General-purpose ranking and selection for computer simulation

Soonhui Lee\textsuperscript{a} and Barry L. Nelson\textsuperscript{b}

\textsuperscript{a}College of Business, Hankuk University of Foreign Studies, Seoul, Republic of Korea; \textsuperscript{b}Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL USA

ABSTRACT

Many indifference-zone Ranking-and-Selection (R&S) procedures have been invented for choosing the best simulated system. To obtain the desired Probability of Correct Selection (PCS), existing procedures exploit knowledge about the particular combination of system performance measure (e.g., mean, probability, variance, quantile) and assumed output distribution (e.g., normal, exponential, Poisson). In this article, we take a step toward general-purpose R&S procedures that work for many types of performance measures and output distributions, including situations where different simulated alternatives have entirely different output distribution families. There are only two versions of our procedure: with and without the use of common random numbers. To obtain the required PCS we exploit intense computation via bootstrapping, and to mitigate the computational effort we create an adaptive sample-allocation scheme that guides the procedure to quickly reach the necessary sample size. We establish the asymptotic PCS of these procedures under very mild conditions and provide a finite-sample empirical evaluation of them as well.

1. Introduction

Although invented in the 1950s and 1960s for application to biostatistics problems, the statistical methods of Ranking and Selection (R&S) have been embraced by the stochastic simulation community as a standard tool for selecting the best of a finite (and relatively small) number of alternative system designs. In fact, R&S procedures are featured in several commercial simulation products. A common characteristic of selection-of-the-best procedures that have been extensively used in simulation is that “best” is defined to be the smallest or largest mean performance. Furthermore, whether Bayesian or frequentist in philosophy, these procedures typically assume that the simulation output data are normally distributed; even procedures that are shown to be asymptotically valid under more general assumptions are derived based on normality. Later we provide a realistic illustration where the output data are highly non-normal.

There is also a literature on R&S problems that differ from the normal-mean case. For instance, there are procedures that define “best” to be the largest or smallest probability, variance, or $q$th quantile. Also, there are procedures designed for output data that are known to be non-normal, including Poisson, Bernoulli, and exponential. The procedures for these situations are customized for the particular performance measure or type of data, exploiting the mathematical–statistical properties of the relevant estimator or distribution.

Suppose, however, that the alternatives under consideration involve distinct technologies: manual versus automated; in-house versus outsourced; or synthetic versus biological. The simulation output data may not only be non-normal but also there is no a priori reason to believe that the alternatives even generate output data from the same distribution family. The R&S literature provides no direct approach for such settings, other than batching the output data to achieve normality.

In this article, we take a big step toward general-purpose R&S, by which we mean procedures that work for many types of performance measures (e.g., means, probabilities, variances, or quantiles) and types of data (discrete- or continuous-valued and almost arbitrary distributions), and not all systems need to have the same output distribution family. We exploit intense computation, via bootstrapping, instead of clever mathematical analysis. To do this we employ a connection between fixed-width Confidence Intervals (CIs) and the desired Probability of Correct Selection (PCS). Our approach is sequential and frequentist in philosophy and incorporates an indifference zone. Finally, our procedures are fixed precision (as measured by the PCS) not fixed budget. Therefore, our first concern is to be able to deliver the desired PCS on virtually any problem, and a secondary (but still important) concern is being able to do so in an efficient manner. This means that we simulate until a specified PCS is achieved and stop, rather than trying to smartly allocate a fixed budget to achieve the best PCS we can attain with it.

As we substitute computation for analysis, our generic procedure will not be competitive when simulation output data are so computationally cheap that we can simulate each alternative system until its point estimator has effectively a zero variance. We will also not beat procedures that directly exploit (correct) distributional information; for instance, if we know that our output data are really Poisson, then we expect that a procedure based on that knowledge should be more efficient than ours, although in...
our empirical studies we are surprisingly close (Lee and Nelson, 2014). On the other hand, we make only very mild assumptions about the output data, and there are only two versions of our procedure: with or without Common Random Numbers (CRNs). When CRNs are employed we require no conservative inequalities to establish the validity of our procedure. The PCS guarantees are proven asymptotically, but because they bootstrap the actual simulation output data our procedures work well in finite samples across a variety of situations.

Preliminary versions of the procedures described in this article and a small empirical evaluation of them were presented in Lee and Nelson (2014). Here we prove their asymptotic validity and supplement the empirical results in Lee and Nelson (2014) with a more carefully chosen set of experiments. Also, we introduce a significant computational speed-up that allows our methods to be applied very rapidly to small problems (say 10 or fewer alternatives) and to be computationally feasible for larger problems.

The article is organized as follows. In Section 2 we describe the related literature that supports our work. Sections 3 and 4 present our R&S procedures and prove their asymptotic validity. A method for significantly reducing the number of sequential steps required to reach a correct selection is introduced in Section 5. Sections 6 and 7 contain experiment results. Conclusions are offered in the final section.

2. Background

Comprehensive treatments of R&S outside of computer simulation can be found in Gupta and Panchapakesan (1979) and Bechhofer et al. (1995). Simulation-focused surveys are provided by Kim and Nelson (2006b) and Kim (2013). Although this background is relevant, our approach to creating an R&S procedure is different in that we exploit a connection between indifference-zone δ selection of the best and fixed-width-δ CIs.

Let $X_{ij}$ represent the $j$th observed output of system $i$, for $i = 1, 2, \ldots, k$, so that $X_i = (X_{i1}, X_{i2}, \ldots, X_{ik})^\top$ is a $k \times 1$ vector representing the $i$th observed output across all systems. Throughout the article we assume that $X_{i1}, X_{i2}, \ldots$ are independent and identically distributed (i.i.d.) with marginal distribution $F_i(x) = Pr[X_{ij} \leq x]$. When we employ CRNs it will be useful to think of $X_1, X_2, \ldots$ as i.i.d. with common joint distribution function $F(x) = Pr[X_{1j} \leq x_1, \ldots, X_{kj} \leq x_k]$, $x = (x_1, \ldots, x_k)^\top \in \mathbb{R}^k$. We neither assume nor fit any specific distribution to the simulation output. We use boldface type as in $X_i$ to indicate a vector of observations across all $k$ systems and underlining as in $X_{nb}$ to indicate a vector of $n$ observations from system $i$. $\hat{\theta}$ is an estimator, and we append $\star$ to quantities defined by bootstrapping.

Let $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)^\top$ be a vector whose $i$th element is a statistical property of the marginal distribution $F_i$, such as its mean, variance, a quantile, or a probability. We are interested in finding the sample size that allows us to select the system with the largest value of $\theta_i$ with a specified PCS by choosing the one with the largest empirical estimate $\hat{\theta}_i$ of $\Theta$. Notice that the objective is finding an (ideally small) sample size that satisfies the PCS constraint.

