*Proceedings of the 1995 Winter Simulation Conference*
*ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman*

# ON BATCH MEANS IN THE SIMULATION AND STATISTICS COMMUNITIES

Michael Sherman

Department of Statistics
Texas A&M University
College Station, TX 77843, U.S.A.

## ABSTRACT

Batching is a well known technique for estimating the variance of point estimators computed from simulation experiments. The batch statistic variance estimator is simply the (appropriately scaled) sample variance of the estimator computed on subsets of data. The simulation and statistics communities seem to be largely unaware of each other's results in this area. Some empirical and theoretical results from the simulation and statistics literature will be discussed and compared.

In particular, we discuss the important issue of selecting batch size and present a new data based method for determining it. The basic idea is to empirically estimate the optimal batch size for a smaller simulation length, and then extrapolate using knowledge of the optimal order of magnitude of batch length for the original simulation length. We provide a small simulation showing the effectiveness of the proposed method.

## 1 INTRODUCTION

Consider the following scenario: We observe the output sequence $\{X_i : 1 \leq i \leq n\}$ from a simulation run, and compute a statistic of interest, $s_n := s_n(X_1, ..., X_n)$, which estimates a real valued parameter, $\theta$. In order to draw inferences from $s_n$ to $\theta$ (e.g., confidence intervals, hypothesis tests) we need an estimate of $Var(s_n)$.

Due to potential serial dependence in the output sequence, estimating $Var(s_n)$ may be quite difficult, particularly if the statistic $s_n$ is complicated. Even for the statistic $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, however, $nVar(s_n) \to \sum_{i=-\infty}^{\infty} Cov(X_0, X_i)$, and it is nontrivial to get accurate estimates of these covariances, or to know at what lag the covariances become negligible.

For this reason, estimators based on reusing the data, variously known as batch, subsampling, resampling, standardized time series estimators have been proposed. We will refer to all these methods as batch methods. The basic idea of all batch methods is the same: compute the statistic of interest on smaller series of consecutive observations as "replicates" of $s_n$, and compute a standardized sample variance of these replicates to estimate $Var(s_n)$.

In Section 2 we formally introduce some of these estimators and present some results from the simulation and statistics communities exploring the efficacy of the estimators. We discuss how the success of all batch methods depends crucially on the choice of batch size (the length of the replicates). Section 3 presents a data based method for determining the optimal batch length for the replicates and presents results from a small simulation. Finally, possible methods of improving the suggested algorithm are discussed.

It should be noted that there are a wide variety of alternatives to batching, e.g., spectral analysis, time series models, and regeneration. These methods are discussed by, e.g., Fishman (1978) and Bratley, Fox, and Schrage (1987). In the sequel we only consider batch variance estimators.

## 2 SOME RESULTS ON BATCHING

Let $\{X_i : 1 \leq i \leq n\}$ denote the simulation output and assume that these observations come from a stationary sequence, i.e., the simulation has reached steady state. Let

$$X_b^m := (X_{m+1}, ..., X_{m+b}), m = 0, ..., (n-b),$$

denote the "subseries" or "batches" of length $b$ starting with the $(m+1)$'th observation, so that in particular, $X_n^0$ denotes the entire output. The important point is that for each $m$, $X_b^m$ maintains the same dependence structure as the original sequence $X_n^0$. Let

$s_b^m$ denote the corresponding subseries replicates, or batch statistics, $s_b^m := s_b(X_b^m)$.

Two natural estimators of $nVar(s_n)$ are $V_N$, based on using all possible nonoverlapping replicates and $V_O$, based on all possible replicates. These are defined as:

$$V_N := b\sum_{i=0}^{k-1}\frac{(s_b^{ib}-\bar{s})^2}{k},$$

where $k$ is the largest integer less than $n/b$, and

$$V_O := b\sum_{i=0}^{n-b}\frac{(s_b^{i}-\bar{s})^2}{(n-b+1)}.$$

It should be noted that various authors use slightly different constants in the definitions of $\hat{V}_N$ and $\hat{V}_O$, but all are reasonably close for small values of the ratio $b/n$.

