Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

BOOTSTRAPPING-BASED FIXED-WIDTH CONFIDENCE INTERVALS FOR RANKING AND SELECTION

Jennifer M. Bekki

Arizona State University at the Polytechnic Campus
7231 E. Sonoran Arroyo Mall, Rm. 230
Mesa, AZ 85212, USA

Barry L. Nelson

Dept. of IE & MS
Northwestern University
Evanston, IL 60208-3119, USA

John W. Fowler

Arizona State University
P.O. Box 878809
Tempe, AZ 85287-8809, USA

ABSTRACT

A ranking and selection (R&S) procedure allowing comparisons between systems to be made based on any distributional property of interest would be useful. This paper presents initial work toward the development of such a procedure. Previously published work gives a method for using bootstrapping to develop fixed-width confidence intervals with a specified coverage probability around a property of interest. Empirical evidence is provided in support of the use of this approach for building fixed-width confidence intervals around both means and quantiles. Additionally, the use of fixed-width confidence intervals for bootstrapped R&S is demonstrated. For two systems, R&S is performed by building a confidence interval around the difference between two systems. Simultaneous fixed-width confidence intervals are used for R&S on more than 2 systems, and the approach is demonstrated for three systems. The technique is shown to be effective for R&S based on both quantiles and means.

1 INTRODUCTION

Ranking and Selection procedures are statistical techniques used to select the best system, or a subset of systems containing the best, from a group of competing alternatives. The procedures guarantee a user specified probability of selecting the best system, and typical assumptions include that the response measure is normally distributed and that the analysis is based on expected values of a single stochastic response. A review of ranking and selection procedures can be found in Kim and Nelson (2006). Traditionally, the response measure is an estimate of the mean, but more recently, increasing attention has been given to developing ranking and selection procedures that allow comparisons to be made more easily on other features of the output distributions. Batur and Choobineh (2010) and Bekki et al. (2007), for example, both look at the use of quantile estimates from independent simulation runs as a means for comparing competing systems. These approaches address some of the restrictions in the traditional ranking and selection procedures, but are still tailored to a specific feature of the distributions under comparison. A flexible ranking and selection procedure would allow comparisons to be made based on any distributional property of interest. This paper presents initial steps toward the development of such a procedure.

1.1 Description of the Problem

Let $\theta_i, i = 1,2,\ldots,k$ represent the same property of interest from distributions $F_i, i = 1,2,\ldots,k$ of k systems. Then assume an empirical estimate of $\hat{F}_i, i = 1,2,\ldots,k$ is generated, and let $\hat{\theta}_i, i = 1,2,\ldots,k$ represent an estimate of θ_i based on $\hat{F}_i$. If N_i represents the number of sampled observations used to generate $\hat{F}_i$, it is of interest to determine the smallest N_i that allows us to identify with a specified probability of correct selection (PCS) the system with the largest value of θ_i.

In the specific case of $k = 2$ systems, if $\hat{\theta}_i$ is the mean, and if $F_i, i = 1,2$ are known in advance to be normal, the problem can easily be solved analytically. In this case, $\hat{\theta}_i = \overline{X}_i, i = 1,2$, $\overline{X}_1$ is

978-1-4244-9864-2/10/$26.00 ©2010 IEEE

$N(\mu_1, \sigma_1^2/N_1)$ and $\overline{X}_2$ is $N(\mu_2, \sigma_2^2/N_2)$. Also, assume that the number of samples used to generate $\hat{F}_1$ and $\hat{F}_2$ are identical ($N_1 = N_2 = N$), and assume $\mu_1 \neq \mu_2$ (where μ_1 and μ_2 are the true means of F_1 and F_2 respectively). Based on normal theory, with Z representing a standard normal random variable, the required sample size, N, which ensures selection of the system with the larger θ_i with probability PCS (=1-α), is shown in (1)

$$N = \lceil z_{1-\alpha}^2 \frac{\sigma_1^2 + \sigma_2^2}{(\mu_2 - \mu_1)^2} \rceil. \tag{1}$$

In the case of real-world situations in which decisions are to be made between k competing systems, however, it is very unlikely that the true distributions, F_i, would be known in advance. Moreover, it is also likely that there would more than two systems under consideration. In such cases, we cannot apply (1) for direct calculation of N, making a non-parametric procedure that does not depend on the F_i distributions desirable. Nonparametric bootstrapping procedures allow estimates of distributional properties to be obtained from samples and do not make any assumptions about the shape of the underlying population distribution. Consequently, an approach for estimating $\theta_i, i = 1, 2, \ldots k$ based on nonparametric bootstrapping would have some attractive features. Swanepoel Swanepoel, van Wyk, and Venter (1983) present a bootstrapping-based approach for determining the sample size required to generate a fixed-width confidence interval around $\theta_i, i = 1, 2, \ldots, k$. The remainder of this paper discusses the Swanepoel Swanepoel, van Wyk, and Venter (1983) procedure, provides empirical support for fixed-width confidence intervals when θ represents both the mean and a quantile from the systems of interest, and then demonstrates how the fixed-width confidence interval approach can be used to perform ranking and selection before concluding with future research directions.