For $k \geq 2$ systems, suppose that we can build fixed-width-δ CIs for all $\theta_i - \theta_j$, $i \neq j$ with simultaneous coverage $1 - \alpha$. As shown in Hsu (1996), if we have

$$Pr\{\hat{\theta}_i - \hat{\theta}_j - (\theta_i - \theta_j) \leq \delta, \forall i \neq j\} \geq 1 - \alpha$$

then with probability greater than or equal to $1 - \alpha$

$$\theta_i - \max_{j \neq i} \theta_j \in \left[\hat{\theta}_i - \max_{j \neq i} \hat{\theta}_j - \delta, \hat{\theta}_i - \max_{j \neq i} \hat{\theta}_j + \delta\right] \tag{1}$$

for $i = 1, 2, \ldots, k$. If $M$ is the index of the system with the largest performance estimate—i.e., $\hat{\theta}_M \geq \hat{\theta}_i$ for all $i \neq M$—then it follows from Equation (1) that with probability at least $1 - \alpha$,

$$\theta_M - \max_{j \neq M} \theta_j \geq \hat{\theta}_M - \max_{j \neq M} \hat{\theta}_j - \delta \geq -\delta$$

since $\hat{\theta}_M - \max_{j \neq M} \hat{\theta}_j \geq 0$. This result implies that if we select the system with the largest performance estimate $\hat{\theta}_M$ as the best system, the selected system will be the best system or a system within $\delta$ of the best system with a probability of at least $1 - \alpha$. Moreover, if the difference between the largest and the second-largest parameter value is strictly greater than $\delta$, then the selected system is the best system with a probability of at least $1 - \alpha$. This is exactly the desired inference for an indifference-zone R&S procedure. In R&S, δ is a user-specified parameter that corresponds to the smallest practically significant difference worth detecting. Thus, if we have a procedure to create fixed-width-δ CIs for $\theta_i - \theta_j$ with overall coverage $\geq 1 - \alpha$, then we also have a selection-of-the-best procedure.

Swanepoel et al. (1983) described a sequential bootstrapping procedure for generating a single fixed-width CI with a specified coverage probability $1 - \alpha$ when $\theta$ is either the mean or median. Given an i.i.d. sample of size $n$, denoted $X_n = (X_1, X_2, \ldots, X_n)$, from a population with marginal distribution $F$ having a distribution property $\theta$, let $\hat{F}_n$ denote the empirical cumulative distribution function (e.c.d.f.) of $X_n$ defined as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} I[X_j \leq x].$$

Let $\hat{\theta}_n$ be the corresponding distributional property of $\hat{F}_n$. Furthermore, let $X_n^* = (X_1^*, X_2^*, \ldots, X_n^*)$ denote a random sample of size $n$ from $\hat{F}_n$, $\hat{F}_n^*$ the implied e.c.d.f. and $\hat{\theta}(X_n^*)$ (also denoted by $\hat{\theta}_n^*$) the corresponding distributional property of $\hat{F}_n^*$. The bootstrap estimator of the probability that $\theta$ is contained within the interval $[\hat{\theta}_n - \delta, \hat{\theta}_n + \delta]$ is

$$P_n^* = Pr\{\hat{\theta}_n \in [\hat{\theta}_n^* - \delta, \hat{\theta}_n^* + \delta]\}. \tag{2}$$

Exact computation of $P_n^*$ is often difficult, but Equation (2) can be estimated given $B$ random samples of size $n$ from $\hat{F}_n$, say, $X_{nb}^* = (X_{1b}^*, X_{2b}^*, \ldots, X_{nb}^*)$, $b = 1, 2, \ldots, B$, by using

$$P_{nb}^* = \frac{1}{B} \sum_{b=1}^{B} I[\hat{\theta}_{nb} \in [\hat{\theta}_{nb}^* - \delta, \hat{\theta}_{nb}^* + \delta]], \tag{3}$$

where $\hat{\theta}_{nb}^*, b = 1, 2, \ldots, B$, is the estimate of the distributional property of interest from the $b$th bootstrap sample.

In their procedure, Swanepoel et al. (1983) sequentially increased the number of observations of $X$ until the stopping time $N^* = \inf\{n \geq n_0 : P_n^* \geq 1 - \alpha\}$, when the desired bootstrap coverage probability is $1 - \alpha$. The asymptotic properties
of $N^*$ were shown when $\theta$ is the mean or median of $X$, as stated in the following theorem.

Theorem of Swanepoel et al. (1983). Under some mild assumptions, as $\delta \downarrow 0$:  
(a) $\hat{\delta} N^* \to c$ a.s.
(b) $\Pr \{ |\theta_{N^*} - \theta| \leq \delta \} \to 1 - \alpha$.

The limit $c$ depends on the distributional property of interest: $c = E(X - \theta)^2 z_{1-\alpha/2}$ when $\theta$ is the mean and $c = (4\theta)^2 z_{1-\alpha/2}$ when $\theta$ is the median, where $f$ is the density function of $X$ and $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

Building from this foundation, we establish similar, but significantly more difficult, asymptotic results for fixed-width-$\delta$ simultaneous CIs for $k(k-1)/2$ pairs of differences $\theta_i - \theta_j$. This, in turn, provides an indifference-zone R&S procedure. Furthermore, we propose a method to more quickly jump to the stopping time $N^*$, avoiding many costly evaluations of $P_{N^*}$.

There are, of course, parametric (in particular, normal-theory) methods to construct fixed-width CIs. See, for instance, Hochberg and Tamhane (1987). More closely related to this research, Aerts and Gijbels (1993) and Háliková (2003) proposed three-stage procedures that use bootstrapping to estimate the critical constant that determines sample size (e.g., $z$-value in the normal-theory case): the first stage provides a crude estimate; the second stage refines it; and the third stage jumps to the stopping time $N^*$. These methods do not directly estimate the PCS but, more important, seem difficult to generalize beyond a single CI.

3. Procedures

In this section we describe algorithms for performing R&S for $k \geq 2$ systems using the bootstrap-based fixed-width CI approach. We present two versions of the algorithm, one that exploits CRNs and one without CRNs. An algorithm that does not use CRNs was presented in Bekki et al. (2010) when the sample size was incremented one at a time. We generalize the algorithm here to allow $\Delta n \geq 1$ additional observations on each iteration; this has the effect of speeding up the algorithm at the possible cost of taking more observations than necessary to guarantee a correct selection. Later we present a method for adaptively choosing $\Delta n$ that provides a substantial computational speed-up without noticeable overshoot of PCS.

