Proceedings of the 1999 Winter Simulation Conference
P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, eds.

EVALUATING THE PROBABILITY OF A GOOD SELECTION

Barry L. Nelson
Souvik Banerjee

Department of Industrial Engineering & Management Sciences
Northwestern University
Evanston, IL 60208-3119, U.S.A.

ABSTRACT

We present a two-stage experiment design for use in simulation experiments that compare systems in terms of their expected (long-run average) performance. This procedure simultaneously achieves the following with a prespecified probability of being correct: (a) find the best system or a near best system; (b) identify a subset of systems that are more than a practically insignificant difference from the best; and (c) provide a lower bound on the probability that the best or near best system has actually been selected. The procedure assume normally distributed data, but allows unequal variances.

1 INTRODUCTION

In this paper we address problems that arise in the *reporting and interpretation of simulation experiments performed to identify the best system*, where best means maximum or minimum expected (long-run average) performance. The procedure we derive allows the simulation analyst to achieve the following goals, all with prespecified probability of being correct:

1. Design their experiment so as to find the best system, or one within a practically insignificant difference from the best system (we refer to this as a "good selection").
2. Bound the difference between each system and the best system, and thereby eliminate all systems that are more than the practically insignificant difference from the best.
3. Revise the prespecified probability of a good selection upward, based on the results of the experiment, and also give a lower bound on the probability that the unique best system has been selected.

Many ranking and selection procedures exist that achieve goal 1 (see, for instance, Bechhofer, Santner and Goldsman 1995, or Goldsman and Nelson 1998). Multiple comparison procedures, specifically multiple comparisons with the best (MCB), can satisfy goal 2 (Hsu 1996). However, the bounds provided by standard MCB procedures are difficult to interpret because they are *constrained* confidence intervals: each interval either contains 0 or has 0 as one end point. A 0 endpoint means that a system can be declared either "no better than the best" or "no different from the best," depending on which end point it is. This subtly is confusing to many analysts. We solve the problem by providing fixed-width, unconstrained MCB intervals, a small extension to existing theory. An important use of such intervals is to eliminate from further consideration all systems that are clearly inferior to the best system.

A more fundamental contribution is made by addressing goal 3. We do this by providing *a lower confidence bound on the achieved probability of a good selection (PGS)* after all of the simulation data have been realized. Since the procedure guarantees a PGS of at least a prespecified nominal level, our lower confidence bound will revise this value upward if the data so indicate. We can also provide a lower confidence bound on the probability that the *unique* best system has been selected, which we refer to as the probability of a correct selection (PCS).

The concept of a revised probability of a good selection is similar in spirit to Hsu's (1984) S-value, which is the smallest confidence level at which the sample best system would be declared to be the true best system. In our case we design the experiment to achieve a given, nominal probability of a good selection that we hope to revise upward if the simulation results are favorable.

The paper is organized as follows: We first describe how lower confidence bounds on PCS and PGS can be obtained in general. These lower confidence bounds on PCS and PGS depend on a lower confidence bound on the difference between the best system and each inferior system; we obtain those bounds in Section 3. Section 4 introduces a specific procedure for reevaluating PCS and PGS after sampling. An illustrative example is provided in Section 5.

2 ASSESSING PGS

Throughout this paper we assume that larger expected performance implies a better system. Let $\mu_1, \mu_2, \ldots, \mu_k$ denote the unknown means of the k systems to be compared, and let $\widehat{\mu}_i$ denote our point estimator of μ_i. In this section we assume only that the distribution of $\widehat{\mu}_i - \widehat{\mu}_\ell - (\mu_i - \mu_\ell)$ does not depend on $\mu_1, \mu_2, \ldots, \mu_k$.

Denote the ordered means by $\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[k-1]} < \mu_{[k]}$, and for the moment suppose that our goal is to find the unique best system $[k]$; later we address the problem of finding either system $[k]$ or a system $[i]$ whose true mean is close enough to $\mu_{[k]}$.