2 FIXED-WIDTH CONFIDENCE INTERVALS

Swanepoel, Swanepoel, van Wyk, and Venter (1983) present an approach for determining the sample size required to generate a fixed-width confidence interval around θ when the samples are i.i.d. Their procedure uses bootstrapping and requires the specification of the desired confidence interval half width, d, the confidence level, α, and an initial sample size, n_0. Starting first with a sample of size $N = n_0$, L new samples, also of size N, are created by sampling with replacement from the original N. An estimate of the property of interest, $\hat{\theta}_{Nl},\ l = 1, 2, \ldots L$, is calculated from each of the L new samples, and P_{NL}^*, the bootstrapped probability that θ is contained within the interval $\left(\hat{\theta}_N - d, \hat{\theta}_N + d\right)$, is estimated using (2) and (3), where $\hat{\theta_N}$ represents the estimate of θ from the original sample.

$$\overline{\hat{\theta}}_{NL} = L^{-1} \textstyle\sum_{l=1}^{L} \hat{\theta}_{Nl} \tag{2}$$

$$P_{NL}^* = L^{-1} \textstyle\sum_{l=1}^{L} I\left(-d \leq \hat{\theta}_{Nl} - \hat{\theta}_N \leq d\right) \tag{3}$$

If $P_{NL}^* \geq 1 - \alpha$, then the sample size N was appropriate. Otherwise, N is increased by 1, an additional sample is obtained, and the process of creating L new samples with replacement and calculating P_{NL}^* based on the new N is repeated until $P_{NL}^* \geq 1 - \alpha$.

2.1 Application of the Procedure when θ is the Mean

Asymptotic properties for the Swanepoel, Swanepoel, van Wyk and Venter (1983) approach are given in their original paper for the cases in which θ represents either the mean or the median. Also presented are results from Monte Carlo simulations, performed to evaluate the small-sample behavior of the procedure. Given that the goal is to incorporate their approach as part of a ranking and selection procedure in which smaller sample sizes are more desirable, this small-sample behavior is critical. Their results when θ represents the mean show that the small-sample performance of the procedure is very sensitive to the choice of L and n_0. Specifically, with smaller L values, the actual coverage probability of confidence intervals based on the initial (not resampled) data were not as high as $1 - \alpha$.

Swanepoel, van Wyk and Venter (1983), however, looked only at values of $L = 25, 100, 200$ and n_0 = 5, 10. Efron and Tibshirani (1993) recommend that L should be at least 1000 for standard error estimates, and at least 10,000 for estimates of confidence intervals. Similarly, they suggest that $n_0 \geq 50$ for most cases. Given present day computing power, performing the Swanepoel, van Wyk and Venter (1983) experimentation with these larger values of L and n_0 is a realistic possibility. To investigate the impact of these parameter changes, the experimentation performed by Swanepoel, van Wyk and Venter (1983) was repeated using values of $L = 200, 10,000$ and $n_0 = 10, 50$. We tested F = a normal distribution, $d = 0.2, 0.3, 0.4$, $\alpha = 0.05, 0.1$, $\hat{\theta}$ = sample mean, and the entire procedure was repeated 100 times for each experiment. The experiments in which $d = 0.3, 0.4$ and $\alpha = 0.1$ represent cases in which the parameters are identical to the parameters in the Swanepoel, van Wyk and Venter (1983) publication, while the experiments in which $d = 0.2$ and $\alpha = 0.05$ give insights for a slightly more severe problem instance. Results of this experimentation are shown in Table 1, which gives the confidence interval half width, desired coverage probability, initial sample size (n_0), and number of times resampling was performed (L).

Table 1: Empirical results to evaluate the small-sample properties of the Swanepoel, van Wyk and Venter (1983) approach with n_0 and L as recommended by Efron and Tibshirani (1993). The results are for the $N(0,1)$ distribution when θ = mean.