First, we describe the procedure without using CRNs. Let $X_{ni} = \{X_{i1}, X_{i2}, \ldots, X_{in} \}$ be a sample of size $n$ from a system with output distribution $F_i$ having distribution property $\theta_i$ and $\hat{F}_i$ the e.c.d.f. based on $X_{ni}$ for system $i = 1, 2, \ldots, k$. Let $\hat{\theta}(X_{ni})$ be an estimate of $\theta_i$ based on $X_{ni}$ for $i = 1, 2, \ldots, k$ and $\hat{\theta}_{ij}(X_{ni}) = \hat{\theta}(X_{ni}) - \hat{\theta}(X_{nj})$ for all $i \neq j$. We want to build simultaneous fixed-width-$\delta$ confidence intervals for all pairs of differences $\theta_i - \theta_j$ for $i \neq j$ by finding $n$ such that

$$\Pr \left\{ |\hat{\theta}_{ij}(X_{ni}) - \delta, \hat{\theta}_{ij}(X_{ni}) + \delta| \right\} \geq 1 - \alpha, \quad (4)$$

where $\hat{\theta}_{ij} = \hat{\theta}_i - \hat{\theta}_j$. The value of $n$ will be the smallest one for which the estimated coverage probability using bootstrapping is at least $1 - \alpha$. Specifically, given $B$ random samples of size $N$ from $F_{N^*}$, $X_{nj} = \{X_{nj1}, X_{nj2}, \ldots, X_{njB} \}$, $b = 1, 2, \ldots, B$, the bootstrap coverage probability is estimated by

$$P_{N^*} = \frac{1}{B} \sum_{b=1}^{B} \prod_{(i,j) \neq (j,k)} \left\{ |\hat{\theta}_{ij}(X_{njb}) - \delta, \hat{\theta}_{ij}(X_{njb}) + \delta| \right\}, \quad (5)$$

where $\hat{\theta}(X_{njb})$ is an estimate of $\theta(X_{njb})$ based on $X_{njb}$ and $\hat{\theta}_{ij}(X_{njb}) = \hat{\theta}(X_{njb}) - \hat{\theta}(X_{nb})$ for all $i \neq j$. The procedure without CRNs described below starts with a sample of size $N = n_0$ from each system $i = 1, 2, \ldots, k$, a desired PCS $1 - \alpha$, a half width (indifference-zone parameter) $\delta$ for the CIs, and a sample-size increment $\Delta n$.

**Bootstrap R&S procedure without CRN**

1. Specify $N = n_0$, set $1/k < 1 - \alpha < 1$, $\delta > 0$, and $\Delta n \geq 1$.
2. Obtain $X_{ni} = \{X_{i1}, X_{i2}, \ldots, X_{in} \}$ a sample of size $N$ from the distribution $F_i$ for $i = 1, 2, \ldots, k$.
3. Compute $\hat{\theta}_{ij}(X_{ni}) = \hat{\theta}(X_{ni}) - \hat{\theta}(X_{nj})$ for all $i \neq j$ where $\hat{\theta}_i$ is a distributional property of $F_i$ and $\hat{\theta}(X_{ni})$ is an estimate of $\theta_i$ based on $X_{ni}$ and form the e.c.d.f. $\hat{F}_i$ for system $i = 1, 2, \ldots, k$.
4. Obtain $B$ bootstrap samples of size $N$ from $\hat{F}_i$ as $X_{nj} = \{X_{nj1}, X_{nj2}, \ldots, X_{njB} \}$ $i = 1, 2, \ldots, k$.
5. Compute $\hat{\theta}_{ij}(X_{njb}) = \hat{\theta}(X_{njb}) - \hat{\theta}(X_{nb})$, $b = 1, 2, \ldots, B$ for all $i \neq j$.
6. Estimate the PCS as

$$P_{N^*} = \frac{1}{B} \sum_{b=1}^{B} \prod_{(i,j) \neq (j,k)} \left\{ |\hat{\theta}_{ij}(X_{njb}) - \hat{\theta}_{ij}(X_{njb})| \leq \delta \right\}.$$
3. Compute \( \hat{\theta}_j(X_N) = \hat{\theta}(X_{N_j}) - \hat{\theta}(X_{N_i}) \) for all \( i \neq j \) where \( \hat{\theta} \) is a distributional property of \( F \) and \( \hat{\theta}(X_{N_j}) \) is an estimate of \( \theta \) based on \( X_{N_j} \) and form the e.c.d.f. \( \hat{F}_N \) based on \( X_N = \{X_1, X_2, \ldots, X_n\} \) as

\[
\hat{F}_N(x) = \frac{1}{N} \sum_{j=1}^{N} I(X_{ij} \leq x),
\]

where \( I \) is the indicator function.

4. Obtain \( B \) bootstrap samples of size \( N \) from \( \hat{F}_N : \{X_{1b}, X_{2b}, \ldots, X_{nb}\} \) for \( b = 1, 2, \ldots, B \), where \( X_{jb} = (X_{1jb}, X_{2jb}, \ldots, X_{nkjb})^\top \) for \( j = 1, 2, \ldots, N \).

5. If \( \hat{F}_{N_b} \geq 1 - \alpha \), report \( \max_{a=1,\ldots,k} \hat{\theta}(X_{N_N}) \) as the best system.

Else

Obtain \( X_{An} = \{X_i, \ j = 1, 2, \ldots, \Delta n\} \) a sample of size \( \Delta n \) from the distribution \( F \).

Set \( X_N = X_N \cup X_{An} \) and \( N = N + \Delta n \).

Go to Step 3.

End If

Remark 1: Although our algorithms state that we form the e.c.d.f.s of the output data, we do not need to actually create the e.c.d.f.s since bootstrapping simply requires random sampling from the data with replacement. However, the e.c.d.f.s are needed for the proofs presented in Section 4.

4. Asymptotic analysis

This section provides theoretical support for the R&S procedures introduced in the previous section. Recall that our goal is to achieve the desired PCS on virtually any problem, taking only as many observations as needed. For an indifference-zone R&S procedure, the most difficult case is when the true differences (which are unknown in practice) are small, and we demand to be able to detect small differences. Recall that \( \delta \) is a fixed quantity that represents the smallest difference that is practically important to the user. Therefore, we employ an asymptotic regime in which we let the indifference-zone parameter \( \delta \) go to 0. This drives the required sample size \( N^* \) (obtained from the procedures) to infinity and, hence, we show that the procedures achieve the desired PCS asymptotically.

The theorems stated below and proved in the Appendix extend Swanepoel et al. (1983) from a CI for a univariate mean or median to simultaneous CIs for the means or any quantile from \( k \geq 2 \) systems. We then show in the corollaries that these results justify our use of bootstrap R&S for either mean or quantile performance measures by providing simultaneous CIs for all pairwise differences.

We first review the key notation. Let \( X_n = \{X_1, X_2, \ldots, X_n\} \) be a random sample of size \( n \) from distribution \( F \) in \( \mathbb{R}^k \) with a \( k \times 1 \) vector of marginal distribution properties \( \Theta \), where \( X_j = (X_{1j}, X_{2j}, \ldots, X_{kj})^\top, \ j = 1, 2, \ldots, n \). Furthermore, let \( \hat{F}_n(x) \) be the e.c.d.f. based on \( X_n \) defined in two different ways for use in the procedure without CRNs, as in Equation (6), and with CRNs, as in Equation (7):

\[
\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} I(X_{ij} \leq x),
\]

\[
\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} I(X_{ij} \leq x) - \frac{1}{n} \sum_{j=1}^{n} I(X_{ij} \leq x) - \frac{1}{n} \sum_{j=1}^{n} I(X_{ij} \leq x),
\]

where \( x = (x_1, x_2, \ldots, x_k)^\top \in \mathbb{R}^k \). The results below are valid in either case.

Let \( X_n^* = \{X_{1n}, X_{2n}, \ldots, X_{nn}\} \) denote a random sample of size \( n \) from \( \hat{F}_n \). Let \( P_r \) and \( P^r \) denote probabilities under \( F \) and \( \hat{F}_n \) and \( E \) and \( E^* \) denote expectations under \( F \) and \( \hat{F}_n \), respectively.