Our rule will be to select the system with the largest performance estimate, $\widehat{\mu}_i$. Therefore, the probability of a (unique) correct selection is

$$\begin{aligned} &\text{PCS} \\ &= \Pr\left\{\widehat{\mu}_{[i]} < \widehat{\mu}_{[k]}, i = 1, 2, \ldots, k-1\right\} \\ &= \Pr\left\{\widehat{\mu}_{[i]} - \widehat{\mu}_{[k]} - (\mu_{[i]} - \mu_{[k]}) \right. \\ &\quad \left. < \mu_{[k]} - \mu_{[i]}, i = 1, 2, \ldots, k-1\right\} \\ &= \Pr\left\{D_i < \mu_{[k]} - \mu_{[i]}, i = 1, 2, \ldots, k-1\right\} \quad (1) \end{aligned}$$

where $D_i, i = 1, 2, \ldots, k-1$ has the same joint distribution as $\widehat{\mu}_{[i]} - \widehat{\mu}_{[k]} - (\mu_{[i]} - \mu_{[k]}), i = 1, 2, \ldots, k-1$. If the values of the differences $\mu_{[k]} - \mu_{[i]}$ were known, as well as the joint distribution of $D_1, D_2, \ldots, D_{k-1}$, then (1) might be evaluated exactly. Since this is impossible in practice, Kim (1986) suggested replacing $\mu_{[k]} - \mu_{[i]}$ in (1) with $(1-\alpha)100\%$ lower confidence bounds on these differences, thereby providing a $(1-\alpha)100\%$ lower confidence bound (LCB) on the achieved probability of correct selection when (1) is evaluated (see also Anderson, Bishop and Dudewicz 1977). Kim (1986) was only able to provide lower confidence bounds on the single difference $\mu_{[k]} - \mu_{[k-1]}$, whereas we will provide bounds on all $k-1$ differences leading to a much tighter LCB on PCS.

One shortcoming of our proposal is that the LCB on PCS can be small when there are one or more systems whose performance is very close to the best, making a unique correct selection unlikely. Thus, it makes sense to provide a lower confidence bound on choosing the best system or a system whose mean is within a practically insignificant difference δ of the best. That is, we want to select a system i such that $\mu_{[k]} - \mu_i \leq \delta$. We call this event a "good" selection, and let PGS denote the probability of a good selection.

Nelson and Banerjee (1999) show that for the procedure we derive

$$\begin{aligned} &\text{PGS} \\ &= \Pr\left\{\widehat{\mu}_{[i]} - \widehat{\mu}_{[k]} - (\mu_{[i]} - \mu_{[k]}) \right. \\ &\quad \left. < \max[\delta, \mu_{[k]} - \mu_{[i]}], i = 1, 2, \ldots, k-1\right\} \quad (2) \\ &= \Pr\left\{D_i < \max[\delta, \mu_{[k]} - \mu_{[i]}], i = 1, 2, \ldots, k-1\right\} \end{aligned}$$

no matter what the configuration of the true means. Our approach will be to substitute LCBs for $\mu_{[k]} - \mu_{[i]}$ in (2) and evaluate this probability numerically.

3 TWO-SIDED MCB

The following lemma will be useful for deriving two-sided, fixed-width MCB confidence intervals:

Lemma 1 (Hsu, 1996, Section 4.2.1) *If*

$$\Pr\{\widehat{\mu}_i - \widehat{\mu}_\ell - (\mu_i - \mu_\ell) < \delta, \forall i \neq \ell\} \geq 1 - \alpha \quad (3)$$

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

$$\mu_i - \max_{\ell \neq i} \mu_\ell \in \left[\widehat{\mu}_i - \max_{\ell \neq i} \widehat{\mu}_\ell \pm \delta\right]$$

for $i = 1, 2, \ldots, k$.

Remark: Notice that (3) will hold when we can form simultaneous two-sided confidence intervals for all-pairwise differences $\mu_i - \mu_\ell$ that take the form $\widehat{\mu}_i - \widehat{\mu}_\ell \pm \delta$. For instance, in the usual one-way analysis of variance model with normally distributed data and equal variances, setting $\delta = q^{(1-\alpha)}_{k,k(n-1)} S_p/\sqrt{n}$—where $q^{(1-\alpha)}_{k,k(n-1)}$ is the $1-\alpha$ quantile of the studentized range distribution of dimension k and degrees of freedom $k(n-1)$, and S_p^2 is the usual pooled variance estimator—achieves (3). This is the procedure given by Hsu (1996, pp. 103–104). Below we propose a two-stage procedure that allows unequal variances across systems, and also allows the value of δ to be specified in advance.