CI Half Width	α	n_0	L	Mean Sample Size (s.e.)	E(N)	Coverage Probability (s.e.)
0.4	0.1	10	200	14.43 (.43)	16.91	0.85 (.04)
0.4	0.1	10	10,000	14.78 (.47)	16.91	0.89 (.03)
0.4	0.1	50	200	50 (0)	16.91	**1.0 (0)**
0.4	0.1	50	10,000	50 (0)	16.91	**1.0 (0)**
0.3	0.1	10	200	23.78 (.88)	30.06	**0.9 (.03)**
0.3	0.1	10	10,000	25.32 (.99)	30.06	0.85 (.04)
0.3	0.1	50	200	50 (0)	30.06	**0.97 (.02)**
0.3	0.1	50	10,000	50 (0)	30.06	**0.97 (.02)**
0.2	0.05	10	200	75.47 (1.56)	96.04	0.91 (.03)
0.2	0.05	10	10,000	91.08 (1.91)	96.04	0.91 (.03)
0.2	0.05	50	200	75.59 (1.32)	96.04	0.86 (.03)
0.2	0.05	50	10,000	92.18 (1.61)	96.04	**0.97 (.02)**

In Table 1, for each experimental configuration, the resulting average sample size requirement and standard error (in parenthesis) across all 100 replications are given, along with the resulting coverage probability and standard error. Bolded results are those in which the resulting coverage probability was at least equal to the specified coverage probability. In the column labeled “E(N)”, Table 1 also gives the theoretical sample size for building a fixed-width confidence interval of half width d with coverage probability $1-\alpha$ around a normal distribution with known σ.

The results in Table 1 demonstrate that when the system's distribution is $N(0,1)$ and when θ is the mean, the Swanepoel, van Wyk and Venter (1983) procedure always returns at least the desired coverage probability when the values of the procedure's parameters take on the values suggested by Efron and Tibshirani (1993). Most notably, when the confidence interval half width and desired coverage probability are the tightest, the experiment done using the Efron and Tibshirani parameters is the only one that returns the desired coverage probability. This implies that the choice of parameters is even more critical when the problem instance is more extreme. Also note that whenever the experimental sample size is greater than the theoretical sample size, the desired coverage probability is obtained. Similarly, the desired coverage probability is less likely to be obtained when the resulting sample size is less than the theoretical value.

2.2 Application of the Procedure when θ is the 0.7 or 0.9 Quantile

The results in Table 1 support to the use of the bootstrapping approach for generating fixed-width confidence intervals around the mean. However, we are interested in a procedure that is useful when θ takes on different types of distributional properties. Consequently, it is of interest to determine whether the Swanepoel, van Wyk and Venter (1983) approach can be used effectively with distributional properties other than the mean. Table 2 shows the application of the procedure for developing confidence intervals around the 0.7 and 0.9 quantiles of the standard normal distribution. Specifically, the impact of the starting sample size and number of times bootstrapping was performed (L) was examined. Desired half widths of 0.2 and 0.3 were examined, along with 200 and 10,000 bootstrapping replications. The entire procedure was repeated 100 times for each set of parameters, and the resulting mean suggested sample size and coverage probability are presented in Table 2.

Table 2: Empirical results to evaluate the small-sample properties of the Swanepoel, van Wyk and Venter (1983) approach with n_0 and L as recommended by Efron and Tibshirani (1993). The results are for the $N(0,1)$ distribution when θ = 0.7 or 0.9 quantile.

Quantile	CI Half Width	α	n_0	L	Mean Sample Size (s.e.)	Coverage Probability (s.e.)
0.7	0.2	0.05	10	10000	145.02 (7.17)	0.84 (.04)
0.7	0.2	0.05	10	200	125.3 (5.89)	0.81 (.04)
0.7	0.2	0.05	50	200	131.49 (5.27)	0.83 (.04)
0.7	0.2	0.05	50	10000	162.33 (6.25)	0.9 (.03)
0.7	0.3	0.1	10	10000	42.67 (2.67)	0.7 (.05)
0.7	0.3	0.1	10	200	39.44 (2.24)	0.73 (.04)
0.7	0.3	0.1	50	10000	60.71 (1.87)	**0.9(0.3)**
0.7	0.3	0.1	50	200	58.29 (1.27)	**0.89 (.03)**
0.9	0.2	0.05	10	10000	225.98 (11.78)	0.86 (.03)
0.9	0.2	0.05	10	200	165.05 (9.92)	0.79 (.04)
0.9	0.2	0.05	50	200	186.58 (8.74)	0.84 (.04)
0.9	0.2	0.05	50	10000	234.84 (10.31)	0.89 (.03)
0.9	0.3	0.1	10	10000	56.72 (4.62)	0.72 (.04)
0.9	0.3	0.1	10	200	54.19 (4.06)	0.75 (.04)
0.9	0.3	0.1	50	10000	82.68 (3.87)	0.88 (.03)
0.9	0.3	0.1	50	200	76.6 (3.28)	0.86 (.03)
0.9	**0.2**	**0.05**	**200**	**10000**	**269.31 (7.85)**	**0.96 (0.02)**

The results in Table 2 demonstrate that the choice of initial sample size and number of times bootstrapping is performed are perhaps even more critical when the confidence interval is being built around quantiles rather than the sample mean. Since quantiles are typically more difficult to estimate than means, this result is not surprising. Additionally, the results indicate that as the quantile being estimated tends to the tails, the accuracy of the approach worsens. Specifically, as the desired coverage probability, confidence interval width, and quantile being estimated get more extreme, the initial sample size needs to get larger and larger. The final row of Table 2 demonstrates that with an initial sample size of 200 and replication of the bootstrapping procedure 10,000 times, the desired coverage probability is returned. Consequently, evidence is given in support of the use of the Swanepoel, van Wyk and Venter (1983) approach for developing fixed-width confidence intervals around quantiles. The results also underscore, however, both the need for having a large enough sample size with which to initiate the process and the fact that this sample size is a function of the specific problem instance.

