Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

BOOTSTRAP RANKING & SELECTION REVISITED

Soonhui Lee

School of Business Administration
UNIST
Ulsan, REPUBLIC OF KOREA

Barry L. Nelson

Dept. of Ind. Engr. & Mgmt. Sci.
Northwestern University
Evanston, IL 60208-3119, USA

ABSTRACT

Many ranking-and-selection (R&S) procedures have been invented for choosing the best simulated system; in this paper we consider indifference-zone procedures that attempt to provide a probability of correct selection (PCS) guarantee. To obtain the PCS guarantee, existing procedures nearly always exploit knowledge about the particular combination of system performance measure (e.g., mean, probability, quantile) and assumed output distribution (e.g., normal, exponential, Poisson). In this paper we take a step toward general-purpose R&S procedures that work for many types of performance measures and output distributions, including situations in which different simulated alternatives have entirely different output distributions. There are only two versions of our procedure: with and without the use of common random numbers, and they can be applied to performance measures that can be expressed as expected values or quantiles. To obtain the desired PCS we exploit intense computation via bootstrapping, and establish the asymptotic PCS under very mild conditions. We also report results of an empirical study to assess the procedures' small-sample properties.

1 INTRODUCTION

The statistical methods of ranking and selection (R&S) have been widely accepted in the stochastic simulation community, so much so that R&S procedures are included in a number of commercial simulation products. In this paper we focus on procedures that attempt to identify the best simulated alternative. A common characteristic of selection-of-the-best procedures that have been used extensively in simulation is that "best" is defined to be smallest or largest *mean* performance. Further, 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.

There is, however, a vast 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 qth quantile (Bechhofer et al. (1995); Gupta and Panchapakesan (1979)). And there are also procedures designed for output data that are known to be non-normal, including Poisson, Bernoulli and exponential. Procedures for these situations are customized for the particular performance measure or type of data, exploiting mathematical-statistical properties of the relevant estimator or distribution.

In this paper we take a step toward general-purpose R&S, by which we mean procedures that work for many types of performance measures (e.g., means, probabilities or quantiles) and types of data (discrete- or continuous-valued and almost arbitrary distributions); in fact, not all systems even 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 probability of correct selection (PCS). Our approach is frequentist in philosophy and incorporates an indifference zone.

Because 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 "to death" (effectively zero variance estimator). We also will not beat procedures that directly exploit distributional information; for instance, if you know your output data really are Poisson, then we expect that a procedure based on that knowledge should be more efficient than ours (e.g., Mulekar and Matejcik (2000)). 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. We prove the asymptotic validity of our procedure, but because it bootstraps the actual simulation output data it works well in finite samples across a variety of situations.

The seed of the idea for this paper is in Bekki et al. (2010), which outlined the basic framework and noted its asymptotic validity for the special case of $k = 2$ systems. We extend the mathematical support to any number of systems and many types of estimators here. We also provide a more extensive empirical study.

2 PCS AND FIXED-WIDTH CONFIDENCE INTERVALS

Let X_{ij} represent the jth observed output of system i, for $i = 1, 2, \ldots, k$, so that $\mathbf{X}_j = (X_{1j}, X_{2j}, \ldots, X_{kj})^T$ is a $k \times 1$ vector representing the jth observed output across all systems. Throughout the paper, 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 common random numbers (CRN), then it will be useful to think of $\mathbf{X}_1, \mathbf{X}_2, \ldots$ as i.i.d. with common joint distribution function $F(\mathbf{x}) = \Pr\{X_{1j} \leqslant x_1, \ldots, X_{kj} \leqslant x_k\}$, $\mathbf{x} = (x_1, \ldots, x_k)^T \in \mathbb{R}^k$. We neither assume nor fit any specific distribution to the simulation output.

Let $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)^T$ be a vector whose ith element is a statistical property of the marginal distribution F_i, such as its mean, 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 θ_i with a specified PCS by choosing the one with the largest empirical estimate $\widehat{\theta}_i$ of it.

For example, suppose there are $k = 2$ systems and we want PCS $\geq 1 - \alpha$. Suppose further that we are willing to assume that $\theta_1 - \theta_2 \geqslant d$ (> 0), where without loss of generality system 1 is the best. Then we want the sample size n such that

$$\Pr\{\ \widehat{\theta}_1 > \widehat{\theta}_2 \mid \theta_1 - \theta_2 \geqslant d\ \} \geqslant 1 - \alpha \tag{1}$$

where $\widehat{\theta}_i$ is an estimator of θ_i based on n outputs from system i, $i = 1, 2$.

Consider now the related problem of choosing n to obtain a fixed-width d CI for $\theta_1 - \theta_2$ with specified coverage probability $1 - \alpha$. Now we want n such that the unconditional

$$\Pr\{\theta_1 - \theta_2 \in [\widehat{\theta}_1 - \widehat{\theta}_2 - d,\ \widehat{\theta}_1 - \widehat{\theta}_2 + d]\} \geqslant 1 - \alpha. \tag{2}$$

Suppose we can form such an interval, and after having done so we select $M = \arg\max_i \widehat{\theta}_i$ as the best system. Then if in fact $\theta_1 - \theta_2 \geq d$, we have

$$\begin{aligned} \Pr\{M = 1\} &= \Pr\{\widehat{\theta}_1 > \widehat{\theta}_2\} \\ &= \Pr\{\ \widehat{\theta}_1 - \widehat{\theta}_2 - (\theta_1 - \theta_2) > -(\theta_1 - \theta_2)\ \} \\ &\geqslant \Pr\{\widehat{\theta}_1 - \widehat{\theta}_2 - (\theta_1 - \theta_2) > -d\ \} \\ &\geqslant 1 - \alpha. \end{aligned} \tag{3}$$

The fixed-width confidence interval approach guarantees that the selected system is the best with probability at least $1 - \alpha$ when $\theta_1 - \theta_2 \geq d$ if we select the system with the largest sample statistic. Therefore, a fixed-width d confidence interval procedure implies an indifference-zone R&S procedure when we have $k = 2$ systems, where the half width d corresponds to the indifference-zone parameter.

