IIE Transactions
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713772245

Variance Reduction in the Presence of Initial-Condition Bias

Barry L. Nelson[a]
[a] Department of Industrial and Systems Engineering, The Ohio State University, Columbus, Ohio

To cite this Article Nelson, Barry L.(1990) 'Variance Reduction in the Presence of Initial-Condition Bias', IIE Transactions, 22: 4, 340 — 350
To link to this Article: DOI: 10.1080/07408179008964188
URL: http://dx.doi.org/10.1080/07408179008964188

Variance Reduction in the Presence of Initial-Condition Bias

BARRY L. NELSON
Department of Industrial and Systems Engineering
The Ohio State University
Columbus, Ohio 43210

Abstract: We consider the application of antithetic variates and control variates variance reduction techniques (VRTs) in the context of steady-state simulation experiments when initial-condition bias is present. We show that, by appropriately modifying the experiment design, incorporating a VRT can improve both point and interval-estimator performance. Guidelines for modifying the experiment design are given.

This paper considers the problem of estimating, via simulation, parameters of the limiting distribution of an ergodic stochastic process, which is sometimes called the "steady-state-simulation problem." We assume that the simulator requires both a point and an interval estimate of one or more parameters of the process. If only a point estimate is required, then the appropriate experiment design is a single simulated realization of the process (see for instance, Cheng [2]). Interval estimation, however, is difficult in single realization designs.

The steady-state-simulation problem is complicated by point-estimator bias, which is introduced by the choice of initial conditions. The initialandition-bias problem is a longstanding one in simulation design and analysis; see Gafarinn, Ancker and Morisaku [5] and Wilson and Pritsker [23, 24] for surveys. Unfortunately, there is no unique solution to the problem since there are many conflicting criteria.

We investigate the consequences of incorporating a variance reduction technique (VRT) into the experiment design when initial-condition bias is present. We show that, under typical assumptions, the VRT can lead to improvement in all standard criteria. Thus, any choice of experiment design or analysis can be improved by incorporating a VRT. This work completes a preliminary investigation of the interaction of output analysis methods (OAMs) and VRTs which emphasizes developing theory and guidelines for simulation experiment design, rather than developing new VRTs or OAMs (Nelson [12, 13, 15, 16]).

We restrict attention to the replicationdeletion and the nonoverlapping-batch-means OAMs for deterministically initialized replications (see for instance, Law and Kelton [11]), and either dependence induction or control variate VRTs (see for instance, Nelson [14], Wilson [22]). Replication-deletion and batch means are the two OAMs most often used in practice. Dependence induction VRTs are the ones that practitioners use, and control variates is a VRT that (we think) should be used.

The paper is organized as follows: In the next section the point and interval estimation problem is described, including the replication-deletion OAM. Then we define the antithetic variates and control variates VRTs, followed by the main results. The fifth section describes some special concerns when the simulation budget is fixed. The next section contains a numerical example, and the final section offers some recommendations and conclusions.

Problem Description

Let Y_{ij} denote the i^{th} simulation output from the j^{th} replication, for $i = 1,2,\ldots,m$ and $j = 1,2,\ldots,k$. We assume that $\lim_{i\to\infty} \mathrm{E}[Y_{ij}] = \theta_j$, but, due to the effect of deterministically selecting the initial state of the stochastic process, $\mathrm{E}[Y_{ij}] \neq \theta_j$ for finite i. Let I denote the initial conditions.

A point estimator of θ_j is

$$\bar{Y}_j(m,d) = (m-d)^{-1} \sum_{i=d+1}^{m} Y_{ij}$$

which is called the truncated sample mean since the first $d < m$ outputs are given weight 0 ("deleted") in hopes of reducing the effect of initial-condition bias. We are interested in the following properties of $\bar{Y}_j(m,d)$ as an estimator of θ_j:

$$\mathrm{Var}[\bar{Y}_j(m,d)|I]$$

$$\mathrm{Bias}[\bar{Y}_j(m,d)|I] = \theta_j - \mathrm{E}[\bar{Y}_j(m,d)|I]$$

Received November 1987; revised February 1989. Handled by the Department of Simulation,

$$\mathrm{MSE}[\bar{Y}_j(m,d)|I] = \mathrm{Bias}^2[\bar{Y}_j(m,d)|I] + \mathrm{Var}[\bar{Y}_j(m,d)|I]$$

which are the variance, the bias, and the **mean** squared error, respectively, of the point estimator, conditioned on I (from here on we drop the condition I for notational convenience).

There are two common cases to consider: (A) When $\theta_j = \theta$ for **all** j and the replications are identically distributed, and (B) when $\theta_{2j-1} = \theta_1$ and the odd numbered replications are identically distributed, while $\theta_{2j} = \theta_2$ and the even numbered replications are identically distributed, for $j = 1,2,\ldots,k/2$, and we are interested in estimating 8 = 8, $- \theta_2$. We study case (A), but comment on case (B) in the final section.

Temporarily dropping the subscript j denoting replications, we assume that the output stochastic process, $\{Y_i;\ i = 1,2,\ldots\}$, can be represented as

$$Y_i = \theta + X_i + b_i \tag{1}$$

where $\{X_i;\ i = 1,2,\ldots\}$ is a zero-mean, finite-variance, stochastic process converging in distribution to a random variable X, and $\{b_i;\ i = 1,2,\ldots\}$ is a deterministic sequence. We can formulate an analogous model for the continuous-time-parameter process $\{Y(t);\ t \geq 0\}$; specifically,

$$Y(t) = \theta + X(t) + b(t) \tag{2}$$

with corresponding assumptions about $X(t)$ and $b(t)$.

Models (1) and (2) have appeared in the simulation literature on initial-condition bias. Schruben [18] assumes that $\{X_i\}$ is stationary and phi mixing, but makes no assumption about $\{b_i\}$. Schruben, Singh and Tierney **[19]** set $b_i = -\theta a_i$, where a_i is a quadratic in i that decreases to 0. Frequently, it is assumed that b_i converges monotonically to 0; e.g., Kelton and Law [9]. Cheng [2] assumes that $|b(t)|$ is continuous and monotonically decreasing.

A special case of (1) that has been used to study the initial-condition-bias problem (Fishman [4], Turnquist and Sussman [21], Kelton and **Law** [10], and Snell and Schruben [20]), and will be used as **an** example below, is the autoregressive order 1 (AR(1)) process. In the notation of (1)

$$X_i = \sum_{j=0}^{i-1} \phi^j \epsilon_{i-j}$$

and

$$b_i = c\phi^i$$

where $\{\epsilon_i;\ i = 1,2,\ldots\}$ is an independent and identically distributed (i.i.d.), zero-mean, finite-variance, stochastic process, and $0 \leq \phi < 1$ and c are constants that determine the dependence structure and bias in the process, respectively (we restrict attention to nonnegative ϕ which implies positive autocorrelations, typical of many queueing simulations). Notice that $|b_i|$ is monotonically decreasing, but $\{X_i\}$ is not stationary.

Returning to the general process **(1)**, let

$$\bar{b}(m,d) = (m-d)^{-1} \sum_{i=d+1}^{m} b_i .$$

We assume that $|\bar{b}(m,d)|$ is a monotonically decreasing function of m for fixed $d < m$. This assumption may not be satisfied by all simulation processes, but it will frequently be the case that there exists m_0 such that $|\bar{b}(m,d)|$ is decreasing for $m > m_0$; Glynn [6] establishes this property for finite-state Markov chains, for example. Implicit in the deletion method is the additional assumption that $|\bar{b}(m,d)|$ is a nonincreasing function of d for $0 \leq d < m$. If all the b_i have the same sign, then these two assumptions imply that

$$(m-d)^{-1} \left| \sum_{i=d+1}^{m} b_i \right| \leq |b_d| \leq (d-1)^{-1} \left| \sum_{i=1}^{d-1} b_i \right| .$$

That is, in absolute value, each b_i is less than the average of the preceding terms, but greater than the average of the remaining terms. Sequences $\{b_i\}$ that go to zero monotonically, such **as** the $AR(1)$ bias process, satisfy this assumption.

