ANALYSIS OF SEQUENTIAL STOPPING RULES

Dashi I. Singham
Lee W. Schruben

Department of Industrial Engineering & Operations Research
University of California, Berkeley
Berkeley, CA 94705, USA

ABSTRACT

Sequential stopping rules applied to confidence interval procedures (CIPs) may lead to coverage that is less than nominal. This paper introduces a method for estimating coverage functions analytically in order to evaluate the potential loss of coverage. This method also provides an estimate for the distribution of the stopping time of the procedure. Knowledge of coverage functions could help evaluate and compare confidence interval procedures while avoiding lengthy empirical testing. Numerical implementation of our method shows that analytical coverage functions approximate those calculated empirically. Analytical coverage functions can be used to explain why many sequential procedures do not provide adequate coverage.

1 INTRODUCTION

Sequential stopping rules are often used to determine the run length of a steady state simulation. These rules, as part of confidence interval procedures, deliver confidence intervals for some parameter of performance having the desired half-width specified by the user. However, the run length required by a stopping rule is unknown, and will most likely vary with multiple replications of the procedure. This variability can make it hard to predict the computational resources required to implement the experiment.

Another disadvantage of sequential stopping rules is that actual coverage of the confidence intervals delivered may not be as high as what was intended. Sequential stopping rules could lead to inappropriately early termination resulting in poor coverage. Procedures have been developed to avoid the problems associated with early stopping and achieve improved coverage (Heidelberger and Welch 1983, Steiger and Wilson 2002). As the number of simulations required by the procedure approaches infinity, coverage approaches nominal (desired) coverage (Chow and Robbins 1965, Glynn and Whitt 1992). By decreasing the maximum allowable half-width of a sequential stopping rule, the number of simulations required can be pushed towards infinity. The coverage function of a procedure applied to data with a known mean parameter value can provide an indication of how well the procedure will cover the true mean when the underlying distribution of the data is different from that assumed (Schruben 1980). Many suggestions have been made regarding the choice of stopping rule that would provide close to nominal coverage for various types of simulations (Law and Kelton 1982). However, the stopping rule choice is often made on a case by case basis.

We develop a method for analytically estimating coverage functions of confidence interval procedures employing sequential stopping rules. This procedure provides information that will help reduce some of the uncertainty pertaining to run length and coverage. In calculating the coverage functions, we estimate the distribution of the stopping time associated with the stopping rule. This provides information on the potential run length of the procedure. The coverage function calculation provides an estimate of the loss of coverage induced by the stopping rule. Schmeiser and Yeh (2002) suggest a single dimensionless criteria for comparing CIPs using coverage functions. Our method can help simulation analysts compare and choose appropriate confidence interval procedures by providing an analytical way to estimate coverage functions.

Sequential stopping rules used in confidence interval procedures can result in a loss of coverage. The extent of the loss depends on the level of output precision required by the procedure. This precision requirement choice is often based on the specific purpose of the simulation model built by the analyst. As a result, the stopping rule may be chosen for the characteristics of a particular simulation model instead of the quality of the rule. The confidence interval procedure could

978-1-4244-5771-7/09/$26.00 ©2009 IEEE 723

deliver output that appears appropriate, but the analyst may not know how much coverage is lost due the choice of stopping rule. We derive a method that quantifies the loss in coverage for a general class of stopping rules and provides some insight into why nominal coverage is not achieved.

Section 2 describes the background relevant to sequential stopping rules. Section 3 describes our procedure for computing coverage functions. Sections 4 and 5 provide numerical examples and discuss the implications.

2 BACKGROUND

We consider simulations that are used to estimate a confidence interval for some scalar mean level, μ, of system performance. The analyst may require that the length of the confidence interval be less than or equal to some pre-specified precision (width) target. Sequential stopping rules are often used to deliver confidence intervals of a fixed length by requiring additional simulation output until an interval small enough to meet the analyst's needs can be delivered from the data. Generally, these types of stopping rules reduce the coverage frequency of the resulting confidence intervals.

