Proceedings of the 1987 Winter Simulation Conference
A. Thesen, H. Grant, W. David Kelton (eds.)

CONTROL VARIATES FOR QUANTILE ESTIMATION

Jason C. Hsu
Department of Statistics
The Ohio State University
Columbus, OH 43210

Barry L. Nelson
Department of Industrial & Systems Engineering
The Ohio State University
Columbus, OH 43210

ABSTRACT

Some new quantile estimators that employ a control variate are introduced. The properties of these estimators do not depend on the usual assumption of joint normality between the random variable of interest and the control. Empirical results are presented.

1. INTRODUCTION

Let Y be a random variable with an unknown distribution F_Y, but for which realizations can be obtained. This paper considers estimating the value y_q such that $P\{Y \leq y_q\} = q$ for prespecified q $(0 < q < 1)$. (We assume Y is absolutely continuous at y_q.) The value y_q is called the *qth* quantile of Y. Much of the literature on simulation output analysis concentrates on estimating E[Y], the long run average of Y. Quantiles provide additional information about the distribution of Y. In fact, in some problems, instead of the expected value of Y, the quantiles of Y are the parameters of primary interest.

For example, Y could be a proposed test statistic whose distribution under the null hypothesis is difficult to evaluate numerically. One might then be interested in estimating the critical values $y_{.90}$, $y_{.95}$, $y_{.99}$ by simulating Y under the null hypothesis.

As a second example, Y might be the delay in queue experienced by a customer arriving to a service system. Then 50% of the customers experience delays less than $y_{.50}$, but 5% of the customers experience delays longer than $y_{.95}$.

Straightforward estimation of y_q is based on the order statistics of the Y's (see Section 2 below). However, sometimes one can observe a second, control, random variable X which is statistically dependent on Y and whose *qth* quantile x_q is known. Section 3 presents improved estimators based on simulated pairs (X,Y) and x_q. Section 4 presents the results of empirical comparisons. Some conclusions are offered in Section 5.

2. THE STANDARD METHOD

Let $Y_1, Y_2, \ldots, Y_n$ be an independent and identically distributed (*i.i.d.*) sample from a distribution F_Y that is absolutely continuous at y_q. Let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the sample Y values ordered from smallest to largest; these are the *order statistics* of the sample. If k = [nq] +1, where [·] is the largest integer function, then $Y_{[k]}$ is a standard estimator of y_q (see David 1981 and Juritz, Juritz and Stephens 1983 for properties of this estimator). In practice, instead of being restricted to a particular order statistic, one may want to interpolate. In this study we utilize the *quantile* function of the S statistical package (Becker and Chambers 1984), in which $Y_{(i)}$ is taken to be the $\frac{i - .5}{n}$*th* sample quantile, and linear interpolation is employed elsewhere (except when nq < 0.5, in which case the estimator is $Y_{(1)}$, or $nq > 1 - \frac{.5}{n}$, in which case the estimator is $Y_{(n)}$). We call this interpolated estimator the "no control variate" (No CV) estimator.

From a different point of view, a uniformly best estimator for y_q among median unbiased estimators based on the Y's that assumes no knowledge of F_Y can be obtained by inverting one-sided sign tests (Lehmann 1986, pp. 94-95 and pp. 120-121). A median unbiased estimator is defined by the property that it is as likely to be greater than the true parameter value as it is likely to be less, i.e., the true parameter is the median of the estimator. This best median unbiased estimator is typically an estimator that randomizes between $Y_{(k)}$ and $Y_{(k-1)}$ or $Y_{(k+1)}$. However, by the Rao-Blackwell Theorem, a nonrandomized version with smaller risk relative to any convex loss function (such as the mean square error) can be obtained by taking the conditional expectation with respect to some sufficient statistic, the set of order statistics in this case. The resulting nonrandomized estimator is then a linear combination of $Y_{(k)}$ and $Y_{(k-1)}$ or $Y_{(k+1)}$. This nonrandomized estimator, which is no longer exactly median unbiased, is typically different from No CV, but probably not by much. Thus, the estimator No CV can be thought of as approximately the best median unbiased estimator based on Y's only.

3. CONTROL VARIATE ESTIMATORS

Control variates (CVs) is a well known variance reduction technique that estimates some characteristic of Y by exploiting knowledge about a random variable X which can be observed simultaneously with Y, and which is statistically dependent on Y. See Bratley, Fox, and Schrage (1987) for an introduction to CVs. We now assume that there exists an X such that (X,Y) has a joint distribution $F_{X,Y}$ which is absolutely continuous at (x_q, y_q), where the *qth* quantile x_q of the marginal distribution of X is known, and

$$(X_1,Y_1), (X_2,Y_2), \dots, (X_n,Y_n),$$

an *i.i.d.* sample of (X,Y), can be observed. In this section we develop estimators of y_q based on simulated pairs (X,Y) and x_q.

Let $X_{(1)} < X_{(2)} < \dots < X_{(n)}$ denote the order statistics of the X's and let $Y_{(1)} < \cdots < Y_{(n)}$ denote the order statistics of Y's. We also let $X_{(0)} = X_{(1)}$, $Y_{(0)} = Y_{(1)}$, $X_{(n+1)} = X_{(n)}$, and $Y_{(n+1)} = Y_{(n)}$, for convenience.

3.1 A Regression-Based Estimate

In addition to the familiar concepts of correlation, there are several other concepts of bivariate dependence that are relevant to our problem. The following definitions and results can be found in Tong (1980).

Definition 3.1. Y is **positively regression dependent** on X if $P\{Y \le y \mid X = x\}$ is nonincreasing in x for all y, i.e., the family of conditional distributions $P\{Y \le y \mid X = x\}$ indexed by x is stochastically increasing in x.

Definition 3.2. X and Y are **associated** if $\mathrm{Cov}(g_1(X,Y), g_2(X,Y)) \ge 0$ for all g_1 and g_2 monotonically nondecreasing in each argument.

Lemma 3.1 If Y is positively regression dependent on X, then X and Y are associated.

Lemma 3.2. Nondecreasing functions of associated random variables are associated.

