Proceedings of the 2001 Winter Simulation Conference
B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.

ON THE MSE ROBUSTNESS OF BATCHING ESTIMATORS

Yingchieh Yeh
Bruce W. Schmeiser

School of Industrial Engineering
Purdue University
West Lafayette, IN 47907–1287, U.S.A.

ABSTRACT

Variance is a classical measure of a point estimator's sampling error. In steady-state simulation experiments, many estimators of this variance—or its square root, the standard error—depend upon batching the output data. In practice, the optimal batch size is unknown because it depends upon unknown statistical properties of the simulation output data. When optimal batch size is estimated, the batch size used is random. Therefore, robustness to estimated batch size is a desirable property for a standard-error estimation method. We consider only point estimators that are a sample mean of steady-state data and consider only mean squared error (mse) as the criterion for comparing standard-error estimation methods. Like previous authors, we measure robustness as a second derivative. We argue that a previous measure—the second derivative of mse with respect to estimated batch size—is conceptually flawed. We propose a new measure, the second derivative of the mse with respect to the estimated center of gravity of the non-negative autocorrelations of the output process. With the previous robustness measure, optimal mse and robustness yielded different rankings of estimation methods. A property of the new robustness measure is that both criteria yield identical rankings.

1 BACKGROUND

We consider simulation experiments that produce steady-state output data $Y_1, Y_2, ..., Y_n$ and estimate the process mean $\mathrm{E}(Y)$ with the sample mean

$$\bar{Y} = \sum_{i=1}^{n} Y_i / n,$$

which has variance

$$\mathrm{Var}(\bar{Y}) = \frac{\mathrm{Var}(Y_i)}{n}[1 + 2\sum_{h=1}^{n}(1 - \frac{h}{n})\rho_h],$$

where $\mathrm{Var}(Y_i)$ is the common variance of individual observations and $\rho_h = \mathrm{corr}(Y_i, Y_{i+h})$ is the autocorrelation at lag h, for $h = 1, 2, ...$

As is typical in analyzing steady-state simulation experiments, we assume that variance $\mathrm{Var}(Y_i)$ and autocorrelations ρ_h are unknown. Therefore, using only the output data $Y_1, Y_2, ..., Y_n$, the variance $\mathrm{Var}(\bar{Y})$ must be estimated. We denote the generic estimator by $\widehat{V}(\bar{Y})$. (The square root of $\widehat{V}(\bar{Y})$ is then an estimator of the standard error.)

Several approaches to estimating $\mathrm{Var}(\bar{Y})$ are based on grouping the output data into batches of size m. We consider here non-overlapping batch means (NBM) estimator (Conway et al. 1959), the standardized-time-series area (STS-area) estimator (Schruben 1983), overlapping batch means (OBM) estimator (Meketon and Schmeiser 1984), and partially overlapping batches (50% OBM, 67% OBM, 75% OBM, and 80% OBM) estimators (Welch 1987). We also consider two linear combinations of NBM and STS-area: Schruben's original, with weights that minimize variance, and our variation, with weights that minimize mse. These estimators can be computed in $O(n)$ time. Song and Schmeiser (1993) discuss these and other batching estimators, including studying their properties via graphical analysis of their quadratic forms.

A fundamental issue is how to choose the batch size, m, as a function of the (known) run length n, the (known) type of estimator, and the (unknown) variance and autocorrelations of the steady-state output process.

Classically, a good batch size has been one that produced good confidence intervals. "Good" in this sense refers to the coverage probabilities and the distribution of interval length, often condensed to the probability of covering the mean and the expected half length. Fishman (1978) doubled NBM batch sizes until the batch means passed a test of independence, an approach that implicitly focuses on the probability of covering the mean. Schmeiser (1982) compared asymptotic coverage probabilities and mean and variance of the half length as a function of number of non-overlapping batches. Schruben (1983) and Goldsman and

Schruben (1984) assumed that different methods would use the same batch sizes in arguing that STS estimators provided more degrees of freedom (that is, smaller variance) than NBM. More recently, Chen and Kelton (2000), Fishman (1996, Sections 6.10–6.11) and Steiger and Wilson (2000) develop procedures for determining an appropriate number of NBM batches and appropriate batch size to obtain good confidence-interval performance.

