2014

Selecting stopping rules for confidence interval procedures

Singham, Dashi I.

ACM

D.I. Singham, 'Selecting stopping rules for confidence interval procedures," ACM Transactions on Modeling and Computer Simulation, Vol. 24, No. 3, Article 18, (July 2014), 18 p.

http://hdl.handle.net/10945/46168

Selecting Stopping Rules for Confidence Interval Procedures

DASHI I. SINGHAM, Naval Postgraduate School

The sample size decision is crucial to the success of any sampling experiment. More samples imply better confidence and precision in the results, but require higher costs in terms of time, computing power, and money. Analysts often choose sequential stopping rules on an ad hoc basis to obtain confidence intervals with desired properties without requiring large sample sizes. However, the choice of stopping rule can affect the quality of the interval produced in terms of the coverage, precision, and replication cost. This article introduces methods for choosing and evaluating stopping rules for confidence interval procedures. We develop a general framework for assessing the quality of a broad class of stopping rules applied to independent and identically distributed data. We introduce coverage profiles that plot the coverage according to the stopping time and reveal situations when the coverage could be unexpectedly low. Finally, we recommend simple techniques for obtaining acceptable or optimal rules.

Categories and Subject Descriptors: I.6.6 [**Simulation Output Analysis**]; I.6.1 [**Simulation Theory**]; I.6 [**Simulation and Modeling**]

General Terms: Confidence interval procedures, Sequential stopping rules

Additional Key Words and Phrases: Finite-sample performance, coverage profiles

ACM Reference Format:
Dashi I. Singham. 2014. Selecting stopping rules for confidence interval procedures. ACM Trans. Model. Comput. Simul. 24, 3, Article 18 (July 2014), 18 pages.
DOI: http://dx.doi.org/10.1145/2627734

1. INTRODUCTION

Any simulation or sampling experiment requires a method for choosing the number of observations to collect. The sampling decision is often determined by budgetary constraints, in that there is limited time and money to execute the experiment. The goal is to balance high quality in the statistical results against increased effort required. Usually, a confidence interval is used to estimate relevant system parameters, though other methods have been suggested [Song and Schmeiser 2009]. Most methods are similar in that they provide some estimate of the mean output quantity, and a level of uncertainty associated with that estimate. Quality of the confidence interval is determined by its achieved confidence level and its precision (or half-width). Generally, a higher sample size is required to achieve higher quality in the results. We refer to procedures designed to generate confidence intervals as confidence interval procedures (CIPs). CIPs can be applied to replications of a computer simulation, or to observations sampled from a physical experiment.

Existing sequential stopping rules attempt to present sampling algorithms that provide the user with the desired quality in the output interval using a minimum number

Author's address: D. Singham, Operations Research Department, Naval Postgraduate School, Monterey CA 93943.

DOI: http://dx.doi.org/10.1145/2627734

of samples. Without knowledge of the distribution of the data, it can be hard to know in advance what would be an appropriate sample size. If the replication cost is low or zero, sequential rules can be used to determine a natural time to stop as otherwise the experiment could run indefinitely. If the replication cost is high, sequential rules can be used to stop as early as possible while still attempting to maintain interval quality. Sampling stops when there is enough output to produce a confidence interval supposedly having the desired confidence level and precision. What is often ignored is the quality of the stopping rule itself. The true coverage of a procedure (the probability that a CIP delivers an interval that contains the true value of the parameter being estimated) is often less than what is intended.

There are many causes for deviation in the coverage. One cause is when the properties of the data are different than those assumed by the method (many methods assume independence and/or normality in the data). However, there is often a loss in the coverage induced by the use of a stopping rule itself, even if the data meet all the assumptions of the procedure. Rules to determine sample sizes of experiments are often chosen on an ad hoc basis, and without knowledge of their quality. If there is a high cost to each replication, the user may design a procedure to stop as early as possible, with an unknown detrimental effect on the results. If the true coverage of the procedure is low, the user has underestimated the risk associated with his or her model. If the true coverage is high, then the procedure has required more replications than necessary. The objective of this article is to introduce graphical and numerical methods for evaluating the coverage of sequential CIPs under a broad context. Additionally, we compare solutions for improving CIP performance (in terms of the coverage and the expected number of observations required).

Approaches to evaluating losses in the coverage and choosing optimal parameters for absolute-precision stopping rules were developed in Singham and Schruben [2012]. That paper derived the coverage when the simulation output was independent and identically distributed (i.i.d.) with a normal distribution, and generated coverage contours to find optimal stopping rule parameters. In this article, we develop a framework for evaluating the coverage for a broad class of CIPs where stopping depends on functions of sample statistics of the data. The framework does not require a specific stopping rule structure (so includes absolute- and relative-precision rules), and also does not make distributional assumptions on the data, although the data must be i.i.d. Additionally, the framework explains how stopping and the coverage are determined by the evolution of the joint distribution of state statistics used in the CIP. Although we often rely on one commonly used CIP to illustrate the ideas in this article, the methods introduced apply to a much broader class of CIPs.

To better understand these types of stopping rules, we discuss parameters used in CIPs and methods for evaluating CIP performance. We present a graphical representation of sequential stopping rules for CIPs and compare it to fixed-sampling representations introduced by Kang and Schmeiser [1990]. Additionally, we introduce coverage profiles that map the coverage according to the stopping time of a rule. Although the coverage of a procedure might be nominal over many replications, particular intervals may have a smaller chance of covering the true mean if the procedure stopped early. Coverage profiles demonstrate how increasing the starting sample size can drastically improve the coverage, because instances where early stopping happens contribute disproportionately to a poor coverage. Many sequential procedures suggest a higher starting sample size to avoid a low coverage, and here we quantify the effects of this strategy.

Most of the current literature addresses the coverage problem by relying on asymptotic results (such as Chow and Robbins [1965]) where a nominal coverage is achieved as the sample size approaches infinity. Here, we focus on finite-sample solutions where we also include the expected stopping time as a measure of quality. Traditionally, small desired precision values have been suggested to push the sample size high enough to obtain a coverage that is close to nominal. We compare this method and new methods with those suggested in Singham and Schruben [2012], and analyze the costs associated with each method.

We describe some background and notation for sequential stopping rules in Section 2. We discuss the relationships between stopping rule parameters and the graphical representations for evaluating the coverage in Section 3. Section 4 describes how the coverage for stopping rules can be calculated in the general case. Section 5 discusses methods for choosing stopping rule parameters to obtain improved or optimal results and Section 6 compares these methods and concludes.

2. BACKGROUND

This section introduces notation and examples of the types of sequential stopping rules we consider and a review of the relevant literature. We emphasize that the rules presented here and in later sections are examples that illustrate the new methods for evaluating stopping rules, and we are not necessarily promoting the use of particular rules.

Sequential CIPs are designed to deliver confidence intervals for simulation output that meet the specifications of the modeler. Two parameters that are often chosen in advance are the desired confidence coefficient (η) and the desired half-width (δ) of the output confidence interval. Let k be the number of observations collected to produce a confidence interval. The value of k can be incremented as observations are added until an interval with confidence coefficient η and half-width smaller than δ can be generated. For sequential rules, the stopping time is random and denoted as k^*.

