Proceedings of the 2003 Winter Simulation Conference
S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, eds.

COMPARISON WITH A STANDARD VIA FULLY SEQUENTIAL PROCEDURES

Seong-Hee Kim

School of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, U.S.A.

ABSTRACT

We develop fully sequential procedures for comparison with a standard. The goal is to find systems whose expected performance measures are larger or smaller than a single system referred to as a standard and, if there is any, to find the one with the largest or smallest performance. Our procedures allow for unequal variances across systems, the use of common random numbers and known or unknown expected performance of the standard. Experimental results are provided to compare the efficiency of the procedure with other existing procedures.

1 INTRODUCTION

Comparison with a standard is one of the general comparison problems we encounter in simulation. For the details of different types of comparison problems in simulation, see Goldsman and Nelson (1998). The goal of comparison with a standard is to find systems whose expected performance measures are larger (smaller) than a standard and, if there is any, to find the one with the largest (smallest) expected performance measure. For this type of problem, each alternative needs to be compared to the standard as well as other alternative systems. Nelson and Goldsman (2001) proposed two-stage procedures for comparison with a standard that account for the known or unknown performance measure of a standard and allow for unequal variances and the use of common random number (CRN). Their procedures provide multiple comparisons with the best (MCB) confidence intervals at the end of procedures so that users can compare how significant observed differences are. Their procedures work well when the number of systems is small, say fewer than 20. Otherwise, they become very conservative. This problem occurs with many two- or three-stage ranking and selection procedures that are developed to find the best system among a number of simulated systems (see Boesel et al. 2002). All of these procedures, including those due to Nelson and Goldsman (2001), employ a special assumption known as a slippage configuration for the proof of the validity of the procedures. The slippage configuration assumes that the performances of inferior systems are all close to the best system, which is rarely true when there are many systems. It is natural to believe that for more than 20 systems, the mean configurations are likely to be spread out rather than all be close to the best.

Many researchers have worked on how to overcome this inefficiency of two- or three-stage procedures. Boesel et al. (2002) and Nelson et al. (2002) introduced an elimination step after the first stage by combining a subset-selection method and two-stage procedures. Chick (1997) and Chick and Inoue (2001ab) proposed completely different procedures from a decision-theoretic point of view, and Chen et al. (1997, 2000) proposed a procedure to find a system that maximizes the probability of correct selection under a budget constraint. Kim and Nelson (2001) proposed a fully sequential procedure that takes only one basic observation at each stage and eliminates a system when there is a clear evidence that it is inferior. All these procedures were shown to be highly efficient compared to classic two- or three-stage procedures. Boesel et al. (2002) and Nelson et al. (2002) showed that their procedures can be successfully applied to very large number of systems, say 500 systems. Among these various remedies, we will take fully sequential approach to develop efficient procedures for comparison with a standard.

In this paper, we propose fully sequential type procedures that find the best of alternatives if there is any system with larger expected performance than a standard and choose the standard otherwise. We also show that the procedures are capable of handling a relatively large number of systems by experiments.

This paper is organized as follows: In Section 2, we define our problem and provide assumptions for simulation output data. Generic and customized procedures for special cases are provided in Section 3. In Section 4, we compare our procedures with the procedures due to Nelson and Goldsman (2001), followed by conclusion in Section 5.

2 PROBLEM

In this section, we define the problem of interest and state assumptions for output data. We have the designated standard denoted as system 0 and k alternative systems. Let X_{ij} be the jth output data from system i and we assume that $X_{i1}, X_{i2}, \ldots$ are independent and identically distributed and normal. As long as X_{ij}'s are either within-replication averages or batch means, the i.i.d. normality assumption is plausible (see Law and Kelton 2000). System i has the expected performance $\mu_i = \mathrm{E}[X_{ij}]$ and variance $\sigma_i^2 = \mathrm{Var}[X_{ij}]$. Without loss of generality, we can assume that

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k.$$

We do not need the assumption of equal variances across system.

Depending on the situation at hand, the expected performance of the standard, μ_0, can be either known or unknown. Nelson and Goldsman (2001) give two examples of known μ_0 cases. When an existing system has been in place so long that its average performance is well known, this is considered as known μ_0. Or when alternative systems are compared to a target value that the existing system is replaced with a new system only when the performance of the new system exceeds the target value, this can be also considered as known μ_0. For known μ_0, we do not simulate the standard but we do need to simulate the standard when its performance is unknown. However, sometimes people might want to simulate the standard even though μ_0 is known for sharper comparison since simulations provide better estimates of relative difference than they do absolute performance since the same simplifications go into all the models.

Note that Goldsman and Nelson (1998) call comparison with unknown μ_0 "comparison with a default" and comparison with known μ_0 "comparison with a standard." However, in this paper we call both cases comparison with a standard. Customized versions for various cases will be presented in Section 3.