To extend to $k > 2$ systems, we consider CIs on all pairs of differences simultaneously. That is, we build fixed-width d CIs for all $\theta_i - \theta_j, i \neq j$ with simultaneous coverage $1-\alpha$. As shown in Hsu (1996), if we have

$$\Pr\{\widehat{\theta}_i - \widehat{\theta}_j - (\theta_i - \theta_j) \leq d, \quad \forall i \neq j\} \geqslant 1-\alpha \tag{4}$$

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

$$\theta_i - \max_{j \neq i} \theta_j \in \left[\widehat{\theta}_i - \max_{j \neq i} \widehat{\theta}_j - d, \widehat{\theta}_i - \max_{j \neq i} \widehat{\theta}_j + d\right] \tag{5}$$

for $i = 1, 2, \ldots, k$. If M is the index of the system with the largest performance estimate, i.e., $\widehat{\theta}_M \geqslant \widehat{\theta}_i$ for all $i \neq M$, then it follows from (5) that with probability at least $1-\alpha$

$$\theta_M - \max_{j \neq M} \theta_j \geqslant \widehat{\theta}_M - \max_{j \neq M} \widehat{\theta}_j - d \geqslant -d$$

as $\widehat{\theta}_M - \max_{j \neq M} \widehat{\theta}_j \geqslant 0$. This result implies that if we select the system with the largest performance estimate $\widehat{\theta}_M$ as the best system, the selected system will be the best system or a system within d of the best system with probability at least $1-\alpha$. Moreover, if the difference between the largest and the second largest parameter value is strictly greater than d, then the procedure guarantees that the selected system is the best system with probability at least $1-\alpha$. Thus, if we have a procedure to create fixed-width d CIs for $\theta_i - \theta_j$ with overall coverage $\geqslant 1-\alpha$, then we also have a selection-of-the-best procedure.

3 BOOTSTRAP FIXED-WIDTH CONFIDENCE INTERVALS

Having established a connection between CIs and R&S, we next present a method for constructing fixed-width d CIs for all pairwise comparisons that depends only weakly on the performance measure or output distributions.

Swanepoel et al. (1983) describe a sequential bootstrapping procedure for generating a single fixed-width CI with a specified coverage probability when θ is either a mean or quantile. Given an i.i.d. sample of size n, denoted $\underline{\mathrm{X}}_n = \{X_1, X_2, \ldots, X_n\}$, from a population with marginal distribution F having a distribution property θ, let $\widehat{F}_n$ denote the empirical cumulative distribution function (ecdf) of $\underline{\mathrm{X}}_n$ defined as

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{I}\{X_j \leqslant x\}.$$

Let $\widehat{\theta}_n$ be the corresponding distributional property of $\widehat{F}_n$. Further, let $\underline{\mathrm{X}}^*_n = \{X^*_1, X^*_2, \ldots, X^*_n\}$ denote a random sample of size n from $\widehat{F}_n$, $\widehat{F}^*_n$ the implied ecdf, and $\widehat{\theta}(\underline{\mathrm{X}}^*_n)$ (also denoted by $\widehat{\theta}^*_n$) the corresponding distributional property of $\widehat{F}^*_n$. The bootstrap estimator of the probability that θ is contained within the interval $[\widehat{\theta}_n - d,\ \widehat{\theta}_n + d]$ is

$$P^*_n = \Pr\left\{\widehat{\theta}_n \in \left[\widehat{\theta}^*_n - d, \widehat{\theta}^*_n + d\right]\right\}. \tag{6}$$

Exact computation of P^*_n is often difficult, but (6) can be estimated given B random samples of size n from $\widehat{F}_n$, say $\underline{\mathrm{X}}^*_{nb} = \{X^*_{1b}, X^*_{2b}, \ldots, X^*_{nb}\}$, $b = 1, 2, \ldots, B$, by using

$$P^*_{nB} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{I}\left\{\widehat{\theta}_n \in \left[\widehat{\theta}^*_{nb} - d,\ \widehat{\theta}^*_{nb} + d\right]\right\} \tag{7}$$

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

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

Theorem 1 (Swanepoel et al. (1983)) Under some mild assumptions, as $d \downarrow 0$,

(a) $d^2 N^* \to c$ a.s.

(b) $\Pr\left\{|\widehat{\theta}_{N^*} - \theta| \leq d\right\} \to 1-\alpha.$

The limit c depends on the distributional property of interest: $c = E(X-\theta)^2 z^2_{1-\alpha/2}$ when θ is the mean and $c = z^2_{1-\alpha/2}/(4f(\theta)^2)$ when θ 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 random distribution.