A basic sequential stopping rule for estimating the sample mean, μ, using chosen parameters η and δ works as follows. We call this particular rule CIP1 and use it as an example throughout this article. Suppose that CIP1 is applied to simulation output that is assumed to be i.i.d. normally distributed. After the user has generated k simulation output samples, the half-width of a confidence interval for the mean can be calculated using the following standard formula:

$$H_{\eta,k} = t_{\eta,k-1}\sqrt{\frac{S_k^2}{k}},$$

where S_k^2 is the sample variance of the k observations and $t_{\eta,k-1}$ is the $(1+\eta)/2$ quantile of the t-distribution with $k-1$ degrees of freedom. If $H_{\eta,k} \leq \delta$, then the procedure stops and returns the interval $[\overline{X}_k - \delta, \overline{X}_k + \delta]$, where $\overline{X}_k$ is the sample mean of the k observations. If $H_{\eta,k} > \delta$, then k is incremented by 1 and the half-width is checked again after including the new observation. The stopping time k^* can be defined according to the following rule:

$$k^* = \min_{k\geq 2} \{k : H_{\eta,k} \leq \delta\} \qquad \text{(CIP1)}.$$

The result of the experiment is a value k^*, and an interval that either covers μ, or fails to cover it. Repeating this experiment multiple times yields an estimate of the coverage of the procedure, which is the proportion of times the resulting interval covers the true parameter. We denote the true coverage of a procedure by the function $\eta^*(\eta, \delta, k_{\min})$, where $k_{\min}$ is the starting sample size, and we will attempt to calculate the value of this function for various input parameters. Many sequential procedures assume that $k_{\min}$ is somewhere between 10 and 30. However, in very expensive experiments (or in clinical trials, where a patient's health is at stake), a smaller value of $k_{\min}$ may be used. A procedure that returns a nominal coverage is one where the actual coverage obtained is equal to the coverage desired. The function $Ek^*(\eta, \delta, k_{\min})$ is the expected stopping time for a rule, and is another measure of the quality of a CIP. We abbreviate the notation for these functions to η^* and Ek^*. For small δ, the values of η^* and Ek^* usually increase as η increases and as δ decreases to 0. Note that the functions η^* and Ek^* apply to any CIP that requires a confidence coefficient, a precision parameter, and a starting sample size, not just CIP1.

We note that many recent CIPs exist and can be applied to data that are not independent or normally distributed, and these may use relative precision stopping rules instead of absolute precision rules. Many of these rules do not check the stopping condition at each observation, but calculate an appropriate number of replications of data to simulate before checking the rule using the batch means method (for an example, see Chen [2012]). We use CIP1 as an example throughout this article because it is simple enough that when applied to i.i.d. normal data, the coverage and distribution of k^* can be derived explicitly as in Singham and Schruben [2012], but the methodology and graphical tools developed can apply to many other types of rules.

This article considers the coverage performance of a general class of CIPs that depend on functions of sample statistics of the data. Any function of the sample mean and sample variance can be used to determine the stopping time. Relative-precision rules are included in this class, where stopping occurs when the half-width is less than or equal to some fraction of the sample mean, assuming the true mean is positive. We define a simple version of this rule as CIP2:

$$k^* = \min_{k \geq 2} \{k : H_{\eta,k} \leq \delta \overline{X}_k\} \qquad \text{(CIP2)}.$$

Relative-precision rules can be useful if the user does not know what amount of precision will be appropriate given the unknown scale of the sample mean. One potential problem with relative-precision rules is that if the sample mean happens to be too large relative to the sample variance, then stopping might occur early, thus hurting the coverage.

Previous analytical results for i.i.d. data state that as δ approaches 0, the coverage of absolute-precision stopping rules approaches η [Chow and Robbins 1965; Glynn and Whitt 1992]. The relative-precision case is studied in Nadas [1969]. This asymptotic validity is often used to justify using sampling rules with small values of δ. Early papers such as Ray [1957] and Law and Carson [1979] document losses in the coverage because of sequential sampling. Law and Kelton [1982] conducted a survey of sequential procedures, and found some rules that performed favorably, although they acknowledged that small sample sizes could lead to a loss in the coverage. Fixed-sample analysis of confidence interval coverage for small samples was investigated in Sargent et al. [1992], where the authors study the importance of having an unbiased variance estimator. Finite-sample analytical results for i.i.d. normally distributed data with known variance were studied in Singham and Schruben [2009], where the distribution of the stopping time and the loss in the coverage were calculated. Because of increased computational power, analysis for larger sample sizes is now readily available in a sequential setting.

3. GRAPHICAL REPRESENTATION OF STOPPING RULES

This section analyzes the effects of various parameters on the quality of stopping rules using graphical methods. Before we discuss these methods, we offer a visual representation of sequential stopping rules that is analogous to the confidence interval coverage plots introduced in Kang and Schmeiser [1990] and motivate the framework presented in Section 4.

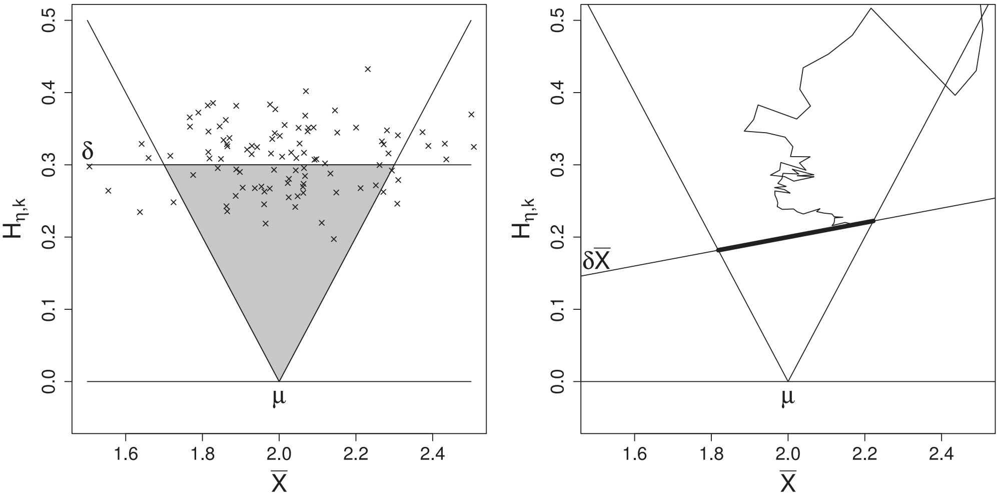

Fig. 1. Left: Intervals generated from multiple realizations of a fixed-sampling experiment. Intervals in the shaded area cover the true mean and meet the precision requirement. Right: The path to stopping for one realization of a relative-precision sequential rule. Intervals that stop in the bold portion of the line succeed in covering μ.

3.1. Visualizing coverage

Graphical representations of confidence interval coverage were introduced in Kang and Schmeiser [1990]. Here, we use the same idea of plotting confidence intervals against the sample mean and half-width in order to estimate CIP coverage. The left plot of Figure 1 shows the results of a fixed-sampling experiment with the crosses marking the center point and half-width of the resulting intervals. We add a horizontal line to mark the desired maximum precision amount of the output interval. The crosses that fall within the lines forming the "V" succeed in covering the true mean. Each line forming the "V" is at a 45-degree angle from the horizontal axis. The crosses that fall below the horizontal line meet the precision requirement. The right plot shows the graphical representation of one realization of a relative-precision sequential sampling rule, and includes a slanted line to mark when the stopping criterion is reached. The jagged line is the path taken by a realization of a CIP as observations are collected. The procedure stops when the stopping line is reached, and the coverage is determined by the proportion of times the path stops in the coverage region. For the fixed-sampling rule, calculating the coverage analytically involves taking the probability-weighted area over the coverage region. For the sequential stopping rule, we derive the coverage in Section 4 by calculating the probability that the path stops in the coverage region.

Next, we discuss graphical methods for evaluating the performance of stopping rules. Many evaluations of the coverage consist of choosing a few test parameters and comparing the actual coverage to nominal. Graphical methods can be used to evaluate the coverage over a range of parameters, thus aiding in the choice of rules. We use these methods to determine the worst-case coverage scenarios, and we design policies to improve the coverage for those cases (leading to conservative coverages in the general case).

Table I lists three graphical tools for evaluating stopping rules. Each tool compares different performance metrics against different parameters. The coverage function, developed by Schruben [1980], compares the true coverage to that intended for all η. Schmeiser and Yeh [2002] use coverage functions to develop a single dimensionless criterion by integrating the deviation of the coverage function from nominal coverage.