It is not unusual for an X to exist such that Y is positively regression dependent on X. Thus, by Lemma 3.1, X and Y are associated. Then by Lemma 3.2, $X_{(k)}$ and $Y_{(k)}$ from a random (*i.i.d.*) sample of (X,Y) are also associated (see Theorem 5.2.2 of Tong 1980). Finally, by the definition of association, $\mathrm{Cov}(X_{(k)}, Y_{(k)}) \ge 0$. Thus, if we assume that, as is sometimes done for simplicity in practice, $E[Y \mid X=x]$ is nearly linear, then we obtain the classical control variate estimate $Y_{(k)} - \beta(X_{(k)} - x_q)$ (e.g. Hammersley and Handscomb 1964, Chapter 5; Bratley, Fox and Schrage 1987, Chapter 2), which under the assumption of positive regression dependence might be expected to do better than the estimator $Y_{(k)}$. In this paper we fix $\beta = 1$, which is equivalent to assuming that the regression has slope 1; estimating the optimal value of β requires partitioning the size n sample into subsamples, and we consider single sample estimators here. We refer to this estimator as the "regression estimator" (Reg).

3.2 A Maximum Likelihood Estimator

Two familiar general methods of estimating an unknown parameter θ are as follows. One is the maximum likelihood method. In this section, it is shown that, even with no knowledge of the joint distribution of X and Y, it is still possible to apply the maximum likelihood method to estimate y_q to some extent. Another general method of estimation is to base the estimator on tests. The Hodges-Lehmann method derives estimators of θ by considering statistical tests for all possible hypothesized values of θ, and, having observed the data, setting the estimator to be the value θ^* for which the observed p-value of the test is maximum, i.e., p-value($\mathbf{H}$: $\theta = \theta^*$) = $\max_{\theta'}$ p-value($\mathbf{H}$: $\theta = \theta'$). The more powerful the family of tests, the more efficient the estimator. In Sections 3.3 and 3.4, using the Hodges-Lehmann method, we derive estimators of y_q by inverting tests for hypothesized values of y_q based on observed pairs (X,Y), and x_q.

We can visualize the observed data in the (X,Y) plane as follows. Each hypothesized value c of y_q corresponds to a horizontal line Y = c which, together with the known vertical line $X = x_q$, divide the (X, Y) plane into four quadrants (see, for example, Figure 1). For notation, let

$N_{00}(c)$ = number of (X, Y) with $X \le x_q$ and $Y \le c$

$N_{01}(c)$ = number of (X, Y) with $X \le x_q$ and $Y > c$

$N_{10}(c)$ = number of (X, Y) with $X > x_q$ and $Y \le c$

$N_{11}(c)$ = number of (X, Y) with $X > x_q$ and $Y > c$.

The $N_{ij}(c)$, i, j = 1, 2, are random variables. Let $n_{00}(c)$, $n_{01}(c)$, $n_{10}(c)$, $n_{11}(c)$ be their realized values. Intuitively, if no knowledge is assumed concerning the joint distribution $F_{X,Y}$ of X and Y, then the essential information concerning y_q is contained in the four numbers N_{00}, N_{01}, N_{10}, and N_{11}.

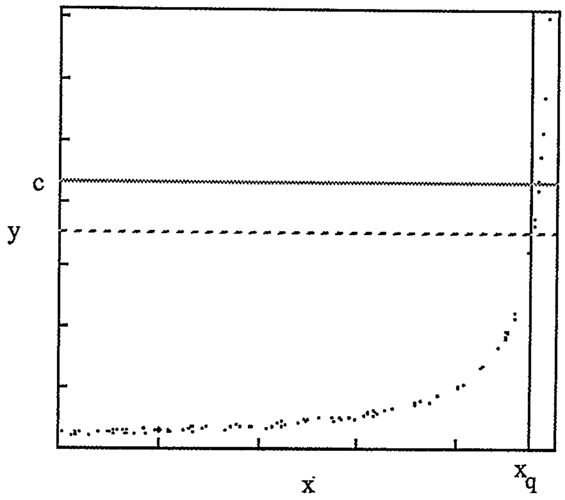

Figure1: Sample Scatter Plot of 100 (x,y) Pairs

For a hypothesized value c of y_q, let

$p_{00}(c) = P\{X \le x_q \text{ and } Y \le c\}$

$p_{01}(c) = P\{X \le x_q \text{ and } Y > c\}$

$p_{10}(c) = P\{X > x_q \text{ and } Y \le c\}$

$p_{11}(c) = P\{X > x_q \text{ and } Y > c\}$.

If c is the true value of y_q, then $p_{01}(c) = p_{10}(c) = p$ (say). Therefore the likelihood function of y_q is

$$\binom{n}{n_{00}(c)\ n_{01}(c)\ n_{10}(c)\ n_{11}(c)} \cdot$$

$$(q-p)^{n_{00}(c)} p^{[n_{01}(c)+n_{10}(c)]} (1-q-p)^{n_{11}(c)}$$

$$= k(n;c)g(n;c,p) \qquad (1)$$

where k(n;c) is the multinomial term and g(n;c,p) is the product of probabilities. Notice that p is a nuisance parameter and c is the parameter of interest.

Let $0 \le p^* \le 1$ and $c^* \in \{1, \ldots, n\}$ be the values of p and c that maximize (1). No closed-form expression for (p^*, c^*) has been found. However, for fixed c, the value of p, say $p^*(c)$, that maximizes (1) is

$$\frac{(q(n-n_{00}(c)) - (1-q)(n-n_{11}(c)))^2}{2n} -$$

$$\frac{[(q(n-n_{00}(c)) - (1-q)(n-n_{11}(c)))^2 + 4q(1-q)n_{11}(c)n_{00}(c)]^{1/2}}{2n}$$

This is easily derived since, for fixed c, k(n;c) is a constant and $\frac{\partial g(n;c,p)}{\partial p} = 0$ is a quadratic in p. An efficient recursion exists for determining k(n;c) from k(n-1;c), which leads to an algorithm that steps through the possible values of c, determines $p^*(c)$ and the corresponding value of (1) for each c, and sets c^* equal to the value that maximizes (1). This estimator, obtained by maximizing this likelihood function with respect to y_q, will be referred to as Maximum Likelihood Estimator 2 (MLE 2).

3.3 An Approximately Median Unbiased Estimator

The hypothesis **H:** $y_q = c$ is the same as **H:** $p_{00}(c) + p_{01}(c) = p_{00}(c) + p_{10}(c)$ (= q) or, equivalently, **H:** $p_{01}(c) = p_{10}(c)$. Thus, estimators of y_q can be obtained by tests of the hypotheses **H:** $p_{01}(c) = p_{10}(c)$.