2 MSE-OPTIMAL BATCH SIZES

An alternative to pursuing good confidence-interval performance is to define a good batch size to be one that produces a small mse for the estimator of the variance of the sample mean, $\widehat{V}(\bar{Y})$. "Good" in this sense means having a small bias and a small variance, as balanced in the well-known result

$$\mathrm{mse}[\widehat{V}(\bar{Y}), \mathrm{Var}(\bar{Y})] = \mathrm{bias}^2[\widehat{V}(\bar{Y}), \mathrm{Var}(\bar{Y})] + \mathrm{Var}[\widehat{V}(\bar{Y})].$$

Goldsman and Meketon (1986), who first used mse to compare estimators of $\mathrm{Var}(\bar{Y})$, defined a bias constant c_b and a variance constant c_v for each type of estimator. Song (1988) showed that their bias and variance constants led to the asymptotic equation

$$\mathrm{mse}[\widehat{V}(\bar{Y}), \mathrm{Var}(\bar{Y})] \approx \mathrm{Var}^2(Y_i)(\frac{c_b^2\,\gamma_1^2}{n^2\,m^2} + \frac{m\,c_v\,\gamma_0^2}{n^3}),$$

where $\gamma_0 = 1 + 2\sum_{h=1}^{\infty} \rho_h$ and $\gamma_1 = 2\sum_{h=1}^{\infty} h\rho_h$.

For sample size n, estimator-type constants (c_b, c_v), and output-process constants (γ_1, γ_0), the asymptotic mse is minimized by

$$m^* \approx [2n\,(\frac{c_b^2}{c_v})\,(\frac{\gamma_1}{\gamma_0})^2]^{1/3} + 1,$$

the asymptotic mse-optimal batch size.

The mse-optimal batch size, m^*, depends upon the type of estimator only through the ratio c_b^2/c_v; comparisons among types of estimators should reflect the different ratios.

The mse-optimal batch size, m^*, depends upon the output process only through the ratio $\tau \equiv \gamma_1/\gamma_0$; comparisons based upon various output processes should reflect a range of ratios. We refer to the ratio τ as the output process's *center of gravity* because

$$\frac{\gamma_1}{\gamma_0} = \frac{2\sum_{h=1}^{\infty} h\rho_h}{1 + 2\sum_{h=1}^{\infty} \rho_h} = \frac{\sum_{h=-\infty}^{\infty} |h|\rho_h}{\sum_{h=-\infty}^{\infty} \rho_h}$$

is the process lag at which the torque of the positive-lag autocorrelations is zero. Different autocorrelograms can yield the same value of τ and therefore the same optimal batch size.

Table 1: Properties of various batching estimators

Estimator	Property			
	bias c_b	variance c_v	m^* $c_b^{2/3} c_v^{-1/3}$	mse(m^*) $c_b^{2/3} c_v^{2/3}$
OBM	1	4/3	0.91	1.21
80% OBM	1	34/25	0.90	1.23
75% OBM	1	11/8	0.90	1.24
67% OBM	1	38/27	0.89	1.26
50% OBM	1	3/2	0.87	1.31
LC-mse	5/3	10/9	1.36	1.51
NBM	1	2	0.79	1.59
LC-var	2	1	1.59	1.59
STS-area	3	2	1.65	3.30

Table 1 is similar to tables in Goldsman and Meketon (1986), Song (1988), and Song and Schmeiser (1995) in comparing estimator types based on mse-optimal performance. For each estimator type, the table columns contain, respectively, the bias constant, the variance constant, the mse-optimal batch size, and the minimal mse. The latter two are scaled, ignoring the effects of the sample size n, the center of gravity τ and the output variance $\mathrm{Var}(Y_i)$.

As is known, of these estimator types the minimal asymptotic mse is obtained with OBM, with complete overlapping being best. The equal-weight linear combination of NBM and STS-area, denoted by LC-var because it minimizes asymptotic variance, reverses the bias and variance constants of NBM and yields the identical optimal mse.

One estimator in the table, shown just before NBM, is new. This estimator, denoted by LC-mse, is a second linear combination of NBM and STS-area, but now using mse-optimal weights rather than Schruben's variance-optimal weights. In particular, the new estimator weights NBM by $2/3$ and STS-area by $1/3$; the mse improvement is about 5%.

3 ROBUSTNESS TO ESTIMATION ERROR

The asymptotic mse-optimal batch size, m^*, would be known approximately if the center of gravity τ were known. Because it is unknown, estimators that depend upon mse-optimal batch size need to estimate τ. Song (1996) estimates τ directly via estimates of many autocorrelations ρ_h. Pedrosa (1994) estimates τ within his 1-2-1 OBM estimator of $\mathrm{Var}(\bar{Y})$. However obtained, the estimator $\widehat{\tau}$ is substituted into the mse-optimal batch-size formula to obtain an estimator $\widehat{m}^*$ of m^*.