Table I. Graphical Methods for Evaluating the Coverage

	Inputs	Outputs
Coverage functions [Schruben 1980]	η	η^*
Coverage contours [Singham and Schruben 2012]	η, δ	η^*, Ek^*
Coverage profiles	k^*	$\eta^*(k^*)$

Coverage contours compare rules over the space of (η, δ) and display metrics such as η^* and Ek^* [Singham and Schruben 2012]. In the next section, we introduce coverage profiles that map the coverage according to the stopping time of a rule.

3.2. Coverage Profiles

We introduce a new way of evaluating stopping rules by comparing the coverage at different stopping times. This allows us to see how good (or bad) the coverage is if a rule happens to stop early. Early stopping can happen when the first few data points are unusually close together, driving the sample variance down. A valid stopping rule is one that delivers intervals that cover the true parameter with probability η. We define the notion of an "optimal stopping rule" for a particular type of rule and data type as a stopping rule using parameters that minimize the expected stopping time while delivering at least a nominal coverage and meeting a precision constraint.

However, in reality, an experiment is typically conducted only once, and a randomly bad confidence interval could mean an imprecise result for decision-making, even if the procedure used delivers a nominal coverage on average. We introduce coverage profiles to explain the relationship between the coverage and the stopping time. Coverage profiles plot the coverage of the stopping rule given different stopping times. The stopping time of a procedure is often an indicator of the quality of the coverage of the interval. A stopping rule requiring many samples is likely to obtain a better coverage because more information has been collected. A run that stops after two or three samples might have a worse coverage because a randomly too-small half-width led to stopping. Coverage profiles provide information on how much to increase the starting sample size to reduce the probability of early stopping and to improve the coverage.

Let $c(k)$ be the coverage of a stopping rule given that $k^* = k$. Figure 2 plots values of $c(k)$ using CIP1 with parameters (0.9,0.3,2) applied to $\mathcal{N}(0, 1)$ data. This rule starts with two samples, so when the stopping rule is met early (say, $k^* = 2$), $c(k)$ is much less than when it meets the stopping rule at $k^* = 60$. We call the values of $c(k)$ a coverage profile, and use them to evaluate the quality of the stopping rule. The stopping rule used in Figure 2 has an overall coverage of 83%, on average, which can be calculated using either the method in Singham and Schruben [2012], or empirical testing. If stopping happens early, the potential undercoverage is even worse. Coverage profiles can be used to determine if there are enough samples to obtain a good interval, or to fix a minimum sample size. The overall probability of coverage can be calculated according to the following:

$$\begin{aligned} P(\text{Cover}) &= \sum_k P(\text{Cover} \mid \text{Stop at } k) P(\text{Stop at } k) \\ &= \sum_k c(k)p(k), \end{aligned} \tag{1}$$

where $p(k)$ is the distribution of the stopping time, or $P(k^* = k)$. It should be noted that $c(k)$ and $p(k)$ are rarely independent, although they are for CIP1 where the data are independent and normally distributed and the stopping rule depends only on the sample variance and not the sample mean. This is because $c(k)$ can be calculated

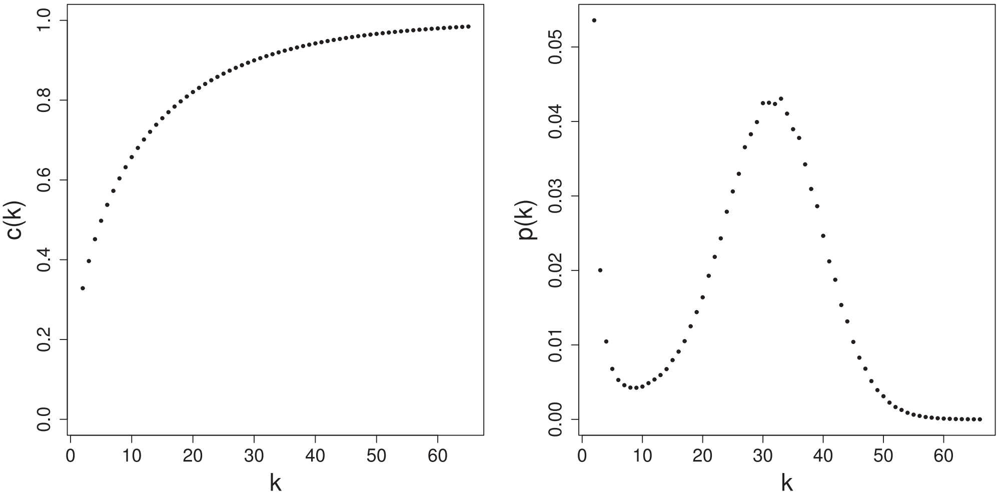

Fig. 2. Coverage profile and stopping time distribution for CIP1 with $\eta = 0.9$, $\delta = 0.3$, and $k_{\min} = 2$ applied to $\mathcal{N}(0, 1)$ data. Left: Coverage profile $c(k)$. Right: Distribution function $p(k)$.

explicitly and does not depend on the historical values of the sample variance (only the current value of the sample variance is needed to determine if stopping should occur). The independence of the sample mean at stopping from the sample variance history is proven in Robbins [1959], and alternate proofs are given in Singham and Schruben [2012]. Let $F_{\overline{X}_k}(x)$ be the cumulative distribution function of the sample mean of k realizations, where $\overline{X}_k \sim \mathcal{N}(\mu, \sigma^2/k)$ and σ^2 is the variance of the data. Then $c(k)$ is the probability that $\overline{X}_k$ is between $\mu+\delta$ and $\mu-\delta$ at stopping, which is $F_{\overline{X}_k}(\mu+\delta)-F_{\overline{X}_k}(\mu-\delta)$.

The left plot of Figure 2 displays the values of $c(k)$ for an implementation of CIP1, so it remains to calculate $p(k)$. For normally distributed data, this distribution can be calculated analytically for absolute-precision rules, as derived in Singham and Schruben [2012], and it can be estimated using simulation for any distribution. The right plot of Figure 2 shows the distribution of the stopping time for the same rule. The mode of the distribution is at $k^* = 2$, but this is also the stopping time with the worst coverage according to $c(k)$. This motivates an exploration into increasing the minimum sample size to see how much the coverage will improve by avoiding early stopping. Many stopping rules improve the coverage by pushing the density of k^* out to higher values. For example, using a smaller value of δ or a higher value of η forces higher stopping times, which reallocates weight to higher values of $c(k)$ (note that $c(k)$ may also change, depending on the rule and distribution of the data). Section 5.4 will discuss specific methods for using coverage profiles to improve the coverage by increasing $k_{\min}$.

4. CALCULATING COVERAGE OF SEQUENTIAL RULES

Before we present solutions to the coverage problem in Section 5, we explain how the structure of stopping rules can be generalized. We evaluate sequential CIPs by seeing how they perform when applied to data with known distributions. Previous work in Singham and Schruben [2012] derived $c(k)$ and $p(k)$ for CIP1 applied to normal data. This section generalizes those results to any distribution, and allows for different stopping rule structures (not limited to absolute-precision or relative-precision rules). We present a numerical integration scheme to calculate the coverage. Simulation methods can easily be employed to obtain the same information, but the integration method demonstrates the underlying components of stopping rule performance. Because we are able to obtain similar results using both methods, we suggest practitioners use simulation, as it is faster to implement and execute.

4.1. Generalized Stopping Rules

The most basic stopping rule (CIP1) uses t-distribution confidence intervals and has an absolute-precision stopping rule as described earlier. For some purposes, understanding the performance of this rule might be sufficient. Whereas normality and independence are rarely valid assumptions to make for real data, these assumptions allow us to isolate the effects of the stopping rule on the coverage results. To calculate performance under broader conditions, we will attempt to generalize stopping rule performance across different metrics and data distributions. We require notation for statistics that make up stopping criteria, and functions that can be applied to these statistics to determine if the stopping rule has been met and if the output interval covers the true parameter being estimated.