Finally, we assume that $\nu(m,d) = \mathrm{Var}[\bar{Y}(m,d)] = \mathrm{Var}[\bar{X}(m,d)]$ is an increasing function of d for fixed m. This is the cost of deletion in terms of point-estimator variance. The assumption will be satisfied by many processes with positive dependence, such as the AR(1).

Returning to the estimation problem, suppose that the k replications are i.i.d. Let **the** point estimator of θ be

$$\bar{Y}(k,m,d) = k^{-1} \sum_{j=1}^{k} \bar{Y}_j(m,d) .$$

Then

$$\mathrm{MSE}[\bar{Y}(k,m,d)] = \mathrm{Bias}^2[\bar{Y}(k,m,d)] + \mathrm{Var}[\bar{Y}(k,m,d)]$$
$$= \bar{b}^2(m,d) + \nu(m,d)/k . \tag{3}$$

Frequently, (3) is an increasing function of d, so that d = 0 appears to be optimal.

The standard interval estimator of θ is the $(1-\alpha)100\%$ confidence interval $\bar{Y}(k,m,d) \pm H(k,m,d)$, where $H(k,m,d) = t_{\alpha/2}(k-1)S(k,m,d)/\sqrt{k}$ is the confidence interval half width, $t_{\alpha/2}(k-1)$ is the $1-\alpha/2$ quantile of the **t** distribution **with** $k-1$ degrees of freedom, and $S^2(k,m,d)$ is the sample variance of the $\{\bar{Y}_j(m,d)\}$. If, as we assume, the $\bar{Y}_j(m,d)$ are normally distributed, then this interval achieves the nominal probability $1-\alpha$ of covering $\mathrm{E}[\bar{Y}(k,m,d)]$, but not of covering 8. Thus, another property of interest is

$$\beta(k,m,d) = \Pr\{|\bar{Y}(k,m,d) - \theta| \le H(k,m,d)\}$$

which is the probability that the interval covers 0. Under the assumption of i.i.d. normally distributed replications,

$$\beta(k,m,d) = \Pr\{|T(k-1, \delta(k,m,d))| \le t_{\alpha/2}(k-1)\}$$

where T is a random variable having a noncentral t distribution with $k - 1$ degrees of freedom and noncentrality parameter

$$\delta(k,m,d) = \frac{\text{Bias}[\bar{Y}(k,m,d)]}{\sqrt{\text{Var}[\bar{Y}(k,m,d)]}}$$

$$= \frac{\sqrt{k}\,\bar{b}(m,d)}{\sqrt{v(m,d)}} .$$

Kelton and Law [10] show that $d > 0$ may be desirable if probability of coverage is an important criterion.

Variance Reduction

In this section we describe antithetic variates (AV) and control variates (CV) VRTs; the development follows Nelson [12], to which the reader is referred for details. We call the standard point and interval estimators described above the "crude" estimators, and search for improved estimators.

Antithetic Variates (AV)

AV induces dependence between pairs of replications, $\{\bar{Y}_{2j-1}, \bar{Y}_{2j}\}$, $j = 1,2,\ldots,k/2$, leaving different pairs independent (we drop the argument (m,d) from $\bar{Y}_j$ for convenience). To define the AV experiment, let $\hat{Y}_j = (\bar{Y}_{2j-1} + \bar{Y}_{2j})/2$ for $j = 1,2,\ldots,k/2$. Let $\hat{Y}$ be the sample mean of the $\hat{Y}_j$, which is the AV point estimator and is algebraically equivalent to $\bar{Y}$. The AV interval estimator is $\hat{Y} \pm H_{av}$, where $H_{av} = t_{\alpha/2}(k/2 - 1)\hat{S}\sqrt{2/k}$ is the half width of the interval, and

$$\hat{S}^2 = \frac{1}{k/2 - 1} \sum_{j=1}^{k/2} (\hat{Y}_j - \hat{Y})^2$$

is the sample variance of the $\{\hat{Y}_j\}$. Let $\varrho = \text{Corr}[\bar{Y}_{2j-1}, \bar{Y}_{2j}]$ be the induced correlation between antithetic pairs of replications. Then the $\text{Var}[\hat{Y}]$ and $\text{Var}[\bar{Y}]$ are different; $\text{Var}[\hat{Y}]/\text{Var}[\bar{Y}] = (1 + \varrho)$, which is called the variance reduction ratio. When $\varrho < 0$, $\hat{Y}$ has smaller variance than $\bar{Y}$. Since AV does not affect the bias of the point estimator, $\hat{Y}$ has smaller MSE as well

$$\text{MSE}[\hat{Y}] = \bar{b}^2(m,d) + (1 + \varrho)v(m,d)/k . \qquad (4)$$

However, superior performance of the AV interval estimator relative to the crude interval estimator is not guaranteed. When no bias is present, the performance of the interval estimator depends on the number of replications, k, the value of the achieved negative correlation, ϱ, and the confidence level α (Nelson [13]). In the presence of initial-condition bias, the probability that the interval covers θ is

$$\hat{\beta}(k,m,d) = \Pr\{|T(k/2 - 1, \hat{\delta}(k,m,d))| \le t_{\alpha/2}(k/2 - 1)\}$$

where

$$\hat{\delta}(k,m,d) = \frac{\delta(k,m,d)}{\sqrt{1+\varrho}} .$$

We compare $\hat{\beta}$ to β for large and small k in a later section.

Control Variates (CV)

CV exploit the dependence between $\bar{Y}_j$ and a $q \times 1$-vector of random variables with known expectation, but that are also observable in each replication. Let the output from the jth replication be the column vector $(\bar{Y}_j, C_{j1},\ldots,C_{jq})'$, where $'$ denotes transpose. Let $\bar{C} = (\bar{C}_1,\ldots,\bar{C}_q)'$ be the sample mean of the control variates, and let $\mu = E[\bar{C}]$. Then the control variate point estimator is

$$\tilde{Y} = \bar{Y} - \tilde{a}'(\bar{C} - \mu)$$

where $\tilde{a}$ is an estimate of the $q \times 1$-vector control multiplier (Wilson [22]). If the $(\bar{Y}_j, C_{j1},\ldots,C_{jq})$ are i.i.d. $(q+1)$-variate normal vectors, then the variance reduction ratio is

$$\frac{\text{Var}[\tilde{Y}]}{\text{Var}[\bar{Y}]} = \frac{k-2}{k-q-2}(1 - R^2)$$

where R^2 is the square of the multiple correlation coefficient of $\bar{Y}$ on $\bar{C}$. The MSE of $\tilde{Y}$ is

$$\text{MSE}[\tilde{Y}] = \bar{b}^2(m,d) + \left(\frac{k-2}{k-q-2}\right)(1 - R^2)v(m,d)/k . \quad (5)$$

Thus, $\tilde{Y}$ has smaller variance and MSE than $\bar{Y}$ if $R^2 > q/(k-2)$. The value of R^2 depends on q and the particular control variates chosen. The factor $(k-2)/(k-q-2)$ is nondecreasing in q, while $1 - R^2$ is nonincreasing in q.