Probabilities and expectations under * are equivalent to estimating them using \( B \rightarrow \infty \) bootstrap samples.

The bootstrap stopping variable \( N^* \) is given by

\[
N^* = \inf \{n \geq n_0 : \Pr \{\sum_{j=1}^{N} \hat{\Theta}(X_{N_j}) - \hat{\Theta}(X_{N_i}) \leq \delta \times I_k \} \geq 1 - \alpha \},
\]

Equation (8) grows as \( \delta \downarrow 0 \), as in Swanepoel et al. (1983).

Before we state our theorems, we introduce the following definitions:

**Definition 1.**

(a) For any \( c > 0 \) and positive-definite covariance matrix \( \Sigma \), let \( \Gamma_\Sigma : \mathbb{R}^+ \mapsto (0, 1) \) be

\[
\Gamma_\Sigma(c) = \int_{[c, \infty]} (2\pi)^{-k/2} |\Sigma|^{-1/2} e^{-y^\top \Sigma^{-1} y/2} dy.
\]

(b) For \( \eta \in (0, 1) \), let \( a_\eta = \Gamma^{-1}_\Sigma(\eta) \); that is,

\[
\Gamma_{\Sigma}(a_\eta) = \eta.
\]

From here on, whenever we raise a vector or matrix to a power, as in \( X^2 \), we mean element-by-element exponentiation. Matrix multiplications are written explicitly, as in \( X^T X \).

**Theorem 1.** Let \( \Theta = E[X] \). Suppose that \( E[|X - \Theta|^3] < \infty \) and that \( \Sigma = E[(X - \Theta)(X - \Theta)^\top] \) is a positive definite matrix. Consider \( N^* \) as defined in Equation (8).

(a) As \( \delta \downarrow 0 \), we have

\[
\delta^2 N^* \to a_1^{2 - \alpha} \quad \text{a.s.}
\]

where \( a_1^{2 - \alpha} = \Gamma_{\Sigma}^{-1}(1 - \alpha) \).

(b) As \( \delta \downarrow 0 \), we have

\[
\Pr \left\{ \sum_{j=1}^{N} \hat{\Theta}(X_{N_j}) - \hat{\Theta}(X_{N_i}) \leq \delta \times I_k \right\} \to 1 - \alpha.
\]

**Proof.** See Appendix A2. \( \square \)

Notice that the assumptions are very weak when \( \Theta = E[X] \) and the statistic is a sample average of i.i.d. outputs. Suppose next that \( \Theta \) is the set of variances of the \( k \) marginal distributions, where the \( i \)th element is \( \theta_i = E(X_i - \mu_i)^2, i = 1, 2, \ldots, k \). Unless \( \mu_i \) is known, \( \theta_i \) is not the expected value of the average of i.i.d. observations. Instead, \( \hat{\Theta}(X_{N_i}) \) and \( \hat{\Theta}(X_{N_i})^* \) are the sample variances based on \( X_{N_i} \) and \( X_{N_i}^* \), respectively, where the \( i \)th element of \( \hat{\Theta}(X_{N_i}) \) (denoted by \( S_{iN_i}^2 \)) is the sample variance of
$X_1, X_2, \ldots, X_m$ and the $i$th element of $\hat{\Theta}(X_n)$ (denoted by $S_n^i$) is the sample variance of $X_{1i}, X_{2i}, \ldots, X_{mi}$; that is,

$$S_n^i = \frac{\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2}{n - 1} \quad \text{and} \quad S_n^2 = \frac{\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2}{n - 1}.$$ 

Nevertheless, we can establish analogous asymptotic results for the sample variance in a manner very similar to Theorem 1, so we omit the proof.

**Corollary 1.** Let $\Theta = [\text{Var}(X_1), \text{Var}(X_2), \ldots, \text{Var}(X_k)]^T$. Suppose $E[|X - E(X)|^4] < \infty$, and let $N^*$ be as defined in Equation (8).

(a) As $\delta \downarrow 0$, we have

$$\delta^2 N^* \rightarrow a_{1-\alpha}^2 \quad \text{a.s.}$$

where $a_{1-\alpha} = \Gamma^{-1}_\Sigma(1 - \alpha)$ with covariance matrix

$$\Sigma = \text{E}\left[\left((X - E(X))^2 - \Theta\right)(X - E(X))^2 - \Theta\right].$$

(b) As $\delta \downarrow 0$, we have

$$\Pr\left(|\hat{\Theta}(X_n^*) - \Theta| \leq \delta \times 1_k\right) \rightarrow 1 - \alpha.$$

Next let $\Theta$ be a set of specific quantiles of the $k$ marginal distributions where the $i$th element is

$$\theta_i = \hat{F}_i^{-1}(q) = \inf\{x : \hat{F}_i(x) \geq q\}, \quad 0 < q < 1, \quad i = 1, 2, \ldots, k.$$ 

Then $\hat{\Theta}(X_n)$ and $\hat{\Theta}(X_n^*)$ are the sample $q$th quantiles based on $X_n$ and $X_n^*$, respectively, where the $i$th element of $\hat{\Theta}(X_n)$ is the sample $q$th quantile of $X_{1i}, X_{2i}, \ldots, X_{mi}$ and the $i$th element of $\hat{\Theta}(X_n^*)$ is the sample $q$th quantile of $X_{1i}, X_{2i}^*, \ldots, X_{mi}^*$. Stronger assumptions on the distributions $F_i$ are required to obtain similar asymptotic results for quantiles.

**Theorem 2.** Let $F_i$ be twice continuously differentiable in a neighborhood of $\theta_i$ and $\xi_i = f_i(\theta_i) > 0$, for $i = 1, 2, \ldots, k$, where $f_i$ is the density associated with $F_i$. Furthermore, let $F_i$ be $(i, j)$th bivariate marginal distribution function. Consider $N^*$ as defined in Equation (8). Suppose that the covariance matrix

$$\Sigma = \begin{pmatrix}
q(1 - q) & \sigma_{12} & \sigma_{1k} \\
\xi_1^2 & \xi_1 \xi_2 & \cdots & \xi_1 \xi_k \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{k1} & \sigma_{k2} & \cdots & q(1 - q) \\
\xi_k \xi_1 & \xi_k \xi_2 & \cdots & \xi_k^2
\end{pmatrix}$$

with

$$\sigma_{ij} = F_{ij}(\theta_i, \theta_j) - q^2$$

is positive definite.

(a) As $\delta \downarrow 0$, we have

$$\delta^2 N^* \rightarrow a_{1-\alpha}^2 \quad \text{a.s.}$$

where $a_{1-\alpha} = \Gamma^{-1}_\Sigma(1 - \alpha)$.