An approximately median unbiased estimator of y_q is derived in this section by inverting uniformly most powerful unbiased tests for **H:** $p_{01}(c) = p_{10}(c)$

A different estimator of y_q is derived in the next section by inverting the likelihood ratio test for **H:** $p_{01}(c) = p_{10}(c)$.

To motivate the median unbiased estimator, consider Figure 1 which is a plot of a random sample of 100 pairs of X and Y; X and Y are strongly dependent. The vertical solid line is X = $x_{.95}$, the known 95th percentile of X. The horizontal wiggled line represents a candidate for $y_{.95}$, the unknown 95th percentile of Y. To estimate $y_{.95}$ based on the Y's **alone**, we would put the estimate somewhere between $Y_{(95)}$ and $Y_{(96)}$. Observe, however, that while the expected number of X's > $x_{.95}$ is 5, in this sample there are 8 X's > $x_{.95}$. Because X and Y are strongly dependent, one would guess that the number of Y's > $y_{.95}$ in this sample is also 8, which would put $y_{.95}$ somewhere between $Y_{(92)}$ and $Y_{(93)}$, as indicated by the dashed horizontal line. More generally, a large difference between the number of X's > x_q and the number of Y's > c is evidence against the candidate value c for y_q. This idea forms the basis for the median unbiased estimator, derived below.

Lemma 3.3. A uniformly most powerful unbiased (UMPU) size-α test ϕ_c^- for **H:** $y_q = c$ versus **K:** $y_q > c$ based on $N_{00}(c)$, $N_{01}(c)$, $N_{10}(c)$, $N_{11}(c)$ exists. It is McNemar's test, which is a conditional test that rejects for small values of $N_{10}(c)$, conditional on $N(c) = N_{01}(c) + N_{10}(c)$. Let b be the number such that

$$\alpha_{b-1} = \sum_{a \le b-1} \binom{n(c)}{a} \left(\frac{1}{2}\right)^{n(c)} < \alpha$$

$$\le \sum_{a \le b} \binom{n(c)}{a} \left(\frac{1}{2}\right)^{n(c)} = \alpha_b,$$

and let $\gamma = (\alpha - \alpha_{b-1})/[\binom{n(c)}{b}\left(\frac{1}{2}\right)^{n(c)}]$ so that $(1-\gamma)\cdot\alpha_{b-1} + \gamma\cdot\alpha_b = \alpha$. Let U be an independent uniform (0,1) random variable. Then the UMPU test rejects if $n_{10}(c) < b$, or if $U < \gamma$ when $n_{10}(c) = b$.

Proof. The hypothesis and alternative **H**: $y_q = c$ versus **K**: $y_q > c$ is the same as **H**: $q = p_{00}(c) + p_{01}(c) = p_{00}(c) + p_{10}(c)$ versus **K**: $q = p_{00}(c) + p_{01}(c) > p_{00}(c) + p_{10}(c)$ or equivalently **H**: $p_{01}(c) = p_{10}(c)$ versus **K**: $p_{01}(c) > p_{10}(c)$, for which the one-sided McNemar's test is UMPU (Lehmann 1986, Section 4.9).

Similarly, the UMPU test ϕ_c^+ for **H**: $y_q = c$ versus **K**: $y_q < c$ based on $N_{00}(c)$, $N_{01}(c)$, $N_{10}(c)$, $N_{11}(c)$ is the one-sided McNemar's test which rejects if $n_{01}(c) < b$, or if $U < \gamma$ when $n_{01}(c) = b$.

Thus, by the usual correspondence between tests and confidence sets (Lehmann 1986, Theorem 3.4), $y_q^- = \inf\{c \mid$ **H**: $y_q = c$ is accepted by $\phi_c^-\}$ is a level $1-\alpha$ lower confidence bound for y_q. Likewise, $y_q^+ = \sup\{c \mid$ **H**: $y_q = c$ is accepted by $\phi_c^+\}$ is a level $1-\alpha$ upper confidence bound for y_q. One way to derive a median unbiased estimator (an estimator which is as likely to be greater than the true parameter value as it is likely to be less) is to look for a common value of y_q^- and y_q^+ when $\alpha = 1/2$ (Lehmann 1986, pp. 94-95). Let $m = n_{00} + n_{01}$ (which does not depend on the hypothesized value c of y_q). When $\alpha = 1/2$, each **H**: $y_q = c$ with $c < Y_{(m)}$ gives $n_{01}(c) > n_{10}(c)$ and hence is rejected by ϕ_c^-. Each **H**: $y_q = c$ with $c > Y_{(m+1)}$ gives $n_{01}(c) < n_{10}(c)$ and is hence rejected by ϕ_c^+. Thus $c \in (Y_{(m)}, Y_{(m+1)})$ are the only candidate estimates. Every **H**: $y_q = c$ with $c \in (Y_{(m)}, Y_{(m+1)})$ gives $n_{01}(c) = n_{10}(c)$. Thus, **H**: $y_q = c$ is accepted or rejected by ϕ_c^- depending on whether its auxiliary random variable U for ϕ_c^- is > 1/2 or not. The same **H**: $y_q = c$ is also accepted or rejected by ϕ_c^+ depending on whether its auxiliary random variable U for ϕ_c^+ is > 1/2 or not. So if the same auxiliary random variable U is employed for all the tests, then y_q^- and y_q^+ have a common value which is either $Y_{(m)}$ or $Y_{(m+1)}$ with probability 1/2 each. This randomized estimator can be expected to have good properties, as it is derived from UMPU tests. However, randomization is not very appealing in practice. Further, according to the Rao-Blackwell Theorem (Lehmann 1983, pp. 50-51), a nonrandomized version with smaller risk (expected loss) relative to any strictly convex loss function (e.g., mean square error) can be obtained by taking the conditional expectation with respect to a sufficient statistic. The Rao-Blackwellized estimator is $(Y_{(m)} + Y_{(m+1)})/2$. In our study, instead of taking the midpoint of $Y_{(m)}$ and $Y_{(m+1)}$, we linearly interpolate between $(X_{(m)}, Y_{(m)})$ and $(X_{(m+1)}, Y_{(m+1)})$ at x_q. (Also, when $m = n_{00} + n_{01} = n$, we take the estimate of y_q to be $Y_{(n)}$. When $m = n_{00} + n_{01} = 0$, we take the estimate of y_q to be $Y_{(1)}$.) In this study we refer to this approximately median unbiased interpolated estimator as Med Unb.