The CV interval estimator is $\tilde{Y} \pm H_{cv}$, where $H_{cv} = t_{\alpha/2}(k-q-1)\tilde{S}$ is the half width of the interval; the calculation of $\tilde{S}$ is discussed in Wilson [22]. In the presence of bias in $\bar{Y}_j$, but not $\bar{C}$, the probability that the interval estimator covers θ is

$$\tilde{\beta}(k,m,d) = \Pr\{|T(k-q-1, \tilde{\delta}(k,m,d))| \le t_{\alpha/2}(k-q-1)\}$$

where

$$\tilde{\delta}(k,m,d) = \frac{\delta(k,m,d)}{\sqrt{1-R^2}}\sqrt{\frac{k-q-2}{k-2}} .$$

(This is a consequence of the distribution of $\tilde{Y}$, which can be found in [16].) We compare $\bar{\beta}$ to β for large and small k in the next section.

Probability of Coverage

When $\delta(k,m,d) \neq 0$, the probability that the crude interval covers θ is less than the nominal value $1-\alpha$, and coverage is monotonically decreasing in $|\delta|$. Deletion improves coverage, but at the expense of increased variance. In this section we examine the coverage of the crude, AV and CV intervals. If k, m and d are the *same* for all three estimators and k is large, then the AV and CV intervals have lower probability of coverage than the crude interval. However, the coverage of the AV and CV intervals can match or surpass the crude interval and still maintain reduced MSE through additional deletion. When k is small, the AV and CV intervals may actually have larger probability of coverage than the crude interval without resorting to further deletion.

Large-Sample Coverage

For the discussion that follows we consider $\delta^2(k,m,d)$ — the ratio of the bias squared to the variance of the point estimator — without loss of generality since degradation in coverage is symmetric in δ. We also fix k and m, so that d is the only controllable factor.

When k is large, equal coverage for all three interval estimators requires that

$$\delta^2(k,m,d) = \hat{\delta}^2(k,m,d) = \bar{\delta}^2(k,m,d) \qquad (6)$$

(see the Appendix). However, if $\varrho < 0$ and $R^2 > q/(k-2)$, then $\hat{\delta}$ and $\bar{\delta}$ are greater than δ in absolute value. To attain equal coverage when k is large, (6) implies that we must find d' and d'' such that

$$\delta^2(k,m,d') = (1+\varrho)\delta^2(k,m,d) \qquad (7)$$

and

$$\delta^2(k,m,d'') = \left(\frac{k-2}{k-q-2}\right)(1-R^2)\delta^2(k,m,d) . \qquad (8)$$

That is, the bias squared to variance ratio of the AV and CV estimators must be reduced, through additional deletion, by an amount equal to the variance reduction ratio. Under our assumptions, $\delta^2(k,m,d)$ is a decreasing function of d. This lends to the following result:

Proposition 1: Consider the crude experiment with design k, m, and d when (2) describes the simulation output process. Suppose that $b(t)$ and $\nu(t,d)$ are continuous functions of t, and each replication is of length τ. If there exists a d' (d'') such that (7) ((8)) is satisfied, then the AV (CV) point estimator has smaller MSE under design k, m, d' (d'') then the crude estimator under design k, m and d.

Proof: Let

$$d' = \inf\{\gamma \in [d,\tau): \delta^2(k,m,\gamma) = (1+\varrho)\delta^2(k,m,d)\},$$

if it exists. Then

$$\frac{(1+\varrho)k\bar{b}^2(m,d)}{\nu(m,d)} = \frac{k\bar{b}^2(m,d')}{\nu(m,d')} \le \frac{k\bar{b}^2(m,d)}{\nu(m,d')}$$

since $|\bar{b}(m,d)|$ is nonincreasing in d. Thus,

$$\mathrm{Var}[\hat{Y}(k,m,d')] = (1+\varrho)\mathrm{Var}[\bar{Y}(k,m,d')] \le \mathrm{Var}[\bar{Y}(k,m,d)]$$

so that $\hat{Y}(k,m,d')$ has smaller MSE. A completely analogous argument proves the result for CV.

In the case of output process (1), it may not be possible to obtain precisely equal coverage because d is discrete. However, if m is large then the difference will be negligible. If no d' exists that satisfies (7), then $(1+\varrho)\nu(\tau,\tau-\epsilon) < \nu(\tau,d)$ for all $\epsilon > 0$ in model (2), or $(1+\varrho)\nu(m,m-1) < \nu(m,d)$ in model (1); this situation seems unlikely when τ and m are large enough to be reasonable. A similar condition applies for d''.

Small-Sample Coverage

When the number of replications, k, is small, the differing degrees of freedom associated with each interval estimator (crude, AV and CV) play a role. This case is important, because a design that specifies a small number of long replications may be desirable when the budget is fixed (see Fixed-Budget Design below) to ensure the validity of the interval estimation procedure (i.e., that $\bar{Y}_i$ is approximately normally distributed) and to further guard against initial-condition bias. It turns out that the confidence interval procedure provides some protection against initial-condition bias when k is small, but there is a cost in terms of the length of the interval.

The key to our results is the following calculation: Let $0 < \eta < 1-\alpha$ be the degradation in probability of coverage due to initial-condition bias. Then we can (numerically) find δ, $\hat{\delta}$, and $\bar{\delta}$ such that

$$\begin{aligned} 1-\alpha-\eta &= \Pr\{|T(k-1,\delta)| \le t_{\alpha/2}(k-1)\} \\ &= \Pr\{|T(k/2-1,\hat{\delta})| \le t_{\alpha/2}(k/2-1)\} \\ &= \Pr\{|T(k-q-1,\bar{\delta})| \le t_{\alpha/2}(k-q-1)\} \end{aligned}$$

(see the Appendix). That is, we can find bias squared to variance ratios such that the crude, AV, and CV interval estimators all have coverage equal to $1-\alpha-\eta$. If all three estimators employ the same design k, m and d, then the AV and CV interval estimators will have equal or better coverage than the crude interval if

$$1 + \varrho \geq (\delta/\hat{\delta})^2$$

and

$$\frac{k - 2}{k - q - 2}(1 - R^2) \geq (\delta/\tilde{\delta})^2 ,$$

respectively. These bounds result from the fact that $\hat{\delta}^2 = \delta^2/(1 + \varrho)$ and $\tilde{\delta}^2 = ((k-q-2)/(k-2))\delta^2/(1 - R^2)$. Thus, if k, m and d are the same for all three estimators, then $1 + \varrho = (\delta/\hat{\delta})^2$ and $(1 - R^2)(k-2)/(k-q-2) = (\delta/\tilde{\delta})^2$. Since coverage decreases if $1 + \varrho$ or $1 - R^2$ decreases, these ratios are lower bounds on the variance reduction such that no degradation in coverage occurs versus the crude estimator.

Figure 1 shows $(\delta/\hat{\delta})^2$ and $(\delta/\tilde{\delta})^2$ as functions of k for $\alpha = \eta = .05$ (i.e., a 95% confidence interval with 5% degradation). When k is small, point estimator variance can be reduced without degrading coverage relative to the crude estimator. However, as k increases these ratios go to 1, as implied by the large-sample result above. Thus, the large-sample result is conservative in the sense that reducing the bias squared to variance ratio (through additional deletion) by an amount equal to the point estimator variance reduction ratio insures equal or superior coverage for all k. The curves in Figure 1 are not particularly sensitive to changes in α and η in the interval [.01, .1].

Unfortunately, when k is quite small the crude interval may beat the **AV** or **CV** interval in other ways. Figure 2 shows the function $\gamma(k)$ (lower curve) such that $1 + \varrho \leq \gamma(k)$ implies $E[H_{av}] \leq E[H]$ for $\alpha = .05$ (Nelson [13]); i.e., the expected half width of the AV interval is shorter than the expected half width of the crude interval. Superimposed on Figure 2 is $(\delta/\hat{\delta})^2$ from Figure 1. Recall that $1 + \varrho \geq (\delta/\hat{\delta})^2$ is required for equal coverage. Unfortunately, the regions of equal coverage and shorter expected length do not intersect, meaning that the protection from initial condition bias with small k is at the expense of larger interval length. The conclusion is that moderate k and additional deletion are usually required for the **AV** and **CV** estimators to outperform the crude estimator.