3 USING FIXED-WIDTH CONFIDENCE INTERVALS FOR RANKING AND SELECTION

Section 2 describes a process for generating fixed-width confidence intervals around a distributional property of interest, θ, from a single system at a time and demonstrates that the same process can be used to generate confidence intervals around multiple types of distributional properties. This section addresses how these confidence intervals could also be used to identify with a specified probability of correct selection the system with the largest (or smallest) $\theta_i, i = 1,2,\ldots,k$.

3.1 Ranking and Selection for Two Systems

When only two systems are under comparison, the fixed-width confidence interval approach can be used to build a confidence interval around the difference in the θ property between the two systems. This two-sided confidence interval around $\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_2$ of width $2d$ (d = confidence interval half width) provides a direct mechanism for selecting the better of the two systems. The half width of the confidence interval is analogous to the indifference zone parameter, δ, typically used in traditional two-stage ranking and selection procedures. Since the confidence interval is constructed as $\hat{\theta} = (\hat{\theta}_1 - \theta_2) \pm d$, if the system with the largest $\hat{\theta}_i$ parameter is chosen as the best, we are guaranteed to have chosen the best system or a system within $d = \delta$ of the best system with the specified confidence. A description of the algorithm for performing ranking and selection of two systems using the bootstrapping-based fixed-width confidence interval approach follows. In the algorithm, $\underline{X}_{Ni} = (X_{1i}, X_{2i}, \ldots, X_{Ni})$ is a sample of size N from the distribution of interest, F_i, in system $i = 1, 2, \ldots, k$. Also, $\hat{F}_{Ni}$ is an empirical cdf of the distribution of interest from system $i = 1, 2, \ldots, k$ generated from $\underline{X}_{Ni}$. And θ_i is a distributional property of F_i, while $\hat{\theta}(\underline{X}_{Ni})$ is an estimate of θ_i based on $\underline{X}_{Ni}, i = 1, 2, \ldots k$.

```
1. $N = n_0$, set 1-$\alpha$ and $d$
2. Obtain sample $\underline{X}_{N1}$ and $\underline{X}_{N2}$, compute $\hat{\theta}(\underline{X}_N) = \hat{\theta}(\underline{X}_{N1}) - \hat{\theta}(\underline{X}_{N2})$, $\hat{F}_{N1}$, and $\hat{F}_{N2}$
3. Obtain L bootstrap samples of size N from $\hat{F}_{N1}$: $\hat{\underline{X}}_{N1}^{(1)}, \ldots, \hat{\underline{X}}_{N1}^{(L)}$ and from $\hat{F}_{N2}$: $\hat{\underline{X}}_{N2}^{(1)}, \ldots, \hat{\underline{X}}_{N2}^{(L)}$
4. Compute $\hat{\theta}(\hat{X}_N^{(l)}) = \hat{\theta}(\hat{X}_{N1}^{(l)}) - \hat{\theta}(\hat{X}_{N2}^{(l)})$, $l = 1, 2, \ldots L$
5. $P_{NL}^* = L^{-1} \sum_{l=1}^{L} I\left(-d \leq \hat{\theta}(\hat{\underline{X}}_N^{(l)}) - \hat{\theta}(\underline{X}_N) \leq d\right)$
6. IF $P_{NL}^* \geq 1 - \alpha$ then
 (a)  Report $\overset{[}{i}= 1,2]\arg\max\{\hat{\theta}_{Ni}\}$
 (b)  Else
        i.   $\underline{X}_{N1}\ U\ \{X_{N+1,1}\}$, $\underline{X}_{N2}\ U\ \{X_{N+1,2}\}$
       ii.   $N = N + 1$
      iii.   Recompute $\hat{\theta}(\underline{X}_N)$, $\hat{F}_{N1}$, and $\hat{F}_{N2}$
       iv.   Go to 3
 (c)  End IF
```

This algorithm was applied to perform ranking and selection on two systems based on both their means and their 0.9 quantiles. The distribution of all systems was normal, and the d parameter was set to either 0.3 or 0.45, indicating no preference between systems for which the θ distributional property is different by less than 0.3 or 0.45. For all three experiments, the bootstrapping procedure was repeated 10,000 times at each sample size, the desired probability of correct selection was set to 0.95, and unique random number streams were used to generate samples from each system, making the resulting confidence intervals independent. The entire procedure was repeated 100 times, and the results from Section 2 guided the selection of the initial sample size, n_0, for each experiment. The results of this empirical analysis are shown in Table 3.