Suppose we receive i.i.d. observations (or simulation replications) $X_1, X_2, \ldots, X_k$. To evaluate the true coverage of a CIP applied to a particular type of data, we assume that the mean μ is known. We define a vector $\mathbf{v}_k$ that contains the statistics of the data at time k that need to be collected to evaluate the stopping rule. We index this vector by k because it will evolve as k increases and samples are collected. For CIP1, the required statistics are the sample mean, $\overline{X}_k$, and the sample variance, S_k^2, so $\mathbf{v}_k = (\overline{X}_k, S_k^2)$. If we suspect our data are non-normal, we can use a modified Cornish-Fisher expansion as is done in Tafazzoli et al. [2011] to obtain confidence intervals that are calculated using the skewness of the data. In this case, one of the statistics in the vector $\mathbf{v}_k$ would be the sample skewness.

Let $T_k(\mathbf{v}_k)$ be the function that returns 1 if the stopping rule is met (not necessarily for the first time) at k when the statistics take values $\mathbf{v}_k$, and 0 otherwise. Again, we index the function T_k by k to denote that the value of k matters in function evaluation. For each sample collected, we evaluate $T_k(\mathbf{v}_k)$ to determine if stopping has occurred. Hence, for a particular sample path, k^* is equal to $\min_{k \geq k_{\min}}\{k : T_k(\mathbf{v}_k) = 1\}$. We also define a function $C_k(\mathbf{v}_k)$ that returns 1 if the interval returned by $\mathbf{v}_k$ includes μ, and 0 otherwise. For a relative-precision rule, $T_k(\mathbf{v}_k) = 1$ if $H_{\eta,k} \leq \delta\overline{X}_k$, and $C_k(\mathbf{v}_k) = 1$ if $|\overline{X}_k - \mu| \leq \delta\overline{X}_k$. For CIP1, the functions $T_k(\mathbf{v}_k)$ and $C_k(\mathbf{v}_k)$ are:

$$T_k\big(\overline{X}_k, S_k^2\big) = \begin{cases} 1 & \text{if } H_{\eta,k} \leq \delta \\ 0 & \text{o.w.} \end{cases} \qquad C_k\big(\overline{X}_k, S_k^2\big) = \begin{cases} 1 & \text{if } |\overline{X}_k - \mu| \leq \delta \\ 0 & \text{o.w.} \end{cases} \tag{2}$$

What will be helpful in numerically calculating the coverage of specific stopping rules is to have a way of updating the values from $\mathbf{v}_{k-1}$ to $\mathbf{v}_k$ for a new observation, X_k. For example, if the stopping rule and the coverage criteria only depend on the sample mean and sample variance, then we can update from $\mathbf{v}_{k-1} = (\overline{X}_{k-1}, S_{k-1}^2)$ to $\mathbf{v}_k$ using the following formulas:

$$\overline{X}_k = \frac{(k-1)\overline{X}_{k-1} + X_k}{k}, \qquad S_k^2 = \frac{k-2}{k-1}S_{k-1}^2 + \frac{(X_k - \overline{X}_{k-1})^2}{k}. \tag{3}$$

The updating formula for S_k^2 is found in Welford [1962]. If a relationship between the time steps exists, it becomes easier to update the conditional probability of stopping at k, given that stopping has not occurred yet. The only information that needs to be maintained are the values of $\mathbf{v}_{k-1}$ and the new observation X_k. Of course, if the stopping rule requires a more-detailed data history, then those values must also be recorded in $\mathbf{v}_k$.

For an interval generated from a given fixed sample size k, we calculate the probability that the stopping rule is met and that the interval covers μ. We do this by taking the probability-weighted integral over the space of $\mathbf{v}_k$ that results in meeting the stopping rule and covering μ. Let $f_{\mathbf{v}_k}$ be the joint density function of the statistics in $\mathbf{v}_k$ for a fixed k, and denote the m components of $\mathbf{v}_k$ as $v_1, v_2, \ldots, v_m$. Then for a fixed sample size k, we calculate the probability that the interval meets the stopping rule, that it covers the true mean, and that it meets both conditions. These three probabilities are:

$$P(\text{Stopping rule met at } k) = \int_{v_1} \cdots \int_{v_m} T_k(\mathbf{v}_k) f_{\mathbf{v}_k}(\mathbf{v}_k) dv_m \cdots dv_1 \tag{4}$$

$$P(\text{Cover at } k) = \int_{v_1} \cdots \int_{v_m} C_k(\mathbf{v}_k) f_{\mathbf{v}_k}(\mathbf{v}_k) dv_m \cdots dv_1$$

$$P(\text{Stop and Cover at } k) = \int_{v_1} \cdots \int_{v_m} T_k(\mathbf{v}_k)\, C_k(\mathbf{v}_k) f_{\mathbf{v}_k}(\mathbf{v}_k)\, dv_m \cdots dv_1. \tag{5}$$

Equation (5) is the probability weighted integral over the values of $\mathbf{v}_k$ for which both the stopping rule is met and μ is covered, as in the shaded area in the left plot of Figure 1. The indicator function for the intersection of both the stopping and coverage events is equal to the product of each event's indicator function. For CIP1 (but using a fixed sample size), Equation (5) becomes:

$$P(\text{Stop and Cover at } k) = \int_{x=-\infty}^{\infty} \int_{y=0}^{\infty} T_k(x, y) C_k(x, y) f_{J_k}(x, y) dy dx,$$

where f_{J_k} is the joint density of $(\overline{X}_k, S_k^2)$ at time k and $T_k(x, y)$ and $C_k(x, y)$ are evaluated according to Equation (2). The probability that the interval covers the true parameter given that it meets the stopping rule is Equation (5) divided by Equation (4). For sequential stopping rules, we need to monitor the movement from the nonstopping region at $k-1$ to the stopping region at k, and calculate the probability that the interval covers μ at stopping. Rather than considering the probability-weighted areas as in the fixed-sample-size case, consider a random stopping time model for the movement of $\mathbf{v}_k$ toward the stopping region. We break down the coverage according to k^*, and for each k calculate the probability that the procedure stops and covers at k (denoted $P_{\text{SC}}(k)$):

$$\eta^* = P(\text{Cover}) = \sum_k P(\text{Stop at } k \text{ and Cover } \mu) = \sum_k P_{\text{SC}}(k),$$

which is a way of rewriting Equation (1). Denote the joint probability distribution function of $\mathbf{v}_k$ with the event that stopping has not occurred before k by $f_{G_k}(\mathbf{v}_k)$. Given that the values of $\mathbf{v}_k$ can take a number of paths as the experiment proceeds, f_{G_k} is the probability it ever reaches $\mathbf{v}_k$ at k. Integrating over possible values of $\mathbf{v}_k$ yields the probability that the procedure ever reaches time k:

$$P(k^* \geq k) = \int_{v_1} \cdots \int_{v_m} f_{G_k}(\mathbf{v}_k) dv_m \cdots dv_1.$$

Calculate $P_{\text{SC}}(k)$ as

$$P_{\text{SC}}(k) = \int_{v_1} \cdots \int_{v_m} T_k(\mathbf{v}_k)\, C_k(\mathbf{v}_k) f_{G_k}(\mathbf{v}_k)\, dv_m \cdots dv_1, \tag{6}$$

where using f_{G_k} incorporates the probability that the procedure reaches time k, and $T_k(\mathbf{v}_k)$ and $C_k(\mathbf{v}_k)$ account for stopping and covering μ at time k. As in Equation (5), we calculate the probability that both the stopping and coverage conditions are met by taking the probability weighted integral over the sample statistic space where both indicator functions are true. In this case, the distribution of the sample statistics (f_{G_k}) is influenced by the fact that we are using a sequential rule and the stopping condition has not been met prior to k. The probability of stopping at k is the probability of the procedure making it to time k and meeting the stopping rule at k:

$$P(k^* = k) = \int_{v_1} \cdots \int_{v_m} T_k(\mathbf{v}_k) f_{G_k}(\mathbf{v}_k)\, dv_m \cdots dv_1.$$