3.4 A Cell-Probability Based Maximum Likelihood Estimator

For fixed c, allowing for the possibility that **H**: $y_q = c$ is false, the likelihood function, as a function of $(p_{00}, p_{01}, p_{10}, p_{11})$ and $(n_{00}(c), n_{01}(c), n_{10}(c), n_{11}(c))$, is proportional to

$$p_{00}^{n_{00}(c)}\, p_{01}^{n_{01}(c)}\, p_{10}^{n_{10}(c)}\, p_{11}^{n_{11}(c)}$$

$$= (q-p_{01})^{n_{00}(c)}\, p_{01}^{n_{01}(c)}\, p_{10}^{n_{10}(c)}\, (1-q-p_{10})^{n_{11}(c)}$$

The maximum likelihood estimators of p_{01} and p_{10} are

$$\hat{p}_{01} = \frac{n_{01}}{n_{00}+n_{01}}\, q$$

$$\hat{p}_{10} = \frac{n_{10}}{n_{10}+n_{11}}\, (1-q)$$

with asymptotic variance-covariance matrix

$$n^{-1}\begin{pmatrix} \frac{p_{01}(q-p_{01})}{q} & 0 \\ 0 & \frac{p_{10}(1-q-p_{10})}{1-q} \end{pmatrix}.$$

The asymptotic likelihood ratio test for **H**: $p_{01}(c) = p_{10}(c)$ therefore rejects if $| \hat{p}_{01} - \hat{p}_{10}|/\sqrt{\frac{\hat{p}_{01}(q-\hat{p}_{01})}{q} + \frac{\hat{p}_{10}(1-q-\hat{p}_{10})}{1-q}} > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)$*th* quantile of the standard normal distribution. The Hodges-Lehmann estimator based on this test is then the value c such that

$$| \hat{p}_{01} - \hat{p}_{10}|/\sqrt{\frac{\hat{p}_{01}(q-\hat{p}_{01})}{q} + \frac{\hat{p}_{10}(1-q-\hat{p}_{10})}{1-q}}$$

is minimum (approximately 0). Note that $\hat{p}_{01} - \hat{p}_{10}$ is a step function which decreases as c increases passed each $Y_{(j)}$. We take our estimate to be the largest c such that $\hat{p}_{01} - \hat{p}_{10}$ is nonnegative, and refer to this estimate as MLE 1.

4. EMPIRICAL RESULTS

In this section we compare the five estimators by simulation using a variety of examples. The results are summarized using boxplots, as well as the more traditional measures of performance: mean square error (MSE), variance (Var), and bias (Bias). The box in a boxplot contains the middle half of the data (i.e., from the *.25th* sample quantile to the *.75th* sample quantile); a horizontal line is drawn through the box at the median of the data; the whiskers extending from the box reach to the most extreme non-outlier; outliers are plotted individually by "*". All numerical studies were done on a Pyramid 90x super mini-computer.

4.1 Linear Dependence Between X and Y

The first example is $Y = X + \varepsilon$, where X is standard normal, and ε is normal with mean 0 and standard deviation 0.2. Thus Y is positively regression dependent on X, and the dependence is linear with a slope of 1. Figure 2 is a sample scatter plot of 1000 pairs of x and y.

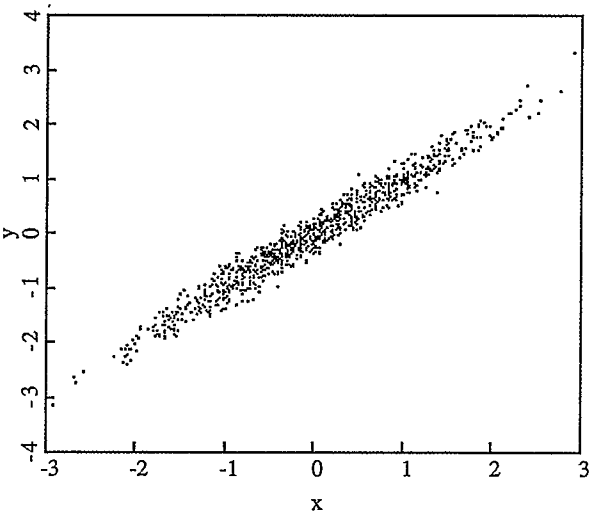

Figure 2: Scatter Plot of 1000 (x,y) Pairs in Example 4.1

Using the statistical package S , 40,000 *i.i.d.* pairs of X and Y were generated. We first divided the 40,000 pairs into 100 sets of samples of size n = 400, and applied the five estimators to each of the samples of 400 to estimate $y_{.95}$. Figure 3 shows the boxplots of the five estimators. In Figure 3, a horizontal line is drawn through the entire plot at $y_{.95}$, the true *.95th* quantile of y, which is $1.645\sqrt{1.04} = 1.67758$.

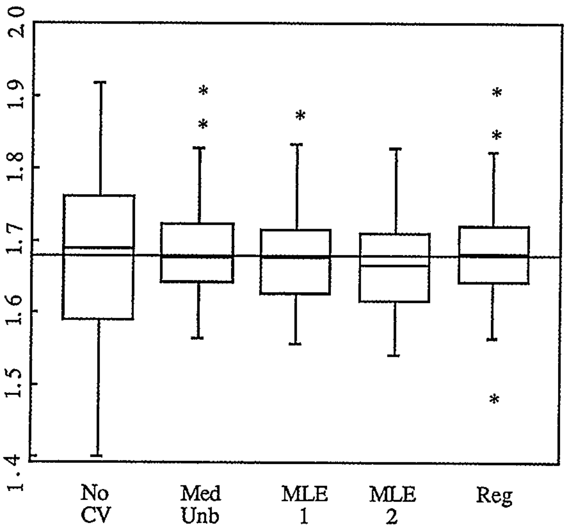

Figure 3: Boxplots of Estimates of $y_{.95}$ (n = 400)

Figure 3 indicates all four control variate methods do better than the no control variate estimate (No CV). The regression estimate (Reg) and the median unbiased estimate (Med Unb) perform about the same, both better than the others. The five estimators are also compared in terms of mean square error in Table 1 below, which further breaks down mean square error into variance and bias (MSE = Var + Bias^2). In terms of MSE, the median unbiased estimator (Med Unb) is somewhat worse than the regression estimator (Reg), but again better than the other estimators.