Table 3: Empirical results to demonstrate the use of fixed-width confidence intervals for ranking and selection between two systems. System and parameter descriptions for each experiment are given.

θ	System 1	System 2	n_0	d	PCS	Coverage Probability (s.e.)	Mean Sample Size (s.e.)
Mean	$N(11,2)$	$N(12,2)$	50	0.3	1.0	0.95 (.02)	331.74 (2.56)
Mean	$N(10,3)$	$N(11,2)$	50	0.45	1.0	0.93 (.03)	238.5 (2.38)
0.9 Quantile	$N(11,2)$	$N(12,2)$	200	0.3	1.0	.94 (.02)	960.98 (20.23)

Column 6 in Table 3 gives the probability of selecting the correct system. Since system 2 has the larger mean or larger 0.9 quantile in all experiments, the probability of correct selection was calculated as the percentage of times in which $\hat{\theta}_2$ was larger than $\hat{\theta}_1$. In all cases, the returned PCS was at least $1 - \alpha$, supporting the use of the approach for performing ranking and selection on both means and quantiles.

Table 3 also provides the coverage probability and standard error (in parenthesis) for the confidence interval on the difference between θ for the two systems. The true 0.9 quantile for the $N(11,2)$ system is 13.56, and the true quantile for the $N(12,2)$ system is 14.56. Consequently, the coverage probability for all three experiments is the probability that the confidence interval contains the value "-1". For all experiments, when the standard error is considered, the desired coverage probability was returned. The average required sample size across all replications for each experiment is also provided in Table 3. The first and third rows show that, not surprisingly, to achieve the desired coverage probability, the required sample size to develop a fixed-width confidence interval around the differences in the 0.9 quantile was more than double the sample size required to create the the fixed-width confidence interval of the same width around the mean for the same two systems.

Finally, Table 3 also demonstrates the impact of the d parameter on both the coverage probability and the sample size requirements. The two experiments in which θ represents the mean both have means different by 1.0 between the two systems. However, the d parameter was increased to 0.45 in the second experiment, making the confidence interval width 0.9. The wider confidence intervals required smaller sample sizes even though the variability in System 1 was also slightly increased.

3.2 Ranking and Selection for More Than Two Systems

Section 3.1 demonstrates that fixed-width confidence intervals can successfully be used to perform ranking and selection on two systems when the θ distributional property represents the mean or the 0.9 quantile. However, when $k > 2$, a single confidence interval on the difference between two systems is clearly not sufficient. This section discusses the extension of the procedure to more than two systems.

Assume first that simultaneous two-sided confidence intervals of fixed-width $2d$ are built with coverage probability $1-\alpha$ around all pairs of differences, $\theta_i - \theta_j \ \forall \ (i,j|i \neq j)$. These simultaneous confidence intervals then imply the following useful result from Section 4.2.1 of Hsu (1996). If

$$Pr\{\hat{\theta}_i - \hat{\theta}_j - (\theta_i - \theta_j) < d \ \forall (i,j|i \neq j)\} \geq 1-\alpha \tag{4}$$

Then,

$$Pr\{\theta_i - \overset{[}{j}\neq i]\max\theta_j \in [\hat{\theta}_i - \overset{[}{j}\neq i]\max\hat{\theta}_j \pm \delta]\} \geq 1-\alpha. \tag{5}$$

Hsu's (1996) result in (5) then implies that if the system with the maximum $\hat{\theta}_i, i = 1,2,\ldots,k$ is selected from all k systems, the selected system is guaranteed with probability $1-\alpha$ to be the best system or a system within δ of the best system. In order to use Hsu's (1996) result to imply ranking and selection, however, simultaneous confidence intervals of fixed-width $2d$ and coverage probability $1-\alpha$ around the differences of all systems must be built. Section 2 shows a suitable approach for developing the fixed-width confidence intervals, but the method for estimating the bootstrapped coverage probability in Section 3.1 does not provide for more than one simultaneous confidence interval of coverage probability $1-\alpha$. Equation (6), provides a necessary modification to the equation, which defines P^*_{NL} to mean simultaneous coverage. In (6), m represents the number of simultaneous confidence intervals being built.

$$P^*_{NL} = L^{-1}\sum_{l=1}^{L}\overset{[}{p}= 1]m\Pi I\{-d \leq \hat{\theta}(\hat{\underline{X}}^{(l)}_{Np}) - \theta(\underline{X}_{Np}) \leq d\} \tag{6}$$

Incorporating this modification to the procedure in Section 3.1, the algorithm for developing bootstrapped simultaneous fixed-width confidence intervals on all pairs of differences is as follows, where k represents the number of systems under comparison. Note that with this approach the number of pairs of differences is $m = \binom{k}{2}$. Also, we define $\hat{\theta}(\underline{X}_{Nij})$ to represent the difference between the $\hat{\theta}(\underline{X}_{Ni})$ and $\hat{\theta}(\underline{X}_{Nj})$ values for $i \neq j$.