As discussed earlier, the fixed-width confidence interval approach can be used for R&S for two systems by building a confidence interval around the difference $\theta = \theta_1 - \theta_2$. Therefore, the results of Swanepoel et al. (1983) can be interpreted as providing R&S on two systems. Bekki et al. (2010) presented empirical results for using just such a procedure for two or three systems, which gave empirical evidence in support of the approach for R&S based on both means and quantiles. This paper presents asymptotic support for bootstrap fixed-width d CIs that parallel those in Swanepoel et al. (1983), but for any number of pairwise differences, thereby providing indifference-zone R&S for $k \geq 2$ systems. First, however, we describe our generic procedure for bootstrap R&S on $k \geqslant 2$ systems in the next section.

4 THE GENERIC PROCEDURE

In this section we describe algorithms for performing R&S for $k \geqslant 2$ systems using the bootstrap-based fixed-width confidence interval approach. We present two versions of the algorithm, one that exploits common random numbers (CRN) and one without CRN. The algorithm without using CRN has been presented in Bekki et al. (2010) when the sample size is 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. We do not yet have a method for adaptively choosing an optimal value for Δn, but empirically we found $\Delta n = 10$ to provide a substantial computational speed-up without noticable loss of statistical efficiency.

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

$$\Pr\left\{\theta_{ij} \in \left[\widehat{\theta}_{ij}(\underline{\mathbf{X}}_n) - d, \widehat{\theta}_{ij}(\underline{\mathbf{X}}_n) + d\right],\ \forall i \neq j\right\} \geqslant 1-\alpha$$

where $\theta_{ij} = \theta_i - \theta_j$. 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 $\widehat{F}_{iN}$, $\underline{X}^*_{iNb} = \{X^*_{i1b}, X^*_{i2b}, \ldots, X^*_{iNb}\}$, $b = 1,2,\ldots,B$, the bootstrap coverage probability is estimated by

$$P^*_{NB} = \frac{1}{B}\sum_{b=1}^{B} \prod_{(i,j|i\neq j)} \mathbf{I}\left\{\widehat{\theta}_{ij}(\underline{\mathbf{X}}_N) \in \left[\widehat{\theta}_{ij}(\underline{\mathbf{X}}^*_{Nb}) - d,\ \widehat{\theta}_{ij}(\underline{\mathbf{X}}^*_{Nb}) + d\right]\right\} \tag{8}$$

where $\widehat{\theta}(\underline{\mathrm{X}}^*_{iNb})$ is an estimate of $\widehat{\theta}(\underline{\mathrm{X}}_{iN})$ based on $\underline{\mathrm{X}}^*_{iNb}$, and $\widehat{\theta}_{ij}(\underline{\mathbf{X}}^*_{Nb}) = \widehat{\theta}(\underline{\mathrm{X}}^*_{iNb}) - \widehat{\theta}(\underline{\mathrm{X}}^*_{jNb})$ for all $i \neq j$. The procedure without CRN 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) d for the CI, and a sample-size increment Δn.

Bootstrap R&S procedure without CRN

1. Specify $N = n_0$, set $1/k < 1 - \alpha < 1$, $d > 0$, and $\Delta n \geq 1$.
2. Obtain $\underline{\mathrm{X}}_{iN} = \{X_{i1}, X_{i2}, \ldots, X_{iN}\}$ a sample of size N from the distribution F_i for $i = 1, 2, \ldots, k$.
3. Compute $\widehat{\theta}_{ij}(\underline{\mathbf{X}}_N) = \widehat{\theta}(\underline{\mathrm{X}}_{iN}) - \widehat{\theta}(\underline{\mathrm{X}}_{jN})$ for all $i \neq j$ where θ_i is a distributional property of F_i and $\widehat{\theta}(\underline{\mathrm{X}}_{iN})$ is an estimate of θ_i based on $\underline{\mathrm{X}}_{iN}$; and form the ecdf $\widehat{F}_{iN}$ of F_i for system $i = 1, 2, \ldots, k$.
4. Obtain B bootstrap samples of size N from $\widehat{F}_{iN} : \underline{\mathrm{X}}^*_{iN1}, \ldots, \underline{\mathrm{X}}^*_{iNB}, i = 1, \ldots, k$.
5. Compute $\widehat{\theta}_{ij}(\underline{\mathbf{X}}^*_{Nb}) = \widehat{\theta}(\underline{\mathrm{X}}^*_{iNb}) - \widehat{\theta}(\underline{\mathrm{X}}^*_{jNb}),\ b = 1, 2, \ldots, B$ for all $i \neq j$.
6. Estimate the PCS as

$$P^*_{NB} = \frac{1}{B} \sum_{b=1}^{B} \prod_{(i,j|i \neq j)} \mathbf{I}\left\{ |\widehat{\theta}_{ij}(\underline{\mathbf{X}}^*_{Nb}) - \widehat{\theta}_{ij}(\underline{\mathbf{X}}_N)| \leqslant d \right\}.$$

7. If $P^*_{NB} \geqslant 1 - \alpha$, report $\arg\max_{i=1,\ldots,k} \widehat{\theta}(\underline{\mathrm{X}}_{iN})$
   Else
   Obtain $\underline{\mathrm{X}}_{i\Delta n}$ a sample of size Δn from the distribution F_i for $i = 1, 2, \ldots, k$.
   Set $\underline{\mathrm{X}}_{iN} = \underline{\mathrm{X}}_{iN} \cup \underline{\mathrm{X}}_{i\Delta n}$ for $i = 1, 2, \ldots, k$ and $N = N + \Delta n$.
   Go to Step 3.
   End If