Fixed-Budget Designs

In any practical problem there is a limit on the budget for a simulation study. Somewhat artificially, we model this limit as a constraint, n, on the total number of outputs observed. The design problem then becomes choosing how to "spend" these n outputs. One strategy is to make k replications of length $m = n/k$, deleting d from each one; this is the fixed-budget analog of the replication-deletion OAM above. Kelton and Law [10], Kelton [7], and **Turnquist** and Sussman [21] have studied the problem of choosing k and d.

When either **AV** or **CV VRTs** are applied in conjunction with the replication-deletion OAM, the results in the previous section hold: Whatever design is chosen, it can be improved by incorporating **AV** or **CV** with additional deletion. The example below illustrates the effect of a fixed budget on the replicationdeletion **OAM**; see also Nelson [15].

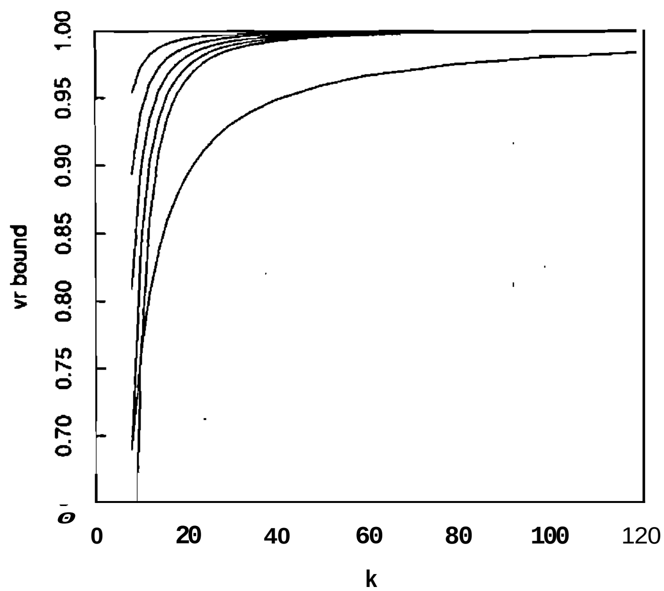

Figure 1. From top to bottom, the functions $(\delta/\tilde{\delta})^2$ for $q = 1, 2, \ldots, 5$ and the function $(\delta/\hat{\delta})^2$.

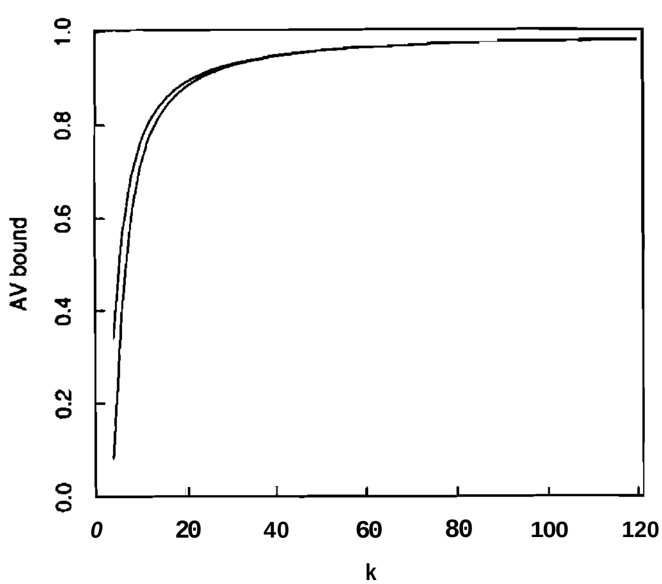

Figure 2. From top to bottom, the functions $(\delta/\hat{\delta})^2$ and $\gamma(k)$.

A second strategy is to make a single replication, deleting d outputs only once. The nonoverlapping-batch-means *OAM* is one method for constructing an interval estimate from a single replication. The basic properties of the batch means interval estimator are described in Schmeiser [17].

The impact of **AV** and **CV VRTs** on the batch means *OAM* when the output process is covariance stationary is derived in Nelson [13, 16]; stationarity implies that initial-condition bias has been eliminated. In the presence of bias, some of the results must be modified. In the remainder of this section we outline some properties of the batch-means OAM when bias is present, leaving the details of applying *AV* and *CV* to the references.

Let $\{Y_1, Y_2, \ldots, Y_n\}$ represent the output from a single replication, and assume that model (1) pertains. Define the jth batch mean to be

$$\bar{Y}_j(k,d;n) = m^{-1} \sum_{i=d+(m-1)j+1}^{jm+d} Y_i$$

for $j = 1,2,\ldots,k$, where $m = (n-d)/k$. The idea behind batch means is to choose k small enough (m large enough) so that the batch means are approximately i.i.d. normally distributed. We suppose that there exists a $k^* \le n-d$ such that these assumptions hold for all $k \le k^*$ (Schmeiser [17]). The batch means play the role of the replication means when constructing an interval estimate or applying a *VRT*.

Under model (1)

$$\bar{Y}_j(k,d;n) = \theta + \bar{X}_j(k,d;n) + \bar{b}_j(k,d;n)$$

where $\bar{X}_j(k,d;n)$ and $\bar{b}_j(k,d;n)$ are the corresponding batch means of the $\{X_i\}$ and $\{b_i\}$ processes, respectively. The assumptions above regarding the batch means will hold if the $\{\bar{X}_j(k,d;n);\ j = 1,2,\ldots,k\}$ are i.i.d. normally distributed for $k \le k^*$.

Let

$$\bar{Y}(1,n,d) = k^{-1}\sum_{j=1}^{k}\bar{Y}_j(k,d;n)$$

and

$$S^2(k,d;n) = (k-1)^{-1}\sum_{j=1}^{k}(\bar{Y}_j(k,d;n) - \bar{Y}(1,n,d))^2$$

be the sample mean and variance, respectively, of the batch means; notice that the sample mean is independent of k. The interval estimator is $\bar{Y}(1,n,d) \pm t_{\alpha/2}(k-1)S(k,d;n)/\sqrt{k}$. For model (1)

$$\mathrm{E}[\bar{Y}(1,n,d)] = \theta + \bar{b}(n,d) .$$

If we also assume that X_t is a covariance stationary process, then if $k \le k^*$

$$\mathrm{E}[S^2(k,d;n)/k] = \mathrm{Var}[\bar{X}(1,n,d)] + s_b^2(k,d;n)/k$$

$$\ge \mathrm{Var}[\bar{Y}(1,n,d)] = \nu(n,d)$$

where

$$\bar{X}(1,n,d) = k^{-1}\sum_{j=1}^{k}\bar{X}_j(k,d;n)$$

and

$$s_b^2(k,d;n) = (k-1)^{-1}\sum_{j=1}^{k}(\bar{b}_j(k,d;n) - \bar{b}(n,d))^2 .$$

The point-estimator bias is independent of the number of batches, k, but the variance-estimator bias is not, even when the assumption of independent batch means is satisfied. Varianceestimator bias is not a concern in replication designs. The following result indicates that, for well-behaved bias processes, $s_b^2(k,d;n)/k$ is a decreasing function of k.