There have been potentially hundreds of papers written about these and related estimators so the following will not be in any way comprehensive. We will examine the basic issues of consistency of the two estimators, efficiency of the two estimators, distributional results related to the estimators, and the important issue of optimal batch size.

### 2.1 Some Preliminary Definitions

In all that follows we assume that the standardized variance of the statistic of interest converges to a nonzero constant, i.e.,

$$V_n := nVar(s_n) \to V \in (0,\infty).$$

We assume that $V_n$ is close to V and thus we take V to be the object of interest in this paper. For any sequence of random variables $X_n$, we will write

$$X_n \xrightarrow{L_2} C$$

if $E(X_n - C)^2 \to 0$ as $n \to \infty$. Practically, $L_2$ convergence of an estimator implies small bias and small variance for long output sequences. $X_n \xrightarrow{D} X$ will denote convergence in distribution to the random variable $X$.

We define strong mixing as follows. For integer $k > 0$, let $F(X_{-\infty}^0)$ denote the $\sigma$-field generated by the observations $(..., X_{-1}, X_0)$, and let $F(X_k^\infty)$ denote the $\sigma$-field generated by the observations $(X_k, X_{k+1}, ...)$. Loosely, $F(X_{-\infty}^0)$ denotes the set of all events depending on observations up to $X_0$. Define:

$$\alpha(k) := sup(|P(AB) - P(A)P(B)|),$$

where the sup is taken over $A \in F(X_{-\infty}^0)$ and $B \in F(X_k^\infty)$.

The underlying process is said to be strongly mixing if $\alpha(k) \to 0$ as $k \to \infty$. Roughly, this means that dependence between observations becomes weak at long lags.

### 2.2 Consistency Results

Several results demonstrate different modes of convergence for the estimators $\hat{V}_N$ and $\hat{V}_O$. For example, Carlstein (1986a) gives the following:

*Result* 1: Assume that the sequence $\{X_i : \infty < i < \infty\}$ is strongly mixing, and let $t_n := n^{1/2}(s_n - Es_n)$. Assume that $b \to \infty, b/n \to 0$. If

$$E|t_n|^{4+\delta} < \infty,$$

for some $\delta > 0$, then

$$\hat{V}_N \xrightarrow{L_2} V.$$

It turns out that under the same conditions we have that

$$\hat{V}_O \xrightarrow{L_2} V.$$

For the special important statistic $\overline{X}$ we can obtain the following corollary:

*Corollary* 1: Assume that $s_n = \overline{X}$. Under the conditions of Result 1 we have: If

$$EX_0^8 < \infty$$

and $\alpha(k) \leq k^{-4}$, for large $k$, then

$$\hat{V}_N \xrightarrow{L_2} V \text{ and } \hat{V}_O \xrightarrow{L_2} V.$$

The consistency in Corollary 1 has also been proven under various assumptions by, e.g., Damerji (1994), Goldsman and Melamed (1992), and for a class of linear statistics by Politis and Romano (1993). Consistency results have been proven for general spatial statistics by, e.g., Possolo (1991) and Sherman (1995).

### 2.3 Comparing $\hat{V}_N$ and $\hat{V}_O$

Both $\hat{V}_N$ and $\hat{V}_O$ are $L_2$ consistent under the same conditions in Result 1. This raises the natural question as to which is preferable. It is not difficult to see that the two have the same expectation so we need to compare their variances. $\hat{V}_O$ has many more summands so it may hope to be less variable, but the replicates contain a great deal of redundancy so it is not clear it is at all advantageous. It is difficult to compare for a general statistic $s_n$, so we now assume that $s_n = \overline{X}$, the sample mean.

Meketon and Schmeiser (1984) show the following:

*Result* 2: Assume that $b \to \infty$ and $b/n \to 0$. Then

$$\frac{Var(\hat{V}_O)}{Var(\hat{V}_N)} \to 2/3.$$

Thus, if one can afford the additional computational burden, it is preferable to use all possible batches of length $b$. This result has also been obtained by Künsch (1989) in his study of the "blockwise bootstrap". His basic idea is to choose $k$ batches of length $b$ from the set of all possible batches of length $b$, "glue" them together, and calculate $s_n^*$, say. If this is done $B$ times then the variance estimator is simply the (unstandardized) sample variance of the $B$ values of $s_n^*$.