We study sequential stopping rules that are designed for simulation output that is assumed to be i.i.d. normal with known variance σ^2 and unknown mean μ. Working with data of this form allows us to estimate the coverage functions analytically. In addition to adding analytical tractability, this assumption has implications for general coverage properties of sequential stopping rules. The concepts learned from our method can be applied, at least empirically, to study other sequential stopping experiments. Assuming independence and normality for our method indicates how well the procedure would perform for data with ideal properties. If coverage is not nominal for data that is i.i.d normal, we probably would have even less information about the coverage for data that is dependent or non-normal.

We focus on the sample mean and variance, which are the sufficient statistics for the normal distribution. These two statistics can be used to construct a confidence interval for μ by centering the interval around $\bar{X}_k$, the sample mean of k observations. Our confidence interval takes the form $[\bar{X}_k - HW_{\eta,k}, \bar{X}_k + HW_{\eta,k}]$ where $HW_{\eta,k}$ is the half-width of the interval with confidence coefficient η based on k samples. The half-width for independent and normally distributed data is defined as

$$HW_{\eta,k} = t_{(1+\eta)/2,k-1}\sqrt{\frac{S_k^2}{k}}. \tag{1}$$

Let S_k^2 be the sample variance of k data points and $t_{(1+\eta)/2,k-1}$ be the $(1+\eta)/2$ quantile of the t-distribution with $k-1$ degrees of freedom. Since the quantity $(1+\eta)/2$ is fixed for a given procedure, we simplify the notation to t_{k-1}. Sequential stopping rule procedures call for two parameters to be specified by the user: the desired confidence coefficient of the output interval, η, and the desired half-width of the interval, δ. Sequential stopping rules specify η and δ at the start of the run, but the number of simulations needed to meet the rule is random and unknown. An absolute precision sequential stopping rule applied to output data involves simulating until $HW_{\eta,k}$ is less than δ. (Relative precision stopping rules are more complicated but would fit into our analysis too.) Define the random variable k^* as the number of simulation observations required in the sequential procedure for a given realization of random output data:

$$k^* = \arg\min_k : HW_{\eta,k} \leq \delta.$$

We will use the distribution of k^* and its relationship to coverage functions to evaluate the performance of a confidence interval procedure. Sequential stopping rules will lead to coverage that is less than desired. Given that a stopping rule may not provide nominal coverage, we would like to find a way to estimate how much coverage is lost, and perhaps compensate for it. We propose an analytical method for estimating some coverage functions of CIPs that employ sequential stopping rules.

In order to estimate the coverage of these CIPs, we calculate the probability that the resulting confidence interval covers the true parameter value μ given that it also satisfies a stopping rule. This probability of covering the mean is compared with the intended coverage of the procedure to form a coverage function. The empirical coverage function for a particular sequential CIP in a specific simulation or data collection context could also be estimated as in (Schruben 1980). As part of the coverage function calculation, we derive the probability distribution function of the stopping time of the sequential stopping rule. This information can be used to compute the expected value of the number of simulation observations the procedure would require. The knowledge of the coverage functions and run times associated with CIPs might help in designing a sequential procedure that meets the coverage needs of the user in a reasonable amount of time.

3   PROCEDURE

The procedure described in this paper estimates the coverage functions of stopping rules with parameters η and δ when the underlying data is i.i.d. normal with known variance σ^2. The probability of coverage for a procedure can be defined by conditioning on the stopping time:

$$P(Cover) = \sum_k P(Cover \,|\, Stop \,at\, k)P(Stop \,at\, k). \tag{2}$$

In a sequential stopping procedure, the value of the half-width is set equal to δ once the procedure terminates. If we assume that the distribution of the sample average given stopping is independent of the stopping time, equation (2) can be written as

$$P(Cover) = \sum_{k=2}^{\infty} P(\mu - \delta \leq \bar{X}_k \leq \mu + \delta)P(k^* = k). \tag{3}$$