1. $N = n_0$, set 1-α and d
2. Obtain sample $\underline{X}_{Ni}$ $i = 1,2,\ldots,k$, compute $\hat{\theta}(\underline{X}_{Nij}) = \hat{\theta}(\underline{X}_{Ni}) - \hat{\theta}(\underline{X}_{Nj}) \ \ \forall \ (i,j|i \neq j)$, and $\hat{F}_{Ni}$ for all $i = 1,2,\ldots k$
3. Obtain L bootstrap samples of size N from $\hat{F}_{Ni}$: $\hat{\underline{X}}^{(1)}_{Ni},\ldots,\hat{\underline{X}}^{(L)}_{Ni}, i = 1,2,\ldots k$
4. Compute $\hat{\theta}(\hat{\underline{X}}^{(l)}_{Nij}) = \hat{\theta}(\hat{\underline{X}}^{(l)}_{Ni}) - \hat{\theta}(\hat{\underline{X}}^{(l)}_{Nj})$, $l = 1,2,\ldots L \ \ \forall \ (i,j|i \neq j)$
5. $P^*_{NL} = L^{-1}\sum_{l=1}^{L}\overset{[}{(}\ i,j)|i \neq j]\Pi I\{-d \leq \hat{\theta}(\hat{\underline{X}}^{(l)}_{Nij}) - \hat{\theta}(\underline{X}_{Nij}) \leq d\}$
6. If $P^*_{NL} \geq 1-\alpha$ then

(a) Report $\overset{[}{i}= 1,2,\ldots,k]\arg\max\{\hat{\theta}_{Ni}\}$

(b) Else

```
        i.   $\underline{X}_{Ni}\ U\ \{X_{N+1,i}\}\ i = 1,2,\ldots,k$
        ii.  $N = N+1$
        iii. Recompute $\hat{\theta}(\underline{X}_{Nij})\ \forall\ (i,j|i \neq j)$ and $\hat{F}_{Ni}, i = 1,2,\ldots,k$
        iv.  Go to 3
    (c) End If
```

This algorithm was used to perform ranking and selection based on both means and 0.9 quantiles for three systems. The distribution of all three systems was normal, and as with the experimentation in Section 3.1, the d parameter was set to either 0.3 or 0.45. For all three experiments, the bootstrapping procedure was repeated 10,000 times at each sample size, the desired probability of correct selection was set to 0.95, and unique random number streams were used to generate samples from each system, making the resulting confidence intervals independent. Also, the entire procedure was repeated 50 times. The results of this empirical analysis are shown in Table 4.

Table 4: Empirical results to demonstrate the use of fixed-width confidence intervals for ranking and selection between three systems. System and parameter descriptions for each experiment are given.

Exp. #	θ	n_0	System 1	System 2	System 3	d	PCS	Mean Sample Size (s.e.)
1	Mean	50	$N(10,2)$	$N(11,2)$	$N(12,2)$	0.3	1.0	483.6 (2.60)
2	Mean	50	$N(10,3)$	$N(11,2)$	$N(12,1)$	0.45	1.0	265.82 (2.03)
3	0.9 Quantile	200	$N(10,2)$	$N(11,2)$	$N(12,2)$	0.3	1.0	1389.44 (29.65)
4	0.9 Quantile	200	$N(10,3)$	$N(12,2)$	$N(14,1)$	0.3	1.0	1666.28 (61.98)

To generate the results in Table 4, simultaneous confidence intervals of coverage probability $1-\alpha$ on all pairs of differences amongst the three systems were generated. In all experiments, system 3 has the largest θ value. Consequently, the reported PCS represents the number of times in which $\hat{\theta}_3$ was the largest of the three $\hat{\theta}_i, i = 1,2,3$ values. The PCS values reported in Table 4 show that the correct system was selected in all replications of all experiments, lending support to the approach of using fixed-width confidence intervals for bootstrapped ranking and selection on multiple types of distributional properties. Of note is that the reported PCS is not the same as whether the confidence intervals on the pairs of differences actually all simultaneously contained the correct values; this value is the simultaneous coverage probability and is reported along with its standard error in Table 5. The experiment numbers in Table 4 (Exp. #) correspond to the experiment numbers in Table 5 and indicate that the desired simultaneous coverage probability is returned in all experiments.