Result 2 was actually obtained by Cox and Lewis (1961) in the case of estimating the intensity of a Poisson process. In fact, they allow any amount of overlap between the batches in their formulation. In particular, let $\hat{V}_H$ and $\hat{V}_T$ denote the variance estimators with half and 3/4 overlap, respectively. Their results show that

$$\frac{Var(\hat{V}_O)}{Var(\hat{V}_H)} \to 8/9$$

and

$$\frac{Var(\hat{V}_O)}{Var(\hat{V}_T)} \to 32/33.$$

Thus, the "partially" overlapping $\hat{V}_H$ and $\hat{V}_T$ may be an attractive alternative to $\hat{V}_O$ if computations are costly.

We mention a related procedure that uses standardized time series (these are weighted batch statistics) as replicates suggested by Schruben (1983). These estimators are sometimes less variable than $\hat{V}_N$, although only nonoverlapping (weighted) batches are used — so the computational burden is similar. Further improvement can sometimes be obtained by combining a standardized time series estimator with $\hat{V}_N$ (Goldsman and Schruben 1984). The covariance between two (combined) estimators has be en studied by Pedrosa and Schmeiser (1993).

## 2.4 Some Distributional Properties

The simulation community often wants to guard against making the batch lengths too small. For this reason, often the number of batches, $k$, is considered fixed at some moderate number, like 10 to 30 (Schmeiser 1982), which often gives reasonable numerical results. Further, this suggests a certain approach to distribution theory based on batching. Consider $\hat{V}_N$. For fixed $k \simeq (n/b)$, this estimator does not converge to $V$. Nevertheless, for large $n$ (and hence large $b$), each batch statistic is approximately normally distributed and the batches are approximately independent. This suggests considering the standardized batch means as a random sample from a normal distribution, and forming the associated $t$-statistic. Towards this end, Glynn and Inglehart (1990) give a continuous time analog to the following (under the assumption that the continous output process can be approximated by Brownian Motion):

*Result* 3: Assume for simplicity that $kb = n$. For $i = 1, ..., k$, let $\overline{X_i} = \sum_{j=(i-1)b+1}^{ib} X_j/b$. Then for fixed $k$, as $b \to \infty$:

$$k^{1/2}(\sum_{i=1}^{k} \overline{X_i}/k - \mu)/(\hat{V}_N/b)^{1/2} \xrightarrow{D} t_{k-1}, \qquad (1)$$

where $t_k$ denotes the student's $t$-distribution with $k$ degrees of freedom. This result is then used to obtain a confidence intervals for $\mu$.

A result analogous to (1) was given for a general statistic, $s_n$, by Carlstein (1986b), and extended to spatial statistics by Sherman (1994). The main assumption in deriving these results is that $Corr(s_n, s_b^m) \simeq (b/n)^{1/2}$. This "mean-like" assumption was also made by Schmeiser, Avramidis, and Hashem (1990) in their study of the variance estimator $\hat{V}_O$. The reason for this terminology is that when $s_n = \overline{X}$ computed from i.i.d. data, $Corr(s_n, s_b^m) = (b/n)^{1/2}$ (exactly).

Distributional results for the variance estimators themselves, as well as an extensive numerical study comparing coverage rates of confidence intervals are given by, e.g., Sargent, Kang, and Goldsman (1992).

## 2.5 Asymptotically Optimal Choice of Batch Size

The previous asymptotic results have been important in justifying the use of batching for variance estimation. However, for either of $\hat{V}_N$ or $\hat{V}_O$, for any given simulation, the important practical question is: How should we select batch size, $b$? In the sequel we define the optimal $b$ to be the one that minimizes $MSE(\hat{V}) := Bias^2(\hat{V}) + Var(\hat{V})$. In this section we discuss two results that give the order of magnitude of $b$ as a function of $n$.