We next present the bootstrap R&S procedure that exploits the use of CRN. The sample size required to attain the desired PCS is expected to be reduced relative to independent sampling. In the algorithm with CRN, a sample will be taken from each of the k systems using CRN across systems to induce a joint distribution on $\{F_1, F_2, \ldots, F_k\}$; we denote that distribution by F. Correspondingly, we draw bootstrap samples from the empirical joint cdf $\widehat{F}_N$, rather than from each marginal ecdf $\widehat{F}_{iN}$. Below we list only the steps that change from the **Bootstrap R&S procedure without CRN**:

Bootstrap R&S procedure with CRN

2. Obtain a sample $\mathbf{X}_j = (X_{1j}, X_{2j}, \ldots, X_{kj})^T$ $j = 1, 2, \ldots, N$ from the joint distribution F.
3. Compute $\widehat{\theta}_{ij}(\underline{\mathbf{X}}_N) = \widehat{\theta}(\underline{\mathrm{X}}_{iN}) - \widehat{\theta}(\underline{\mathrm{X}}_{jN})$ for all $i \neq j$ where θ_i is a distributional property of F_i, and $\widehat{\theta}(\underline{\mathrm{X}}_{iN})$ is an estimate of θ_i based on $\underline{\mathrm{X}}_{iN}$; and form the ecdf $\widehat{F}_N$ based on $\underline{\mathbf{X}}_N = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\}$ as

$$\widehat{F}_N(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{I}\{X_{1j} \leqslant x_1,\ X_{2j} \leqslant x_2,\ \ldots,\ X_{kj} \leqslant x_k\}.$$

4. Obtain B bootstrap samples of size N from $\widehat{F}_N : \{\mathbf{X}^*_{1b}, \mathbf{X}^*_{2b}, \ldots, \mathbf{X}^*_{Nb}\}$ for $b = 1, 2, \ldots, B$, where $\mathbf{X}^*_{jb} = (X^*_{1jb}, X^*_{2jb}, \ldots, X^*_{kjb})^T$ for $j = 1, 2, \ldots, N$.
7. If $P^*_{NB} \geqslant 1 - \alpha$, report $\arg\max_{i=1,\ldots,k} \widehat{\theta}(\underline{\mathrm{X}}_{iN})$
   Else
   Obtain $\underline{\mathbf{X}}_{\Delta n} = \{\mathbf{X}_j,\ j = 1, 2, \ldots, \Delta n\}$ a sample of size Δn from the distribution F.
   Set $\underline{\mathbf{X}}_N = \underline{\mathbf{X}}_N \cup \underline{\mathbf{X}}_{\Delta n}$ and $N = N + \Delta n$.
   Go to Step 3.
   End If

Later we report the results from experiments with and without CRN to illustrate the impact of CRN on the sample size required to attain the desired PCS.

5 ASYMPTOTIC RESULTS

The theorems stated below extend Swanepoel et al. (1983) from a single CI to simultaneous CIs for multiple means and quantiles; proofs can be found in Lee and Nelson (2014). These asymptotic results support our use of bootstrap R&S for $k \geqslant 2$ systems and either mean or quantile performance measures, as shown in the corollaries. We first review the key notation.

Let $\underline{\mathbf{X}}_n = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\}$ be a random sample of size n from distribution F (in $\mathbb{R}^k$) with a $k \times 1$ vector of marginal distribution properties Θ, where $\mathbf{X}_j = (X_{1j}, X_{2j}, \ldots, X_{kj})^T,\ j = 1, 2, \ldots, n$. Further, let $\widehat{F}_n(\mathbf{x})$ be the ecdf based on $\underline{\mathbf{X}}_n$ defined in two different ways for use in the procedure without CRN, as in (9), and with CRN, as in (10):

$$\widehat{F}_n(\mathbf{x}) = \prod_{i=1}^{k} \left(\frac{1}{n} \sum_{j=1}^{n} \mathbf{I}\{X_{ij} \leqslant x_i\} \right) \tag{9}$$

$$\widehat{F}_n(\mathbf{x}) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{I}\{X_{1j} \leqslant x_1,\ X_{2j} \leqslant x_2,\ \ldots,\ X_{kj} \leqslant x_k\} \tag{10}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_k)^T \in \mathbb{R}^k$.

Let $\underline{\mathbf{X}}_n^* = \{\mathbf{X}_1^*, \mathbf{X}_2^*, \ldots, \mathbf{X}_n^*\}$ denote a random sample of size n from $\widehat{F}_n$. The bootstrap stopping variable N^* is given by

$$N^* = \inf\left\{ n \geqslant n_0 : \Pr\left\{ |\widehat{\Theta}(\underline{\mathbf{X}}_n^*) - \widehat{\Theta}(\underline{\mathbf{X}}_n)| \leqslant d \cdot \mathbf{1} \right\} \geqslant 1 - \alpha \right\} \tag{11}$$

where $\mathbf{1}$ is the k-dimensional column vector of ones. When $\Theta = \mathrm{E}[\mathbf{X}]$, then $\widehat{\Theta}(\underline{\mathbf{X}}_n)$ and $\widehat{\Theta}(\underline{\mathbf{X}}_n^*)$ are the sample mean vectors based on $\underline{\mathbf{X}}_n$ and $\underline{\mathbf{X}}_n^*$, respectively. That is $\widehat{\Theta}(\underline{\mathbf{X}}_n) = \bar{\mathbf{X}}_n = \sum_{j=1}^{n} \mathbf{X}_j / n$ and $\widehat{\Theta}(\underline{\mathbf{X}}_n^*) = \bar{\mathbf{X}}_n^* = \sum_{j=1}^{n} \mathbf{X}_j^* / n$. Notice that the "mean" case includes probabilities as they are expected values of indicator functions.