Table 1: MSE, Var, and Bias of Estimates of $y_{.95}$ (n = 400)

	No CV	Med Unb	MLE 1	MLE 2	Reg
MSE	0.01277	0.00472	0.00481	0.00488	0.00398
Var	0.01276	0.00471	0.00481	0.00479	0.00395
Bias	0.00155	0.00310	0.00167	-0.00958	0.00495

To check how the sample size n affects the relative performance of the estimators, we then divided the 40,000 (x,y) pairs into 400 sets of samples of size n = 100, and applied the five estimators to each sample of 100 to estimate $y_{.95}$. Figure 4 shows the boxplots of the 400 sets of estimates that resulted. Figure 4 indicates that, when the sample size n is smaller, the performance of the median unbiased estimator (Med Unb) and the regression estimator (Reg) remain roughly the same between them, but become even better relative to the other estimators. This is also true in terms of MSE, given in Table 2 below. Both MLE 1 and MLE 2 exhibit median bias in Figure 4.

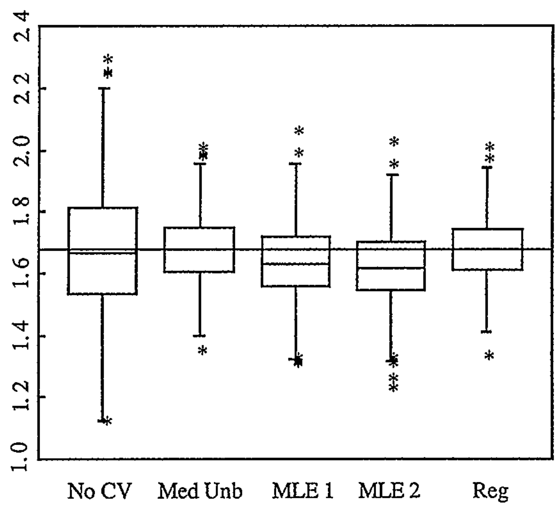

Figure 4: Boxplots of Estimates of $y_{.95}$ (n = 100)

Table 2: MSE, Var, and Bias of Estimators of $y_{.95}$ (n = 100)

	No CV	Med Unb	MLE 1	MLE 2	Reg
MSE	0.04638	0.01120	0.01686	0.01939	0.00981
Var	0.04638	0.01120	0.01532	0.01636	0.00980
Bias	0.00126	0.0022	-0.0392	-0.05509	0.00143

At least in this example, in which the linear functional relationship between X and Y assumed by the regression estimate (Reg) is exactly correct, not much is lost by using the median unbiased estimate (Med Unb) which does not assume a known functional relationship between X and Y. The next example compares the estimators when the relationship between X and Y has a large curvature around (x_q, y_q). In particular, it shows that whereas the regression estimator (Reg) can behave very badly, the median unbiased estimator (Med Unb) continues to do well.

4.2 Nonlinear Dependence Between X and Y

The second example is $Y = (\frac{1}{1.01-X} - \varepsilon)/100$ where X is Uniform (0,1) and ε is Uniform (0,1/2). Again, Y is positively regression dependent on X, but the dependence, especially around the quantiles $(x_{.95}, y_{.95})$ = (0.95, 0.16418), is highly nonlinear. Figure 5 is a sample scatter plot of 1000 (x,y) pairs.

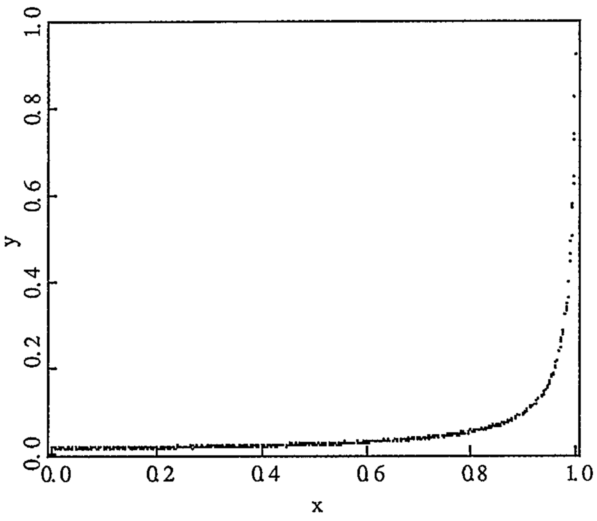

Figure 5: Scatter Plot of 1000 (x,y) Pairs in Example 4.2

Using the statistical package S, a random (*i.i.d.*) sample of 40,000 pairs (X,Y) was generated. We then divided the 40,000 pairs (x,y) into 100 sets of samples of size n = 400, and applied the five estimators to each sample of size n = 400 to estimate $y_{.95}$. Figure 6 shows the boxplots of the resulting 100 sets of estimates.

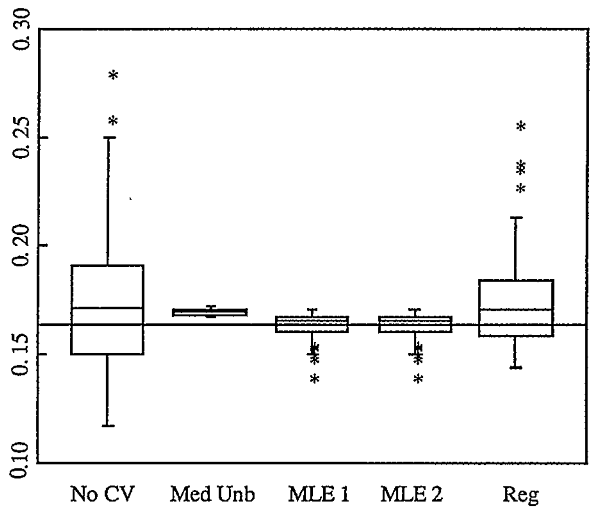

Figure 6: Boxplots of Estimates of $y_{.95}$ (n = 400)

In Figure 6, a horizontal line is drawn through y = 0.16418, the true .95*th* quantile of y. The boxplots show that the regression estimator (Reg) does relatively poorly in this nonlinear setting, while the approximately median unbiased estimator (Med Unb) is outstanding. A similar conclusion can be drawn from Table 3, which displays the mean square error (MSE), variance (Var), and bias of the five estimators.