Carlstein (1986a) adresses the question of batch size by examining the statistic $X$ in the special case where the output sequence is generated by the AR(1) process:

$$X_i = \rho X_{i-1} + \epsilon_i, \qquad (2)$$

where $\epsilon_i$'s are independent standard normal random variables. He shows that in this situation

$$Bias(\hat{V}_N) = \frac{-2\rho}{(1-\rho)^3(1+\rho)}(1/b) + o(1/b),$$

$$Var(\hat{V}_N) = \frac{2}{(1-\rho)^4}(b/n) + o(b/n),$$

and thus that the asymptotically optimal $b$ is

$$b_{opt} = \left(\frac{2\rho}{1-\rho^2}\right)^{2/3} n^{1/3}. \qquad (3)$$

This shows that larger batch sizes are needed for stronger correlations in the sequence, which is intuitively reasonable. Equation (3) also suggests a method for obtaining $b$ in any given situation. Assume (temporarily) that the sequence is generated by an AR(1) process, estimate $\rho$ (e.g., by Least Squares), and plug the resulting $\hat{\rho}$ into Equation (3).

Song and Schmeiser (1988) discuss a more general result for the estimator $\hat{V}_O$: Assume that

$$Bias(\hat{V}_O) = -(1/b)c_b\gamma_1 + o(1/b)$$

and

$$Var(\hat{V}_O) = (b/n)c_v\gamma_0^2 + o(b/n).$$

Then

$$b_{opt} = \frac{2c_b^2\gamma_1^2}{c_v\gamma_0^2}n^{1/3}, \qquad (4)$$

where $c_b$, $c_v$ are constants depending on the process and $\gamma_i$ depends on an infinite sum of covariances in the process.

These results giving the order of magnitude for the optimal $b$ aid in the derivation of a model free (e.g., without assuming an AR(1) structure) data based determination of appropriate batch size $b$ presented in Section 3.

## 3 A MODEL FREE, DATA BASED CHOICE OF BATCH SIZE

The statistics and simulation communities have long sought an effective method of determining an appropriate batch size without making assumptions on the mechanism generating the output (e.g., AR(1)). Noting just two comments from the recent literature, Sargent, Kang, and Goldsman (1992) say that "a good batch size estimation procedure would be of tremendous importance" while Damerji (1994) says that "batch-size selection is still an unresolved problem".

We will focus on the estimator $\hat{V}_O$ although the proposed method is generally applicable to the other batch variance estimators discussed. We consider only $s_n = \overline{X}$, although the method may well be applicable to other "mean-like" statistics. The basic idea is to empirically estimate the best $b$ for a sequence of smaller length, $m$, and then extrapolate to obtain the best $b$ for the entire sequence of length $n$. The basis of this is as follows: Note from the results of Section 2.5 (equations (3) and (4)) that the optimal $b$ is of the form $b_n = Cn^{1/3}$ where the constant $C$ depends only on the underlying process and not on $n$. For any shorter sequence of length $m$, say, we have $b_m = Cm^{1/3}$ and thus $b_n = (n/m)^{1/3}b_m$. If we can estimate $b_m$ by $\hat{b}_m$ then our estimate of $b_n$ will simply be $\hat{b}_n = (n/m)^{1/3}\hat{b}_m$. Towards this end we give:

*The Algorithm for Estimating $b_n$*

1) Choose a pilot value for $b$, $b = b^*$, say, and calculate $\hat{V}_O$ using $b = b^*$.

2) For some $m$, consider $X_m^i, i = 1, ..., (n-m+1)$, all possible series of length $m$. For the $i$'th series of length $m$, let $\hat{V}_{m^*}^i$ denote the batch variance estimator computed from the series $X_m^i$ using a batch size of $m^*$, and define:

$$\hat{b}_m = argmin_{m^*} \sum_{i=1}^{n-m+1} \left(\hat{V}_{m^*}^i - \hat{V}_O\right)^2/(n-m+1).$$

This is the empirical estimate of the $b_m$ that minimizes $MSE(\hat{V}_O)$ for a sequence of length $m$.

3) Compute $\hat{b}_n = (n/m)^{1/3}\hat{b}_m$

4) Set $b^* = \hat{b}_n$ and repeat steps 1) to 3).

The algorithm could be iterated if desired. In the simulations in Section 3.1 the algorithm converged in the single iteration approximately 60 percent of the time, and in the simulation experiment no further iteration was performed. We note that a similar algorithm for the purpose of estimating the distribution function or the bias of an estimator has been suggested by Hall and Jing (1994).