At least two replications are needed to calculate a confidence interval, though the minimum value of k can be set to a greater value to avoid stopping early. The next steps outline a recursive method for calculating $P(k^* = k)$. For a fixed value of k, if stopping occurs, the sample variance S_k^2 must be less than $\delta^2 k / t_{k-1}^2$ (this can be derived by setting (1) to be less than or equal to δ and solving for solving for $\sqrt{S_k^2}$). For a normally distributed sample, the distribution of the sample variance is related to the distribution of a chi-squared random variable. We re-write the distribution of $HW_{\eta,k}$ as

$$
\begin{aligned}
P(HW_{\eta,k} \leq \delta) &= P\left(S_k^2 \leq \frac{\delta^2 k}{t_{k-1}^2}\right) \\
&= P\left(\chi_{k-1}^2 \leq \frac{\delta^2 k(k-1)}{t_{k-1}^2 \sigma^2}\right).
\end{aligned}
\tag{4}
$$

Equation (4) can be used to calculate the probability of stopping at k=2. For greater values of k, the probability of stopping must be calculated conditional on not stopping at lower values of k. The probability of stopping at three samples is:

$$P(k^* = 3) = P(k^* > 2)P(k^* = 3 | k^* > 2).$$

For greater values of k, the recursive calculation for the probability of stopping is written as

$$P(k^* = k) = P(k^* > 2) \left[\prod_{i=3}^{k-1} P(k^* > i | k^* > i-1)\right] P(k^* = k | k^* > k-1). \tag{5}$$

The next step is to calculate the conditional probabilities $P(k^* = k | k^* > k-1)$ and $P(k^* > k | k^* > k-1)$ for $k \geq 3$. The probability of stopping at k given that the rule is not met by time $k-1$ is

$$P(k^* = k | k^* > k-1) = 1 - P(k^* > k | k^* > k-1).$$

So it is sufficient to calculate the conditional probabilities $P(k^* = k | k^* > k-1)$ for $k \geq 3$. These are calculated by integrating over the possible ranges of $\bar{X}_{k-1}$ and S_{k-1}^2, and determining the values of the next observation (X_k) that result in stopping. The random variable $\bar{X}_{k-1}$ can take values from $[-\infty, \infty]$, but the values of S_{k-1}^2 that need to be considered are limited.

To see this, consider the fact that the sample variance for the first $k-1$ observations must be large enough to avoid meeting the stopping rule at $k-1$, but small enough that the next observation X_k could lower the variance enough to meet the stopping rule at time k. Consider the lower bound on the sample variance that ensures the stopping rule is not met at $k-1$ (call this V_{min}). For k^* to be greater than $k-1$, S_{k-1}^2 must be at least as big as V_{min}. Rearranging the terms in (1) and

725

requiring $HW_{\eta,k-1}$ to be greater than δ yields the value of V_{min},

$$V_{min} := \frac{\delta^2(k-1)}{t_{k-2}^2}, \tag{6}$$

which bounds from below the values of S_{k-1}^2 that allow stopping at k. In order to stop at time k, the variance S_k^2 must be low enough to meet the stopping rule at time k, meaning $S_k^2 \leq \delta^2 k / t_{k-1}^2$. To relate the variances S_{k-1}^2 and S_k^2, the following recursion is helpful:

$$S_k^2 = \frac{k-2}{k-1}S_{k-1}^2 + \frac{(X_k - \bar{X}_{k-1})^2}{k}.$$

Setting S_k^2 less than $\delta^2 k / t_{k-1}^2$ and solving for $|X_k - \bar{X}_{k-1}|$ yields a value, defined as X_k^b in (7), that is the maximum absolute difference between X_k and $\bar{X}_{k-1}$ that would allow stopping. If the next observation is close to the previous observations (which are characterized by their sample mean), the total variance will be lower than if the next observation is far away from the current mean. The variable X_k^b, defined as

$$X_k^b := \sqrt{\frac{\delta^2 k^2}{t_{k-1}^2} - \frac{(k-2)kS_{k-1}^2}{k-1}}, \tag{7}$$