Proposition 2: For any sequence of constants $b_1, b_2, \ldots, b_n$, $s_b^2(k,d;n)/k \le s_b^2(n,d;n)/n$ if and only if

$$s_b^2(n,d;n) \le (k(m-1))^{-1}\sum_{j=1}^{k}\sum_{i=d+(j-1)m+1}^{jm+d}(b_i - \bar{b}_j(k,d;n))^2 .$$

Remark: The term on the right-hand side is the average variance within a batch. The proposition states that batching decreases the bias only if the average variance within the batches exceeds the total variance of the sequence before batching. This condition will not hold for well-behaved bias processes ($b_i = c\phi^i$, for example); that is, the average variance within the batches is typically much less than the total variance in the sequence. The result applies if n is replaced by $k_1 < n$, and k is replaced by $k_2 = k_1/m_2$, for some integer m_2 that divides k_1, showing that typically the bias continues to increase as k decreases.

Proof: Without loss of generality, we prove the result for $d = 0$. For convenience, we drop the dependence of the quantities on d and n. Using the standard sum of squares decomposition

$$\sum_{i=1}^{n}(b_i - \bar{b})^2 = m\sum_{j=1}^{k}(\bar{b}_j(k) - \bar{b})^2 + \sum_{j=1}^{k}\sum_{i=(j-1)m+1}^{jm}(b_i - \bar{b}_j(k))^2$$

Thus,

$$s_b^2(k)/k = (km(k-1))^{-1}\left(\sum_{i=1}^{n}(b_i - \bar{b})^2 - \sum_{j=1}^{k}\sum_{i=(j-1)m+1}^{jm}(b_i - \bar{b}_j(k))^2\right)$$

The condition for $s_b^2(k)/k \le s_b^2(n)/n$ is immediate after collecting terms.

Thus, a larger number of batches (smaller batch size) typically means smaller bias in the variance estimator. However, there may be some benefit from the bias, since it inflates the variance estimate, widening the interval estimate, and Compensating for the degradation in coverage due to point-estimator bias. Quantifying exactly how much compensation occurs seems quite complex. And, of course, we want to maintain $k \leq k^*$.

When AV or CV is applied, probability of coverage declines since the VRT shrinks the interval around a biased point estimator. If we delete $d' > d$ to improve coverage, then we assume that the remaining outputs are batched so that $m' = (n - d')/k$ is the batch size. A potential problem occurs if the additional deletion causes the independence assumption to be violated. In the final section we find that the additional deletion required is small, so that the properties of the batch means should not be significantly altered. Thus, we continue to assume that the batch means are independent.

Under our assumptions, $\bar{b}^2(n,d') \leq \bar{b}^2(n,d)$ and $\nu(n,d') > \nu(n,d)$, the same effects as in the replication case, so coverage should improve similarly. We expect that $s_b^2(k,d';n) \leq s_b^2(k,d;n)$, since we are discarding more extreme terms near the beginning of the bias process; thus the variance estimator is also less biased.

Examples

When a design exists for which the AV and CV interval estimators have probability of coverage as large as the crude interval, we have shown that the corresponding point estimators retain the advantage of smaller MSE under that design. However, this result gives no insight into the actual value of the MSE after additional deletion. As an illustration, consider the AR(1) process described above and the AV point and interval estimator. We choose the AR(1) process as an example because all of the quantities of interest can be directly calculated (see the Appendix).

A more traditional formulation of the AR(1) process is

$$Y_i = \theta + \phi(Y_{i-1} - \theta) + \epsilon_i, \; i = 1,2,\ldots,m$$

where Y_i is the ith output in a simulated replication. We represent the choice of initial condition by y_0, which is a constant and the same for all replications. Kelton and Law [10] provide expressions for the variance, bias and MSE of $\bar{Y}(k,m,d)$. From these results we can show that

$$\delta^2(k,m,d) = kc^2\left[\frac{(\phi^{d+1}(1-\phi^{m-d}))^2}{(m-d)(1-\phi^2) - \phi(1-\phi^{m-d})(2+\phi^{2d+1}(1-\phi^{m-d}))}\right] \qquad (9)$$

where $c = (y_0 - \theta)\sqrt{(1-\phi^2)/\sigma_\epsilon^2}$ is the difference between y_0 and θ in units of the steady-state $(m \to \infty)$ standard deviation of Y_i, and $\sigma_\epsilon^2 = \mathrm{Var}[\epsilon_i]$. For a given k, m and d, we find d' such that $\hat{\delta}(k,m,d') \approx \delta(k,m,d)$ by direct search. We choose d to be the asymptotically $(m \to \infty)$ optimal truncation point in terms of minimum MSE for the crude estimator (Snell and Schruben [20])

$$d = 0, \quad \text{if ?} < \frac{1-\phi^2}{@}$$

$$\frac{-\ln(\phi^2(c^2-1))}{2\ln\phi}, \quad \text{otherwise}$$

(this expression for d corrects a typographical error in Snell and Schruben). Thus, we begin with the best design for the crude estimator and improve it using variance reduction.

We examine the effect of dependence in the output process by varying ϕ from 0 (independent) to .99 (highly positively dependent). Let $\varrho = -.5$ be the value of the negative correlation induced by AV, which is optimistic in practical problems. However, $\varrho = -.5$ provides a good illustration since the smaller ϱ is the more additional deletion is required. We reserve "AV" for the application of antithetic variates without additional deletion; and AV+ for antithetic variates with $d' \geq d$ deleted to obtain probability of coverage equal to the crude estimator.

For $k = 10$ replications of length $m = 128$ outputs per replication, Figure 3 shows the MSE, expected half width, and probability of coverage for the crude, AV, and AV+ point and interval estimators when $c = 1$. As expected, all measures for all estimators are progressively degraded as ϕ increases. The important observation is that AV+ has only slightly larger MSE and expected half width than AV, yet has probability of coverage equal or superior to the crude estimator.

Next we fix the budget at n outputs divided into k replications of length $m = n/k$, with d outputs deleted from each replication. Specifically, let $\phi = .9$, $\varrho = -.5$, and $n = 15000$. Figures 4(a-c) and 5(a-c) show the MSE, expected half width, and probability of coverage as a function of k for the crude, AV and AV+ point and interval estimators.

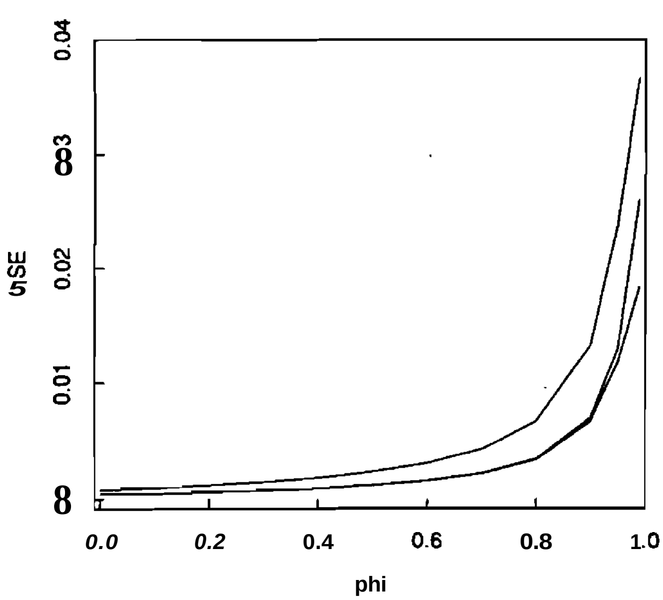

Figure 3a. For $k = 10$, $m = 128$ and $c = 1$: From top to bottom, the MSE of the crude, AV+, and AV point estimators in units of the steady-state variance.