(b) As $\delta \downarrow 0$, we have

$$\Pr\left(|\hat{\Theta}(X_n^*) - \Theta| \leq \delta \times 1_k\right) \rightarrow 1 - \alpha.$$ 

**Proof.** See Appendix A3.

The results above justify simultaneous fixed-width-$\delta$ CIs for $k$ means or quantiles based on bootstrapping. They establish the asymptotic validity of our generic R&S procedure when we extend them to all pairs of difference estimates using the linear transformation $A$ defined as

$$A = [a_{ij}], \quad i = 1, 2, \ldots, \left(\begin{array}{c}k \\ 2 \end{array}\right); \quad j = 1, 2, \ldots, k,$$

where

$$a_{ij} = \begin{cases}
1, & (j - 1) \left(\frac{k - j}{2}\right) + 1 \leq i \leq j \left(\frac{k - j + 1}{2}\right); \\
-1, & i = h(k + 1) - (k - j); \\
0, & \text{otherwise.}
\end{cases}$$

The linear transformation $A$ transforms vectors in $\mathbb{R}^k$ into vectors that correspond to all pairs of differences in $\mathbb{R}^\left(\begin{array}{c}k \\ 2 \end{array}\right)$. We are now prepared to state the asymptotic validity of our generic R&S procedures in Corollaries 2 and 3. The stopping time used in our procedures is

$$N_A^* = \inf\left\{n \geq n_0 : \Pr^*\left(|A\hat{\Theta}(X_n^*) - A\Theta| \leq \delta \times 1_k\right) \geq 1 - \alpha\right\}.$$ 

**Corollary 2.** Under the same assumptions as in Theorem 1, consider $N_A^*$ as defined in Equation (10).

(a) As $\delta \downarrow 0$, we have

$$\delta^2 N_A^* \rightarrow a_{1-\alpha}^2 \quad \text{a.s.}$$

where $a_{1-\alpha} = \Gamma^{-1}_{A\Sigma A^T}(1 - \alpha)$ and $\Sigma$ is defined as in Theorem 1.

(b) As $\delta \downarrow 0$, we have

$$\Pr\left(|A\hat{\Theta}(X_n^*) - A\Theta| \leq \delta \times 1_k\right) \rightarrow 1 - \alpha.$$ 

**Proof.** See Appendix A4

Next let $\hat{\Theta}$ be the sample quantiles defined in Theorem 2.

**Corollary 3.** Under the same assumptions as in Theorem 2, consider $N_A^*$ as defined in Equation (10).

(a) As $\delta \downarrow 0$, we have

$$\delta^2 N_A^* \rightarrow a_{1-\alpha}^2 \quad \text{a.s.}$$

where $a_{1-\alpha} = \Gamma^{-1}_{A\Sigma A^T}(1 - \alpha)$ and $\Sigma$ is defined as in Theorem 2.

(b) As $\delta \downarrow 0$, we have

$$\Pr\left(|A\hat{\Theta}(X_n^*) - A\Theta| \leq \delta \times 1_k\right) \rightarrow 1 - \alpha.$$ 

**Proof.** See Appendix A5.
The assumptions that support our theorems are comparatively weaker than those that support other \( \delta \rightarrow 0 \) analysis in R&S. For instance, Robbins et al. (1968) assumed that output data from each system are independent and have common, finite variance. Dameedji et al. (1996) also assumed that output data from each system are independent and further that their variances are known, along with other conditions on the distribution mean, variance, and absolute centered moments. Similarly, Kim and Nelson (2006a) assumed that the output data from each system are mutually independent and that the standardized partial sum of each system’s output converges to a Brownian motion process.

**Remark 2**: Notice that means (which includes probabilities), variances, and quantiles cover much of what is used in practice. One possible extension of our asymptotic results is when the performance measure \( \theta \) is a function, like a cost, associated with (say) the mean performance measure \( \theta = g(\mu) \), and we can estimate \( \mu \). Then our procedure can still be shown to be asymptotically valid for continuously differentiable \( g \) using a delta-method argument.

### 5. Jump-ahead procedure

In the generic procedures described in Section 3, the number of observations \( N \) obtained from each simulated system is increased until the estimated PCS is at least \( 1 - \alpha \). The sample-size increment on each iteration, \( \Delta n \), can be made larger than one to speed-up the procedure, and this is important because estimating the PCS is a non-trivial calculation for large \( k \). In Lee and Nelson (2014) we suggested \( \Delta n = 10 \), but this could be too aggressive in some problems and unnecessarily conservative (small) when the stopping time \( N^* \) is in the hundreds or thousands. Recall that our primary goal is to provide a PCS guarantee, but we also want to attain it efficiently in the sense of keeping both the required sample size \( N^* \) and the computational overhead from bootstrapping small. Bootstrapping occurs after we collect each increment \( \Delta n \) of additional output data. When variances are unknown, the minimal number of stages or “jumps” that an indifference-zone R&S procedure can have is two, and this is only possible if we know the distributions of the outputs. Since we assume no distributional information, the best we can hope for is an approximation that keeps the number of jumps small.

Here we outline an adaptive method for selecting \( \Delta n \) that is aggressive when \( P^*_{NB} \) is far from \( 1 - \alpha \) but cautious as we approach it. The idea is straightforward: From any intermediate observed \( (N, P^*_{NB} < 1 - \alpha) \) pair, fit a simplified normal-theory approximation for PCS as a function of \( N \) and use it to project what the stopping time \( N^* \) will be; call this projection \( \hat{N}^* \). Then set \( \Delta n = c(\hat{N}^* - N) \), where \( c \) is a factor such as 0.8, to avoid overshoot. Now simulate the additional observations, estimate the new PCS and new projection, and repeat until \( P^*_{NB} \geq 1 - \alpha \).

To simplify notation, let \( \hat{\theta}_i = \hat{\theta}(X_{UN}) \) be the point estimator of \( \theta_i, i = 1, 2, \ldots, k \), where the sample size \( N \) will be clear from the context. Our approximation will assume that \( (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k) \) are jointly normally distributed with means \( (\theta_1, \theta_2, \ldots, \theta_k) \), common variance \( \sigma^2/N \), and common correlation \( \rho \) (which is zero unless CRNs are employed). We would like to approximate the sample size \( N^* \) for which

\[
\Pr(|\hat{\theta}_i - \hat{\theta}_j| \leq \delta, \forall i \neq j) \geq 1 - \alpha
\]

by fitting this simplified model to the observed PCS.

Let \( Z_{ij} = (\hat{\theta}_i - \hat{\theta}_j - (\theta_i - \theta_j)) \). Then \( Z = \{Z_{ij}, \forall i \neq j\} \) is appropriately organized and under our simplified model, has a multivariate normal distribution with mean vector 0 and variance–covariance matrix

\[
\frac{\sigma^2(1 - \rho)}{N} \mathbf{AIA}^\top,
\]

where \( \mathbf{A} \) is as defined in Equation (9) and \( \mathbf{I} \) is the \( k \times k \) identity matrix. Notice that only the term \( \sigma^2(1 - \rho) \) is unknown, since \( \mathbf{A} \) and \( \mathbf{I} \) depend only on \( k \), and \( N \) is given. For \( 0 < \beta < 1 \), let \( h(\beta) = \Gamma_{\mathbf{AIA}}^{-1}(\beta); h(\beta) \) is the \( \beta \) quantile of the maximum of the absolute values of the components of a mean-zero multivariate normal distribution with covariance matrix \( \mathbf{AIA}^\top \). Given \( \beta \), \( h(\beta) \) is also only a function of \( k \). Then for any such \( \beta \),

\[
\Pr\left(|Z_{ij}| \leq \frac{\sigma^2(1 - \rho)}{N} h(\beta), \forall i \neq j\right) = \beta.
\]

Suppose we have run our procedure for \( N \) observations and obtained estimated PCS \( P^*_{NB} < 1 - \alpha \). Recall that \( P^*_{NB} \) is also the estimated coverage probability; that is, for the current value of \( N \),

\[
\Pr(|Z_{ij}| \leq \delta, \forall i \neq j) \approx P^*_{NB}.
\]

We can calibrate our approximation to this observed probability by selecting \( \sigma^2(1 - \rho) \) so that

\[
\sqrt{\frac{\sigma^2(1 - \rho)}{N} h(P^*_{NB})} = \delta.
\]

Here we are treating the entire term \( \sigma^2(1 - \rho) \) as a single parameter to match the observed PCS and not worrying about separating \( \sigma^2 \) and \( \rho \).