The results in Table 4 also give the mean and standard error of the sample size required across all replications for each experiment. These results demonstrate again that the estimation of quantiles requires a greater sample size than the estimation of means. Additionally, the results demonstrate that as the θ values get closer to each other, the required sample size increases. This is evidenced by the fact that in Experiment 4, the 0.9 quantiles are 0.718 units apart (the 0.9 quantile from system 2 is 0.718 units larger than the same quantile from system 1, and the quantile from system 3 is 0.718 units larger than the quantile in system 2), and the required sample size is 1666. Comparatively, in Experiment 3, where the 0.9 quantiles are 1.0 units apart, the required sample size is only 1389.

Table 4 also illustrates that as more systems are involved in the ranking and selection (and, therefore, a greater number of simultaneous confidence intervals must simultaneously have correct coverage in order to make $P^*_{NL} \geq 1-\alpha$), the required sample sizes increase. For example, when θ represents the mean and $k=2$ systems, the mean sample size given in Table 3 was only 331, but when a third system is introduced (as in Experiment 1 in Table 4), the sample size increases to 483. Similarly large increases in sample size are noted for the case in which θ represents the 0.9 quantile. These extra sample sizes increase the run-time for the procedure, but also may contribute to the high returned PCS. Empirical evidence into the impact on sample size and PCS when $k > 3$ systems should be obtained as part of future work.

4 CONCLUSIONS AND FUTURE WORK

A ranking and selection procedure that allows comparisons to be made based on any distributional property of interest from the populations under investigation would be of great use. This paper presented initial work toward the development of such a procedure. Swanepoel, van Wyk and Venter

Table 5: Simultaneous coverage probability for all pairs of differences in the experiments detailed in Table 4.

Exp. #	Simultaneous Coverage Probability (s.e.)
1	0.96 (.01)
2	0.96 (.01)
3	0.96 (.01)
4	0.96 (.01)

(1983) give a method for using bootstrapping to develop fixed-width confidence intervals with a specified coverage probability around a distributional property of interest. Empirical evidence was provided in support of the use of their approach. Specifically, results showed that when resampling was performed enough times and the initial sample size was large enough, the approach was effective for generating fixed-width confidence intervals around both the mean and the 0.7 and 0.9 quantiles.

The fixed-width confidence interval approach was then used to perform bootstrapped ranking and selection. For two systems, the approach required only building a fixed-width confidence interval around the difference between the θ in the two systems. The half width of the confidence interval is analogous to the indifference zone parameter of traditional ranking and selection procedures. When bootstrapped ranking and selection is performed on more than two systems, it was shown that simultaneous two-sided confidence intervals of coverage probability $1-\alpha$ and fixed-width $2d$ should first be built around all pairs of differences, $\theta_i - \theta_j \ \forall \ (i,j|i \neq j)$. The system with the largest $\hat{\theta}_i, i = 1,2,\ldots k$ is then selected as the best system and is guaranteed to either be the best system or a system within d of the best. For both $k = 2$ and $k = 3$ systems, bootstrapped ranking and selection based on fixed-width confidence intervals was shown to be appropriate for making selections based both on means and on 0.9 quantiles.

This paper, however, represents only an initial step toward the development of a procedure for determining the earliest point at which a sampling experiment (e.g., a single replication of each of k DES models, each corresponding to a competing system) can be stopped so that the correct system (the one with the largest θ) is chosen at least as often as the specified PCS. To reach that objective, much still needs to be done. Specifically, additionally work should be done with regard to determining the initial sample size required by the bootstrapping procedure used to develop the fixed-width confidence intervals. This work will help ensure that $\hat{\theta}$ is a good estimator of θ so that the confidence intervals actually return the desired coverage probabilities. Additionally, the approach presented by Swanepoel, van Wyk and Venter (1983) samples one additional observation at a time. This solution is not practical, and investigations should be made to determine a more efficient method for increasing the sample size between bootstrapping replications. Also of note is that the systems under investigation in this paper had only mild variability and typically had θ values far apart from each other when compared to the d parameter. Consequently, future work should also include validation of the approach under more extreme conditions. Lastly, the choices of the L and n_0 parameters were shown in Section 2 to be important. The investigated values of L, 200 and 10,000, are very different; it is conceivable that a value between those two extremes would produce acceptable results and would provide results more quickly, making it a more useful procedure in practice. Additionally, an appropriate choice of n_0 was shown to vary based on the specific system under investigation as well as the d parameter. Therefore, further investigations in to the n_0, L, and d parameters, as well as the relationship between them, is an important component of future work on the procedure.