In comparing types of estimators, Song (1988) and Song and Schmeiser (1995) recognized that robustness to batch-size error, $\widehat{m}^* - m^*$, is an important property. They measured robustness by the second derivative

$$\frac{\partial^2 \text{mse}[\widehat{V}(\bar{Y}), \text{Var}(\bar{Y})]}{\partial \widehat{m}^{*2}}|_{\widehat{m}^*=m^*} \approx \frac{3\,\text{Var}^2(Y_i)\,(\gamma_0^8\, c_v^4)^{1/3}}{(2\,\gamma_1^2\, n^{10}\, c_b^2)^{1/3}},$$

which for constant values of sample size n, center of gravity τ and observation variance $\text{Var}(Y_i)$ is proportional to $c_b^{-2/3} c_v^{4/3}$.

We argue that the robustness measure is flawed. The idea of using the second derivative is fine, being both intuitive and traditional. Evaluating the second derivative at m^* is also fine, allowing each type of estimator to use its own mse-optimal batch size. The flaw arises from the implicit assumption that the batch-size error is the same for each type of estimator; in reality, batch-size error $\widehat{m}^* - m^*$ is proportional to c_b^2/c_v.

Rather than focusing on batch size, which differs depending upon the type of estimator, the robustness measure should focus on the common output process, whose effect on the batching estimators considered here is measured via its center of gravity, τ. The effect of τ then easily is traced to its effect on $\widehat{m}^*$, whose effect easily is traced to $\text{mse}[\widehat{V}(\bar{Y}), \text{Var}(\bar{Y})]$. Therefore, we suggest that a more-appropriate robustness measure is

$$\frac{\partial^2 \text{mse}[\widehat{V}(\bar{Y}), \text{Var}(\bar{Y}))]}{\partial \widehat{\tau}^2}|_{\widehat{\tau}=\tau} \approx \frac{4\,\text{Var}^2(Y_i)\,(2\,\gamma_0^{10}\, c_b^2\, c_v^2)^{1/3}}{3\,(n^8\,\gamma_1^4)^{1/3}}.$$

The derivation is straightforward. The estimated mse-optimal batch size is

$$\widehat{m}^* \approx [2\,n\,(\frac{c_b^2}{c_v})\,\widehat{\tau}^2]^{1/3} + 1,$$

where $\widehat{\tau}$ is the estimated center of gravity. For tractability, ignore the additive constant 1, which is important only when autocorrelations are negligible. Substitute $\widehat{m}^*$ for m into the equation for $\text{mse}[\widehat{V}(\bar{Y}), \text{Var}(\bar{Y})]$ so that the mse is a function of $\widehat{\tau}$. Take the second derivative of the resulting mse equation with respect to $\widehat{\tau}$. Evaluate the second derivative at $\widehat{\tau} = \gamma_1/\gamma_0$ to obtain the new robustness measure.

For fixed run length n and output-process constants γ_0, γ_1 and $\text{Var}(Y_i)$, the new robustness measure is proportional to $c_b^{2/3} c_v^{2/3}$. Also, the new measure is proportional to the optimal mse, as shown in Table 1.

Table 2, whose structure is similar to Table 1, shows the new and previous robustness measures for each type of estimator. For all batch-means estimators, the new robustness measure is smaller than the previous measure. For the STS-area estimator, the new robustness measure is more than two and and half times larger than the previous measure. The new measure is also larger for both linear-combination estimators.

Table 2: Robustness of some batching estimators

Estimator	Robustness	
	previous measure $c_b^{-2/3}\,c_v^{4/3}$	new measure $c_b^{2/3}\,c_v^{2/3}$
OBM	1.47	1.21
80% OBM	1.51	1.23
75% OBM	1.53	1.24
67% OBM	1.58	1.26
50% OBM	1.72	1.31
LC-mse	0.82	1.51
NBM	2.52	1.59
LC-var	0.63	1.59
STS-area	1.21	3.30

The ratio of the new measure to the previous measure is $c_b^{4/3} c_v^{-2/3}$. Therefore the decrease for the batch-means estimators, which have $c_b = 1$, is due only to the variance constant. The increase in STS-area is due to its large bias constant, $c_b = 3$.

The new robustness measure favors OBM in that OBM is now optimal in terms of both mse and mse robustness. As was previously known, partial overlapping provides a good alternative to OBM, with fifty-percent overlapping having only an 8% mse penalty. The mse penalty for using NBM rather than OBM is 31%.

The results of Table 2 indicate that one should use as much overlapping as is possible. Complete overlapping is not always possible, such as with dynamic batch means (Yeh 1999). When no overlapping is possible, the new estimator LC-mse is asymptotically mse-optimal.

4 CONCLUSIONS

We consider estimating the variance of the sample mean of observations from a steady-state simulation experiment, with emphasis on comparing types of estimators based on mse. Five conclusions arise.

First, we conclude that robustness to batch-size error should not be measured as the second derivative with respect to estimated batch size.

Second, we conclude that robustness to batch-size error should be measured as the second derivative with respect to the estimated center of gravity of the output process.

