\sqrt{q(1-q)/(p(1-p))}$ is the linear regression slope of $\mathcal{I}(Y \le y_q)$ against $\mathcal{I}(X \le x_p)$. We will now show that β_{∞}^{*} can be estimated without the need for subsamples.

First, we propose an estimator for R. Of course, $\mathcal{I}(Y \le y_q)$ is unobservable, but an asymptotically valid estimator of R can be obtained by replacing y_q with any consistent estimator of it. A convenient choice is $\hat{y}_q$, the crude estimator, because $\#\{i : Y_i \le \hat{y}_q\} = nq$. Then using unbiased estimators for the covariance and variance, the sample correlation simplifies to

$$\hat{R} = \frac{N_{00}(\hat{y}_q)(1-q) - qN_{01}(\hat{y}_q)}{\sqrt{N_0 N_1 q(1-q)}}. \tag{29}$$

Multiplying this by $\sqrt{q(1-q)/(p(1-p))}$, which is known, provides the estimator of $\hat{\gamma}$ of γ.

To complete the estimator of β_{∞}^{*}, we need to estimate the ratio $f_X(x_p)/f_Y(y_q)$. A crude estimator of $f_X(x_p)$ is

$$\hat{f}_X(x_p) = \frac{\#\{i : X_i \in \mathcal{A}\}}{n|\mathcal{A}|},$$

where $\mathcal{A}$ is a small neighborhood around x_p, and $|\mathcal{A}|$ is the measure of $\mathcal{A}$. If we choose $\mathcal{A}$ to be the interval $[\min\{x_p, \hat{x}_p\}, \max\{x_p, \hat{x}_p\}]$, then

$$\hat{f}_X(x_q) = \frac{np - N_0}{n(\hat{x}_p - x_p)}.$$

This estimator may be inaccurate when the interval is narrow. But since $\hat{\beta}_{\infty}^{*}$ is multiplied by the interval width in (27), the inaccuracy occurs when it matters little.

Similarly, a crude estimator of $f_Y(y_q)$ is

$$\hat{f}_Y(y_q) = \frac{\#\{i : Y_i \in \mathcal{B}\}}{n|\mathcal{B}|}.$$

We let $\mathcal{B} = [\min\{\hat{y}_q^{\infty}, \hat{y}_q\}, \max\{\hat{y}_q^{\infty}, \hat{y}_q\}]$, where $\hat{y}_q$ is the crude estimator and $\hat{y}_q^{\infty}$ is the (yet to be determined) direct estimator.

Substituting all of the individual terms into (27), and cancelling terms where possible, gives

$$\hat{y}_q^{\infty} = \hat{y}_q - \hat{\gamma}\,\frac{(np - N_0)}{\#\{i : Y_i \in \mathcal{B}\}}|\hat{y}_q^{\infty} - \hat{y}_q|.$$

Then isolating $\#\{i : Y_i \in \mathcal{B}\}$ yields

$$\mathrm{sgn}(\hat{y}_q^{\infty} - \hat{y}_q)\#\{i : Y_i \in \mathcal{B}\}$$
$$= -\hat{\gamma}(np - N_0) = \hat{\gamma}(N_0 - np).$$

Since $\#\{i : Y_i \in \mathcal{B}\}$ is simply the number of order statistics separating $\hat{y}_q = Y_{(nq)}$ and $\hat{y}_q^{\infty}$, we can approximate $\hat{y}_q^{\infty}$ as an order statistic:

$$\hat{y}_q^{\infty} \doteq Y_{(nq + \mathrm{sgn}(\hat{y}_q^{\infty} - \hat{y}_q)\#\{i : Y_i \in \mathcal{B}\})} = Y_{(nq + \hat{\gamma}(N_0 - np))} = \hat{y}_q^{\mathrm{dqe}},$$

which is (26). Notice that we do not explicitly find $\hat{f}_X$, $\hat{f}_Y$ or $\hat{\beta}_{\infty}^{*}$. If $k = nq + \hat{\gamma}(N_0 - np)$ is not an integer, then we interpolate between $Y_{(\lfloor k \rfloor)}$ and $Y_{(\lfloor k \rfloor + 1)}$.