Let $T_1, T_2, \ldots, T_k$ be independent t random variables, each with ν degrees of freedom, and define the random variable $R = \max_i T_i - \min_i T_i$. Let $r_\nu^{(1-\alpha)}$ be the $1-\alpha$ quantile of R. The quantity r will be the critical constant in our two-stage procedure. See Tables 1, 2 and 3 for numerical values; these values were obtained via simulation and almost all are accurate to the second decimal place.

Table 1: Critical Values $r_\nu^{0.9}$

$\nu \backslash k$	3	4	5	6	7	8	9	10	11	12	13	14	15
1	19.8	27.1	34.5	41.1	48.5	55.9	62.5	68.8	75.8	81.5	90.8	96.8	103.70
2	6.23	7.54	8.60	9.50	10.40	11.21	11.84	12.51	13.29	13.77	14.48	15.03	15.59
3	4.65	5.35	5.92	6.48	6.93	7.30	7.73	8.03	8.36	8.64	8.91	9.08	9.45
4	4.01	4.61	5.10	5.47	5.78	6.07	6.35	6.60	6.79	6.98	7.21	7.37	7.53
5	3.73	4.27	4.66	4.99	5.25	5.49	5.73	5.92	6.10	6.26	6.40	6.54	6.67
6	3.55	4.05	4.42	4.69	4.96	5.17	5.35	5.51	5.65	5.81	5.93	6.06	6.19
7	3.45	3.91	4.23	4.52	4.74	4.94	5.11	5.28	5.40	5.53	5.65	5.75	5.86
8	3.37	3.83	4.13	4.38	4.61	4.80	4.96	5.10	5.22	5.34	5.45	5.56	5.65
9	3.32	3.74	4.07	4.29	4.50	4.69	4.82	4.96	5.09	5.21	5.31	5.40	5.48
10	3.26	3.68	3.99	4.21	4.42	4.60	4.73	4.86	4.99	5.09	5.19	5.28	5.36
11	3.23	3.64	3.94	4.16	4.36	4.54	4.65	4.79	4.90	5.01	5.10	5.19	5.28
12	3.19	3.59	3.89	4.12	4.30	4.48	4.60	4.72	4.84	4.93	5.02	5.11	5.19
13	3.16	3.57	3.84	4.09	4.26	4.41	4.56	4.68	4.79	4.87	4.98	5.06	5.14
14	3.16	3.55	3.83	4.04	4.22	4.38	4.51	4.62	4.74	4.84	4.92	4.99	5.08
15	3.13	3.52	3.79	4.02	4.20	4.36	4.48	4.59	4.71	4.79	4.88	4.96	5.04
16	3.11	3.51	3.77	4.00	4.17	4.33	4.46	4.56	4.67	4.76	4.83	4.93	4.98
17	3.12	3.47	3.75	3.97	4.15	4.29	4.42	4.54	4.63	4.73	4.81	4.89	4.96
18	3.09	3.47	3.74	3.96	4.12	4.27	4.39	4.51	4.61	4.71	4.77	4.85	4.93
19	3.08	3.44	3.72	3.93	4.10	4.27	4.38	4.50	4.57	4.68	4.74	4.82	4.91
20	3.08	3.45	3.71	3.91	4.09	4.24	4.36	4.47	4.56	4.66	4.74	4.81	4.87
30	3.01	3.36	3.63	3.82	3.99	4.12	4.27	4.34	4.44	4.53	4.60	4.66	4.72
40	2.98	3.35	3.59	3.77	3.94	4.08	4.19	4.29	4.38	4.46	4.53	4.60	4.66
50	2.97	3.32	3.57	3.76	3.91	4.06	4.17	4.26	4.34	4.42	4.50	4.56	4.62
60	2.96	3.30	3.56	3.75	3.90	4.02	4.14	4.23	4.32	4.42	4.48	4.54	4.60
70	2.95	3.30	3.55	3.73	3.89	4.02	4.12	4.23	4.30	4.38	4.45	4.51	4.57
80	2.95	3.30	3.53	3.73	3.88	4.00	4.10	4.21	4.30	4.37	4.44	4.51	4.56
90	2.94	3.28	3.53	3.71	3.88	3.99	4.10	4.20	4.29	4.36	4.44	4.50	4.55
100	2.93	3.28	3.52	3.72	3.86	3.99	4.11	4.20	4.28	4.36	4.43	4.49	4.54