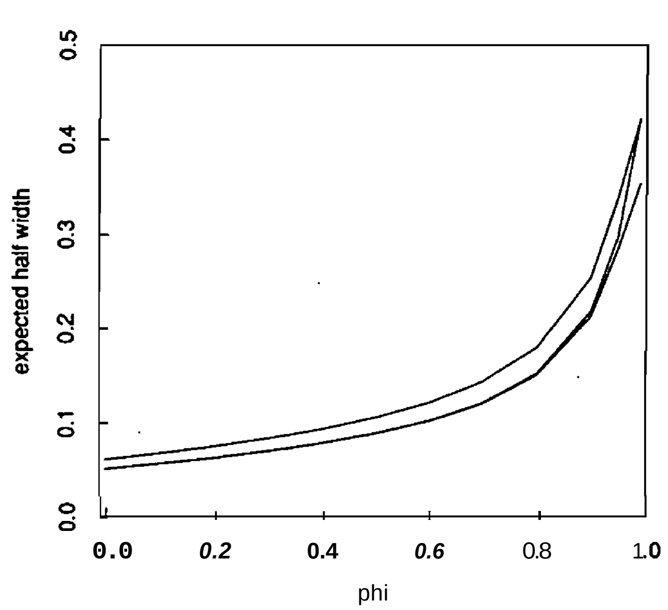

Figure 3b. For $k = 10$, $m = 128$ and $c = 1$: From top to bottom, the expected half width of the crude, **AV +**, and **AV** interval estimators in units of the steady-state standard deviation.

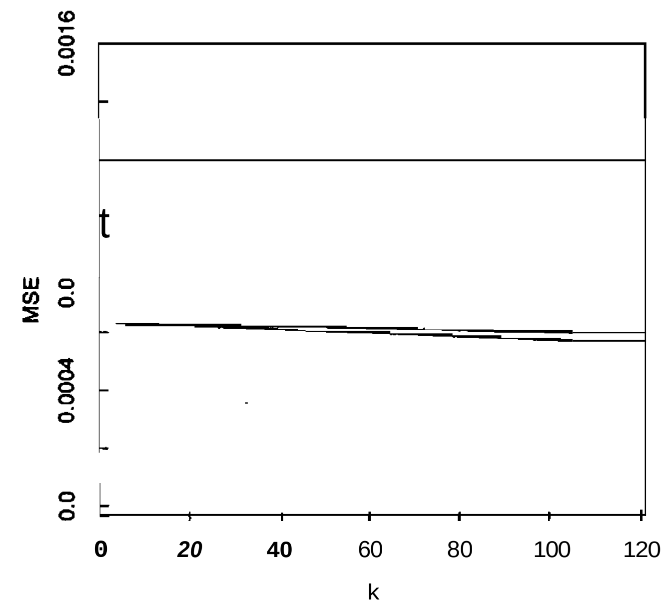

Figure 4a. For $\phi = .9$, $\rho = -.5$ and $c = 1$: From top to bottom, the MSE of the crude, **AV +**, and **AV** point estimators in units of the steady-state variance.

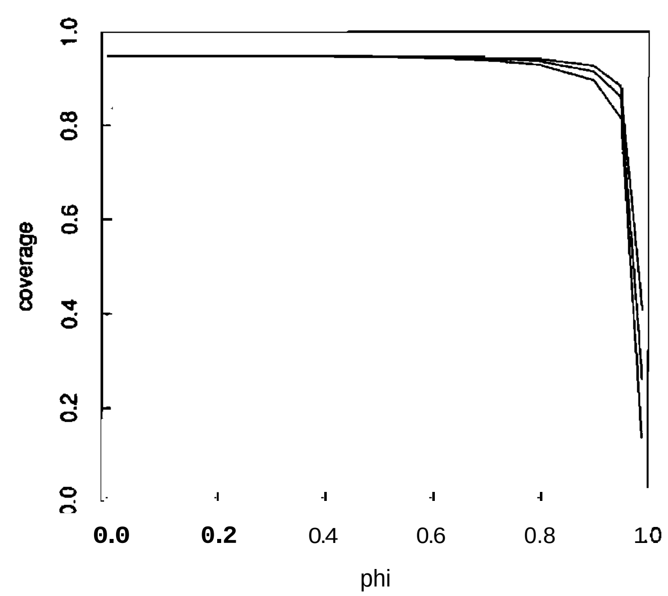

Figure 3c. For $k = 10$, $m = 128$ and $c = 1$: From top to bottom, the probability of coverage for $a = .05$ **of the AV +**, crude, and **AV** interval estimators.

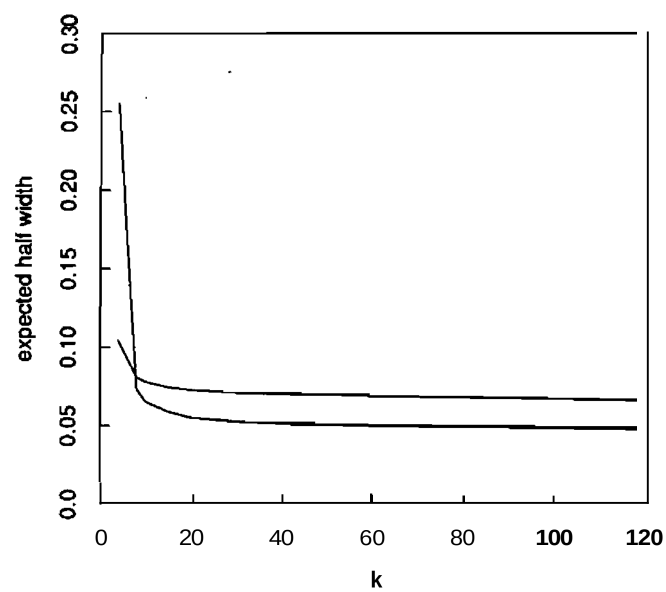

Figure 4b. For $\phi = .9$, $\rho = -.5$ and $c = 1$: From top to bottom, **the** expected half width of the crude, **AV+**, and **AV** interval estimators in units of the steady-state standard deviation; **AV +** and **AV** are indistinguishable.

From Snell and Schruben [20], the asymptotically optimal truncation point is $d = 0$ when $c = 1$, and $d = 3$ when $c = 3$, and we determined that $d' = 4$ and 7, respectively.

When c is small (Figure **4a**), **MSE** is fairly constant for all values of k and the additional deletion **has** little effect on the **AV** estimator. When c is large (Figure **5a**), **MSE** increases **as** k increases, and the bias is **so** large that **AV+** actually has smaller **MSE** than **AV** when the replications are very short. For **small** k, $E[H_{a\prime}] > E[H]$, **but as** k increases the **AV** interval has shorter expected length, and the additional deletion has little effect. Finally, the probability of coverage for the **AV+** interval is **superior** for **all** values of k, **as** guaranteed by our choice of d'.

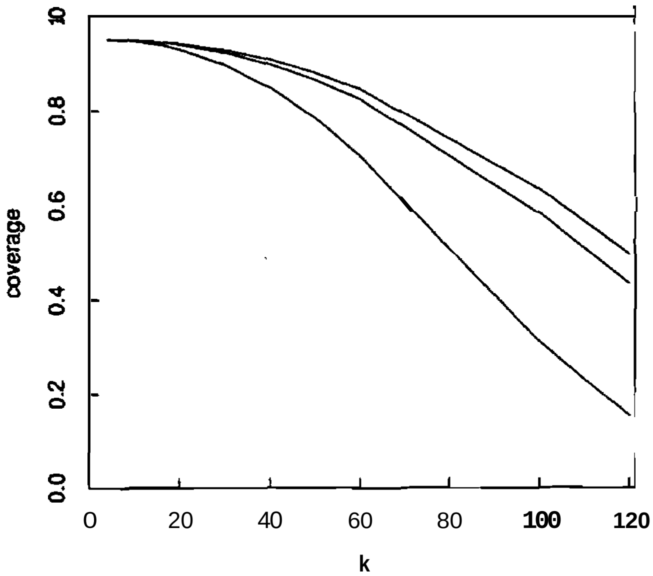

Figure 4c. For $\phi = .9$, $\varrho = -.5$ and $c = 1$: From top to bottom, the probability of coverage for $\alpha = .05$ of the AV+, crude, and AV interval estimators.