Theorem 2 Let $\Theta = \mathrm{E}_F[\mathbf{X}]$. Suppose that $E_F[|\mathbf{X} - \Theta|^3] < \infty$ and $\Sigma = E_F\left[(\mathbf{X} - \Theta)(\mathbf{X} - \Theta)^T\right]$ is a positive definite matrix. Consider N^* as defined in (11).

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

$$d^2 N^* \to a^2 \qquad a.s.$$

where a solves the k-dimensional integral equation

$$\int_{[-a,a]^k} (2\pi)^{-k/2} |\Sigma|^{-1/2} e^{-\mathbf{y}^T \Sigma^{-1} \mathbf{y}/2} d\mathbf{y} = 1 - \alpha. \tag{12}$$

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

$$\Pr\left\{ |\bar{\mathbf{X}}_{N^*} - \Theta| \leq d \cdot \mathbf{1} \right\} \to 1 - \alpha.$$

Next let Θ be a set of specific quantiles of the k marginal distributions where the ith element is defined as

$$\theta_i = F_i^{-1}(q) = \inf\{x : F_i(x) \geqslant q\}, \qquad 0 < q < 1 \qquad i = 1, 2, \ldots, k.$$

Then $\widehat{\Theta}(\underline{\mathbf{X}}_n)$ and $\widehat{\Theta}(\underline{\mathbf{X}}_n^*)$ are the sample qth quantiles based on $\underline{\mathbf{X}}_n$ and $\underline{\mathbf{X}}_n^*$, respectively, where the ith element of $\widehat{\Theta}(\underline{\mathbf{X}}_n)$ is the sample qth quantile of $X_{i1}, X_{i2}, \ldots, X_{in}$ and the ith element of $\widehat{\Theta}(\underline{\mathbf{X}}_n^*)$ is the sample qth quantile of $X_{i1}^*, X_{i2}^*, \ldots, X_{in}^*$.

Theorem 3 Let F_i be twice continuously differentiable in a neighborhood of θ_i and $\delta_i = f_i(\theta_i) > 0$, for $i = 1,2,\ldots,k$, where f_i is the density associated with F_i. Further, let F_{ij} be (i,j)th bivariate marginal distribution function. Consider N^* as defined in (11).

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

$$d^2 N^* \to a^2 \qquad a.s.$$

where a solves Equation (12) with covariance matrix

$$\Sigma = \begin{pmatrix} \frac{q(1-q)}{\delta_1^2} & \frac{\sigma_{12}}{\delta_1\delta_2} & \cdots & \frac{\sigma_{1k}}{\delta_1\delta_k} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\sigma_{k1}}{\delta_k\delta_1} & \frac{\sigma_{k2}}{\delta_k\delta_2} & \cdots & \frac{q(1-q)}{\delta_k^2} \end{pmatrix}$$

where

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

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

$$\Pr\left\{ |\widehat{\Theta}(\underline{\mathbf{X}}_{N^*}) - \Theta| \leq d \cdot \mathbf{1} \right\} \to 1 - \alpha.$$

The theorems stated above provide the basis for the asymptotic validity of our generic procedure for R&S on any number of systems by extending them to all pairs of difference estimates using the linear transformation $\mathbf{A}$ defined as

$$\mathbf{A} = [a_{ij}],\ i = 1,2,\ldots,m;\ j = 1,2,\ldots,k;\ m = \binom{k}{2} \tag{13}$$

where

$$a_{ij} = \begin{cases} 1, & (j-1)\left(k - \frac{j}{2}\right) + 1 \leqslant i \leqslant j\left(k - \frac{j+1}{2}\right);\ j = 1,2,\ldots,k-1 \\ -1, & i = hk - \frac{h(h+1)}{2} - (k-j);\ j = 2,3,\ldots,k;\ 1 \leqslant h \leqslant j-1 \\ 0, & \text{otherwise.} \end{cases}$$

We are now prepared to state the asymptotic validity of our generic R&S procedures in Corollaries 4 and 5. The stopping time used in our procedures can be defined as

$$N_{\mathbf{A}}^* = \inf\left\{ n \geqslant n_0 : \Pr\{|(\mathbf{A}\widehat{\Theta}(\underline{\mathbf{X}}_n^*) - \mathbf{A}\widehat{\Theta}(\underline{\mathbf{X}}_n)| \leqslant d \cdot \mathbf{1}\} \geqslant 1 - \alpha \right\}. \tag{14}$$

Corollary 4 Under the same assumptions as in Theorem 2, consider $N_{\mathbf{A}}^*$ as defined in (14).

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

$$d^2 N_{\mathbf{A}}^* \to a^2 \qquad a.s.$$

where a solves Equation (12) with covariance matrix $A\Sigma A^T$, where Σ is defined as in Theorem 2.