We assume that we want to find a system with the largest expected performance measure. Then, our goal is to provide selection procedures that guarantee the following:

$$\Pr\{\text{select system } 0\} \geq 1-\alpha \text{ whenever } \mu_0 \geq \mu_k \qquad (1)$$

and

$$\Pr\{\text{select system } k\} \geq 1-\alpha$$

$$\text{whenever } \mu_k \geq \mu_0 + \delta \text{ and } \mu_k \geq \mu_{k-1} + \delta. \qquad (2)$$

We first provide a generic procedure then provide customized versions when μ_0 is known or unknown, and when systems are simulated independently or with CRN.

3 PROCEDURES

In this section, we provide a generic fully sequential procedure for comparison with a standard followed by customized versions.

The following procedure finds a system with the largest expected performance measure but multiplying each observation X_{ij} by -1 will solve a minimization problem.

3.1 Generic and Customized Procedures

Generic Procedure

Setup: Select confidence level $1-\alpha$, indifference zone δ and first-stage sample size $n_0 \geq 2$. Calculate η and c as described below in **Constants**.

Initialization: Let $I = \{0, 1, 2, \ldots, k\}$ be the set of systems still in contention.
Obtain n_0 observations X_{ij}, $j = 1, 2, \ldots, n_0$ from each system $i = 0, 1, 2, \ldots, k$.
For all $i \neq \ell$, $i, \ell = 0, 1, 2, \ldots, k$ compute $S^2_{i\ell}$, the sample variance of the difference between system i and system ℓ, and let

$$a_{i\ell} = \frac{\eta(n_0-1)S^2_{i\ell}}{\delta_{i\ell}} \text{ and } \lambda_{i\ell} = \frac{\delta_{i\ell}}{2c}$$

where

$$\delta_{i\ell} = \begin{cases} \delta/2, & \text{if } i = 0 \text{ or } \ell = 0 \\ \delta, & \text{otherwise.} \end{cases}$$

Screening: For each $i \neq \ell$, $i \in I$, and $\ell \in I$,

$$\text{if } \sum_{j=1}^{r}(\mathcal{X}_{ij} - \mathcal{X}_{\ell j}) < \max\{0, -a_{i\ell} + \lambda_{i\ell}\, r\},$$

then eliminate i from I, where

$$\mathcal{X}_{qj} = \begin{cases} X_{qj} + \delta/2, & \text{if } q = 0 \\ X_{qj}, & \text{otherwise.} \end{cases}$$

Stopping Rule: If $|I| = 1$, then stop and select the system whose index is in I.
Otherwise, set $r = r + 1$ and take one additional observation $X_{i,r}$ from each system $i \in I$.

Constants: The constant c may be any nonnegative integer. The constant η is the solution to the equation

$$\sum_{\ell=1}^{c}(-1)^{\ell+1}\left(1-\frac{1}{2}\mathcal{I}(\ell=c)\right)\times \left(1+\frac{2\eta(2c-\ell)\ell}{c}\right)^{\frac{-(n_0-1)}{2}} = \beta$$

where $\mathcal{I}$ is the indicator function and β is selected so that the overall confidence is $1-\alpha$.

The generic procedure can be easily applied to various situations by adjusting some parameters of the procedure. Here are some examples.

Case A: When μ_0 is known and systems are simulated independently, $S^2_{0\ell} = S^2_\ell$, $\mathcal{X}_{0j} = \mu_0 + \delta/2$, and $\beta = 1-(1-\alpha)^{1/k}$ where S^2_ℓ is the usual sample variance of system ℓ.

Case B: When μ_0 is unknown and systems are simulated independently, use the procedure as is with $\beta = 1-(1-\alpha)^{1/k}$.

Case C: When μ_0 is unknown and CRN is used, use the procedure as is with $\beta = \alpha/k$.

We do not provide a customized version for the case when μ_0 is known and CRN is used. Nelson and Goldsman (2001) showed that using CRN only for alternative systems when μ_0 is known can be counterproductive since all inferior alternatives will tend to show a good performance when an inferior alternative happens to show good performance, which increases the chance of incorrect selection. Therefore, it is safer to simulate all systems independently when μ_0 is known.

We do not provide the statistical validity of the generic procedure but Kim (2002) proves that if X_{ij}, $j = 1, 2, \ldots,$ are i.i.d. normally distributed and X_{0j}, $j = 1, 2, \ldots,$ are either constant or i.i.d. normally distributed, then the Generic Procedure guarantees (1) and (2) with or without CRN.