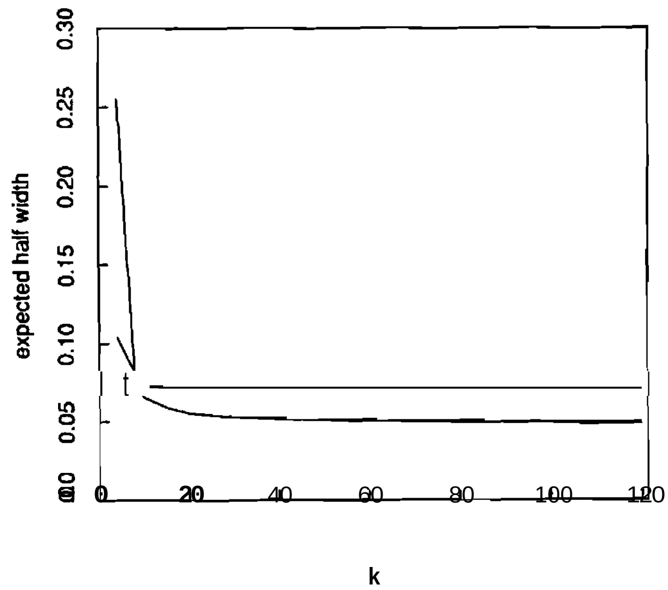

Figure 5b. For $\phi = .9$, $\varrho = -.5$ and $c = 3$: From top to bottom, the expected half width of the crude, AV+, and AV interval estimators in units of the steady-state standard deviation; AV+ and AV are Indistinguishable.

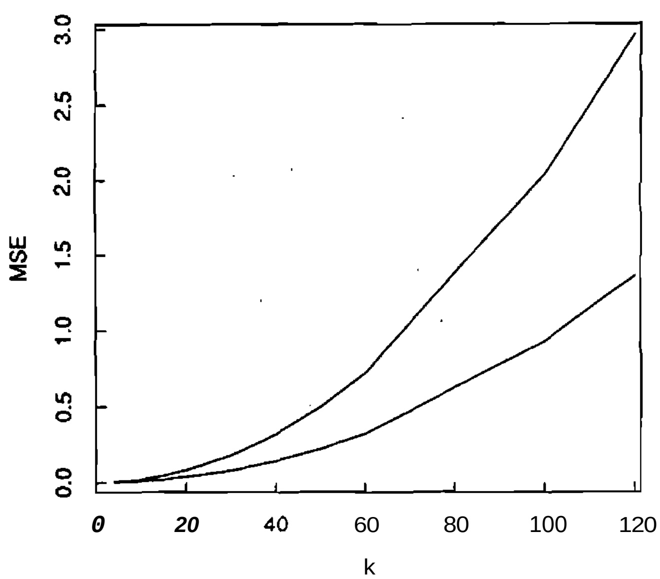

Figure 5a. For $\phi = .9$, $\varrho = -.5$ and $c = 3$: From top to bottom, the MSE of the crude, AV, and AV+ point estimators in units of the steady-state variance; crude and AV are indistinguishable.

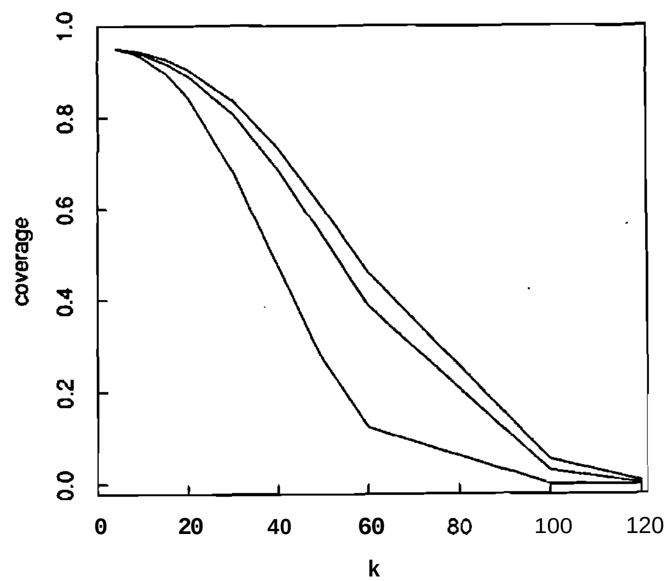

Figure 5c. For $\phi = .9$, $\varrho = -.5$ and $c = 3$: From top to bottom, the probability of coverage for $a = .05$ of the AV+, crude, and AV interval estimators.

Conclusions and Recommendations

The results above show that AV and CV VRTs can improve the performance of point and interval estimators in the presence of bias. However, the improvement requires additional deletion, leaving the difficult question of how much additional deletion. We propose the following approximation:

We assume that for $i > d$, $b_i \approx c\phi^i$ for some constants c and $0 < \phi < 1$; that is, the bias coverges to zero geometrically fast. We approximate $v(m,d)$ by $\sigma^2/(m-d)$, where

$$\sigma^2 = \lim_{m \to \infty} m \operatorname{Var}[\bar{Y}_j(m,d)].$$

Then

$$\delta^2(k,m,d) \approx \frac{kc^2\phi^{2d+2}}{(1-\phi)^2\sigma^2/(m-d)} . \qquad (10)$$

Equation (10) implies that $\delta^2(k,m,d+\Delta) \approx \nu\delta^2(k,m,d)$, where $0 < \nu < 1$ is the variance reduction ratio, when Δ satisfies

$$\left[\frac{\phi^{\Delta} - \phi^{m-d}}{1 - \phi^{m-d}}\right]^2 = \nu\left[1 - \frac{\Delta}{m-d}\right]. \qquad (11)$$

Letting m go to infinity on both sides of (11) and solving for Δ yields

$$\Delta = \frac{\ln\nu}{2\ln\phi} .$$

This result is independent of m and d, and will be a reasonable approximation when m is large, and much larger than d; it applies to either single or multiple replication designs.

In practical problems ν can be estimated by the ratio of the estimated variances of the crude and VRT point estimators. Both AV and CV VRTs allow the variance of the crude estimator to be estimated without performing a second experiment. The parameter ϕ can either be estimated by approximating the output process by an $AR(1)$ process, or can be chosen close to 1 to be conservative. Table 1 shows the additional deletion increment Δ for different values of ν and ϕ. These values can be used as a guide, and they indicate that little additional deletion is needed unless the variance reduction ratio is quite small and/or the dependence in the output process is very strong.

We have emphasized obtaining coverage equal to the crude interval estimator. However, it is sometimes possible to achieve improved coverage and smaller MSE by incorporating a VRT, as shown in the examples. Nelson [15] describes a modification of AV, called antithetic-variate splitting (AVS), for use in replication deletion designs. AVS "splits" k/s replications at the point of truncation (Y_d) into $s \geq 2$ dependent subreplications. AVS can reduce point-estimator variance below classical AV, and it reduces bias by using the savings $(kd/s$ deleted rather than $kd)$ to make longer replications. Thus, the AVS estimator may have coverage superior to the crude estimator even without additional deletion.

Table 1. Additional Deletion Increment Δ

ν	ϕ 0.20	0.50	0.80	0.90	0.99
0.90	0	0	0	1	5
0.80	0	0	1	1	11
0.70	0	0	1	2	18
0.60	0	0	1	2	25
0.50	0	1	2	3	34
0.40	0	1	2	4	46
0.30	0	1	3	6	60
0.20	1	1	4	8	80
0.10	1	2	5	11	115

We have only considered deterministically initialized replications, since that is the most common approach. However, procedures that randomly select initial conditions in a way that reduces bias have been proposed (Deligonul [3], Kelton [8]). VRTs might also be incorporated into these designs.