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

$$\Pr\left\{ |\mathbf{A}\widehat{\Theta}(\underline{\mathbf{X}}_{N_{\mathbf{A}}^*}) - \mathbf{A}\Theta| \leq d \cdot \mathbf{1} \right\} \to 1 - \alpha.$$

Next let $\widehat{\Theta}$ be the sample quantiles defined in Theorem 3.

Corollary 5 Under the same assumptions as in Theorem 3, consider $N_{\mathbf{A}}^*$ as defined in (14).

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

$$d^2 N_{\mathbf{A}}^* \to a^2 \qquad a.s.$$

where a solves Equation (12) with covariance matrix $A\Sigma A^T$, where Σ is defined as in Theorem 3.

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

$$\Pr\left\{|\mathbf{A}\widehat{\Theta}(\underline{\mathbf{X}}_{N_{\mathbf{A}}^*}) - \mathbf{A}\Theta| \le d \cdot \mathbf{1}\right\} \to 1-\alpha.$$

6 EXPERIMENT RESULTS

Section 4 described the generic procedure for R&S using bootstrap-based fixed-width CIs for various types of output distributions and for any number of systems. In this section we present empirical results from applying the procedure with $k = 10$ systems having normal or Poisson output distributions when θ is the mean. We also present preliminary results for R&S based on the 0.5 or 0.8 quantiles when the output is normally distributed. Finally, we revisit the Poisson case to compare the efficiency of our procedure to procedures designed specifically for Poisson output data.

6.1 Bootstrap R&S for the Mean

All results presented here are averaged over 100 macro-replications of the entire experiment, and in all cases a sample-size increment of $\Delta n = 10$ was used.

Tables 1–4 contain results for selecting the system with the largest mean with or without CRN when varying the initial sample size n_0, the half width (indifference-zone parameter) d, and the number of bootstrap resamples B; the desired confidence level is $1-\alpha = 0.95$ for all experiments.

For the case of normally distributed output, the empirical results in Table 1 are without using CRN. The true mean vector is $\mu = (1, 1.5, 2, 2.1, 3, 3.2, 4, 5, 5.2, 5.5)$ and the covariance matrix is $\Sigma = \mathbf{I}_{10}$, where $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. The PCS was calculated as the fraction of the 100 macroreplications in which a system whose mean is within d of the true best mean (which is 5.5) was selected. The estimated coverage probability is P_{NB}^* from Step 6 of the algorithm, and the true coverage probability is computed as the percentage of the time that the 45 CIs simultaneously cover all pairwise differences $\theta_i - \theta_j$ for all $i \neq j$.

Table 1 shows that the required sample size N^* increases as the half width (indifference zone parameter) d decreases, as expected. The PCS values are greater than or equal to 0.95 in all cases, and in fact are conservative since achieving simultaneous coverage is more stringent than simply selecting the best.

With the initial sample size n_0 fixed, the required sample size increases slightly as B increases, as noted in Swanepoel et al. (1983), and the true coverage tends to improve. In the experiments with $B = 50$ the coverage probabilities are less than the desired coverage probabilities when $d = 0.1$ and $d = 0.3$, although the desired PCS is still achieved.

To implement the algorithm with CRN, we consider a common base random variable Z which is $N(0,1)$ and set $X_i = \sigma_z Z + W_i$, where W_i's are $N(\mu_i, \sigma_{W_i}^2)$ random variables for $i = 1, 2, \ldots, k$. Then the correlation between X_i and X_j for $i \neq j$ is

$$\mathrm{Corr}(X_i, X_j) = \frac{\sigma_z^2}{\sqrt{\sigma_z^2 + \sigma_{W_i}^2}\sqrt{\sigma_z^2 + \sigma_{W_j}^2}}.$$

By letting $\sigma_{W_i}^2 = 1 - \sigma_z^2$, $0 < \sigma_z^2 < 1$, we have $\mathrm{Var}(X_i) = 1$ for $i = 1, 2, \ldots, k$ and $\mathrm{Corr}(X_i, X_j) = \sigma_z^2$. In Table 2, $\sigma_z = \sqrt{0.9}$ and $\sqrt{0.5}$ were used; therefore, $\mathrm{Corr}(X_i, X_j) = 0.9$ and 0.5.

The results in Table 2 show that the required sample size is indeed reduced as the correlation between systems increases. For instance, with $n_0 = 50$, $B = 200$ and $d = 0.3$, the required sample sizes per system are 114.5 and 50 when the correlation is 0.5 and 0.9, respectively, as compared to 220.8 in Table 1.