It may be possible to improve on this direct estimator by using wider intervals to obtain more accurate estimates of f_X and f_Y, at the cost of losing the cancellation that allows the estimator to be expressed in terms of order statistics. The estimator might also be improved by replacing $\hat{y}_q$ with a more accurate estimator of y_q in the computation of $\hat{R}$; one such estimator is $\hat{y}_q^{\text{dqe}}$ obtained from a first iteration of this procedure.

6. Empirical Evaluation

In this section we present a portion of an extensive empirical evaluation of the following four probability estimators and five quantile estimators.

CRUDE: This is the usual empirical cdf for Y; it is inverted and interpolated to obtain quantile estimators.

MLE: This is the probability estimator (8), which is based on the single control variate $C = \mathcal{I}(X \le x_q)$; it is inverted to obtain quantile estimators. This quantile estimator was called ILRT by Hsu and Nelson (1990), and was presented in §2.2.

MCV: This is the poststratified probability estimator based on s control variates $C^{(\ell)} = \mathcal{I}(X \le b_\ell)$ for $\ell = 1, 2, \ldots, s$; it is inverted to obtain quantile estimators. We consider $s = 2, 3$ and we use the approximate cutpoints described in §4.4.

EOPT: This is the probability estimator based on the estimated optimal control variate $\hat{C}^* = \Phi(\hat{a}_0 + \hat{a}_1 X)$ described in §4.1; it is inverted to obtain quantile estimators. The parameters a_0 and a_1 are estimated via a nonlinear least-squares regression of $\mathcal{I}(Y_i \le \hat{y}_q^{\text{ml}})$ on $\Phi(a_0 + a_1 X_i)$.

DIRECT: This is the quantile estimator (26) based on estimating the asymptotically optimal multiplier β_∞^* for the control variate $C = X$; it was presented in §5. There is no corresponding probability estimator.

To obtain the MLE, MCV and EOPT quantile estimators, we used the weighted-average interpretation of control variates (§3) to invert the corresponding probability estimator. The MLE is actually a special case of MCV with the single cutpoint $b_1 = x_q$.

Results for the following data models are presented.

BVN: (X, Y) are standard bivariate normal with correlation ρ. For this case, the functional relationship assumed by EOPT is correct. This model also allows us to vary the dependence between X and Y systematically, although we present results only for $\rho = 0.95$ here. It is worth noting that a correlation of 0.95 between X and Y seems quite large, but it only implies correlations of 0.75, 0.73, and 0.67 between $\mathcal{I}(Y \le y_q)$ and $\mathcal{I}(X \le x_q)$ when $q = 0.90$, 0.95, and 0.99, respectively. This illustrates why naive use of control variates to estimate probabilities and quantiles is typically less successful than using them to estimate means.

SAN1: Y is the time to complete a stochastic activity network, and X is the length of the path with the longest expected length. For this example,

$$Y = \max\{A_1 + A_2, A_1 + A_3 + A_5, A_4 + A_5\},$$

and $X = A_1 + A_3 + A_5$, where $A_1, A_2, \ldots, A_5$ are i.i.d. exponentially distributed random variables with common mean 1. This example was used by Hsu and Nelson (1990).

SAN2: Similar to the previous example,

$$Y = \max\{A_1 + A_2, A_1 + A_3 + A_5, A_4 + A_5\},$$

but in this case the A_i are i.i.d. gamma distributed with common shape parameter 2 and scale parameter 1. The control variate is

$$X = \max\{B_1 + B_2, B_1 + B_3 + B_5, B_4 + B_5\},$$

where the B_i are i.i.d. exponentially distributed with common mean 1, and each pair (A_i, B_i) is generated using the inverse cdf method and common random numbers. In other words, SAN1 is used as an external control variate for SAN2.

MM1: Y is the delay in queue of the tenth arrival to an $M/M/1$ queue that is initially empty, and X is the sum of the first nine service times. Specifically, let $G_1, G_2, \cdots$ be i.i.d. exponentially distributed random variables with mean $(0.9)^{-1}$, representing the interarrival-time gaps; and let $S_1, S_2, \cdots$ be i.i.d. exponentially distributed random variables with mean 1, representing the service times. Define $D_0 = 0$ and $S_0 = 0$. Then D_i, the delay in queue of the ith arriving customer, is defined by the recursive relationship

$$D_i = \max\{0, D_{i-1} + S_{i-1} - G_i\},$$

for $i = 1, 2, \ldots$. Therefore, $Y = D_{10}$ and $X = \sum_{i=1}^{9} S_i$. This example was also used by Hsu and Nelson (1990).