It remains to calculate f_{G_k}. For CIPs that rely on the sample statistics $\overline{X}_k$ and S_k^2, we have functions that update values of $\overline{X}_{k-1}$ and S_{k-1}^2 based on a new observation X_k to calculate $\overline{X}_k$ and S_k^2 according to Equation (3). We invert these functions so that given $\overline{X}_k$, S_k^2, and X_k, we find the corresponding values of $\overline{X}_{k-1}$ and S_{k-1}^2. Define these inverting functions for general $\mathbf{v}_k$ as I_k, where $\mathbf{v}_{k-1} \leftarrow I_k(\mathbf{v}_k, X_k)$. Then calculate $f_{G_k}(\mathbf{v}_k)$ as

$$f_{G_k}(\mathbf{v}_k) = \int_z [1 - T_{k-1}(I_k(\mathbf{v}_k, z))]\, f_{G_{k-1}}(I_k(\mathbf{v}_k, z))\, f_{X_k}(z) dz, \tag{7}$$

where f_{X_k} is the probability distribution function of the observation X_k. Equation (7) is derived by integrating over all possible values of X_k, and for each X_k inferring the value of $\mathbf{v}_{k-1}$ that, when updated with X_k, leads to $\mathbf{v}_k$. We integrate the probability-weighted areas over $\mathbf{v}_{k-1}$, where stopping did not happen (hence allowing the process to make it to $\mathbf{v}_k$). Singham and Schruben [2012] derive a related conditional distribution explicitly for CIP1 applied to normally distributed data.

4.2. Numerical Procedure

The main goal is to calculate $P_{\text{SC}}(k)$ for each $k \geq k_{\min}$. To do this, we develop a numerical integration scheme to calculate f_{G_k}. We discretize the space along each dimension of $\mathbf{v}_k$ and discretize the support of X_k. The first step is to calculate the joint distribution of $\mathbf{v}_{k_{\min}}$ to initialize the procedure. We use a Newton-Cotes method with a rectangular rule approximation. Then we use a discretized version of Equation (5) to calculate $P_{\text{SC}}(k_{\min})$. The problem becomes a matter of calculating probabilities for a discrete set of states.

To continue to step $k_{\min} + 1$, consider the transition from time $k - 1$ to k. We project the distribution of $\mathbf{v}_k$ forward using known relationships among $\mathbf{v}_{k-1}$, $\mathbf{v}_k$, and X_k. In order for the system to be in state $\mathbf{v}_k$ and for stopping to occur after time $k - 1$, we project the distribution of the system at states $\mathbf{v}_{k-1}$ (where stopping does not occur) forward by integrating over the possible values of X_k that lead to the resulting states $\mathbf{v}_k$. For example, if we were starting at time $k - 1 = k_{\min} = 2$, we would begin with the possible values of $\mathbf{v}_2$ where $T_2(\mathbf{v}_2) = 0$. Initialize $f_{G_3}(\mathbf{v}_3) = 0$ for all states in $\mathbf{v}_3$. For each $\mathbf{v}_2$ where $T_2(\mathbf{v}_2) = 0$, we loop over the discretized possible values of X_3, calculate the appropriate values of $\mathbf{v}_3$, and increment $f_{G_3}(\mathbf{v}_3)$ by $f_{G_2}(\mathbf{v}_2) f_{X_3}(X_3)$. In this way, we calculate the probability that the system reaches a particular set of values $\mathbf{v}_3$ without stopping beforehand. At each stage after calculating f_{G_k}, we calculate the probability of stopping at k and covering the true mean using Equation (6).

The main input to this procedure is a discretization of the support of the statistics of $\mathbf{v}_k$ and X_k and the associated probability weights at each increment of X_k. We use the probability weights of X_k to calculate the joint densities of statistics such as $\overline{X}_k$ and S_k^2 using the updating functions. Error is introduced because $T_k(\mathbf{v}_k)$ and $C_k(\mathbf{v}_k)$ are evaluated at the centers of each rectangle of the discretization, but those values are applied over the entire rectangle. Increasing the number of discretization points improves the solution, but there appears to be a negative bias. The coverage delivered by the integration will generally be less than the true coverage, but will approach the true value as the discretization is refined. The left plot of Figure 1 shows the regions for which $T_k(\mathbf{v}_k)$ and $C_k(\mathbf{v}_k)$ evaluate to 1 or 0. It is important for the discretization to be fine enough at the corresponding values of $\overline{X}_k$ and S_k^2 on the boundary of the shaded triangle for the solution to converge. Note also that the value of S_k^2 that corresponds to the half-width threshold required for stopping increases as k increases.

4.3. Numerical Results

Simulation can be used to estimate the distribution of k^* and the probability of covering μ by repeatedly applying the stopping rule to random data simulated from a known distribution. Both numerical integration and simulation lead to approximate results, with longer computation times required to get better results. Generally, we recommend simulation because the implementation and run time are faster than for numerical integration. Our routine involved discretization of X_k and the statistics in $\mathbf{v}_k$ into at least 2,000 increments. The simulation results use 10 million replications for each experiment.

We briefly highlight examples illustrating the types of rules that the procedure just described can be applied to, and we compare the performance with simulation methods. The first example is CIP1 using $\eta = 0.9$, $\delta = 0.3$ applied to $\mathcal{N}(0, 1)$ data. Using the method in Section 4.2, we were able to numerically derive the values of $P_{\mathrm{SC}}(k)$ and the overall coverage as approximately 83% and the expected stopping time as approximately 28. The code was written in C and ran on a single processor with 2GB of memory in 7.5 hours. A simulation applying the stopping rule 10 million times returned values for each $P_{\mathrm{SC}}(k)$ that were within 0.3% of the numerical results in 15 minutes, with an overall difference in the coverage of under 0.5%.

Consider the relative precision rule CIP2 applied to data that are exponentially distributed with mean 1. We find a difference in the coverage delivered by the analytical and simulation methods to be less than 0.15%, and the maximum difference in $P_{\mathrm{SC}}(k)$ was less than 0.04%. Simulation results were computed in 26 minutes, whereas the numerical results took 15 hours to compute.

As a final example, consider a situation where a performance measure of a system is compared to a fixed value, c, and we are interested in estimating if the system performance is better or worse than c. A naive sequential experiment involves sampling until an interval with confidence coefficient η can be generated that does not include c. Once this interval is generated, the sample mean is compared to c to provide an estimate of whether the true performance μ is greater or less than c. Modifying the functions $T_k(\mathbf{v}_k)$ and $C_k(\mathbf{v}_k)$ allows us to run this experiment, and we find that using $\eta = 0.9$, $c = 0.3$, with data that are distributed as $\mathcal{N}(0, 1)$, both the numerical and simulation methods suggest this procedure will deliver the correct answer approximately 93% of the time. Of course, more-sophisticated methods exist for comparing the performance of two systems using sequential methods (see Kim and Nelson [2001]).

5. SOLUTIONS FOR IMPROVED COVERAGE

In this section, we analyze different reasons for low coverages of CIPs and present possible solutions. Some of these solutions are the result of optimization models, and some are the result of worst-case analysis. Because the underlying distribution of simulation output is rarely known, we encourage modelers to use these policies to obtain conservative rules, rather than rules optimized for a particular distribution. Asymptotically, stopping rules of type CIP1 (and other similar types) obtain a nominal coverage as δ approaches 0. This section presents methods for obtaining an improved coverage in a finite-sample setting. We vary the parameters η, δ and $k_{\min}$ and assume that we desire some fixed coverage η^o and precision δ^o. The results presented are calculated using simulation unless otherwise indicated.