Although only mentioned briefly, an important problem is estimating the difference between the parameters of two stochastic processes (case (B) in Background above). Nelson [13] describes the impact of the common-random-numbers VRT on single and multiple replication designs in the absence of initial-condition bias. Modification of these results to account for the presence of bias would follow the same lines as the development for AV here, and the conclusions would be similar.

Acknowledgements

This research was partially supported by the National Science Foundation under Grant No. ECS-8707634. The author benefited from discussions with John B. Neuhardt, Marc E. Posner, and Charles H. Reilly of The Ohio State University.

REFERENCES

[1] Bickel, P. J., and K. A. Doksum (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden-Day, Inc., San Francisco.

[2] Cheng, R.C.H. (1976). "A Note on the Effect of Initial Conditions on a Simulation Run," *Operational Research Quarterly, 27*, 2, 467-470.

[3] Deligonul, Z. S. (1987), "Antithetic Bias Reduction for Discrete-Event Simulations." *Journal of the Operational Research Society. 38*, 5, 431-437.

[4] Fishman, G. S. (1972), "Bias Considerations in Simulation Experiments," *Operations Research,* 20, 785-790.

[5] Gafarian, A. V., C. J. Ancker, Jr., and T. Morisaku (1978). "Evaluation of Commonly Used Rules for Detecting 'Steady State' in Computer Simulation," *Naval Research Logistics Quarterly, 25*, 3, 511-529.

[6] Glynn, P. W. (1984), "Some Asymptotic Formulas for Markov Chains with Applications to Simulation," *Journal of Statistical Computation and Simulation. 19,* 97-112.

[7] Kelton, W. D. (1986). "Replication Splitting and Variance for Simulating Discrete-Parameter Stochastic Processes," *Operations Research Letters, 4,* 6, 275-279.

[8] Kelton, W. D., (1989), "Random Initialization Methods in Simulation," *IIE Transactions, 21,* 4, 355-367.

[9] Kelton, W. D., and A. M. Law (1983), "A New Approach for Dealing with the Startup Problem in Discrete Event Simulation," *Naval Research Logistics Quarterly, 30.* 641-658.

[10] Kelton, W. D., and A. M. Law (1984), "An Analytical Evaluation of Alternative Strategies in Steady-State Simulation," *Operations Research.* 32, 1, 169-184.

[11] Law, A. M., and W. D. Kelton (1982), *Simulation Modeling and Analysis*, McGraw-Hill, N.Y.

[12] Nelson, B. L. (1987), "Variance Reduction for Simulation Practitioners," *Proceedings of the 1987 Winter Simulation Conference,* 43-51.

[13] Nelson, B. L. (1987), "Some Properties of Simulation Interval Estimators Under Dependence Induction," *Operations Research Letters, 6*, 4, 169-176.

[14] Nelson, B. L. (1987), "A Perspective on Variance Reduction in Dynamic Simulation Experiments," *Communications in Statistics, B16*, 2, 385-426.
[15] Nelson, B. L. (1988), "Antithetic-Variate Splitting for Steady-State Simulations," *European Journal of Operational Research, 36*, 360-370.
[16] Nelson, B. L. (1989). "Batch Size Effects on the Efficiency of Control Variates in Simulation," *European Journal of Operational Research*, 43, 2, 184-1%.
[17] Schmeiser, B. (1982), "Batch Size Effects in the Analysis of Simulation Output," *Operations Research*, 30, 556-568.
[18] Schruben, L. W. (1982), "Detecting Initialization Bias in Simulation Output," *Operations Research*. 30, 3, 569-590.
[19] Schruben, L.. H. Singh, and L. Tierney (1983), "Optimal Tests for Initialization Bias in Simulation Output," *Operations Research, 31*, 6, 1167-1178.
[20] Snell, M., and L. Schruben (1985), "Weighting Simulation Data to Reduce Initialization Effects," IIE *Transactions, 17*, 4, 354-363.
[21] Turnquist, M. A., and J. M. Sussman (1977), "Toward Guidelines for Designing Experiments in Queueing Simulation," *Simulation, 28*, 5. 137-144.
[22] Wilson, J. R. (1984). "Variance Reduction Techniques for Digital Simulation," *American Journal of Mathematical and Management Sciences, 4, 3* & 4, 277-312.
[23] Wilson, J. R., and A. A.B. Pritsker (1978), "A Survey of Research on the Simulation Startup Problem," *Simulation, 31*, 2, 55-58.
[24] Wilson, J. R., and A.A. B. Pritsker (1978), "Evaluation of Startup Policies in Simulation Experiments," *Simulation, 31*, 3, 79-89.

Appendix

Results are derived and calculations explained in this section.

Large-Sample Coverage

As the number of replications, k, goes to infinity, $\beta(k,m,d)$ goes to 0, since

$$\lim_{k\to\infty} \delta^2(k,m,d) = \lim_{k\to\infty} \frac{\sqrt{k}\,\bar{b}^2(m,d)}{\nu(m,d)}$$

is infinite. Thus, the crude, AV, and CV intervals all have probability of coverage 0, asymptotically, which is not a useful result. However, suppose instead that $\delta(k,m,d)$ remains constant as k increases. Then

$$T(k-1,\delta(k,m,d)) \xrightarrow{D} N(\delta(k,m,d),\ 1)$$

(Bickel and Doksum [1]), where $N(a,$ b) denotes a normal distribution with mean a and variance b, and D denotes convergence in distribution. An analogous result holds for noncentrality parameters $\hat{\delta}(k,m,d)$ and $\tilde{\delta}(k,m,d)$. This shows that, for k large, probability statements about noncentral-t random variables become statements about normal random variables with mean equal to the noncentrality parameter and unit variance. Thus, equal probability of coverage requires equal noncentrality parameters.

Small-Sample Coverage

The values of 6, $\hat{\delta}$, and $\tilde{\delta}$ for different k were found using IMSL routine MDSTI to provide quantiles of the t distribution, IMSL routine MDTN to evaluate the cumulative distribution function of the noncentral t distribution, and a bisection search to find δ, $\hat{\delta}$, and $\tilde{\delta}$ satisfying the probability statements. An absolute error tolerance of .000001 was specified.

AR(1) Example

From Kelton and Law [10]

$$\bar{b}(m,d) = \frac{(y_0 - \theta)\phi^{d+1}(1 - \phi^{m-d})}{(m - d)(1 - \phi)}$$

and

$$\nu(m,d) = \frac{\sigma_\epsilon^2}{(m-d)(1-\phi)^2}\left(1 - \frac{\phi(1-\phi^{m-d})(2+\phi^{2d+1}(1-\phi^{m-d}))}{(m-d)(1-\phi^2)}\right).$$

Substituting these results into the definitions yields $\delta(k,m,d)$ and $\mathrm{MSE}[\bar{Y}(k,m,d)]$. Also from Kelton and Law,

$$\mathrm{E}[H(k,m,d)] = t_{\alpha/2}(k-1)\sqrt{\frac{2}{k-1}}\,\frac{\Gamma(k/2)}{\Gamma((k-1)/2)}\sqrt{\frac{\nu(m,d)}{k}}$$

where Γ is the gamma function. The probability of coverage, $\beta(k,m,d)$, was calculated using the IMSL routine MDTN to evaluate the noncentral-t cumulative distribution function. Results for AV are immediate by modifying the variance and degrees of freedom.

Barry L. Nelson is an associate professor in the Department of Industrial and Systems Engineering at The Ohio State University. His Ph.D. is from the School of Industrial Engineering at Purdue University. His research interests are 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, and TIMS. Dr. Nelson is an Associate Editor for *Naval Research Logistics*, and is Secretary-Treasurer of the TIMS College on Simulation.