is the maximum value of $|X_k - \bar{X}_{k-1}|$ that would allow stopping to occur at k. If S_{k-1}^2 is too large, the term under the radical becomes negative. The maximum value of S_{k-1}^2 that allows X_k^b to be real-valued is the maximum value of S_{k-1}^2 that allows for the possibility of stopping. Values of the sample variance higher than this amount (call this V_{max}) do not need to be considered. Equation (8) defines V_{max},

$$V_{max} := \frac{\delta^2 k(k-1)}{t_{k-1}^2(k-2)}, \tag{8}$$

which bounds from above the values of S_{k-1}^2 that allow stopping, according to what is allowed by (7). The variance at step $k-1$ needs to be close enough to the stopping variance in order to allow for the possibility of the next sample bringing the procedure to a stop. Equations (6) and (8) provide bounds on the values of S_{k-1}^2 that could lead to stopping at observation k. Equation (7) provides the range of values of X_k that could lead to stopping, given $\bar{X}_{k-1}$ and S_{k-1}^2. X_k must be within X_k^b away from $\bar{X}_{k-1}$ so that the variance will fall low enough for stopping to occur at k.

The ranges $S_{k-1}^2 \in [V_{min}, V_{max}]$, and $X_k \in [\bar{X}_{k-1} - X_k^b, \bar{X}_{k-1} + X_k^b]$ can be used to construct an integral representation for the conditional probability of stopping. A probability weighted integral over the variables $\bar{X}_{k-1}$ and S_{k-1}^2 gives the probability of stopping at k given that stopping has not yet occurred at time $k-1$, as shown in (9). Let the notation $f_{\bar{X}_{k-1}}$ refer to the probability density function of the random variable $\bar{X}_{k-1}$, and $f_{S_{k-1}^2|k^*>k-1}$ be the density function for the sample variance conditional on the stopping rule not being met by time $k-1$. Then the conditional probability of stopping at k is

$$P(k^* = k | k^* > k-1) = \int_x \int_{y=V_{min}}^{V_{max}} P(\bar{X}_{k-1} - X_k^b \leq X_k \leq \bar{X}_{k-1} + X_k^b) f_{S_{k-1}^2|k^*>k-1}(y) f_{\bar{X}_{k-1}}(x) dy dx. \tag{9}$$

Note that the value of X_k^b depends on S_{k-1}^2. The conditional probabilities calculated in (9) can be used to compute the probability of stopping at a particular time k as given in (5). These probabilities in turn are used to calculate the coverage probabilities in (3). The distribution of S_{k-1}^2 given that stopping has not yet occurred is different from that of the unconditional distribution of S_{k-1}^2. It is possible for the output data to meet the stopping rule at some time less than $k-1$, but additional output could drive the half-width to be greater than δ at time $k-1$. We derive the density function for S_{k-1}^2 given $k^* > k-1$ in the Appendix.

4   NUMERICAL RESULTS

The representation described in Section 3 was implemented numerically to estimate coverage functions for various stopping rules. The program takes the values of η, δ, and σ^2 as inputs and delivers the distribution of the stopping time and the resulting coverage function for the case where the underlying data was normally distributed with variance σ^2. Coverage functions can also be calculated empirically by applying CIPs to data generated from a particular distribution. By repeatedly applying the CIP to these random number streams, the probability of coverage can be estimated by recording the proportion of intervals that cover the known mean of the distribution. In order to assess the accuracy of our analysis, we estimated empirical coverage functions and compared them to our analytical results. Without loss of generality, we set μ equal to zero for our computations.