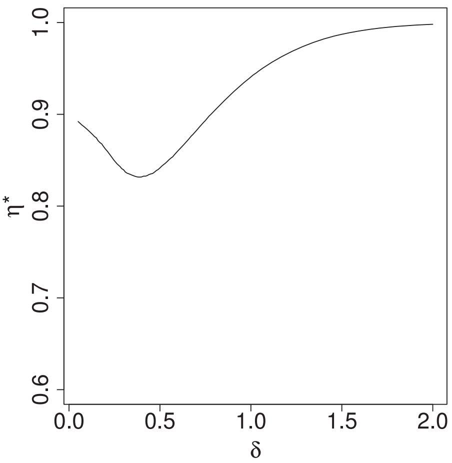

Fig. 3. The function $\eta^*(0.9, \delta, 2)$ applied to $\mathcal{N}(0, 1)$ data, plotting the coverage at both limits and the worst case.

5.1. Choosing a Precision

A common way to address the coverage problem is to choose a value of δ small enough so that the asymptotic conditions approximately hold. Singham and Schruben [2012] calculate coverage contours over the space of η and δ, and for all examples considered, the coverage converges from below as δ approaches 0. By choosing a positive value of δ, the stopping time is finite but the coverage is subnominal, which means the output confidence interval is underestimating risk. Figure 3 plots the typical effect of δ on the coverage. As δ decreases to 0, the coverage approaches nominal, and as δ increases, it approaches 1 as the intervals become wide. There is a value of δ where the coverage is worst, and we will use this value in later experiments to explore worst-case scenarios.

We can choose values of δ that are small enough to give a coverage that is close to nominal, but the appropriate choice of δ depends on the underlying variance of the data. Data with a higher variance require more samples on average to meet a precision requirement than data with a lower variance, and hence has a better coverage for a given value of δ. Data with a lower variance are likely to meet the precision requirement earlier, and too few samples may lead to poor coverage values.

Generally, the value of δ may be prespecified by the user because some amount of precision is desired. The stopping rule should be tested against data with the approximate distribution of the simulation data. If the coverage level is too low, then the value of δ can be decreased and the coverage tested until an acceptable level is found, because the coverage generally increases as δ decreases. The following minimization formulation shows how δ is often selected:

$$\begin{aligned} \min_{\delta} \quad & Ek^*(\eta^o, \delta, k_{\min}) \\ \text{s.t.} \quad & \eta^*(\eta^o, \delta, k_{\min}) \geq \eta^o - \epsilon \\ & \delta \leq \delta^o. \end{aligned} \tag{8}$$

Because a nominal coverage cannot be obtained using this policy (except in the limit), we must accept some lower coverage $\eta^o - \epsilon$, where ϵ is a small positive number. Table II lists values of δ required to achieve a subnominal coverage for different values of the variance σ^2 (for normally distributed data) used in CIP1 and CIP2, where $\epsilon = 0.01$. For CIP1, we use $\mathcal{N}(0, \sigma^2)$ data, and for CIP2, we use $\mathcal{N}(5, \sigma^2)$ data. In addition to the disadvantage of undercoverage, decreasing δ can be quite expensive. The following

Table II. Values of δ Required for $\mathcal{N}(0, \sigma^2)$ Data for CIP1 and $\mathcal{N}(5, \sigma^2)$ for CIP2
For the four right-most columns, the first value is the value of δ required, the second value is the corresponding Ek^*

	Intended coverage	Actual coverage	$\sigma^2 = 1/4$	$\sigma^2 = 1$	$\sigma^2 = 4$	$\sigma^2 = 100$
CIP1	0.90	0.89	0.03, 743.6	0.06, 743.7	0.12, 743.5	0.63, 674.1
	0.95	0.94	0.05, 381.4	0.11, 315.0	0.22, 315.0	1.14, 293.4
	0.99	0.98	0.18, 53.0	0.37, 50.3	0.75, 49.0	3.74, 49.2
CIP2	0.90	0.89	<0.01, 743.4	0.01, 632.8	0.03, 684.8	0.10, 1067
	0.95	0.94	0.01, 264.6	0.02, 288.1	0.05, 301.2	0.15, 674.1
	0.99	0.98	0.04, 50.2	0.07, 55.9	0.14, 54.26	0.37, 191.8

policy describes how to choose δ when the coverage behaves as in Figure 3 and Ek^* is increasing as δ decreases. For CIP2, we note that the particular values generated depend on our choice of $\mu = 5$ as relative precision rules use the sample mean to determine stopping. However, we still observe similar asymptotic effects as δ decreases.

POLICY 1. *Start with $\delta = \delta^o$. Decrease δ until $\eta^*(\eta^o, \delta, k_{\min}) \geq \eta^o - \epsilon$, where $\epsilon > 0$ is the acceptable loss in the coverage.*

Policy 1 is the approach implied by Chow and Robbins [1965] and the resulting asymptotic literature, but we do not recommend it because it fails to achieve a nominal coverage and encourages high sample sizes. The results in the next sections suggest that modifying η and $k_{\min}$ might be more effective.

5.2. Choosing a Confidence Coefficient

Singham and Schruben [2012] present coverage contours for evaluating this loss in the coverage exactly when the data are normally distributed. The main result is that for a given stopping rule and type of output data, the coverage can be estimated and optimal stopping parameters found. Coverage contours were introduced to optimize η^* and Ek^* over η and δ. A similar optimization to Equation (8) with η as an additional decision variable, is formulated as

$$\begin{aligned} \min_{\eta,\delta} \quad & Ek^*(\eta, \delta, k_{\min}) \\ \text{s.t.} \quad & \eta^*(\eta, \delta, k_{\min}) \geq \eta^o \\ & \delta \leq \delta^o, \end{aligned} \tag{9}$$

and the coverage is required to be η^o rather than $\eta^o - \epsilon$. Optimizing over the space of (η, δ) (minimizing Ek^* while achieving $\eta^* \geq \eta^o$) delivers a solution that η should be inflated to some value $\eta' > \eta^o$ while the stopping rule is in use, with the knowledge that the coverage will be η^o at stopping, resulting in Policy 2. The inflated value of η' required to achieve the coverage for values of δ that result in the worst coverage appears insensitive to the value of σ^2 for normally distributed data. Thus, if the data are normally distributed and t-confidence intervals are used, Table 1 in Singham and Schruben [2012] gives values of η' that will achieve a coverage of at least η^o (for any value of δ). In general, it is relatively cheap to increase the value of η used rather than use a smaller δ.

POLICY 2. *Start with $\eta = \eta^o$. Increase η until $\eta^*(\eta, \delta, k_{\min}) \geq \eta^o$.*

5.3. Adjusting for the Sample Skewness

One of the most common potential problems with sequential stopping rules in CIPs is that the data are not independent or normally distributed. If the distribution can be estimated, then the coverage contours calculated over the parameter space can be used to choose an appropriately inflated value of η. If the half-width calculation

Table III. Coverage of Stopping Rules Using the Cornish-Fisher Adjustment for Skewness (CIP1 and CIP2)

		(0.90,0.30,2)		(0.95,0.15,2)	
	$(\eta, \delta, k_{\min})$	No adj	CF	No adj	CF
CIP1	$\mathcal{N}(0, 1)$	0.839	0.866	0.936	0.947
	Exp(1)	0.697	0.743	0.891	0.916
	Gamma(2,2)	0.868	0.888	0.943	0.948
		(0.90,0.30,2)		(0.95,0.15,2)	
	$(\eta, \delta, k_{\min})$	No adj	CF	No adj	CF
CIP2	$\mathcal{N}(5, 1)$	0.981	0.982	0.938	0.943
	Exp(1)	0.826	0.837	0.932	0.936
	Gamma(2,2)	0.824	0.833	0.923	0.929

assumes normality, then η must be significantly inflated in order to compensate for nonnormality [Singham and Schruben 2012]. A different CIP that takes into account the distribution of the data can reduce the bias of the procedure. Section 4 describes how the coverage can be calculated for various distributions and stopping rules, so the effect of the mismatch between the data and assumptions of the rule can be calculated directly.