INVT: Y is the average cost for 30 periods in an (s, S) inventory system, and X is the standardized average demand for the same 30 periods. See Koenig and Law (1985) for a detailed description of the model, and

Wilson and Pritsker (1984) for the use of standardized averages as control variates. An important feature of this example is that the exact distribution of X is not known. However, because X is defined as

$$X = \frac{\sum_{i=1}^{m} (D_i - E[D])}{\sqrt{m \operatorname{Var}[D]}},$$

where D_i is the demand in period i, its distribution as $m \to \infty$ is $N(0, 1)$. Therefore, $\Pr\{X \le x\} \doteq \Phi(x)$ when m is large. Standardized averages are one way to avoid the requirement that we know the precise distribution of the control variate. In this example each D_i is Poisson with mean 25 and $m = 30$.

For BVN, SAN1, MM1, and INVT, the true values of the quantiles can be computed analytically, allowing us to evaluate the mean squared error (MSE) of the estimators, and also to use SAN1 as an external control variate. For SAN2, the true values of the quantiles were estimated via a simulation with $n = 100000$ replications. All simulations and analysis were performed in *S-PLUS*.

Tables 1 and 2 report the experiment results for probability and quantile estimation, respectively. In the probability experiments we estimated $\Pr\{Y \le y_q\}$, and in the quantile experiments we estimated y_q, for $q =$ 0.90, 0.95, and 0.99. Table 3 gives the sample sizes used for each basic experiment,[2] and all estimators were applied to the same data set. The basic experiment was then repeated 1000 times to estimate the MSE of each point estimator. Tables 1 and 2 report the ratio of the MSE of CRUDE to that of each of the improved estimators; therefore, a ratio greater than 1 shows an MSE reduction. Only significant digits (based on the simulation standard error) are reported. The same information is displayed graphically in Figure 5.

The overall conclusions from both tables are the same: All estimators achieve MSE reductions, with OPT achieving the greatest reductions. However, MCV with $s = 3$ is nearly as good as OPT while being much easier to apply. In fact, the nonlinear regression required for OPT implies so much additional computation that it may be inferior to MCV if computation time is consid-

Table 1 Ratio of the MSE of the Crude Probability Estimator to the MSE of Each Control-Variate Probability Estimator

Model	q	MLE	MCV $s = 2$	MCV $s = 3$	EOPT
BVN	0.90	2.4	2.7	2.9	3.0
$\rho = 0.95$	0.95	2.2	2.6	2.7	2.8
	0.99	2.0	2.2	2.3	2.4
SAN1	0.90	1.6	1.8	1.9	1.9
	0.95	1.5	1.7	1.8	1.8
	0.99	1.3	1.5	1.6	1.6
SAN2	0.90	3.1	3.7	3.9	4.0
	0.95	3.4	4.0	4.2	4.4
	0.99	3.5	4.5	4.7	5.1
MM1	0.90	1.3	1.3	1.3	1.4
	0.95	1.2	1.2	1.2	1.2
	0.99	1.2	1.2	1.2	1.3
INVT	0.90	1.7	2.0	2.0	2.3
	0.95	1.4	1.5	1.6	1.5
	0.99	1.7	1.8	1.8	1.9

Table 2 Ratio of the MSE of the Crude Quantile Estimator to the MSE of Each Control-Variate Quantile Estimator