Empirical coverage functions were calculated by applying sequential stopping rules to simulated normal random variables. The sequential stopping rules were constructed by choosing different combinations of values for η and δ. A given stopping rule was applied repeatedly to simulated random variables. The proportion of confidence intervals that covered the true mean at stopping was recorded as an estimate of the coverage for the particular stopping rule.
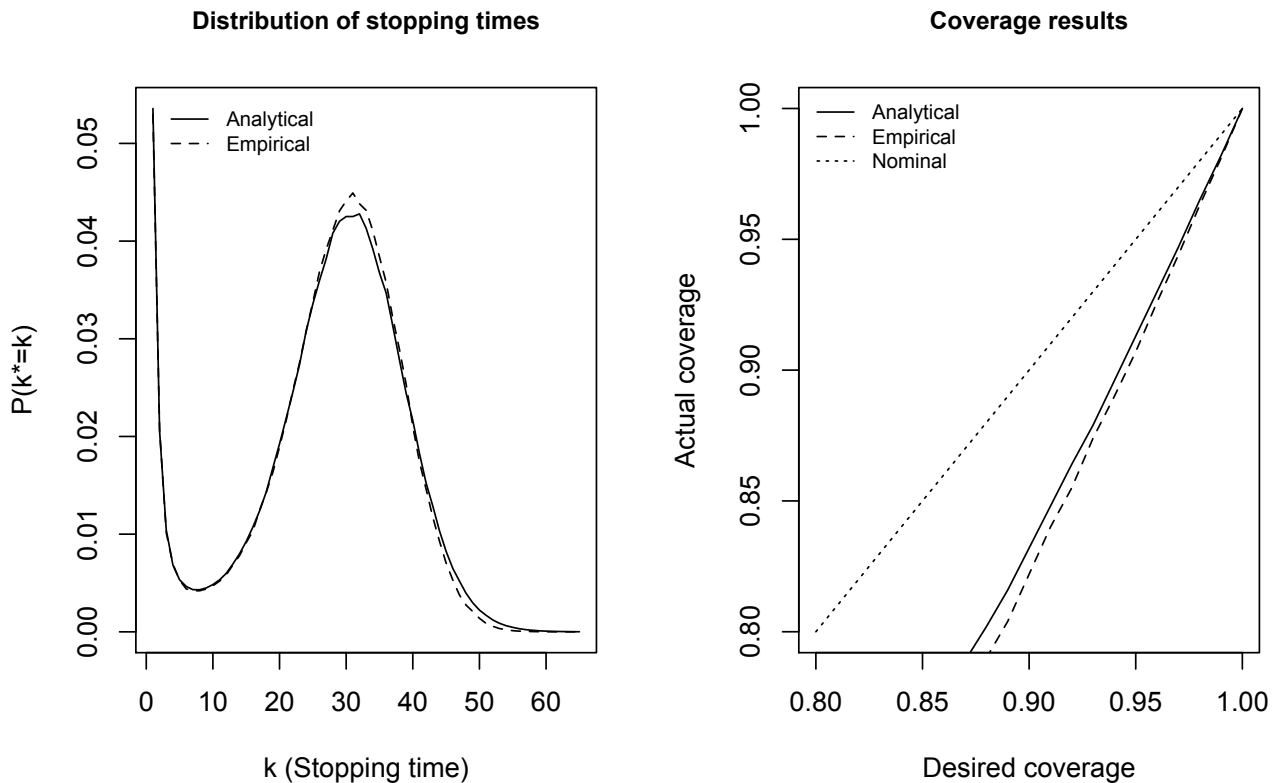

Figure 1: Distribution of stopping time of $\eta = 0.9$, and $\delta = 0.3$ (left). Coverage results for various values of η and $\delta = 0.3$ (right). Underlying data is assumed to be i.i.d $\mathscr{N}(0,1)$. Probability distribution information was computed using the GNU Scientific Library.

Figure 1 shows the results for the stopping rule with parameters $\eta = 0.9$ and $\delta = 0.3$ applied to standard normal data. The left plot shows the distribution of the stopping times for rules calculated both according to our framework and by simulation. The solid line corresponds to our analytical method while the dashed line is the empirical result. There appears to be a relatively high probability of stopping at very low values of k, resulting in a probability distribution function that is not unimodal. If the first few observations happen to be close together, the stopping rule could be met before a sample that represents the variance of the data can be collected. Early stopping contributes to a loss in coverage due to the inadequate sample size, and knowledge of the probability of stopping early gives an indication of how often this problem could occur.