### 3.1 A Simulation Experiment

We performed a small simulation experiment to study the effectiveness of the proposed algorithm. The model generating the output sequence is an AR(1) sequence as described in Equation (2). From Equation (3) we know the asymptotically optimal $b$, and we take this to be the correct value. Simulation experiments have shown that for all cases considered the asymptotic values coincide approximately with finite sample optimality. The optimal values of $b$ for each setting considered are given in Table 1.

Table 1: Optimal Batch Sizes for the AR(1) Process

| | $\rho$ | |
|---|---|---|
| $n$ | .5 | .8 |
| 200 | 8 | 16 |
| 1000 | 13 | 28 |

Ten output sequences of length 200 were generated from Equation (2) with $\rho = .5$. In each case the pilot value in Step 1 was $b^* = 20$ and $m = 10$. In some cases the algorithm chooses the largest possible $m^*$ in step 2. For this reason we take the *argmin* over a restricted set of $m^*$ values which rules out unreasonable estimates of $\hat{b}_n$ (this seems to be necessary to avoid occasionally bad estimates, see Table 2). In this case the search set for $\hat{b}_m$ is $(1, ..., 8)$.

The detailed results are given in Table 2. From Table 1, we see that the correct value is $b = 8$. Columns 2 and 3 give the values from steps 2 and 3 in the algorithm, while columns 4 and 5 give the values on the next interation. We take the last column to be the final estimate of $b$.

Table 2: Determination of Optimal Block Size, $b = 8$

| Sim. | $\hat{b}_m$ | $\hat{b}_n$ | $\hat{b}_m$ | $\hat{b}_n$ |
|---|---|---|---|---|
| 1 | 3 | 8 | 3 | 8 |
| 2 | 8 | 22 | 8 | 22 |
| 3 | 2 | 5 | 3 | 8 |
| 4 | 6 | 16 | 4 | 10 |
| 5 | 1 | 3 | 2 | 5 |
| 6 | 4 | 10 | 4 | 10 |
| 7 | 8 | 22 | 8 | 22 |
| 8 | 4 | 10 | 3 | 8 |
| 9 | 4 | 10 | 4 | 10 |
| 10 | 5 | 14 | 4 | 10 |

In 5 of 10 trials the algorithm converged in a single iteration. The revised estimate moved closer to the correct value in all 5 trials for which the algorithm did not converge in one iteration. On the other hand, in two of the trials the algorithm chose the largest possible value $\hat{b}_m = 8$.

We extended the simulations as follows: For each of two output lengths $n = 200$ and $n = 1000$, for each of two values of the AR(1) parameter $\rho = .5$ and $\rho = .8$, and for each of two pilot values $b^*$, we generated 20 output sequences, and applied the algorithm to each sequence to obtain $\hat{b}_n$. In all cases $m = n/20$ and for $n = 1000$ the search set for $\hat{b}_m$ was $(1, ..., 25)$. In Table 3 for each setting we give $\bar{b}$, the average of the 20 $\hat{b}_n$'s, and its estimated standard error.

Table 3: Determining Optimal Block Size

| Setting | pilot $b^*$ | $\bar{b}$ | st. error |
|---|---|---|---|
| $n = 200$, $\rho = .5$ | 20 | 12.7 | 1.5 |
| optimal $b = 8$ | 10 | 8.90 | .52 |
| $n = 200$, $\rho = .8$ | 20 | 19.3 | .91 |
| optimal $b = 13$ | 10 | 13.8 | 3.9 |
| $n = 1000$, $\rho = .5$ | 50 | 30.7 | 4.2 |
| optimal $b = 16$ | 25 | 20.3 | 2.5 |
| $n = 1000$, $\rho = .8$ | 50 | 34.8 | 2.7 |
| optimal $b = 28$ | 25 | 30.3 | 1.4 |

The results are promising but mixed. In all cases the algorithm brings the user closer to the correct $b$ than was the pilot value. Also, the algorithm seems to work better when the pilot value, $b^*$ is close to the correct value. This is reasonable, as in Step 2) of the algorithm, the estimated MSE is closest to the true when $\hat{V}_N$ is closest to $V$.