Table 2: Critical Values $r_\nu^{0.95}$

$\nu \backslash k$	3	4	5	6	7	8	9	10	11	12	13	14	15
1	39.4	53.2	66.5	81.4	96.2	110.7	122.0	132.7	149.0	158.4	176.2	189.2	201.40
2	8.72	10.37	11.84	13.03	14.17	15.30	16.30	17.00	18.15	18.66	19.55	20.36	21.10
3	5.91	6.71	7.42	8.05	8.56	9.03	9.51	9.84	10.22	10.60	10.91	11.06	11.58
4	4.90	5.58	6.10	6.53	6.89	7.19	7.48	7.74	8.02	8.18	8.47	8.64	8.79
5	4.48	5.09	5.49	5.83	6.11	6.37	6.62	6.83	7.02	7.20	7.35	7.49	7.63
6	4.23	4.75	5.13	5.42	5.69	5.89	6.10	6.26	6.40	6.57	6.70	6.84	6.98
7	4.05	4.53	4.87	5.18	5.40	5.59	5.76	5.96	6.09	6.20	6.33	6.45	6.54
8	3.95	4.43	4.72	4.99	5.21	5.41	5.57	5.69	5.86	5.97	6.07	6.19	6.27
9	3.89	4.30	4.61	4.84	5.06	5.25	5.39	5.53	5.67	5.76	5.88	5.98	6.08
10	3.79	4.21	4.52	4.75	4.96	5.12	5.29	5.43	5.51	5.63	5.74	5.82	5.91
11	3.76	4.16	4.45	4.67	4.88	5.03	5.19	5.32	5.41	5.51	5.62	5.70	5.80
12	3.70	4.10	4.40	4.63	4.79	4.97	5.10	5.22	5.33	5.44	5.53	5.61	5.69
13	3.67	4.07	4.33	4.59	4.76	4.90	5.05	5.15	5.27	5.36	5.46	5.53	5.62
14	3.67	4.05	4.31	4.52	4.71	4.87	4.99	5.10	5.20	5.32	5.38	5.47	5.55
15	3.62	4.00	4.27	4.50	4.66	4.83	4.95	5.04	5.16	5.26	5.34	5.41	5.50
16	3.60	3.97	4.25	4.47	4.63	4.78	4.93	5.02	5.12	5.21	5.28	5.38	5.44
17	3.60	3.96	4.22	4.42	4.61	4.74	4.88	4.99	5.08	5.18	5.25	5.32	5.40
18	3.56	3.93	4.20	4.41	4.57	4.72	4.84	4.95	5.05	5.14	5.22	5.30	5.36
19	3.55	3.90	4.17	4.37	4.54	4.70	4.81	4.94	5.00	5.12	5.18	5.25	5.33
20	3.54	3.90	4.16	4.36	4.51	4.69	4.80	4.90	4.98	5.07	5.15	5.23	5.30
30	3.45	3.80	4.04	4.26	4.40	4.53	4.66	4.73	4.83	4.93	4.99	5.05	5.11
40	3.43	3.77	4.01	4.17	4.34	4.47	4.58	4.66	4.77	4.83	4.90	4.98	5.03
50	3.40	3.74	3.97	4.17	4.30	4.44	4.55	4.64	4.72	4.79	4.87	4.93	4.99
60	3.38	3.72	3.97	4.14	4.29	4.41	4.50	4.60	4.70	4.78	4.85	4.90	4.94
70	3.37	3.71	3.95	4.12	4.28	4.40	4.48	4.60	4.67	4.74	4.81	4.87	4.92
80	3.36	3.71	3.93	4.12	4.26	4.37	4.47	4.57	4.66	4.73	4.80	4.86	4.91
90	3.35	3.68	3.93	4.10	4.25	4.37	4.47	4.57	4.64	4.71	4.78	4.85	4.90
100	3.36	3.68	3.91	4.09	4.23	4.35	4.47	4.55	4.64	4.72	4.79	4.84	4.88