Table 3: MSE, Var, and Bias of Estimators of $y_{.95}$ (n = 400)

	No CV	Med Unb	MLE 1	MLE 2	Reg
MSE	1.0618e-3	2.8469e-5	4.7839e-5	4.7839e-5	5.3414e-4
Var	9.7551e-4	1.5672e-6	4.5427e-5	4.5427e-5	4.4012e-4
Bias	9.2903e-3	5.1867e-3	-1.5531e-3	-1.5531e-3	9.6967e-3

To check whether the relative performance of the estimators is much affected by a smaller sample size n, we then divided the 40,000 pairs (x,y) into 400 sets of samples of size n = 100, and applied the five estimators to each sample of size n = 100 to estimate $y_{.95}$. Figure 7 displays the boxplots of the 400 sets of estimates that resulted.

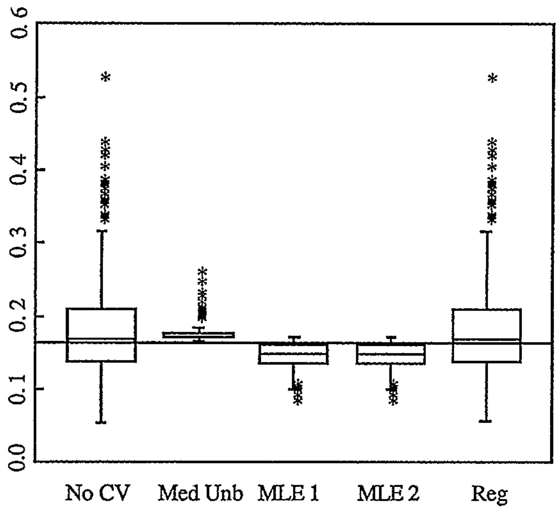

Figure 7: Boxplots of Estimators of $y_{.95}$ (n = 100)

The horizontal line is again drawn through y = 0.16418, the true .95*th* quantile of y. The relative performance of the estimators is similar to the n = 400 case, a conclusion that can be drawn from Table 4 below, also.

Table 4: MSE, Var, and Bias of Estimators of $y_{.95}$ (n = 100)

	No CV	Med Unb	MLE 1	MLE 2	Reg
MSE	4.5942e-3	1.7372e-4	6.5116e-4	6.5116e-4	4.5703e-3
Var	4.3064e-3	7.4569e-5	3.3023e-4	3.3023e-4	4.2813e-3
Bias	1.6965e-2	9.9576e-3	-1.7915e-2	-1.7915e-2	1.7002e-2

In the course of our study, we found that the two likelihood function based estimators, MLE 1 and MLE 2, tend to behave similarly and, to a lesser extent, the no control variate estimator (No CV) and the regression estimator (Reg) tend to behave similarly. This phenomenon is very pronounced in this nonlinear example, as shown in Figure 8, noticeable but less pronounced in the other examples

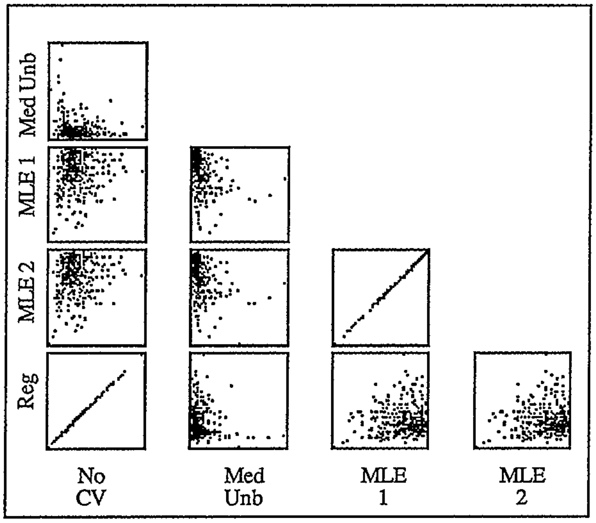

Figure 8: Pairwise Scatter Plots of Estimators of $y_{.95}$ (n = 100)

4.3 Systems Simulation Example

The M/M/1 queue is a single server, first-come-first-served service system in which customers arrive according to a Poisson process and service times are *i.i.d.* negative exponential random variables. Let Y be the delay in queue (not including service) experienced by the l th ($l > 0$) customer to arrive to an M/M/1 queue that had h ≥ 0 customers present at time 0. The control variate X is the sum of the service times of the first l+h-1 customers; i.e. the customers arriving before the lth customer. The distribution of X is Erlang, and the distribution of Y is a mixture of Erlangs (Kelton and Law, 1985).

Observations (X,Y) were generated by a FORTRAN simulation using IMSL subroutines ggamr and ggexn to generate interarrival and service times. The value of $x_{.95}$ was obtained from the S function qgamma. The cdf of Y was evaluated using the algorithm of Kelton and Law (1985). The examples below are an M/M/1 queue with arrival rate .9 customers/unit time, service rate 1 customer/unit time, and h = 0 customers present at time 0. We consider the delay in queue of the 10*th* arriving customer.

Figure 9 shows a plot of 1000 pairs (x,y). While Cor(X,Y) is unknown, the sample correlation based on 40,000 pairs was .713, which seems to indicate strong dependence. However, the boxplots in Figures 10 and 11, for 100 size n = 400 samples and 400 size n = 100 samples, respectively, shows little improvement for the control variate estimators over the No CV estimator. The corresponding Tables 5 and 6 show how small the improvement is in terms of MSE. Clearly, large correlation by itself is not enough to insure improved estimator performance for these control variate estimators. On the other hand, it is encouraging that the control variate estimators do not seem to do worse than No CV.

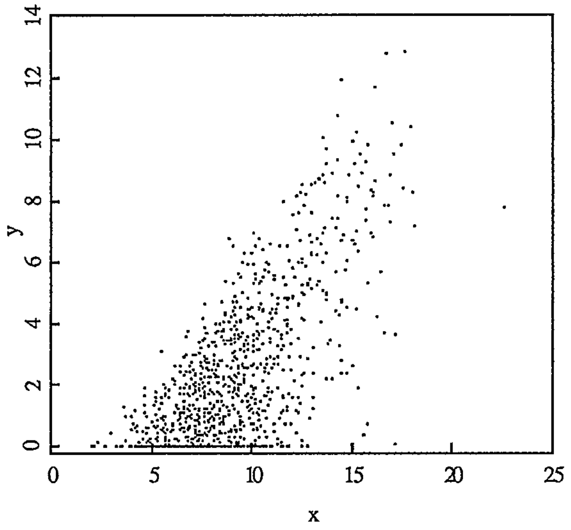

Figure 9: Scatter Plot of 1000 (x,y) Pairs (h = 0) in Example 4.3

Table 5: MSE, Var, and Bias of Estimates of $y_{.95}$ (h = 0, n = 400)