Third, we propose an mse-optimal linear combination of NBM and STS-area estimators, analogous to Schruben's variance-optimal linear combination.

Fourth, we show that, for each type of batching estimator, the new robustness measure is proportional to the estimator's mse at the mse-optimal batch size. Therefore, ranking types of estimators based on optimal mse is equivalent to ranking based on the new robustness measure.

Fifth, the single ranking obtained with the new robustness measures favors OBM, then partially overlapped batches, then NBM, and then STS-area. The mse-optimal linear combination, which by definition ranks higher than the variance-optimal linear combination, is a bit better than NBM alone.

Finally, we comment that the issues of mse-optimal estimation and batch-size robustness apply to all steady-state point estimators whose variances are estimated using batch statistics (Schmeiser, Avramidis and Hashem 1990, and Wood 1995). Few results, however, exist beyond estimation of the output-process mean.

REFERENCES

Chen, E. J. and W. D. Kelton 2000. A stopping procedure based on phi-mixing conditions. *Proceedings of the Winter Simulation Conference*, ed. J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick. 617–626. Piscataway, New Jersey: IEEE.

Conway, R. W., B. M. Johnson, and W. L. Maxwell 1959. Some problems of digital systems simulation. *Management Science* 6, 92–110.

Fishman, G. S. 1978. Grouping observations in digital simulation. *Management Science* 24, 510–521.

Fishman, G. S. 1996. *Monte Carlo: Concepts, Algorithms, and Applications.* New York: Spring-Verlag.

Goldsman, D. M. and M. S. Meketon. 1986. A comparison of several variance estimators. Technical Report J-85-12, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

Goldsman, D. M. and L. W. Schruben. 1984. Asymptotic properties of some confidence interval estimators for simulation output. *Management Science* 30, 1217–1225.

Meketon, M. S. and B. W. Schmeiser. 1984. Overlapping batch means: Something for nothing? *Proceedings of the Winter Simulation Conference*, ed. S. Sheppard, U. Pooch, and C. D. Pegden, 227–230. Piscataway, New Jersey: IEEE.

Pedrosa, A. 1994. *Automatic Batching in Simulation Output Analysis.* Ph.D. Dissertation, Purdue University, West Lafayette, Indiana.

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

Schmeiser, B. W., T. Avramidis and S. Hashem. 1990. Overlapping batch statistics. *Proceedings of the Winter Simulation Conference*, ed. O. Balci, R. P. Sadowski, and R. E. Nance, 395–398. Piscataway, New Jersey: IEEE.

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

Song, W. T. 1988. *Estimators of the Variance of the Sample Mean: Quadratic Forms, Optimal Batch Sizes, and Linear Combinations.* Ph.D. Dissertation, Purdue University, West Lafayette, Indiana.

Song, W. T. 1996. On the estimation of optimal batch sizes in the analysis of simulation output. *European Journal of Operational Research* 88, 304–319.

Song, W. T. and B. W. Schmeiser. 1993. Variance of the sample mean: Properties and graphs of quadratic-form estimators. *Operations Research* 41, 501–517.

Song, W. T. and B. W. Schmeiser. 1995. Optimal mean-squared-error batch sizes. *Management Science* 41, 110–123.

Steiger, N. M. and J. R. Wilson. 2000. Experimental performance evaluation of batch means procedures for simulation output analysis. *Proceedings of the Winter Simulation Conference*, ed. J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick. 627–636. Piscataway, New Jersey: IEEE.

Welch, P. D. 1987. On the relationship between batch means, overlapping batch means, and spectral estimation. *Proceedings of the Winter Simulation Conference*, ed. A. Thesen, H. Grant and W. D. Kelton, 320–323. Piscataway, New Jersey: IEEE.

Wood, D. C. 1995. *Variances and Quantiles in Dynamic-System Performance: Point Estimation and Standard Errors.* Ph.D. Dissertation, Purdue University, West Lafayette, Indiana.

Yeh, Y. 1999. *Steady-State Simulation Output Analysis via Dynamic Batch Means.* M.S. Thesis, Purdue University, West Lafayette, Indiana.

AUTHOR BIOGRAPHIES

YINGCHIEH YEH is a Ph.D. student in the School of Industrial Engineering at Purdue University. In 1999, he received an M.S. degree in Industrial Engineering from Purdue University. His primary research interests are the probabilistic and statistical aspects of stochastic simulation, especially simulation output analysis.

BRUCE W. SCHMEISER is a professor in the School of Industrial Engineering at Purdue University. His interests lie in applied operations research, with emphasis in stochastic models, especially the probabilistic and statistical aspects of stochastic simulation. He is an active participant in the Winter Simulation Conference, including being Program Chair in 1983 and chairing the Board of Directors during 1988–1990.