Table 3: Critical Values $r_\nu^{0.99}$

$\nu \backslash k$	3	4	5	6	7	8	9	10	11	12	13	14	15
1	200.8	271.3	330.1	392.6	478.7	536.2	585.0	638.8	709.0	738.0	858.8	954.4	967.10
2	18.89	21.50	24.47	26.45	29.42	33.07	33.49	35.02	37.79	38.08	39.70	42.23	43.82
3	9.84	10.88	12.00	13.01	13.43	14.27	15.15	15.61	16.06	16.75	17.27	17.54	18.19
4	7.22	8.14	8.88	9.49	9.98	10.22	10.57	10.89	11.40	11.47	11.93	12.04	12.29
5	6.34	7.07	7.49	7.95	8.30	8.59	8.88	9.15	9.38	9.64	9.76	10.11	10.13
6	5.80	6.42	6.77	7.15	7.45	7.68	7.98	8.12	8.28	8.45	8.64	8.76	8.89
7	5.45	6.01	6.39	6.67	6.92	7.13	7.33	7.55	7.68	7.84	7.93	8.13	8.18
8	5.27	5.76	6.07	6.34	6.62	6.78	7.01	7.10	7.29	7.36	7.56	7.65	7.73
9	5.11	5.56	5.87	6.08	6.34	6.52	6.72	6.84	6.98	7.09	7.19	7.34	7.38
10	4.99	5.41	5.71	5.93	6.15	6.32	6.51	6.64	6.76	6.81	6.98	7.04	7.13
11	4.91	5.27	5.58	5.78	6.00	6.15	6.35	6.46	6.58	6.61	6.77	6.84	6.99
12	4.80	5.19	5.49	5.71	5.87	6.04	6.22	6.31	6.42	6.53	6.61	6.73	6.80
13	4.76	5.12	5.38	5.66	5.81	5.93	6.11	6.20	6.34	6.41	6.50	6.61	6.67
14	4.69	5.06	5.34	5.55	5.71	5.93	5.97	6.10	6.22	6.35	6.41	6.48	6.57
15	4.65	5.01	5.28	5.50	5.65	5.83	5.96	6.04	6.14	6.23	6.31	6.41	6.48
16	4.63	4.94	5.25	5.42	5.59	5.78	5.91	5.98	6.09	6.17	6.21	6.33	6.38
17	4.61	4.94	5.17	5.39	5.57	5.68	5.82	5.92	6.02	6.12	6.16	6.28	6.31
18	4.56	4.91	5.17	5.37	5.51	5.65	5.76	5.91	5.97	6.04	6.15	6.21	6.31
19	4.53	4.88	5.12	5.32	5.48	5.63	5.73	5.85	5.94	6.03	6.06	6.14	6.23
20	4.51	4.85	5.04	5.28	5.45	5.58	5.70	5.79	5.88	5.97	6.00	6.11	6.19
30	4.37	4.67	4.91	5.09	5.22	5.40	5.47	5.55	5.64	5.74	5.82	5.87	5.90
40	4.32	4.59	4.84	4.98	5.15	5.27	5.38	5.46	5.53	5.61	5.66	5.74	5.79
50	4.28	4.58	4.79	4.96	5.08	5.23	5.31	5.38	5.46	5.54	5.62	5.67	5.72
60	4.25	4.55	4.76	4.93	5.07	5.17	5.23	5.35	5.44	5.51	5.55	5.61	5.67
70	4.24	4.51	4.75	4.92	5.04	5.16	5.23	5.33	5.40	5.45	5.57	5.57	5.62
80	4.19	4.53	4.70	4.88	5.02	5.11	5.24	5.29	5.39	5.45	5.51	5.57	5.62
90	4.20	4.50	4.70	4.85	5.01	5.13	5.21	5.30	5.37	5.42	5.48	5.57	5.60
100	4.20	4.50	4.70	4.86	4.97	5.09	5.20	5.25	5.36	5.42	5.49	5.54	5.59

3.1 Procedure

Consider the following algorithm for producing fixed-width confidence intervals for all pairwise comparisons when the data are normally distributed and independent:

3.1.1 Fixed-Width, All-Pairwise Comparisons

1. Specify confidence level $1-\alpha$, halfwidth $\delta > 0$, and initial sample size $n_0 \geq 2$.
2. Sample i.i.d. observations $Y_{i1}, Y_{i2}, \ldots, Y_{in_0}$ from all systems $i = 1, 2, \ldots, k$, and compute the sample variances

$$S_i^2 = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} (Y_{ij} - \bar{Y}_i)^2$$

for $i = 1, 2, \ldots, k$.