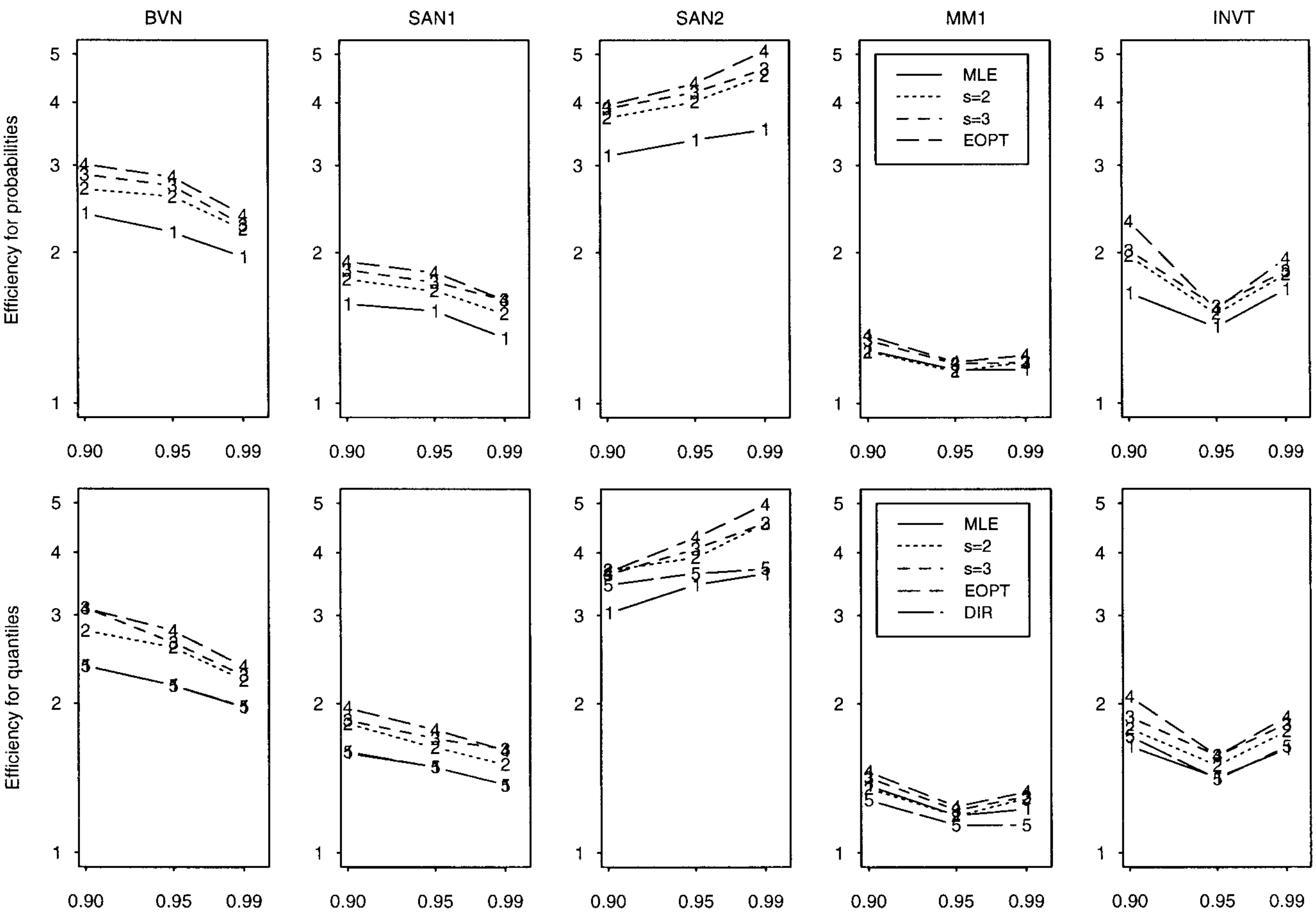
Model	q	MLE	MCV $s = 2$	MCV $s = 3$	EOPT	DIRECT
BVN	0.90	2.4	2.8	3.1	3.1	2.4
$\rho = 0.95$	0.95	2.2	2.6	2.6	2.8	2.2
	0.99	2.0	2.2	2.3	2.4	2.0
SAN1	0.90	1.6	1.8	1.9	2.0	1.6
	0.95	1.5	1.6	1.7	1.8	1.5
	0.99	1.4	1.5	1.6	1.6	1.4
SAN2	0.90	3.0	3.7	3.6	3.7	3.5
	0.95	3.5	3.9	4.1	4.3	3.6
	0.99	3.6	4.6	4.6	5.0	3.7
MM1	0.90	1.4	1.4	1.4	1.5	1.3
	0.95	1.2	1.2	1.2	1.2	1.1
	0.99	1.2	1.3	1.3	1.3	1.1
INVT	0.90	1.6	1.8	1.9	2.1	1.7
	0.95	1.4	1.5	1.6	1.6	1.4
	0.99	1.6	1.8	1.8	1.9	1.6

Table 3 Sample Size n for Each Value of q

q	n
0.9	1000
0.95	5000
0.99	10000

[2] The lone exception was the INVT experiments where the sample sizes were $n = 1000$, 1000 and 5000 for $q = 0.90$, 0.95, and 0.99, respectively. These sizes were chosen only because of the slow execution speed of the INVT experiments.