If issues of dependence and nonnormality can be resolved by modifying the CIP, then the bias induced by the sequential rule is the main bias remaining. However, in many cases, the nature of the distribution and dependence will be unknown before the experiment begins. One strategy, presented by Tafazzoli et al. [2011], is to estimate the dependence and deviation from normality as batches of data are collected and to adjust the t-value used for the confidence intervals accordingly. The authors are able to achieve an improved coverage for many scenarios involving nonnormal distributions by using a Cornish-Fisher adjustment suggested by Johnson [1978], and then using the von Neumann test for randomness to determine when batch means are approximately independent. We focus on the nonnormality adjustment in Policy 3.

POLICY 3. *Use an adjustment for nonnormality to compute the half-width as the sequential procedure progresses.*

We use the adjustment for the skewness of the data as implemented by Tafazzoli et al. [2011] as part of CIP1 applied to different data types. Table III lists the results for two stopping rules, $(0.90, 0.30, 2)$ and $(0.95, 0.15, 2)$, where the second rule requires intervals with a higher confidence coefficient and a smaller precision. We see that adjusting for skewness does improve the coverage for both stopping rules when the data have a skewed distribution, but not enough to entirely account for the bias in the stopping rule. The improvement from using the adjustment is better when the rules have a worse coverage, as in the rule $(0.90, 0.30, 2)$. Adjusting the CIP to account for nonnormality in the data can result in wider intervals even when the data are normal, as small samples of normal data may exhibit skewness. Table III also shows the effect of the skewness adjustment on relative precision rules. We see similar small increases using the skewness adjustment for the relative precision rule as for the absolute precision rule. We note that for $\mathcal{N}(5, 1)$, the stopping rule $(0.90, 0.30, 2)$ may have large values of $\delta \overline{X}_k$, leading to early stopping with wide intervals resulting in high coverage.

5.4. Increasing the Starting Sample Size

Perhaps one of the easiest ways to improve the coverage is by increasing the starting sample size. Many simulation experiments use a sequential rule automatically starting with more than two samples. The choice of this starting sample size often depends on the resources available. In this section, we describe how coverage profiles can be used

Table IV. Optimal $k_{\min}$ Values for Worst-Case Stopping Rules Using Increased Values of $k_{\min}$ for CIP1 and CIP2 Expectations are rounded to the nearest integer. $Ek^*(\eta')$ refers to the expected stopping time using optimal policies from Section 5.2.

		$\eta^o = 0.90$			$\eta^o = 0.95$			$\eta^o = 0.99$		
Rule	Distribution	$k^*_{\min}$	Ek^*	$Ek^*(\eta')$	$k^*_{\min}$	Ek^*	$Ek^*(\eta')$	$k^*_{\min}$	Ek^*	$Ek^*(\eta')$
CIP1	$\mathcal{N}(0, 1)$	12	19	24	14	22	29	16	27	33
	Exp(1)	22	32	64	22	33	72	24	38	88
	Gamma(2,2)	17	25	39	18	27	46	21	33	59
CIP2	$\mathcal{N}(5, 1)$	11	18	23	12	21	28	15	27	33
	Exp(1)	17	19	26	19	20	24	22	22	26
	Gamma(2,2)	14	18	24	18	21	32	21	22	28

to choose a starting sample size to obtain a better coverage. We also describe how this method can be cheaper than changing parameters η and δ. In general, a higher value of Ek^* implies a better coverage, but in a finite-sample environment, we want to keep Ek^* at reasonable values.

Recall that for a given stopping rule, $p(k)$ is the probability of stopping at k, $c(k)$ is the coverage of the rule given that it stops at k, and an overall coverage is calculated according to Equation (1). Because $c(k)$ is often less than nominal for small values of k, using a larger $k_{\min}$ will likely improve the coverage by avoiding stopping too early. Coverage profiles can provide some sense of what the starting sample size should be. If $p(k)$ suggests that there is a significant probability of stopping at a k where $c(k)$ is low, stopping rules should be designed to reduce $p(k)$ at that value of k.

POLICY 4. *Increase* $k_{\min}$ *until* $\eta^*(\eta, \delta, k_{\min}) \geq \eta^o$.

If η^* and Ek^* are increasing in $k_{\min}$, this policy will be optimal over possible values of $k_{\min}$. A heuristic is to increase $k_{\min}$ and recalculate $p(k)$ and $c(k)$ under the new stopping rule until the coverage is at the desired level. The formulation is the same as Equation (9), except the optimization occurs only over $k_{\min}$. For many distributions, as $k_{\min}$ approaches infinity, the coverage for CIP1 approaches 1 as the interval half-width δ remains fixed and the variance of $\overline{X}_k$ decreases with larger values of k. We experiment using the stopping rules that deliver the worst coverage for a given η (by choosing the value of δ that has the worst coverage according to Section 5.1). For these rules, we determine the smallest value of $k_{\min}$ required to deliver the coverage η^o. We call this value $k^*_{\min}$, and we also report the expected stopping time of the rule, which is compared to that using inflated values of η as described in Section 5.2. The results for both CIP1 and CIP2 are in Table IV. It appears that for the worst-case scenarios, it is cheaper to increase $k_{\min}$ than to inflate η. For CIP1 applied to nonnormal distributions, the difference between the Ek^* values of the two methods is particularly high. This difference is less extreme for CIP2, possibly because the mean and variance are correlated for the exponential and gamma distributions. A sample with an unusually high sample mean is likely to have a high sample variance, reducing the chance of early stopping under a relative precision rule. A comparison of CIP2 to CIP1 reveals that similar values of $k_{\min}$ are needed to avoid a loss in coverage for the worst-case choices of δ.

5.4.1. Lower Bound on the Coverage for Normally Distributed Data. Estimating the coverage for stopping rules using different values of $k_{\min}$ can be done numerically or using simulation. Either method can be time-consuming depending on the desired accuracy of the results. In this section, we describe a method for obtaining a lower bound on the coverage for values of $k_{\min} > 2$ using just the information from the estimation of $p(k)$ and $c(k)$ for $k_{\min} = 2$. This lower bound is exact for CIP1 applied to normally distributed data and is approximate for other distributions. Using these lower bound estimates for the coverage, we choose a value of $k_{\min}$ that will deliver at least a nominal coverage.

Table V. Values of $k^*_{\min}$ for Policy 5, and the Corresponding Ek^* and η^* Values
Expectations are rounded to the nearest integer.

	$\eta^o = 0.90$			$\eta^o = 0.95$			$\eta^o = 0.99$		
	$k^*_{\min}$	Ek^*	η^*	$k^*_{\min}$	Ek^*	η^*	$k^*_{\min}$	Ek^*	η^*
$\mathcal{N}(0,1)$	15	20	0.909	16	22	0.954	19	28	0.991

We denote $p_i(k)$ as the probability of stopping at k using a rule with $k_{\min} = i$, and $c_i(k)$ as the corresponding probability of covering given that stopping occurs at k. We assume that $c(k)$ is independent of the stopping rule and $p(k)$, which is true for CIP1 applied to normal data. This assumption generally does not hold, but as increasing the starting sample size generally increases the overall coverage for the types of rules we study, we expect that leaving $c(k)$ the same will be a conservative approximation of the coverage profile. Next, we construct an approximation for $p_i(k)$. Define $\hat{p}_i(k)$ as the modified distribution of $p_2(k)$ where we take the weight from all stopping times less than i and add it to the weight at i:

$$\hat{p}_i(k) = \begin{cases} p_2(k) & k > i \\ \sum_{j=2}^{i} p_2(j) & k = i \\ 0 & k < i. \end{cases}$$

Now assume that $c(k)$ is increasing in k. This is true for CIP1 applied to normally distributed data and often holds for other distributions as well. Then:

$$\sum_k p_2(k)\, c_2(k) \le \sum_k \hat{p}_i(k)\, c_2(k) \le \sum_k p_i(k)\, c_2(k). \tag{10}$$

The coverage using $\hat{p}_i(k)$ has at least the coverage of a rule starting with only two samples. Next, we show that $\sum_k \hat{p}_i(k) c_2(k) \le \sum_k p_i(k) c_2(k)$. Starting with i samples means that stopping can no longer occur at $k < i$, so the mass must be distributed along the points $k \ge i$. If $c(k)$ is increasing in k, the lowest coverage can be calculated by adding all the mass at i, which is how $\hat{p}_i(k)$ is constructed. Hence, the true coverage of starting at i is at least as good as that calculated using $\hat{p}_i(k)$.