Given this parameter, the target sample size to obtain PCS of at least \( 1 - \alpha \) satisfies

\[
\hat{N}^* \geq \left(\frac{\sqrt{\sigma^2(1 - \rho) \times h(1 - \alpha)}}{\delta}\right)^2 = N \times \left(\frac{h(1 - \alpha)}{h(P^*_{NB})}\right)^2
\]

and the \( \sigma^2(1 - \rho) \) terms cancel. The intuition is that the critical value \( h(1 - \alpha)^2 \) is the scale-free sample-size multiplier to attain PCS \( 1 - \alpha \). Thus, the ratio \( h(P^*_{NB})/h(1 - \alpha)^2 \) is the fraction of the curve we have climbed if the current sample size \( N \) corresponds to a PCS of only \( P^*_{NB} \).

Unfortunately, calculating \( h(\beta) = \Gamma_{\mathbf{AIA}}^{-1}(\beta) \) is a difficult root-finding problem for large \( k \), but we have a straightforward empirical solution: Let \( M = \max_{i \neq j} |Z_{ij}| \), where \( \{Z_{ij}, \forall i \neq j\} \) have a mean-zero multivariate normal distribution with variance \( \mathbf{AIA}^\top \). Then \( h(\beta) \) is the \( \beta \) quantile of \( M \), whose distribution depends only on \( k \). Therefore, from an i.i.d. Monte Carlo sample \( M_1, M_2, \ldots, M_k \), \( h(\beta) \) can be estimated by the order statistic \( M_{(\lceil \beta k \rceil)} \) and only one sample is needed for any fixed \( k \), as \( M_{(\lceil \beta k \rceil)} \) is a function of \( \beta \) once the data are sorted.
The jump-ahead approximation moves our procedure from theoretically interesting to practically useful because, as we illustrate in Section 6, it allows us to find \( N^* \) in a small number of jumps, thereby avoiding the computational overhead of repeatedly evaluating \( P^*_{NB} \), while not overshooting and being inefficient.

### 6. Empirical evaluation

In this section we present an empirical evaluation of the algorithms described in Section 3 combined with the jump-ahead procedure of the previous section. We consider \( k = 2, 4, 10, \) and 20 systems. For one set of results, the output data have a mix of exponential and normal marginal distributions, and the performance measures include the mean and the 0.8-quantile. For a second set of experiments the output data have negative binomial distributions and the performance measure is the mean.

In preliminary work, favorable empirical results for a single distribution family (normal or Poisson), \( \theta \) being the mean or various quantiles and \( k \) up to 10, were presented in Bekki et al. (2010) and Lee and Nelson (2014). Lee and Nelson (2014) also contains results when varying the initial sample size \( n_0 \), the indifference-zone parameter \( \delta \), and the number of bootstrap resamples \( B \); they also compared the efficiency of the algorithm with Poisson outputs to procedures designed specifically for Poisson data. In all cases the algorithms performed very well. Therefore, we focus here on the mixed distribution case, on a second discrete case (both highly skewed and less-skewed instances), larger numbers of systems \( k \), and the use of the jump-ahead procedure. We show below and in Appendix A6 that the algorithms achieve the desired PCS in five to ten jumps provided \( n_0 \) is large enough. They can be more or less efficient than the normal-theory procedure of Rinott (1978), depending on the initial sample size, but they retain their PCS guarantee for non-normal data, whereas Rinott’s procedure may not.

#### 6.1. Selecting the best mean

All results presented here are averaged over 100 macroreplications of the entire experiment. Tables 1 to 3 contain results for selecting the system with the largest mean when using CRNs with an induced correlation \( \rho = 0.9 \). We varied the initial sample size \( n_0 \), the number of systems \( k \), and the configuration of the means. The number of bootstrap samples was \( B = 200 \), the desired confidence level was \( 1 - \alpha = 0.95 \) and the indifference-zone parameter was \( \delta = 0.1 \) for all experiments. We present only CRN results, as we recommend using CRNs whenever possible.

### Table 1. Empirical results for means from 100 macroreplications for half-exponential and half-normal distributions in slippage configuration with CRNs (\( \rho = 0.9 \))

| \( k \) | \( n_0 \) | \( Ave \ N^* \) | PCS | Est. coverage | True coverage | Ave jumps |
|---|---|---|---|---|---|---|
| 2(n) | 100 | 7186 | 0.96 | 0.96 | 0.93 | 4.83 |
| 2(e) | 100 | 7696 | 0.95 | 0.96 | 0.94 | 4.86 |
| 4(n) | 100 | 8207 | 0.97 | 0.96 | 0.94 | 6.64 |
| 4(e) | 100 | 8447 | 0.99 | 0.96 | 0.95 | 6.54 |
| 10(n) | 100 | 9974 | 0.98 | 0.96 | 0.96 | 8.92 |
| 10(e) | 100 | 9994 | 0.95 | 0.96 | 0.92 | 8.56 |
| 20(n) | 100 | 11221 | 0.97 | 0.96 | 0.91 | 10.05 |
| 20(e) | 100 | 11437 | 1.00 | 0.96 | 0.96 | 10.13 |

Corresponding results without CRNs are presented in Tables A1 to A3, respectively, of Appendix A6, and we compare them to the results with CRN at the end of this section.

In Tables 1 and 2, the simulation outputs have a mix of distributions; specifically, half of the systems have a normally distributed output and the other half have an exponentially distributed output with the means either in a slippage configuration or a monotone-decreasing means configuration. The true mean vector in Table 1 is the slippage configuration \( \Theta = (5, 5, \ldots, 5, 5 + \delta)^\top \), so system \( k \) is the best. The true mean vector used in Table 2 has \( \theta_i = 5 + (i - 1) \delta \) for \( i = 1, 2, \ldots, k \), so system \( k \) is again the best system. In the monotone-decreasing means configuration, we alternated exponentially distributed systems and normally distributed systems; for example, when \( k = 4 \), systems 1 and 3 have normally distributed outputs, whereas systems 2 and 4 have exponentially distributed outputs; we then reverse the assignments. The normally distributed outputs have variance 1, whereas the exponentially distributed outputs have variance \( \delta^2 \). In the tables, “n” or “e” indicate that the best system has normally or exponentially distributed outputs, respectively.