The right plot shows the coverage results calculated according to both methods for $\delta = 0.3$ and various values of η (desired coverage). The dotted line plots nominal coverage. This figure shows the loss in coverage for stopping rules using various values of η and a value of $\delta = .3$ when the underlying data has a variance of one.

Figure 2 shows the results when $\delta = .15$. For the left plot, the distribution of the stopping time is calculated when $\eta = .9$. We can see that the expected stopping time is greater than when $\delta = .3$. The right plot of Figure 2 shows the coverage results when η is varied and $\delta = .15$. The coverage is greater here than when $\delta = .3$, and the analytical estimates closely match the empirical coverage values. The probability of stopping early is lower than in the case where $\delta = .3$, and helps explain why coverage is better when more strict stopping rules are employed.

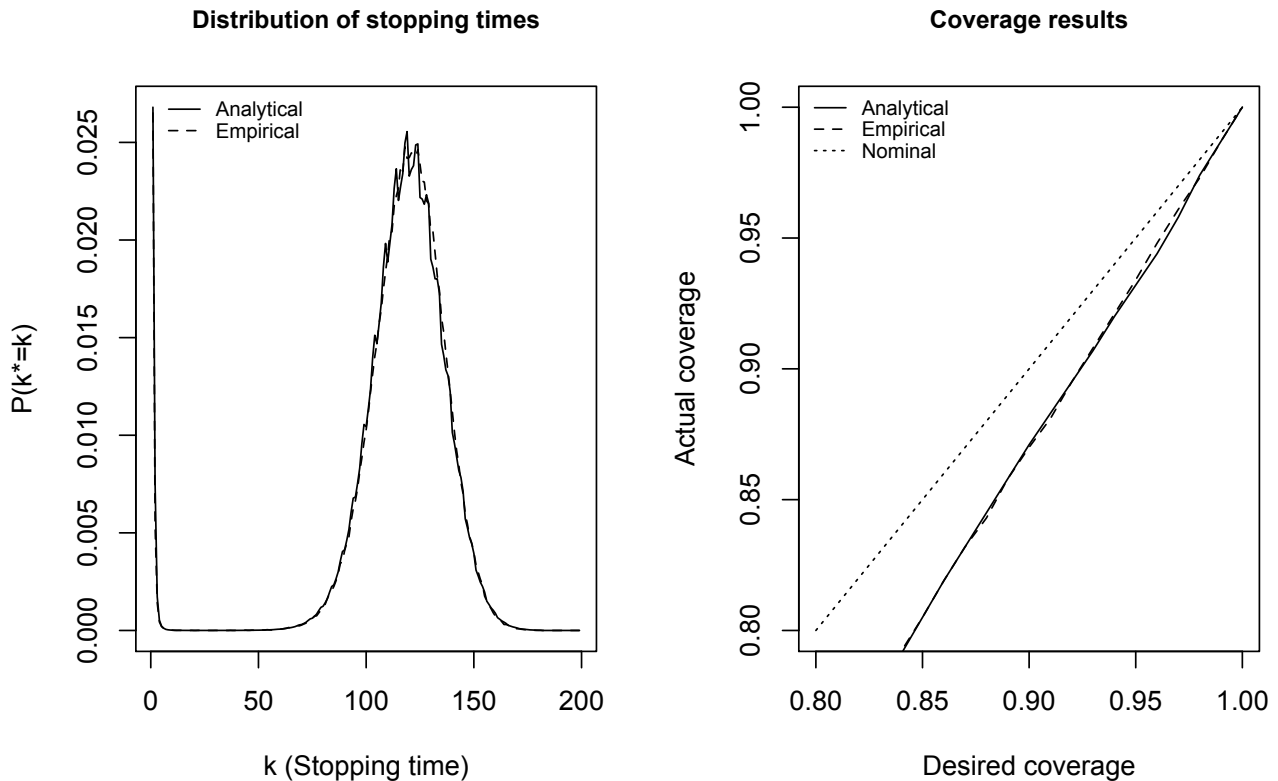**Distribution of stopping times**   **Coverage results**

Figure 2: Distribution of stopping time of $\eta = 0.9$, and $\delta = 0.15$ (left). Coverage results for various values of η and $\delta = 0.15$ (right). Underlying data is assumed to be i.i.d $\mathcal{N}(0,1)$.