Figure 5 **Ratio of MSE of Crude Probability Estimator or Quantile Estimator to MSE of Each Control-Variate Estimator (Estimator** 1 = **MLE** = **MCV (***s* = 1), 2 = **MCV** (*s* = 2), 3 = **MCV** (*s* = 3), 4 = **OPT, and** 5 = **DIRECT)**



ered. However, by including OPT without discounting for computation we can compare the other estimators to the best that we can hope to achieve, at least for BVN data. DIRECT—which is available only for quantile estimation—performs about the same as the MLE, and therefore is not worth the additional work required to compute it.

Substantial reductions are possible when the correlation is strong, as in the SAN2 example, where the MSE of the crude estimator is three or more times greater than the MSE of the control-variate estimator. Recall that SAN2 employs an external control variate. The control mimics the response so closely that *Y* and *X* nearly always correspond to the same path through the network. On the other hand, very modest gains were achieved for the MM1 example. The results for SAN1 and SAN2 are similar to the results obtained by Avramidis and Wilson (1998) when they applied Latin Hypercube Sampling to estimate quantiles of the completion time of stochastic activity networks. They achieved MSE ratios ranging from 2.3–3.9 for their best estimator applied to estimate the 0.95 quantile.

7. Conclusions

We have presented new estimators that unify and extend previous control-variate estimators for probabilities and quantiles. These estimators either approximate the optimal transformation of a given control variate to maximize its correlation with the response, or they provide a method for (implicitly) estimating the optimal multiplier for a given control variate.

Gratifyingly, the simple discrete approximations to the optimal transformation—for which both probability and quantile estimators are easily obtained using the weighted-average interpretation described in §3—perform very well.

One important prerequisite for our estimators is the availability of a control variate whose distribution, or at least some quantile of it, is known. This is typically a more substantial requirement than finding a control variate with known mean. The use of a standardized average, as in the INVT example, is one way to circumvent the problem. However, the normal approximation may not be sufficiently accurate for estimating tail quantiles or probabilities of X if the number of terms in the sum is not large or if the random variables have highly skewed distributions. When the cumulant generating function of X is known, then we suggest using the saddlepoint approximation of Lugannani and Rice (see Daniels 1987 for estimating probabilities, or Hesterberg 1994 for estimating quantiles). When the cumulant generating function of X is not known but the skewness is, then we suggest using a translated gamma approximation $a + bG$, where G is gamma distributed with the desired $|$skewness$|$, and a and b are constants with b negative if the skewness is negative. An alternative is to use Edgeworth approximations, but they are known to perform poorly in the tails of distributions. Unfortunately, we do not have a satisfactory approximation that uses kurtosis or higher moments; the general saddlepoint approximations of Easton and Ronchetti (1986) or Wang (1992) should perform somewhat better than Edgeworth approximations, but they also have some difficulty with tail behavior.

Finally, we note that the weighted-average methods described in this paper can be combined with smoothing and interpolation (see, for instance, Dielman et al. 1994) and with importance sampling (Hesterberg 1993, 1996) for further variance reductions.[3]

Appendix

In this section we present the analysis that supports the recommendation in §4.4 for specifying multiple cutpoints for the poststratified probability and quantile estimator. This is based on approximately optimizing (24).

Suppose that the following hold in a neighborhood of x_q: X has a continuous, nonzero density; $Y = d_0 + d_1 X + \epsilon$ where $E[\epsilon | X = x] = o(x - x_q)$ and $d_1 > 0$; the distribution of ϵ is approximately independent of X; and ϵ is small (small first absolute moment). Our analysis is as σ_ϵ^2 approaches 0 (where σ_ϵ is any scale parameter for ϵ such as the standard deviation or $E[|\epsilon|]$), implying that Corr$[Y, X]$ approaches 1 in this neighborhood.