Finally, the methods presented thus far also assume that the sampling is being performed on i.i.d. data. However, when the procedure is applied to DES based on queueing systems, the data used to generate $\hat{F}_i$ will not be i.i.d. Suggestions for how to implement bootstrapping procedures on weakly dependent stationary data are found within the literature. Künsch (1989) and Liu and Singh (1992) first introduced the moving block bootstrap, for example. In this approach, blocks of a fixed number of successive observations are resampled (vs. the individual observations themselves). The drawback of such an approach is that both the bias and the variance of the estimators are very dependent on the size of the blocks (Lahiri, 1999), making the selection of block size very important. Additionally, the resampled data series generated from the moving block bootstrap no longer maintains the stationarity of the original data set (Politis and Romano, 1994). Politis and Romano (1994) present the stationary bootstrap, which guarantees stationarity in the resampled data.

In the stationary bootstrapping approach, in each bootstrapping replication blocks of a random number of consecutive observations are wrapped together to create pseudo time series, from which the estimate of θ is taken. Further investigation into how/whether these bootstrapping approaches can be into the procedure should be performed as part of future work.

REFERENCES

Batur, D. and Choobineh, F. 2010. A quantile-based approach to system selection, *European Journal of Operational Research* 3: 764-772.

Bekki, J.M., Fowler, J.W., Mackulak, G.T., and Nelson, B.L. 2007. Using quantiles in ranking and selection procedures. In *Proceedings of the 2007 Winter Simulation Conference*, ed. S. G. Henderson, B. Biller, M.-H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 1722-1728. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Efron, B. and Tibshirani, R.J. 1993. *An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability)*, New York, New York: Chapman & Hall.

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

Kim, S.-H. and Nelson, B.L. 2006. Selecting the best system. In *Handbooks in Operations Research and Management Science, Volume 13: Simulation*, ed. S. Henderson and B.L. Nelson, Oxford, UK: Elseiver.

Künsch, Hans R. 1989. The jackknife and the bootstrap for general stationary observations, *The Annals of Statistics* 17(9):1217-1241.

Lahiri, S.N. 1999. Theoretical comparisons of block bootstrap methods, *The Annals of Statistics* 27(1):386 - 404.

Liu, R.Y. and Singh, K. 1992. Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap*, ed. R. LePage and L. Billard, New York, New York: John Wiley.

Politis, Dimitris N. and Romana, Joseph P. 1994. The stationary bootstrap, *Journal of the American Statistical Association* 89(428):1303-1313.

Swanepoel, J.W.H., van Wyk, J.W.J., and Venter, J.H. 1983. Fixed width confidence intervals based on bootstrap procedures, *Sequential Analysis* 2(4):289-310.

AUTHOR BIOGRAPHIES

JENNIFER M. BEKKI is an Assistant Professor in the Department of Engineering at Arizona State University. She received her BSE in Bioengineering and MSE and PhD in Industrial Engineering from Arizona State University. As a faculty member in the Department of Engineering, she is taking part in the continued development and refinement of an innovative, multi-disciplinary undergraduate engineering program that uses research-based approaches for pedagogy and curricular design. Dr. Bekki's research interests are in discrete event simulation methodology, applied operations research (particularly in the semiconductor industry), and, more recently, the application of research-based assessment and pedagogy to topics in engineering education. Her email address is <jennifer.bekki@asu.edu>.

BARRY L. NELSON is the Charles Deering McCormick Professor and Chair of the Department of Industrial Engineering & Management Sciences at Northwestern University. His research interests are the design and analysis of stochastic simulation experiments, particularly issues of multivariate input modeling, optimization via simulation and metamodeling. He has been Editor in Chief of *Naval Research Logistics*, has served on the Board of Directors of the Winter Simulation Conference, and is a member of IIE (senior member), INFORMS (fellow), ACM, and ASA. His email address is <nelsonb@northwestern.edu>.

JOHN W. FOWLER is the Avnet Professor of Supply Networks and a Professor of Industrial Engineering at Arizona State University (ASU). His research interests include modeling, analysis, and control of manufacturing and service systems. His research has been supported by the National Science Foundation, Semiconductor Research Corp., International SEMATECH, Asyst, IBM, Intel, Motorola, Infineon Technologies, ST Microelectronics, and the Mayo Clinic. Dr. Fowler is an author on over 70 journal publications, 100 conference papers, and 10 book chapters. He is the founding editor of the new journal *IIE Transactions on Healthcare Systems Engineering*. He is also an Area Editor for *SIMULATION: Transactions of the Society for Modeling and Simulation International*, an Associate Editor for *IEEE Transactions on Semiconductor Manufacturing*, and on the Editorial Board for the *Journal of Simulation*. He is a Fellow of the Institute of Industrial Engineers, was recently the INFORMS Vice President for Chapters/Fora, and is on the Winter Simulation Conference Board of Directors. His email address is <john.fowler@asu.edu>.