These plots can be recalculated for different values of η, δ and σ^2 to calculate the loss in coverage for various stopping rules. The coverage functions calculated by our proposed method appear to be close to those generated empirically. One advantage of our analytical framework is that the computation time required is generally much less than the time it takes to calculate the coverage functions empirically by simulation. The empirical method requires a decision on the number of replications to run in order to produce a smooth coverage function. This number can be quite high depending on the stopping rule parameters. Our analytical method provides estimates of the coverage level and distribution of the stopping time with relatively less computational effort.

5 CONCLUSION

Analytical coverage functions can help determine the potential loss of coverage of a confidence interval procedure. Knowledge of the coverage functions of a procedure can signal potential problems with insufficient coverage and allow stopping rules to be compared. Additionally, the availability of the distribution of the stopping time could help determine if enough computational resources are available. All of these aspects of our method help remove some of the uncertainty associated with sequential procedures. The analyst might be willing to sacrifice coverage in order to obtain shorter run times. Our

method allows for a quick view of the relevant information allowing the analyst to make a decision on which stopping rule to use prior to generating any simulations.

We use i.i.d normal random variables as the underlying data for our procedure, since many CIPs attempt to convert simulation data to this form (say, through batching). Calculating coverage functions according to the assumptions of independence and normality show how CIPs can fail to provide adequate confidence intervals for parameters even when the underlying data has all the properties assumed by the method. In other words, these coverage functions show how a procedure can provide inadequate coverage even when applied to data with ideal properties.

The fact that our results closely match coverage functions calculated empirically has implications for explaining the bias associated with termination of a simulation. Although we only consider a special case for the simulation output distribution, the intuition we have gained could be used to explain the loss of coverage under other circumstances. We hope that our procedure will remove some of the guesswork from the stopping rule choice and provide a new way of evaluating confidence interval procedures. Future research will include determining the relationship between the distribution of the sample mean at stopping and the distribution of the stopping time. We also are working on determining the effect of sequential stopping rules on data that is dependent, non-normal, or contains an initialization bias.

APPENDIX

We wish to calculate the density function of S^2_{k-1} given $k^* > k-1$. This is the density function of the sample variance at time $k-1$ given that the stopping rule has not yet been met. Consider the distribution of the sample variance of a set of $k-1$ normally distributed random variables with variance σ^2. Let Z^2_i be a squared standard normal random variable. The following terms are equivalent in distribution:

$$S^2_{k-1} \stackrel{d}{=} \frac{\sigma^2 \chi^2_{k-2}}{k-2} \stackrel{d}{=} \frac{\sigma^2(Z^2_1 + Z^2_2 + \ldots + Z^2_{k-2})}{k-2}.$$

Since we are calculating the variance sequentially, consider the value of the variance by collecting squared normal random variables. To calculate S^2_k given S^2_{k-1}, use the values of $Z^2_1, Z^2_2 \ldots Z^2_{k-2}$ used to generate S^2_{k-1}. We then relate the distribution of S^2_k to S^2_{k-1} by

$$S^2_k \stackrel{d}{=} \left[\frac{S^2_{k-1}(k-2)}{\sigma^2} + Z^2_{k-1} \right] \frac{\sigma^2}{(k-1)}. \tag{10}$$