Two key consequences of these assumptions are that $y_q \doteq d_0 + d_1 x_q$ and that the cutpoints $b_\ell \doteq x_q$ for $\ell = 1, 2, \cdots s$. The optimal values of $b_1, b_2, \ldots, b_s$ approach x_q as σ_ϵ approaches 0 because

$$\Pr\{Y \le y_q | X \ll x_q\} \doteq 1,$$

$$\Pr\{Y \le y_q | X \gg x_q\} \doteq 0,$$

under these conditions. Therefore, strata well away from x_q would have little value in reducing estimator variance.

Based on these approximations

$$\begin{aligned} p_{\ell 0}(y_q, \mathbf{b}) &= \int_{b_\ell}^{b_{\ell+1}} f_X(x) \Pr\{d_0 + d_1 X + \epsilon \le y_q | X = x\} dx \\ &= \int_{b_\ell}^{b_{\ell+1}} f_X(x) F_{\epsilon|x}(y_q - (d_0 + d_1 x)) dx \\ &\doteq f_X(x_q) \int_{b_\ell}^{b_{\ell+1}} F_{\epsilon|x_q}(-d_1(x - x_q)) dx \qquad (30) \\ &= f_X(x_q) \frac{\sigma_\epsilon}{d_1} \int_{c_\ell}^{c_{\ell+1}} F_{\epsilon|x_q}(-\sigma_\epsilon u) du, \qquad (31) \end{aligned}$$

for $\ell = 1, 2, \ldots, s$, where $u = d_1(x - x_q)/\sigma_\epsilon$, $c_\ell = d_1(b_\ell - x_q)/\sigma_\epsilon$, and $F_{\epsilon|x}$ is the conditional distribution of ϵ given $X = x$. For $\ell = 1, 2, \ldots, s - 1$ the approximation in (30) follows by approximating f_X and $F_{\epsilon|x}$ by their values at $x = x_q$, since $x \doteq x_q$ for $b_1 \le x \le b_s$. For $\ell = s$, we note that the integrand $F_{\epsilon|x}(-d_1(x - x_q))$ is near zero except where $x \doteq x_q$, since $d_1 > 0$ and σ_ϵ is very small.

The approximation can be made rigorous by taking the limit of $p_{\ell 0}(y_q, \mathbf{b})/\sigma_\epsilon$ as $\sigma_\epsilon \to 0$ and $b_\ell \to x_q$ for $\ell = 1, 2, \ldots, s - 1$, with c_ℓ fixed. The result for $\ell = s$ follows by splitting the integral over c_s to $c_{s+1} = \infty$ into two pieces, say at $c_s + \sigma_\epsilon^{-1/2}$. The rigorous result requires certain technical conditions on how $F_{\epsilon|x}$ depends on x.

Similarly,

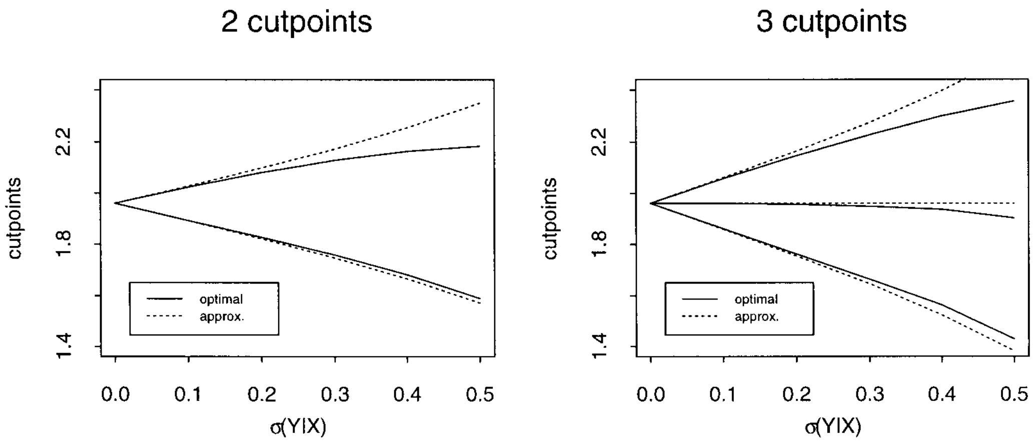
$$p_{\ell 1}(y_q, \mathbf{b}) \doteq f_X(x_q) \frac{\sigma_\epsilon}{d_1} \int_{c_\ell}^{c_{\ell+1}} (1 - F_{\epsilon|x_q}(-\sigma_\epsilon u)) du, \qquad (32)$$

for $\ell = 0, 1, \ldots, s - 1$. Finally, $p_{00}(y_q, \mathbf{b}) \doteq q$ and $p_{s1}(y_q, \mathbf{b}) \doteq 1 - q$ as σ_ϵ approaches 0 and b_1 and b_s approach x_q.