Table 1: Empirical results from 100 macroreplications for normal distributions without CRN.

n_0	d	B	Average N^*	PCS	Est. Coverage	True Coverage
10	0.1	50	1656.3	1	0.9658	0.89
50	0.1	50	1661.5	1	0.9672	0.89
100	0.1	50	1656.5	1	0.9658	0.87
10	0.1	100	1734	1	0.9562	0.87
50	0.1	100	1743	1	0.9589	0.91
100	0.1	100	1752.6	1	0.9570	0.88
10	0.1	200	1832.1	1	0.9557	0.93
50	0.1	200	1829.2	1	0.9558	0.94
100	0.1	200	1830.2	1	0.9560	0.94
10	0.3	50	213.1	1	0.9718	0.93
50	0.3	50	214.3	0.99	0.9706	0.92
100	0.3	50	210.1	1	0.9722	0.94
10	0.3	100	215.5	0.99	0.9631	0.97
50	0.3	100	213.1	1	0.9601	0.93
100	0.3	100	214.2	1	0.9610	0.92
10	0.3	200	218.2	1	0.9602	0.92
50	0.3	200	220.8	1	0.9594	0.96
100	0.3	200	219.3	1	0.9595	0.95
10	0.5	50	83.8	1	0.9736	0.97
50	0.5	50	82.4	1	0.9756	0.97
100	0.5	50	100.8	1	0.9842	0.99
10	0.5	100	82.2	1	0.9680	0.96
50	0.5	100	82.7	1	0.9669	0.93
100	0.5	100	100	1	0.9852	0.98
10	0.5	200	85.4	1	0.9659	0.98
50	0.5	200	83.9	1	0.9651	0.99
100	0.5	200	100.0	1	0.9855	0.99

Table 2: Empirical results from 100 macroreplications for normal distributions with CRN.

Corr	n_0	d	B	Average N^*	PCS	Est. Coverage	True Coverage
0.5	50	0.3	200	114.5	1	0.9621	0.94
0.9	50	0.3	200	50.0	1	0.9995	1
0.5	50	0.1	200	936.4	1	0.9567	0.90
0.9	50	0.1	200	198.8	1	0.9607	0.96

Table 3: Empirical results from 100 macroreplications for Poisson distributions without CRN.

n_0	d	B	Mean N^*	PCS	Est. Coverage	True Coverage
50	0.1	200	6389.8	1	0.9544	0.87
100	0.1	200	6381.7	1	0.9546	0.94
50	0.1	400	6718.3	1	0.9539	0.89
100	0.1	400	6711.8	1	0.9536	0.94
50	0.3	200	761.3	1	0.9569	0.94
100	0.3	200	767.1	0.97	0.9569	0.91
50	0.3	400	779.5	0.99	0.9551	0.91
100	0.3	400	786.8	1	0.9558	0.96
50	0.5	200	288.3	0.99	0.9605	0.91
100	0.5	200	289.1	0.99	0.9595	0.91
50	0.5	400	292.2	1	0.9576	0.96
100	0.5	400	290.3	0.99	0.9568	0.97

Table 4: Empirical results from 100 macroreplications for Poisson distributions with CRN.

λ_W	n_0	d	B	Average N^*	PCS	Est. Coverage	True Coverage
0.5	50	0.3	200	669.8	1	0.9568	0.97
0.9	50	0.3	200	594.2	1	0.9558	0.95
0.5	50	0.1	200	5600.7	1	0.9539	0.94
0.9	50	0.1	200	5037.4	1	0.9555	0.9

For the Poisson output distribution, the true means are again $\lambda = (1, 1.5, 2, 2.1, 3, 3.2, 4, 5, 5.2, 5.5)$. The empirical results in Table 3 are obtained without using CRN. The PCS values are all greater than 0.95, but the CI simultaneous coverage probability can be significantly less that 0.95 when $B = 50$, again emphasizing that the number of bootstrap samples cannot be too small. For most cases the required sample size increases and the coverage probability tends to improve as B increases.

To implement the algorithm with CRN, we consider a common base Poisson random variable W with parameter λ_W and set $X_i = W + W_i$ where W_i is a Poisson random variable with parameter λ_{W_i}. Then the correlation between X_i and X_j for $i \neq j$ is

$$\text{Corr}(X_i, X_j) = \frac{\lambda_W}{\sqrt{\lambda_W + \lambda_{W_i}}\sqrt{\lambda_W + \lambda_{W_j}}}.$$

In the results shown in Table 4, $\lambda_W = 0.9$ and 0.5 are used. Unlike the normal case, the correlations between the systems are not equal to each other; when $\lambda_W = 0.9$ the $\text{Corr}(X_i, X_j)$ ranges from 0.17 to 0.73, and when $\lambda_W = 0.5$ it ranges from 0.09 to 0.41.

The results in Table 4 show that the required sample size is reduced compared to the results without CRN. Notice that the required sample sizes for the Poisson case are much larger than the required sample sizes for the normal case with the same configuration of the means because the variances in the Poisson case are much higher than the normal case.

6.2 Bootstrap R&S for Quantiles

Here we present some preliminary results for selecting the system with the largest q quantile when we have $k = 10$ systems, normally distributed output, with means $\mu = (1, 1.5, 2, 2.1, 3, 3.2, 4, 5, 5.2, 5.5)$ and all standard deviations being 1, as before. We consider the $q = 0.5$ (median) and $q = 0.8$ quantiles. Of course, the median is the same as the mean, while the 0.8 quantiles are $(1.84, 2.34, 2.84, 2.94, 3.84, 4.04, 4.84, 5.84, 6.04, 6.34)$.