3. Determine the total sample size needed by letting

$$N_i = \max \left\{ n_0 + 1, \left\lceil \left(\frac{r_{n_0-1}^{(1-\alpha)} S_i}{\delta} \right)^2 \right\rceil \right\}. \qquad (4)$$

4. From system i, for $i = 1, 2, \ldots, k$, take additional samples $Y_{i,n_0+1}, Y_{i,n_0+2}, \ldots, Y_{i,N_i}$.

5. Compute the generalized sample mean

$$\widehat{\mu}_i = \sum_{j=1}^{N_i} \beta_{ij} Y_{ij}$$

where for each i the β_{ij} are chosen such that $\beta_{i1} = \beta_{i2} = \cdots = \beta_{in_0}$, $\sum_{j=1}^{N_i} \beta_{ij} = 1$, and

$$S_i^2 \sum_{j=1}^{N_i} \beta_{ij}^2 = \left(\frac{\delta}{r_{n_0-1}^{(1-\alpha)}} \right)^2.$$

6. Report the simultaneous confidence intervals

$$\mu_i - \mu_\ell \in [\widehat{\mu}_i - \widehat{\mu}_\ell \pm \delta]$$

for all $i \neq \ell$.

A proof that the intervals in step 3.1.1 are indeed simultaneous $(1-\alpha)100\%$ confidence intervals when the simulation output data are normally distributed can be found in Hochberg and Tamhane (1987, pp. 200-201). In fact, this procedure produces simultaneous confidence intervals for all contrasts involving $\mu_1, \mu_2, \ldots, \mu_k$. Application of

lemma 1 immediately yields the desired two-sided MCB confidence intervals

$$\mu_i - \max_{\ell \neq i} \mu_\ell \in \left[\widehat{\mu}_i - \max_{\ell \neq i} \widehat{\mu}_\ell \pm \delta \right] \tag{5}$$

for $i = 1, 2, \ldots, k$.

3.2 Inference

What inference is possible at the end of the procedure defined above? The MCB intervals (5) imply that $\mu_i - \max_{\ell \neq i} \mu_\ell \leq \widehat{\mu}_i - \max_{\ell \neq i} \widehat{\mu}_\ell + \delta$. Therefore, if this upper bound is less than 0 we can infer that system i is inferior to the best; if this upper bound is less than $-\delta$, and we have chosen δ so that differences greater than δ are practically significant, then we can eliminate all such systems from further consideration.

Let $B = \text{argmax } \widehat{\mu}_i$; that is, B is the index of the system selected as best. Notice that we can also claim, with probability $\geq 1 - \alpha$, that

$$\mu_B - \max_{\ell \neq B} \mu_\ell \geq \widehat{\mu}_B - \max_{\ell \neq B} \widehat{\mu}_\ell - \delta \geq -\delta.$$

Thus, with confidence level $1 - \alpha$ we are assured that we have made a good selection in that the mean of the selected system is within δ of the true best mean. Stated differently, the event

$$\mathcal{A} = \{\widehat{\mu}_i - \widehat{\mu}_\ell - (\mu_i - \mu_\ell) < \delta, \forall i \neq \ell\}$$

implies a good selection will be made. In fact, even less is required. Nelson and Goldsman (1998) show that

$$\mathcal{B} = \left\{\widehat{\mu}_{[i]} - \widehat{\mu}_{[k]} - (\mu_{[i]} - \mu_{[k]}) < \delta, i = 1, 2, \ldots, k-1\right\}$$

is sufficient to guarantee a good selection. Clearly $\mathcal{A} \Rightarrow \mathcal{B}$.

Now consider the event

$$\begin{aligned} \mathcal{B}' \quad = \quad & \left\{\widehat{\mu}_{[i]} - \widehat{\mu}_{[k]} - (\mu_{[i]} - \mu_{[k]}) < \max[\delta, \mu_{[k]} - \mu_{[i]}],\right. \\ & \left. i = 1, 2, \ldots, k-1\right\}. \end{aligned}$$

Clearly $\mathcal{B} \Rightarrow \mathcal{B}'$, so $\Pr\{\mathcal{B}'\} \geq \Pr\{\mathcal{B}\}$. It is also the case the $\mathcal{B}'$ implies a good selection will be made:

- If $\mu_{[i]} < \mu_{[k]} - \delta$, so that $[i]$ is not a good selection, then $\mathcal{B}'$ implies that $\widehat{\mu}_{[i]} < \widehat{\mu}_{[k]}$ and system $[i]$ will not be selected.
- If $\mu_{[i]} \geq \mu_{[k]} - \delta$, so that $[i]$ is a good selection, then $\mathcal{B}'$ implies that $\widehat{\mu}_{[i]} - \widehat{\mu}_{[k]} - (\mu_{[i]} - \mu_{[k]}) \leq \delta$ for all such $[i]$. This is precisely the event $\mathcal{B}$ that guarantees a good selection will be made.