This follows from setting $S^2_{k-1}(k-2)/\sigma^2 \stackrel{d}{=} Z^2_1 + Z^2_2 + \ldots + Z^2_{k-2}$. Using (10), we can write the distribution of $S^2_k | S^2_{k-1}$ in terms of Z^2_{k-1}:

$$
\begin{aligned}
P(S^2_k \leq x \,|\, S^2_{k-1}) &= P\left(\left[\frac{S^2_{k-1}(k-2)}{\sigma^2} + Z^2_{k-1} \right] \frac{\sigma^2}{(k-1)} \leq x \,|\, S^2_{k-1} \right) \\
&= P\left(Z^2_{k-1} \leq \frac{1}{\sigma^2}[x(k-1) - S^2_{k-1}(k-2)] \,|\, S^2_{k-1} \right).
\end{aligned}
$$

Here, S^2_{k-1} and $S^2_k | S^2_{k-1}$ are formed from the same stream of squared normal random variables $Z^2_1, Z^2_2, \ldots, Z^2_{k-2}$, with $S^2_k | S^2_{k-1}$ including an additional random variable Z^2_{k-1}. Recall that in order to have $k^* > k-1$, S^2_{k-1} must be bounded from below by V_{min}. By integrating over the possible values of S^2_{k-1} according to its distribution conditional on not having stopped yet, we recursively calculate the conditional distribution of the variance as

$$f_{S^2_k | k^* > k-1}(x) = \int_{y=V_{min}}^{\infty} f_{Z^2_{k-1}} \left(\frac{1}{\sigma^2}[x(k-1) - y(k-2)] \right) f_{S^2_{k-1} | k^* > k-1}(y) dy, \tag{11}$$

where f is again used to represent the density function of the random variable in its subscript. Since the variance is calculated sequentially, at time k we have the value of S^2_{k-1}. We calculated (11) numerically and included the results in (9) to estimate the distribution of the stopping time.

REFERENCES

Chow, Y., and H. Robbins. 1965. On the Asymptotic Theory of Fixed-Width Sequential Confidence Intervals for the Mean. *The Annals of Mathematical Statistics* 36 (2): 457–462.

Glynn, P., and W. Whitt. 1992. The Asymptotic Validity of Sequential Stopping Rules for Stochastic Simulations. *The Annals of Applied Probability* 2 (1): 180–198.

Heidelberger, P., and P. Welch. 1983. Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research* 31 (6): 1109–1144.

Law, A., and W. Kelton. 1982. Confidence Intervals for Steady-State Simulations, II: A Survey of Sequential Procedures. *Management Science* 28 (5): 550–562.

Schmeiser, B., and Y. Yeh. 2002. On choosing a single criterion for confidence-interval procedures. In *Proceedings of the 2002 Winter Simulation Conference*, ed. J. S. ed. E. Yucesan, C.-H. Chen and J. Charnes, 345–352. Piscataway, New Jersey: Institute of Electrical and and Electronics Engineers.

Schruben, L. 1980. A Coverage Function for Interval Estimators of Simulation Response. *Management Science* 26 (1): 18–27.

Steiger, N., and J. Wilson. 2002. An Improved Batch Means Procedure for Simulation Output Analysis. *Management Science* 48 (12): 1569–1586.

AUTHOR BIOGRAPHIES

DASHI I. SINGHAM is a Ph.D. candidate in Industrial Engineering & Operations Research at the University of California at Berkeley. She has a Masters in Statistics from Berkeley, and holds a bachelors degree in Operations Research & Financial Engineering from Princeton University. Dashi's research interests include simulation modeling and analysis with applications in transportation. Her email address for these proceedings is <dsingham@berkeley.edu>.

LEE W. SCHRUBEN is a Chancellor's Professor and past Chairman in the Department of Industrial Engineering and Operations Research at the University of California at Berkeley. Prior to joining the Berkeley faculty, he was on the Operations Research and Industrial Engineering faculty at Cornell where he held the Schultz Endowed Professorship in Engineering. He received his PhD from Yale and is a Fellow of the Institute for Operations Research and Management Science. Professor Schruben's research interests are in simulation modeling and analysis methodologies with a broad range of applications, most recently focusing on biopharmaceutical production and supply chains. His email address is <lees@berkeley.edu>.