	No CV	Med Unb	MLE 1	MLE 2	Reg
MSE	0.25373	0.20469	0.17396	0.16887	0.19937
Var	0.25310	0.20381	0.17394	0.16885	0.19885
Bias	0.02498	0.02957	-0.00456	0.00363	0.02278

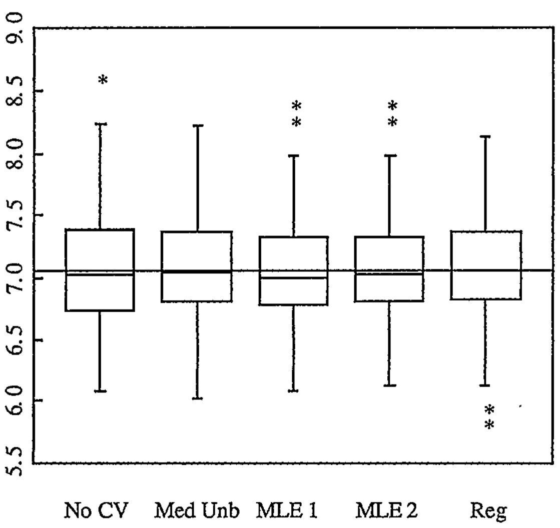

Figure 10: Boxplots of Estimates of $y_{.95}$ (h = 0, n = 400)

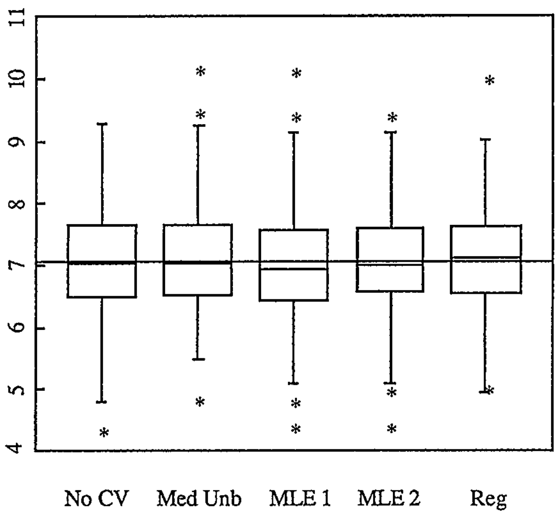

Figure 11: Boxplots of Estimates of $y_{.95}$ (h = 0, n = 100)

Table 6: MSE, Var, and Bias of Estimates of $y_{.95}$ (h = 0, n = 100)

	No CV	Med Unb	MLE 1	MLE 2	Reg
MSE	0.74743	0.66530	0.63485	0.61128	0.62466
Var	0.74665	0.66330	0.63218	0.61088	0.62252
Bias	0.02780	0.04476	-0.05163	0.01998	0.04627

4.4. Bivariate Gamma Example

The pairs (X,Y) have the bivariate gamma distribution of Schmeiser and Lal (1982). This distribution allows any gamma marginals with any feasible correlation. Schmeiser and Lal's bivariate gamma generator, gbiv, was coded in FORTRAN using IMSL functions dcadre for the numerical integration needed to determine parameters, and ggamr for univariate gamma generation. The bivariate gamma provides another example where E[Y|X=x] is nonlinear. The values of $x_{.95}$ and $y_{.95}$ were obtained from the S function qgamma.

Figure 12 shows a scatter plot of 1000 pairs (x,-y), where X and Y have identical univariate gamma marginal distributions with shape parameter 5 and scale parameter 1. The Cor(X,Y) = -.8, so that Cor(X,-Y) = .8. Based on the scatter plot this example appears to be in between the linear and nonlinear examples 4.1 and 4.2, respectively; that is, nonlinear dependence between X and Y, but not so pronounced.

We estimated the .95*th* quantile of -Y, which is equivalent to estimating the negative of the .05*th* quantile of Y. Figures 13 and 14 are the boxplots of the distributions of the five estimators for 100 size n = 400 samples and 400 size n = 100 samples, respectively. The corresponding Tables 7 and 8 quantify the performance in terms of mean square error, variance, and bias. As in example 4.2, nonlinearity caused the regression estimator (Reg) to perform badly. The other three control variate estimators are superior to No CV, with the approximately median unbiased estimator (Med Unb) possibly somewhat better considering both bias and variability.

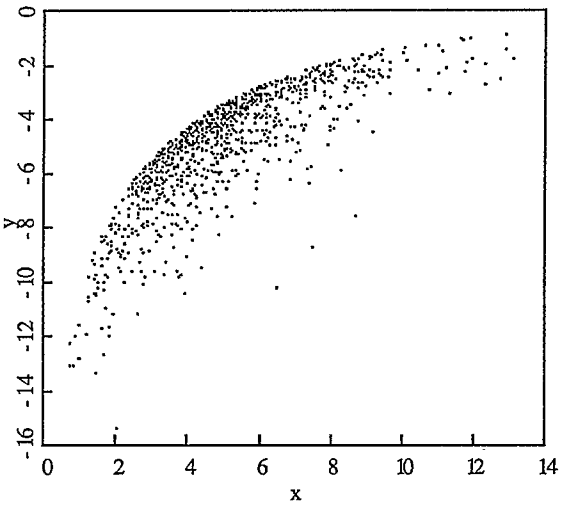

Figure 12: Scatter Plot of 1000 (x,y) Pairs in Example 4.4

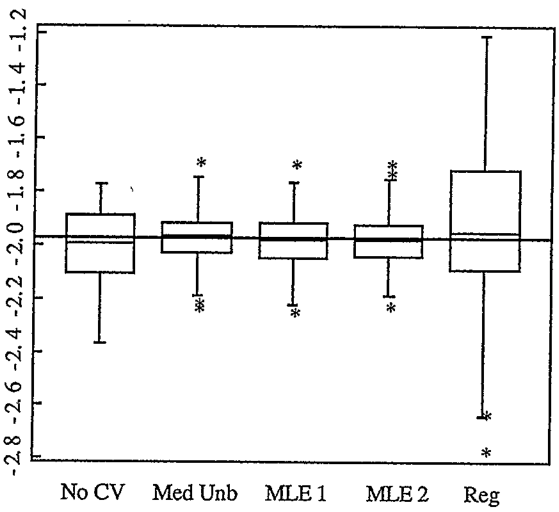

Figure 13: Boxplots of Estimates of $y_{.95}$ (n = 400)