Thus, PGS $\geq \Pr\{\mathcal{B}'\}$ and our goal becomes obtaining a lower confidence bound on $\Pr\{\mathcal{B}'\}$. We achieve this by substituting $(1 - \alpha)100\%$ simultaneous lower confidence bounds of the form

$$\mu_{[k]} - \mu_i \geq \widehat{\mu}_B - \widehat{\mu}_i - \delta$$

for all $i \neq B$ into $\mathcal{B}'$, giving a $(1-\alpha)100\%$ lower confidence bound on PGS of

$$\Pr\{D_i < \max[\delta, \widehat{\mu}_B - \widehat{\mu}_i - \delta], i = 1, 2, \ldots, k-1\}. \tag{6}$$

These bounds are based on the fact that the two-sided MCB intervals (5) imply that

$$\max_{\ell \neq i} \mu_\ell - \mu_i \geq \max_{\ell \neq i} \widehat{\mu}_\ell - \widehat{\mu}_i - \delta$$

for $i = 1, 2, \ldots, k$ with probability $\geq 1 - \alpha$.

4 COMPUTING A LCB ON PGS

In this section we show to evaluate the right-hand side of (6) for the two-stage procedure. To obtain a lower confidence bound on PCS we simply replace $\max[\delta, \mu_{[k]} - \mu_{[i]}]$ by $\mu_{[k]} - \mu_{[i]}$ in the derivation.

Let $\xi = \delta / r_{n_0-1}^{(1-\alpha)}$. Then

$$\begin{aligned} & \text{PGS} \\ & \geq \Pr\left\{\widehat{\mu}_{[i]} - \widehat{\mu}_{[k]} - (\mu_{[i]} - \mu_{[k]}) < \right. \\ & \qquad \left. \max[\delta, \mu_{[k]} - \mu_{[i]}], i = 1, 2, \ldots, k-1\right\} \\ & = \Pr\left\{\frac{\widehat{\mu}_{[i]} - \mu_{[i]}}{\xi} \leq \frac{\widehat{\mu}_{[k]} - \mu_{[k]}}{\xi} + \right. \\ & \qquad \left. \frac{\max[\delta, \mu_{[k]} - \mu_{[i]}]}{\xi}, i = 1, 2, \ldots, k-1\right\} \\ & = \Pr\left\{T_i \leq T_k + \frac{\max[\delta, \mu_{[k]} - \mu_{[i]}]}{\xi}, \right. \\ & \qquad \left. i = 1, 2, \ldots, k-1\right\} \\ & = \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} F_{n_0-1}\left(t + \max[\delta, \mu_{[k]} - \mu_{[i]}]/\xi\right) \\ & \qquad \times dF_{n_0-1}(t) \end{aligned} \tag{7}$$

where $T_1, T_2, \ldots, T_k$, are independent t random variables, each with $n_0 - 1$ degrees of freedom, and F_{n_0-1} is the cdf of the t distribution with $n_0 - 1$ degrees of freedom. The fact that $(\widehat{\mu}_{[i]} - \mu_{[i]})/\xi, i = 1, 2, \ldots, k$ are independent t random variables follows from Stein (1945).

To obtain a $(1 - \alpha)100\%$ lower confidence bound on PGS we substitute $(1 - \alpha)100\%$ lower confidence bounds on $\mu_{[k]} - \mu_{[i]}$ in (7) and evaluate the integral numerically.

5 EXAMPLE

To illustrate the performance of the procedure, consider $k = 5$ independent systems represented by normal distributions with means and variances as shown in Table 4. Suppose we take $n_0 = 10$ initial replications from each and apply the procedure with $1 - \alpha = 0.9$ and $\delta = 0.5, 1$ or 2. The last two columns of Table 4 shows the average of the lower confidence bounds on PGS and PCS from 100 replications of the entire procedure.

The primary feature to notice in these results is that PGS always remains larger than 0.9, while PCS can be larger or small depending on how large or small our indifference zone δ is. A large value of δ implies little second-stage sampling and a small chance of selecting the unique best system (remember that these are lower confidence bounds on PCS, which is why the estimated PCS can be below $1/k = 0.2$); while a small value of δ delivers precise estimates and a larger chance of selecting the unique best.