#### Table 2. Empirical results for means from 100 macroreplications for half-exponential and half-normal distributions in the monotone-decreasing means configuration with CRNs (\( \rho = 0.9 \))

| \( k \) | \( n_0 \) | \( Ave \ N^* \) | PCS | Est. coverage | True coverage | Ave jumps |
|---|---|---|---|---|---|---|
| 4(n) | 100 | 8647 | 0.98 | 0.96 | 0.95 | 6.53 |
| 4(e) | 100 | 9079 | 1.00 | 0.96 | 0.97 | 6.55 |
| 10(n) | 100 | 12086 | 1.00 | 0.96 | 0.96 | 8.84 |
| 10(e) | 100 | 12575 | 1.00 | 0.96 | 0.93 | 8.98 |
| 20(n) | 100 | 17745 | 0.99 | 0.96 | 0.96 | 10.82 |
| 20(e) | 100 | 18421 | 1.00 | 0.96 | 0.95 | 10.80 |

In Table 3 negative binomial output data are used; there are two parameters, the target number of successes \( r \) and the probability of success \( p \); these imply that the mean is \( \theta = r(1 - p)/p \). We used \( p = 0.5 \) (therefore, \( \theta = r \)) for all experiments. The true mean vector used in Table 3 is the slippage configuration \( (\theta, \theta, \ldots, \theta, \theta + \delta)^\top \) with \( \theta = 1 \) or 10 implying that system \( k \) is the best system. The negative binomial distribution is highly skewed when \( \theta = 1 \), whereas it is less skewed when \( \theta = 10 \).

To induce correlation representing the effect of CRNs for the mixed-distribution case, we used the NORTA method described in Nelson (2013), inducing a common correlation of 0.9 between all pairs of systems. To represent CRNs for negative binomial (NB) distributions, we set \( X_i = Z + W_i \) where the \( W_i \)’s are independent NB(\( \beta_i, p \)) for \( i = 1, 2, \ldots, k \), and \( Z \) is NB(\( \beta_0, p \)). Then

#### Table 3. Empirical results for means from 100 macroreplications for negative binomial distributions in the slippage configuration with CRNs (\( \rho = 0.9 \))

| \( k \) | \( \theta \) | \( \beta_0 \) | \( Ave \ N^* \) | PCS | Est. coverage | True coverage | Ave jumps |
|---|---|---|---|---|---|---|---|
| 2 | 1 | 0.9 | 240 | 0.94 | 0.93 | 0.97 | 3.21 |
| 2 | 10 | 1 | 1680 | 0.99 | 0.96 | 0.96 | 4.27 |
| 4 | 1 | 0.9 | 336 | 0.93 | 0.96 | 0.93 | 4.28 |
| 4 | 10 | 9 | 2815 | 1.00 | 0.96 | 0.94 | 5.36 |
| 10 | 1 | 0.9 | 482 | 1.00 | 0.96 | 0.98 | 5.12 |
| 10 | 10 | 9 | 4144 | 1.00 | 0.96 | 0.95 | 6.27 |
| 20 | 1 | 0.9 | 590 | 0.99 | 0.96 | 0.97 | 5.48 |
| 20 | 10 | 9 | 5147 | 0.99 | 0.96 | 0.95 | 7.02 |
the correlation between \( X_i \) and \( X_j \) for \( i \neq j \) is

\[
\text{Corr}(X_i, X_j) = \frac{\beta_0}{\sqrt{\beta_0 + \beta_i} \sqrt{\beta_0 + \beta_j}}.
\]

With \( p = 0.5 \), when \( \theta_i = \theta_j = 1 \), we have \( \text{Corr}(X_i, X_j) = \beta_0 \); when \( \theta_i = \theta_j = 10 \), we have \( \text{Corr}(X_i, X_j) = \beta_0/10 \). In Table 3, \( \beta_0 = 0.9 \) is used to induce a common correlation 0.9 when \( \theta = 1 \); similarly, we set \( \beta_0 = 9 \) when \( \theta = 10 \).

The PCS is the fraction of the 100 macroreplications where a system whose mean is within \( \delta \) of the true best mean was selected. The estimated coverage probability is \( P_{NB}^* \) from step 6 of the algorithm, and the true coverage probability is the fraction of the macroreplications where the \( \binom{n}{k} \) CIs simultaneously cover all pairwise differences \( \theta_i - \theta_j \) for all \( i \neq j \). The average number of jumps that the jump-ahead procedure generates until \( P_{NB}^* \geq 1 - \alpha \) is denoted by “Ave jumps.”

We quickly discovered that \( P_{NB}^* \) is often 0 when \( N \) is small and \( k \) is large, such as when \( N = n_0 \) and \( k \geq 10 \), which provides no information to the jump-ahead procedure. Therefore, we adjusted the procedure to use \( \max(1/k, P_{NL}^*) \), since pure guessing provides PCS 1/\( k \). Figure 1 shows the sample sizes generated by the jump-ahead procedure with this adjustment for a case with \( k = 20 \) systems in the slippage configuration and having a mix of distribution types; the left-hand figure is the result of one macrorepliation, whereas the right-hand figure is the result of 10 macroreplications. Notice that the observed PCS values are zero initially, but after a couple of jumps the procedure is able to make rapid progress without significant overshoot. For the left-hand plot in Fig. 1, each dot denotes a pair of sample size \( N \) (starting from \( N = n_0 \)) and its corresponding estimated PCS \( P_{NB}^* \) at each jump; for this example, the procedure stops at the seventh jump as \( P_{NB}^* \geq 0.95 \), which results in \( N^* = 50824 \). Recall that in the best case with unknown variance R&S requires two jumps and that fully sequential procedure uses \( \Delta n = 1 \) (i.e., \( N^* \) jumps). We observe that five to ten jumps were made for \( c = 0.8 \) in all experiments; however, if the user wants to make the procedure less conservative (fewer jumps), the constant \( c \) can be set closer to one.

The results from all tables show that the number of jumps and the required terminal sample size \( N^* \) increase as the number of systems increases, as expected. The desired PCS is achieved in all but one case, and the desired CI coverage is attained except in a few cases when \( n_0 = 50 \). It seems clear that using CRNs implies that we should start with a larger initial sample size \( n_0 \). This makes sense, as when we employ CRNs we bootstrap entire vectors, rather than each marginal c.d.f. individually, and 50 vectors is a small number to represent a joint distribution, particularly when \( k \) is large. As we have observed in previous papers, \( n_0 \) should be larger for bootstrap R&S than for distribution-specific procedures, as we need the bootstrap distribution, even at \( n_0 \), to be a good representation of the true distribution.

Notice that when the best system has exponentially distributed outputs the required sample size \( N^* \) tends to be larger than when the best system has normally distributed outputs, due to the variances of the exponentially distributed outputs being much higher than the normal case. The required sample size \( N^* \) for the monotone-decreasing means configuration is larger than the slippage configurations for the same reason.

For both output distribution cases, the use of CRNs greatly reduces the required sample size, with savings of as much as 90%. As a specific instance, in the case of half-normal and half-exponential distributions when \( k = 20 \) and the best system is exponentially distributed, the required sample size obtained from the algorithm with CRNs is 11 437 (Table 1), whereas the required sample size without CRN is 50 257 (Table A1 in Appendix A6); when \( k = 2 \), the required sample size obtained from the algorithm with CRN is 7466 (Table 1), whereas the required sample size without CRNs is 11 061 (Table A1 in Appendix A6). For the case of negative binomial distributions, the sample size savings obtained from the use of CRNs are all close to 90% (compare Tables 3 and A3).