Table 7: MSE, Var, and Bias of Estimates of $y_{.95}$ (n = 400)

	No CV	Med Unb	MLE 1	MLE 2	Reg
MSE	0.01974	0.01173	0.01156	0.01182	0.08610
Var	0.01901	0.01170	0.01134	0.01157	0.08447
Bias	-0.02711	-0.00556	-0.01511	-0.01599	0.04035

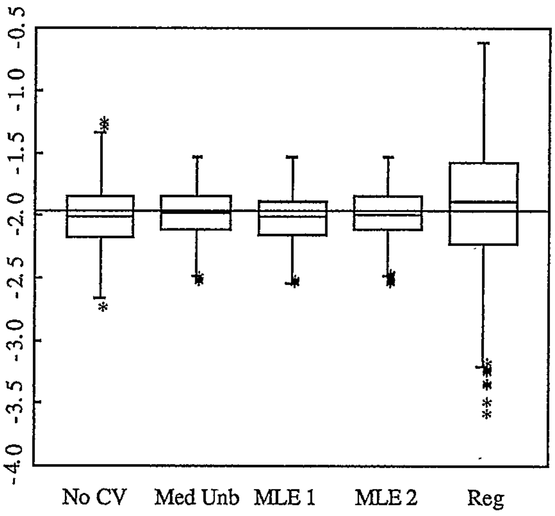

Figure 14: Boxplots of Estimates of $y_{.95}$ (n = 100)

Table 8: MSE, Var, and Bias of Estimates of $y_{.95}$ (n = 100)

	No CV	Med Unb	MLE 1	MLE 2	Reg
MSE	0.06314	0.03847	0.04032	0.04044	0.28204
Var	0.06206	0.03801	0.03788	0.03990	0.28164
Bias	-0.03286	-0.02157	-0.04938	-0.02324	0.02016

5. CONCLUSIONS

Variance reduction research has concentrated on estimating population means and variances, which are but two of the characteristics of the population (see Nelson 1987a for a survey). Quantiles provide additional information about the population, and can in fact be the parameters of primary interest in certain problems. Thus, it is important to develop good techniques for estimating quantiles.

Techniques based on regression have been the primary focus of control variate research (Glynn and Whitt 1987, Nelson 1987b, and Rothery 1982 are some exceptions). Our viewpoint is that regression techniques are unnatural for estimating quantiles, because it is unnatural to think of quantiles as expected values. We propose quantile estimators that are based on estimating the joint probablistic behavior of the variable of interest and the control variate instead.

The empirical study presented here shows the three new control variate estimators to be promising. Further, the study shows that, in quantile estimation, the Pearson correlation between the control variate and the variable of interest is not a good predictor of success or failure in variance reduction. For this problem, other concepts of bivariate dependence, such as regression dependence and association, may be at least as relevant. Finally, in evaluating the performance of estimators, simple graphical techniques such as boxplots add valuable information to the usual measures of mean square error, variance, and bias. Boxplots allows one to assess median biases, concentration of the middle 50% of the distributions, and tendencies for outliers.

ACKNOWLEDGEMENT

Jason Hsu's research is made possible in part by Grant No. 1R01 CA41168, awarded by the National Cancer Institute. Barry Nelson's research was partially supported by National Science Foundation Grant No. ECS-8707634. Computing resources were provided by the Mathematical Sciences Computing Laboratory of the Ohio State University.

REFERENCES

Becker, Richard A. and Chambers, John M. (1984). *S - An Interactive Environment for Data Analysis and Graphics.* Wadsworth, Belmont, California.

Bratley, P., Fox, B.L., and Schrage, L.E. (1987). *A Guide to Simulation, 2nd edition.* Springer-Verlag, N.Y.

David, H. A. (1981). *Order Statistics, 2nd edition.* Wiley, New York.

Glynn, P.W. and Whitt, W. (1986). Indirect estimation via L = λW. Dept. of Industrial Engineering, University of Wisconsin-Madison.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods.* Chapman and Hall, London.

Juritz, J.M., Juritz, J.W.F. and Stephens, M.A. (1983). On the accuracy of simulated percentage points. *Journal of the American Statistical Association* **78**, 441-444.

Kelton, W.D. and Law, A.M. (1985). The transient behavior of the M/M/s queue, with implications for steady-state simulation. *Operations Research* **33**, 378-396.

Lehmann, E. L. (1983). *Theory of Point Estimation.* Wiley, New York.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses, 2nd edition.* Wiley, New York.

Nelson, B.L. (1987a). A perspective on variance reduction in dynamic simulation experiments. *Communications in Statistics* **B16**, 385-426.

Nelson, B.L. (1987b). On control variate estimators. *Computers & Operations Research* **14**, 219-225.

Rothery, P. (1982). The use of control variates in Monte Carlo estimation of power. *Applied Statistics* **31**, 125-129.

Schmeiser, B.W. and Lal, R. (1982). Bivariate gamma random vectors. *Operations Research* **30**, 355-374.

Tong, Y. L. (1980). *Probability Inequalities in Multivariate Distributions.* Academic Press, New York.

AUTHORS' BIOGRAPHIES

JASON C. HSU is an Associate Professor in the Department of Statistics of the Ohio State University. He received a Ph.D. in Statistics from Purdue University in 1977. His research interests include multiple comparisons, simultaneous statistical inference, ranking and selection, and statistical software development. He is a member of IMS and ASA.

Jason C. Hsu
Department of Statistics
The Ohio State University
1958 Neil Ave.
Columbus, OH 43210, USA.
(614)292-7663

BARRY L. NELSON is an Assistant Professor in the Department of Industrial and Systems Engineering at the Ohio State University. He is also associated with the Ohio State University Statistical Consulting Service. His Ph.D. is from the School of Industrial Engineering at Purdue University, and his research interests center on design and analysis of simulation experiments, particularly methods for statistically efficient simulation. He teaches courses in simulation and stochastic processes. Continuing memberships include ASA, IIE, ORSA, TIMS, and SCS. Dr. Nelson is an active member of the TIMS College of Simulation and Gaming, and is editor of its Newsletter.

Barry L. Nelson
Department of Industrial and Systems Engineering
The Ohio State University
Columbus, OH 43210, USA
(614)292-0610