Table 4: Average LCB on PGS and PCS over 100 Replications of the Procedure with $n_0 = 10$, $1 - \alpha = 0.9$ and $r_9^{(0.9)} = 4.07$ when the Data are Normally Distributed

$\mu_1, \ldots, \mu_5$	$\sigma_1, \ldots, \sigma_5$	δ	PGS	PCS
0, 0, 0, 0, 1	1,2,3,4,5	0.5	0.98	0.91
		1.0	0.97	0.21
		2.0	0.97	0.03
1, 2, 3, 4, 5	1, 1, 1, 1, 1	0.5	0.99	0.96
		1.0	0.99	0.47
		2.0	0.98	0.15

6 CONCLUSIONS

The procedure presented in this paper provides quite a bit of useful information to the experimenter, much more than others that have been proposed for this problem. Similar procedures derived in Nelson and Banerjee (1999) allow for the use of common random numbers and exploiting equal variances across systems when this is believed to be the case. However, such wide-ranging inference comes at a price that indicates areas for further research.

The procedure in this paper is conservative if the primary interest is in selecting a good system (recall that the procedure also provides inference on all-pairwise comparisons). In other words, PGS $> 1-\alpha$ for *all* configurations of the means $\mu_1, \mu_2, \ldots, \mu_k$. Procedures that are much tighter (require less sampling) while still providing the same inference are desirable.

The method for constructing the LCBs on PCS and PGS are highly parametric, meaning that they depend strongly on the assumption of normally distributed data. Nonparametric versions, based on bootstrapping for instance, would be extremely valuable since the LCB could account for both the possible increase in confidence due to encountering a favorable configuration of the means, and also the possible degradation in confidence due to violation of the assumptions of the procedure.

ACKNOWLEDGMENTS

This research was partially supported by National Science Foundation Grant number DMI-9622065, and by Symix Corporation/Pritsker Division and Systems Modeling Corporation.

REFERENCES

Anderson, P. O., T. A. Bishop and E. J. Dudewicz. 1977. Indifference-zone ranking and selection: Confidence intervals for true achieved P(CD). *Communications in Statistics* A6:1121–1132.

Bechhofer, R. E., T. J. Santner and D. M. Goldsman. 1995. *Design and analysis for statistical selection, screening and multiple comparisons*. New York: John Wiley.

Goldsman, D. M. and B. L. Nelson. 1998. Comparing systems via simulation. In *Handbook of Simulation*, ed. J. Banks, 273–306. New York: John Wiley.

Hochberg, Y., and A. C. Tamhane. 1987. *Multiple comparison procedures*. New York: John Wiley.

Hsu, J. C. 1984. Ranking and selection and multiple comparisons with the best. In *Design of experiments: Ranking and selection*, ed. T. J. Santner and A. J. Tamhane, 23–33. New York: Marcel Dekker.

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

Kim W-C. 1986. A lower confidence bound on the probability of a correct selection. *Journal of the American Statistical Association* 81:1012–1017.

Nelson, B. L. and D. M. Goldsman. 1998. Comparisons with a standard in simulation experiments. Technical Report, Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, Illinois.

Nelson, B. L. and S. Banerjee. 1999. Estimating the probability of a good selection. Technical Report, Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, Illinois.

Stein, C. 1945. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics* 16:243–258.

AUTHOR BIOGRAPHIES

BARRY L. NELSON is a Professor in the Department of Industrial Engineering and Management Sciences at Northwestern University. His research centers on the design and

analysis of computer simulation experiments on models of stochastic systems. Nelson has served the profession as the Simulation Area Editor of *Operations Research* and President of the INFORMS (then TIMS) College on Simulation. He has held many positions for the annual Winter Simulation Conference, including Program Chair in 1997.

SOUVIK BANERJEE is a graduate student in the Department of Industrial Engineering & Management Sciences at Northwestern University. He obtained his Bachelors degree in Industrial Engineering from the Indian Institute of Technology, India and his Masters degree in Industrial Engineering and Operations Research from the University of Massachusetts at Amherst. His research interests include strategic resource allocation and agile workforce systems. He has also conducted research on global optimization and has written commercial software for automatic code generation for system dynamics modeling. He is currently a student member of INFORMS.