One disquieting feature is that the procedure seems to work no better for $n = 1000$ than for for $n = 200$. At least partially, this is due to the two different search sets for $\hat{b}_m$. In the $n = 200$ case the upper bound in the search set for $\hat{b}_m$ is closer to the true than for $n = 1000$. Another issue is how effectiveness depends on the choice of $m$. Hopefully, further simulations with refinements to the algorithm can improve this method for estimating batch size.

## ACKNOWLEDGMENT

The author thanks Bruce Schmeiser for organizing this session, and for inviting the author to present this paper and mix with the simulation community.

## REFERENCES

Bratley, P., Fox, B.L., and Schrage, L.E. 1987. *A guide to simulation*. Springer Verlag, New York.

Carlstein, E. 1986a. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Annals of Statistics* 14:1171–1179.

Carlstein, E. 1986b. Asymptotic normality for a general statistic from a stationary sequence. *Annals of Probability* 14:1371–1379.

Cox, D.R. and Lewis, P. 1961. *The statistical analysis of series of events.* London: Methuen.

Damerji, H. 1994. Mean-square consistency of the variance estimator in steady-state simulation output analysis. *Operations Research* 43:282–291.

Fishman, G.S. 1978. *Principles of Discrete Event Simulation.* John Wiley & Sons, New York.

Glynn, P.W. and Inglehart, D.L. 1990. Simulation output analysis using standardized time series. *Mathematics of Operations Research* 15:1–16.

Goldsman, D. and Melamed, B. 1992. A large-sample result for the method of batch means. *Technical Report.* School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia.

Goldsman, D. and Schruben, L. 1984. Asymptotic properties of some confidence interval estimators for simulation output. *Management Science* 30:1217–25.

Hall, P. and Jing, B.Y. 1994. On sample re-use methods for dependent data. *Research Report #SR8-94.* Centre for Math. and its Appl., Australian National Univ., Canberra.

Künsch, H. 1989. The jackknife and bootstrap for general stationary observations. *Annals of Statistics* 17:1217–1241.

Meketon, M.S. and Schmeiser, B. 1984. Overlapping batch means: something for nothing? In *Proc. of the 1984 Winter Simulation Conference*, 227–230.

Pedrosa, A.C. and Schmeiser, B.W. 1993. Asymptotic and finite-sample correlations between OBM estimators. In *Proc. of the 1993 Winter Simulation Conference*, 481–488.

Politis, D.N. and Romano, J.P. 1993. On the sample variance of linear statistics derived from mixing sequences. *Stochastic Processes and Appl.* 45:155–167.

Possolo, A. 1991. Subsampling a random field. In *Spatial Statistics and Imaging*, Ed. by A. Possolo, IMS Lecture Notes. 20: 286–294.

Sargent, R.G., Kang, K., and Goldsman, D. 1992. An investigation of finite-sample behavior of confidence interval estimators. *Operations Research* 40:898–913.

Schmeiser, B. 1982. Batch size effects in the analysis of simulation output. *Operations Research.* 30:556–568.

Schmeiser, B., Avramidis, T., and Hashem, S. 1990. Overlapping batch statistics. In *Proc. of the 1990 Winter Simulation Conference*, 395–398.

Schruben, L. 1983. Confidence interval estimation using standardized time series. *Operations Research* 31:1090–1108.

Sherman, M. 1994. Asymptotic normality for a general statistic computed from a random field. *Mathematical Methods of Statistics* 3:326–345.

Sherman, M. 1995. Variance estimation for statistics computed from spatial lattice data. To appear in *Journal of the Royal Stat. Soc. B.*

Song, W.M.T. and Schmeiser, B. 1988. Estimating standard errors: empirical behavior of asymptotic MSE-optimal batch sizes. In *Comp. Science and Stat.: Proc. of the 20th Symposimum on the Interface*, 575–580.

## AUTHOR BIOGRAPHY

**MICHAEL SHERMAN** is an Assistant Professor in the Department of Statistics at Texas A&M University. He received his Ph.D. in Statistics from the University of North Carolina at Chapel Hill. His research interests include resampling methods for dependent data, spatial statistics, and survival analysis. Dr. Sherman is a member of the American Statistical Association and the Institute of Mathematical Statistics.