POLICY 5. *Fixing η^o and δ^o, choose the smallest $k_{\min}$ value such that the lower bound on the coverage, $\sum_k \hat{p}_i(k)\, c_2(k)$, is at least η^o.*

Policy 5 will have a coverage that is at least nominal but also may require a higher starting sample size and expected number of replications than Policy 4, as seen in Table V. For CIP1 applied to normal data, this method will result in a quick lower bound on the coverage without needing to recalculate $c(k)$ and $p(k)$. Generally, $c(k)$ is increasing in k for k larger than some finite integer. For other distributions and relative-precision rules, $c(k)$ is not necessarily independent of the stopping rule, so it is possible that $c(k)$ will decrease when a larger $k_{\min}$ is used, though in general this is unlikely to happen. If $c(k)$ is decreasing in k, then early stopping is less of an issue.

A final option is to design stopping rules so that $c(k) \ge \eta^o$ for all k. This would imply that any interval would have the appropriate level of risk, preventing the option of early stopping leading to a particularly bad interval. This could involve increasing $k_{\min}$ to be large enough that enough samples will be generated to obtain a nominal coverage for any stopping time. If the user was willing to compromise on δ, a procedure could be designed that would inflate the interval depending on the stopping time. The main point is that the user should consider the interplay between $c(k)$ and $p(k)$ in designing stopping rules in order to determine where the low coverage is coming from and how it can be prevented.

The first policy of decreasing δ appears to be the most costly because it only achieves nominal coverage in the limit. Choosing a smaller value of δ pushes the density of the stopping time out to higher values where coverage is higher. The policy of increasing $k_{\min}$ also works in this way by reducing the probability of low coverage at early stopping. The policy of increasing η also involves reallocating the distribution of the stopping time to higher values and inflating values of $c(k)$ by using a higher confidence coefficient. The third policy of adjusting for sample skewness also leads to increased values of $c(k)$ by generating intervals that better match the skewness in the data. One could use a combination of these rules (by inflating the confidence coefficient, choosing a larger starting sample, and adjusting for skewness) to achieve improved results.

6. CONCLUSIONS

This article analyzes the parameters of a class of stopping rules for CIPs and gives a framework for assessing their quality. We exploit the tradeoffs between these parameters to provide simple rules that either reduce the risk of a poor coverage, or minimize the expected stopping time while meeting performance constraints. Optimal rules can be determined if certain assumptions are made about the data, but in reality the distribution of simulation output is usually not known. We describe general guidelines motivated by worst-case analyses to suggest conservative rules that can be used in the absence of concrete knowledge of stopping rule performance. Most of the methods work by increasing the expected number of replications to prevent early stopping, but we can modify the stopping time distribution in an efficient way to prevent overly large sample sizes.

Even if the data meet all the assumptions underlying a rule, there is usually a bias associated with stopping rules that leads to a reduction in the coverage and underestimation of risk for decision making. Graphical tools such as coverage functions, contours, and profiles are available to visualize the effect of a CIP on the coverage. Numerical techniques and simulation methods can be used to estimate CIP performance, but we recommend using simulation methods because they are faster and easy to implement.

We describe simple solutions for improving the coverage and results to guide users in their choice of stopping rule parameters. Adjustments to the confidence coefficient, precision parameter, and starting sample size can improve the coverage, and the CIP itself can be changed by taking into account the distribution of the underlying data. The simplest solutions we can offer are to increase the starting sample size as high as is computationally reasonable and to use an inflated value of η to determine when stopping occurs. These techniques seem to provide the greatest increase in the coverage relative to the increased computational cost. We note that for extremely expensive experiments, increasing the starting sample size may not be an option. The user may need to compromise on the confidence or precision in order to obtain an acceptable coverage with few samples.

Because optimal rules require assumptions on the data, we emphasize the importance of formulating conservative rules. In the absence of information about the distribution of the simulation output, the best advice we can give is to estimate stopping rule performance for ideal data and for extreme data. For example, if the CIP assumes normality, stopping rule performance should be tested on normal data to isolate the bias associated with the rule, but it should also be tested on a distribution such as the exponential to determine how poor the coverage might be if the distribution is different than anticipated. The solutions suggested here can be applied to these test distributions with the intent of deriving a conservative rule.

ACKNOWLEDGMENTS

The author is grateful to Lee Schruben and Dave Goldsman for many valuable discussions on this topic and to the associate editor and two anonymous referees for many useful comments that improved the quality and presentation of the article.

REFERENCES

E. J. Chen. 2012. A stopping rule using the quasi-independent sequence. *Journal of Simulation* 6, 2, 71–80.

Y. S. Chow and H. Robbins. 1965. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Annals of Mathematical Statistics* 36, 2, 457–462.

P. W. Glynn and W. Whitt. 1992. The asymptotic validity of sequential stopping rules for stochastic simulations. *Annals of Applied Probability* 2, 1, 180–198.

N. J. Johnson. 1978. Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association* 73, 363, 536–544.

K. Kang and B. Schmeiser. 1990. Graphical methods for evaluating and comparing confidence-interval procedures. *Operations Research* 38, 3, 546–553.

S. H. Kim and B. L. Nelson. 2001. A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on Modeling and Computer Simulation* 11, 3, 251–273.

A. M. Law and J. S. Carson. 1979. A sequential procedure for determining the length of a steady-state simulation. *Operations Research* 27, 5, 1011–1025.

A. M. Law and W. D. Kelton. 1982. Confidence intervals for steady-state simulations, II: A survey of sequential procedures. *Management Science* 28, 5, 550–562.

A. Nadas. 1969. An extension of a theorem of Chow and Robbins on sequential confidence intervals for the mean. *Annals of Mathematical Statistics* 40, 2, 667–671.

W. D. Ray. 1957. Sequential confidence intervals for the mean of a normal population with unknown variance. *Journal of the Royal Statistical Society. Series B (Methodological)* 19, 1, 133–143.

H. Robbins. 1959. Sequential estimation of the mean of a normal population. In U. Grenander (ed.), *Probability and Statistics; The Harald Cramér Volume.* Almquist & Wiksell and John Wiley and Sons, New York, 235–245.

R. G. Sargent, K. Kang, and D. Goldsman. 1992. An investigation of finite-sample behavior of confidence interval estimators. *Operations Research* 40, 5, 898–913.

B. Schmeiser and Y. Yeh. 2002. On choosing a single criterion for confidence-interval procedures. In *Proceedings of the 2002 Winter Simulation Conference*. IEEE, 345–352.

L. Schruben. 1980. A coverage function for interval estimators of simulation response. *Management Science* 26, 1, 18–27.

D. I. Singham and L. W. Schruben. 2009. Analysis of sequential stopping rules. In *Proceedings of the 2009 Winter Simulation Conference.* IEEE, 724–730.

D. I. Singham and L.W. Schruben. 2012. Finite-sample performance of absolute precision stopping rules. *INFORMS Journal on Computing* 24, 4, 624–635.

W. T. Song and B. W. Schmeiser. 2009. Omitting meaningless digits in point estimates: The probability guarantee of leading-digit rules. *Operations Research* 57, 1, 109–117.

A. Tafazzoli, N. M. Steiger, and J. R. Wilson. 2011. N-Skart: A nonsequential skewness-and autoregression-adjusted batch-means procedure for simulation analysis. *IEEE Transactions on Automatic Control* 56, 1, 1–11.

B. P. Welford. 1962. Note on a method for calculating corrected sums of squares and products. *Technometrics* 4, 3, 419–420.

Received August 2012; revised July 2013, January 2014; accepted May 2014