Table 5: Empirical results from 100 macroreplications for normal distributions without CRN and $q = 0.5, 0.8$.

q	n_0	d	B	Average N^*	PCS	Est. Coverage	True Coverage
0.5	100	0.3	200	390.9	1	0.9585	0.97
0.5	200	0.3	200	388.0	1	0.9593	0.96
0.5	200	0.3	400	400.5	1	0.9574	0.98
0.8	100	0.3	200	508.8	1	0.9580	1
0.8	200	0.3	200	508.7	0.99	0.9581	0.99
0.8	200	0.3	400	523.1	1	0.9581	0.99

Table 6: Empirical results from 100 macroreplications for normal distributions with CRN and $q = 0.5, 0.8$.

q	Corr	n_0	d	B	Average N^*	PCS	Est. Coverage	True Coverage
0.5	0.5	100	0.3	200	285.8	1	0.9598	0.98
0.5	0.5	200	0.3	200	286.1	1	0.9600	0.99
0.5	0.5	200	0.3	400	279.2	1	0.9586	0.98
0.5	0.9	100	0.3	200	151.9	1	0.9633	1
0.5	0.9	200	0.3	200	200.7	1	0.9871	1
0.5	0.9	200	0.3	400	200.4	1	0.9857	1
0.8	0.5	100	0.3	200	383.1	1	0.9587	1
0.8	0.5	200	0.3	200	387.2	1	0.9590	0.99
0.8	0.5	200	0.3	400	383.3	1	0.9582	0.97
0.8	0.9	100	0.3	200	211.1	1	0.9620	0.99
0.8	0.9	200	0.3	200	220.3	1	0.9649	1
0.8	0.9	200	0.3	400	219.7	1	0.9653	0.99

Therefore, the difference between the best and the second best is 0.3. We again use $\Delta n = 10$ in all experiments, and simulate with and without CRN as before.

The results in Tables 5–6 show that both the desired PCS and coverage are attained, although again somewhat conservatively. CRN again reduces the sample size needed to attain the PCS and coverage. Notice that we started with larger values of n_0 so as to obtain a decent estimate of the 0.8 quantile even in the first stage.

6.3 Comparison Against a Customized Procedure

We revisit the Poisson case and compare our bootstrap R&S algorithm without CRN with Rinott's procedure (which is for normally distributed output data) and R&S procedures specifically designed for Poisson data from Mulekar and Matejcik (2000). Recall that there are $k = 10$ systems with mean vector $\lambda = (1, 1.5, 2, 2.1, 3, 3.2, 4, 5, 5.2, 5.5)$, and the desired PCS is $1 - \alpha = 0.95$. We set $n_0 = 10$ to facilitate using the tables available for these other methods.

Since Rinott's procedure allows the sample size for each system to be different, we compare the total sample size required by each method across all ten systems. The total for Rinott's procedure was 5558. Mulekar and Matejcik (2000) proposed both an exact method and a normal approximation method; the total for the exact method and the normal approximation for our example were 7600 and 6910, respectively. The total sample size for bootstrap R&S, averaged over 100 macroreplications, was 7620, which is close to the values suggested by the customized Poisson procedures, showing little if any loss in efficiency.

7 CONCLUSIONS

In this paper we demonstrated empirically, and provided asymptotic support for, general-purpose R&S procedures based on bootstrapping for performance measures that can be expressed as expected values or quantiles. By "general purpose" we mean that the procedure need not be tailored to the specific performance measure of interest or assumed distribution of the simulation output.

At least two challenges remain: Reducing the computational overhead to implement bootstrap R&S, and deriving procedures that are less conservative (which, by reducing the sample size, will also reduce computation). In this paper we employed a sample-size increment $\Delta n > 1$ to reduce the number of times that the bootstrap coverage probability needs to be estimated, but we have not yet provided any guidelines for choosing Δn in an optimal way. In addition, our selection of the best is based on simultaneous confidence intervals for all pairs of differences, which is stronger inference than needed for selecting the best (see, for instance, Hsu (1996)). Therefore, we believe it might be possible to tighten the procedure by using bootstrapping to estimate PCS directly. Finally, we simulate all k systems until the best is selected, but multistage R&S procedures that eliminate inferior systems along the way tend to be more efficient.

ACKNOWLEDGMENTS

This work was supported by the 2013 Research Fund (1.120084.01) of the Ulsan National Institute of Science and Technology.

REFERENCES

Bechhofer, R. E., T. J. Santner, and D. M. Goldsman. 1995. *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*. New York: Wiley.

Bekki, J. M., B. L. Nelson, and J. W. Fowler. 2010. "Bootstrapping-based fixed-width confidence intervals for ranking and selection". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1024–1033. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Gupta, S. S., and S. Panchapakesan. 1979. *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations*, Volume 44. Siam.

Hsu, J. C. 1996. *Multiple Comparisons: Theory and Methods*. Chapman & Hall/CRC.

Lee, S., and B. L. Nelson. 2014. "Bootstrap Ranking and Selection for Simulation". Technical report, Northwestern University, Department of Industrial Engineering and Management Sciences.

Mulekar, M. S., and F. J. Matejcik. 2000. "Determination of sample size for selecting the smallest of k Poisson population means". *Communications in Statistics-Simulation and Computation* 29 (1): 37–48.

Swanepoel, J., J. V. Wyk, and J. Venter. 1983. "Fixed width confidence intervals based on bootstrap procedures". *Communications in Statistics-Sequential Analysis* 2 (4): 289–310.

AUTHOR BIOGRAPHIES

SOONHUI LEE is an Assistant Professor in the School of Business Administration at UNIST. 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. Her email address is shlee@unist.ac.kr.

BARRY L. NELSON is the Walter P. Murphy Professor and Chair of 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*, from Springer. His e-mail address is nelsonb@northwestern.edu.