### 6.2. Selecting the best 0.8-quantile

Next we present results for selecting the system with the largest 0.8-quantile, with or without CRNs when we have a mix of output distributions. In all cases we used the slippage configuration.

The best 0.8-quantile was set as the 0.8-quantile of a normal distribution with mean \( 5 + \delta \) and variance 1; that is, \( \theta_k = 5 + \delta + \Phi^{-1}(0.8) \). The inferior systems had 0.8-quantile \( \theta_i = 5 + \Phi^{-1}(0.8) \). When the distribution was exponential, we solved for the distribution’s mean so that it had the desired value \( \theta_k \) or \( \theta_i \) for its 0.8-quantile. To incorporate CRNs, we used the NORTA method cited earlier. The results with CRNs (with induced correlation \( \rho = 0.9 \)) are listed in Table 4, and are without CRNs in Table A4 of Appendix A6.
The PCS values are all greater than or equal to 0.95, and the CI simultaneous coverage probabilities are mostly greater than or equal to 0.95. As expected, CRNs greatly reduce the required sample size.

6.3. Comparison with Rinott’s procedure

In this section we consider four different output distributions, normal, Poisson, negative binomial, and a mix of half-normal and half-exponential distributions to compare the PCS and efficiency of our bootstrap R&S algorithm without CRNs to Rinott’s procedure. Rinott’s procedure is a two-stage, normal-theory procedure for which the sample size from each system is proportional to its sample variance; it does not exploit CRNs but takes several stages to estimate the PCS and stops as soon as it is achieved.

| Table 4. | Empirical results for quantiles \( \alpha = 0.8 \) from 100 macroreplications for half-exponential and half-normal distributions in the slippage configuration with CRN \( \mu = 0.9 \) |
|-------------------|-------------|-------------|----------|-------------|
| \( k \) | \( n_0 \) | Ave \( N^* \) | PCS | Est. coverage | True coverage | Ave jumps |
| 2(n) | 100 | 16.936 | 0.97 | 0.71 | 0.94 | 5.24 |
| 2(e) | 100 | 17.584 | 0.96 | 0.96 | 0.95 | 5.12 |
| 4(n) | 100 | 22.408 | 0.99 | 0.96 | 0.97 | 6.60 |
| 4(e) | 100 | 22.904 | 0.97 | 0.96 | 0.91 | 6.48 |
| 10(n) | 100 | 30.454 | 1.00 | 0.96 | 1.00 | 8.03 |
| 10(e) | 100 | 30.703 | 0.99 | 0.95 | 0.95 | 8.32 |
| 20(n) | 100 | 38.102 | 0.99 | 0.96 | 0.95 | 9.37 |
| 20(e) | 100 | 38.553 | 0.99 | 0.96 | 0.96 | 9.36 |

Rinott’s procedure is more sensitive to the initial sample size than our bootstrap R&S procedure; for example, when \( n_0 = 10 \) and \( 50 < N_B < N_R \) for all distributions, whereas \( N_B > N_R \) when \( n_0 = 5 \), even when the output data are normal as assumed by Rinott’s procedure. For instance, the sample size obtained from applying our procedures for normal distributions when \( n_0 = 5 \) is 2030, which is much smaller than the sample size 3494 from Rinott’s procedure. Moreover, our procedure achieves the desired PCS in this case. This is because Rinott’s procedure is a two-stage procedure and hence with a small initial sample size in the first stage, the procedure becomes conservative in the second stage to guarantee the desired PCS. However, our jump-ahead procedure takes several stages to estimate the PCS and stops as soon as it is achieved.

7. Illustration

This section illustrates using our procedure in a difficult R&S problem that Goldsman et al. (1991) employed to demonstrate methods for selecting the best system. The goal is to find the airline-reservation system with the largest expected time-to-failure (TTF). In the experimental setup, each system works for all systems, and we varied the repair rates \( \lambda \) for all systems, and we varied the repair rates \( \mu \) as shown in Table 6. Clearly, the system with the fastest repair rate will be the best, but we applied the procedures as if we do not know this point.

Note that computer failure is rare, repair times are fast, and hence the resulting \( E[\text{TTF}] \) is large. The replication outputs are highly variable and non-normal (closer to exponential), so much so that Goldsman et al. (1991) batched the data before applying the two-stage R&S procedure created for normally distributed output data from Rinott (1978). We applied our procedure directly without any pre-processing of the data.

Table 6 shows the results without and with CRNs; the CRN cases are indicated by * . Notice that the desired PCS of 0.95 is attained, provided \( n_0 \) is not too small, and the required sample sizes are significantly reduced by using CRNs.

8. Conclusions

In this article we have provided essentially generic R&S procedures for computer simulation. Under very mild conditions on the simulation output data they provide asymptotically correct selection for all of the typical performance measures: means,
probabilities, variances, and quantiles. The distribution families of the simulation outputs need not be known, or even the same across systems, and there are only two versions of the procedure: without or with CRNs.

We achieved this generality by using bootstrapping to estimate the PCS at the current sample size, and we sequentially increase the sample size until the bootstrap PCS is ≥ 1 − α. Thus, the generality comes at the cost of data storage (we need to retain the output data to bootstrap it) and calculation of the estimated PCS. To mitigate the computational expense, particularly as the number of systems k increases, we provide a procedure that quickly jumps to the required sample size $N^*$.

To derive and prove the correctness of our procedures, we exploited a connection between PCS and the coverage of simultaneous CIs for all pairs to differences. Since coverage of all pairwise differences is a more demanding objective than correct selection of the best (see, for instance, Hsu (1996)), there is clearly room to make the procedures more efficient. Also, whereas our procedure is sequential, it is not an eliminating procedure, meaning that all k systems receive $N^*$ replications. This provides another opportunity for savings, as reducing k reduces the required effort for both the simulation and bootstrap computations.

Acknowledgements

The authors thank the Associate Editor and two referees for their feedback.

Funding

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT&Future Planning 2014R1A1A3049955, and by Hankuk University of Foreign Studies Research Fund. Portions of this work were published in the Proceedings of the 2014 Winter Simulation Conference as Lee and Nelson (2014).

Notes on contributors

Sooenhui Lee is an Assistant Professor in the College of Business at Hankuk University of Foreign Studies. She received her B.S. at KAIST, M.S. at Georgia Institute of Technology, and Ph.D. in Industrial Engineering and Management Sciences at Northwestern University. Her research interests include stochastic optimization and its application.

Barry L. Nelson is the Walter P. Murphy Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. He is a Fellow of INFORMS and IIE. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems, and he is the author of Foundations and Methods of Stochastic Simulation: A First Course, published by Springer.

References

Aerts, M. and Gijbels, I. (1993) A three-stage procedure based on bootstrap critical points. Sequential Analysis, 12, 93–113.

Anscombe, F.J. (1952) Large-sample theory of sequential estimation. Mathematical Proceedings of the Cambridge Philosophical Society, 48, 600–607.