Now let $V = \epsilon/\sigma_\epsilon$ be a standardized version of ϵ, and let $F_{V|x_q}$ be the distribution of V given $X = x_q$. Then $F_{\epsilon|x_q}(-\sigma_\epsilon u) = F_{V|x_q}(-u)$. Substituting this result into (31) and (32), we see that (24) is approximately proportional to an expression which depends solely on integrals of $F_{V|x_q}$ between standardized cutpoints c_ℓ for $\ell = 0, 1, \ldots, s + 1$. The first order term of this expression (corresponding to the $O(n^{-1})$ term of (24)) is

[3] The authors thank the associate editor, referee, and James R. Wilson for their help in revising the paper. Nelson's work was partially supported by National Science Grant No. DMI-9622065.

Figure 6 **Optimal Cutpoints $\{b_\ell\}$ and Asymptotic Approximation to Them When X and Y are Bivariate Normal as a Function of the Conditional Standard Deviation $\sigma_\epsilon = \sqrt{1 - p^2}$ of Y given X**



$$\int_{-\infty}^{c_1} (1 - F_{V|x_q}(-u))du$$

$$+ \sum_{\ell=1}^{s-1} \frac{\int_{c_\ell}^{c_{\ell+1}} F_{V|x_q}(-u)du \int_{c_\ell}^{c_{\ell+1}} (1 - F_{V|x_q}(-u))du}{c_{\ell+1} - c_\ell}$$

$$+ \int_{c_s}^{-\infty} F_{V|x_q}(-u)du. \tag{33}$$

Key to (33) are the approximations

$$\Pr\{X < b_1\} \doteq q \doteq p_{00}(y_q, \mathbf{b}),$$

$$\Pr\{b_\ell < X \leq X_{\ell+1}\} \doteq (b_{\ell+1} - b_\ell) f_X(x_q),\ \ell = 1, 2, \ldots, s - 1,$$

$$\Pr\{X > b_s\} \doteq 1 - q \doteq p_{s1}(y_q, \mathbf{b}),$$

as $b_1, b_2, \ldots, b_s$ approach x_q.

When ϵ is normally distributed and σ_ϵ is its standard deviation, so that $F_{V|x_q}(u) = \Phi(u)$, the values of c_ℓ that minimize (33) are

$$c_\ell = \begin{cases} -0.674, 0.674 & \text{for } s = 2, \\ -1, 0, 1 & \text{for } s = 3. \end{cases} \tag{34}$$

For two cutpoints, this is the same as choosing cutpoints so that the conditional probability $\Pr\{Y \leq y_q \mid X \leq b_\ell\}$ equals 0.25 for b_1 and 0.75 for b_2 (asymptotically, as $\sigma_\epsilon \to 0$). For three cutpoints the conditional probabilities are 0.16, 0.5, and 0.84, respectively. Figure 6 shows the optimal cutpoints as a function of σ_ϵ in the bivariate normal problem, together with curves that correspond to (34). The agreement between the optimal cutpoints and the recommended approximation is very close. All optimizations were performed using standard Newton optimization in *S-PLUS*. Integrals were evaluated using $\int_a^b \Phi(u)du = b\Phi(b) - a\Phi(a) + \phi(b) - \phi(a)$ where ϕ is the standard normal density, and using the *S-PLUS* built-in functions for Φ and ϕ.

This analysis leads to the method for choosing cutpoints of the form $b_\ell = x_q + c_\ell \hat{\sigma}_\epsilon / \hat{d}_1$ described in §4.4. The parameters $\hat{d}_1$ and $\hat{\sigma}_\epsilon$ can be chosen to be the slope and residual standard deviation from a global nonlinear or local linear regression of Y on X. The values of c_ℓ can be those that are optimal for a normal distribution, or they can be chosen by fitting a nonnormal distribution to the residuals and optimizing for that distribution. However, the simpler criterion of choosing equally spaced cs with mean 0 and sample variance 1 should be adequate in practice to capture most of the available variance reduction.

References

Avramidis, A. N., "Variance Reduction for Quantile Estimation via Correlation Induction," in *Proc.* 1992 *Winter Simulation Conf.*, J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson (eds.), IEEE, Piscataway, NJ, 572–576, 1992.

—— and J. R. Wilson, "Correlation-Induction Techniques for Estimating Quantiles in Simulation Experiments," *Operations Res.* forthcoming.

Breiman, L. and J. H. Friedman, "Estimating Optimal Transformations for Multiple Regression and Correlation," *J. American Statistical Association*, 80 (1985), 580–619.

Chambers, J. M. and T. J. Hastie, *Statistical Models in S*, Wadsworth, Pacific Grove, CA, 1985.

Cochran, W. G., *Sampling Techniques*, Third Edition, John Wiley, New York, 1977.

Daniels, H. E., "Tail Probability Approximations," *International Statistical Review*, 55 (1987), 37–48.

David, H. A., *Order Statistics*, Second Edition, John Wiley, New York, 1981.

Davidson, R. and J. G. MacKinnon, "Efficient Estimation of Tail-Area Probabilities in Sampling Experiments," *Economics Letters*, 8 (1981), 73–77.

—— and ——, "Regression-Based Methods for Using Control Variates in Monte Carlo Experiments," *J. Econometrics*, 54 (1992), 203–222.

Dielman, T., C. Lowry, and R. Pfaffenberger, ''A Comparison of Quantile Estimators,'' *Comm. in Statistics*, B23 (1994), 355–371.

Dobson, A. J., *An Introduction to Generalized Linear Models*. Chapman and Hall, New York, 1990.

Efron, B. and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.

Easton, G. S. and E. Ronchetti, ''General Saddlepoint Approximations with Applications to *L* Statistics,'' *J. American Statistical Association*, 81 (1986), 420–429.

Fieller, E. C. and H. O. Hartley, ''Sampling with Control Variables,'' *Biometrika*, 41 (1954), 494–501.

Hastie, T. J. and R. J. Tibshirani, *Generalized Additive Models*, Chapman and Hall, New York, 1990.

Hesterberg, T. C., ''Control Variates and Importance Sampling for the Bootstrap,'' in *Proc. Statistical Computing Section of the American Statistical Association* (1993), 40–48.

——, ''Saddlepoint Quantiles and Distribution Curves, with Bootstrap Applications,'' *Computational Statistics*, 9 (1994), 207–212.

——, ''Control Variates and Importance Sampling for Bootstrap Tail Estimation,'' *Statistics and Computing*, (1996),

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

Koenig, L. and A. M. Law, ''A Procedure for Selecting a Subset of Size *m* Containing the ℓ Best of *k* Independent Normal Populations with Applications to Simulation,'' *Comm. in Statistics*, B14 (1985), 719–734.

Nelson, B. L., ''Control Variate Remedies,'' *Oper. Res*., 38 (1990), 974–992.

Rao, C. R., *Linear Statistical Inference and Its Applications*, John Wiley, New York, 1973.

Ressler, R. L. and P. A. W. Lewis, ''Variance Reduction for Quantile Estimates in Simulations via Nonlinear Controls,'' *Comm. in Statistics*, B19 (1990), 1045–1077.

Rothery, P., ''The Use of Control Variates in Monte Carlo Estimation of Power,'' *Applied Statistics*, 31 (1982), 125–129.

Wang, S., ''General Saddlepoint Approximations in the Bootstrap,'' *Statistics and Probability Letters*, 13 (1992), 61–66.

Weiss, L., ''On the Asymptotic Joint Normality of Quantiles from a Multivariate Distribution,'' *J. National Bureau of Standards*, 68B (1964), 65–66.

Wilson, J. R. and A. A. B. Pritsker, ''Variance Reduction in Queueing Simulation Using Generalized Concomitant Variables,'' *J. Statistical Computation and Simulation*, 19 (1984), 129–153.

Accepted by Pierre L'Ecuyer; received October 1995. *This paper has been with the authors* 11 *months for* 2 *revisions*.