4 EXPERIMENTAL RESULTS

In this section we summarize the results of experiments performed to compare the following procedures:

1. A two-stage procedure due to Nelson and Goldsman (2001): their generic procedure (NG) allows for unknown and unequal variances across systems and the use of CRN. They proposed two versions for the case where CRN is employed: one is based on the assumption of sphericity and the other uses the Bonferonni inequality. We tested the version established under the assumption of sphericity since it is shown to be more efficient than the other. However, there could be a severe degradation in PCS if the sphericity assumption does not hold, which is often the case when there are many alternative systems.
2. The fully sequential procedure (FSP) proposed in Section 3, both with and without CRN.

The systems were represented by various configurations of k normal distributions; either system 0 or system 1 was the true best (had the largest true mean). We evaluated each procedure on different variations of the systems, examining factors including the number of systems, k; the correlation between systems, ρ; the true means, $\mu_0, \mu_1, \mu_2, \ldots, \mu_k$; and the true variances, $\sigma_0^2, \sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$. The configurations, the experiment design, and the results are described below.

4.1 Configurations and Experiment Design

To ensure the first-stage sample size is not too small, we chose the first-stage sample size to be $n_0 = 10$. The number of systems in each experiment varied over $k = 2, 5, 10, 25, 100$.

The indifference zone, δ, was set to $\delta = 1/\sqrt{n_0}$ and we set the variance of the best system (either system 0 or 1) to one, with the exception for known μ_0 case where the variance of system 0 is zero. Roughly, we can interpret δ as the standard deviation of the first-stage sample mean of the best system.

4.1.1 Mean Configurations

Two configurations of the true means were used for each case when the standard (system 0) is the best and an alternative (system 1) is the best.

When system 1 is the best, the Slippage Configuration (SC) and the monotonic decreasing means (MDM) configuration were used.

- To test the statistical guarantee of proposed procedures, the SC configuration is used in which μ_1 was set to δ, while $\mu_0 = \mu_2 = \mu_3 = \cdots = \mu_k = 0$. Since all of the inferior systems are close to the best, it will be hard to detect the true best and inferior systems so this is a difficult configuration.
- To investigate the effectiveness of the procedures in eliminating non-competitive systems, the MDM were also used. In the MDM configuration, the means of all systems were spaced evenly apart according to the following formula: when system 1 is the best, $\mu_i = \mu_1 - \delta|i-1|$, for $i = 0, 2, 3, \ldots, k$.

When system 0 is the best, the equal mean configuration (EMC) and the MDM configuration were used.

- The EMC, in which $\mu_0 = \mu_1 = \cdots = \mu_k = 0$, will be the most difficult configuration instead of SC.
- For MDM, μ_0 is set to δ and $\mu_i = \mu_0 - \delta|i|$.

Table 1: Example of Mean Configuration when $k = 5$

best	Model	$\mu_0, \mu_1, \mu_2, \mu_3, \mu_4, \mu_5$
system 1	SC	$0, \delta, 0, 0, 0, 0$
	MDM	$0, \delta, 0, -\delta, -2\delta, -3\delta$
system 0	EMC	$0, 0, 0, 0, 0, 0$
	MDM	$\delta, 0, -\delta, -2\delta, -3\delta, -4\delta$

Table 2: Example of Variance Configuration when $k = 5$

best	Model	$\sigma_0, \sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5$
system 1	constant	$1, 1, 1, 1, 1, 1$
	increasing	$1+\delta, 1, 1+\delta, 1+2\delta, 1+3\delta, 1+4\delta$
	decreasing	$\frac{1}{1+\delta}, 1, \frac{1}{1+\delta}, \frac{1}{1+2\delta}, \frac{1}{1+3\delta}, \frac{1}{1+4\delta}$
system 0	constant	$1, 1, 1, 1, 1, 1$
	increasing	$1, 1+\delta, 1+2\delta, 1+3\delta, 1+4\delta, 1+5\delta$
	decreasing	$1, \frac{1}{1+\delta}, \frac{1}{1+2\delta}, \frac{1}{1+3\delta}, \frac{1}{1+4\delta}, \frac{1}{1+5\delta}$

4.1.2 Variance Configurations

For each configuration of the means we examined the effect of both equal and unequal variances. In the equal-variance configuration σ_i was set to one. In the unequal-variance configuration the variance of the best system was set both higher and lower than the variances of the other systems. In the MDM configurations, experiments were run with the variance directly proportional to the mean of each system, and inversely proportional to the mean of each system. Specifically, $\sigma_i^2 = |\mu_i - \delta| + 1$ to examine the effect of increasing variance as the mean decreases, and $\sigma_i^2 = 1/(|\mu_i - \delta| + 1)$ to examine the effect of decreasing variances as the mean decreases. In addition, some experiments were run with means in the SC and the EMC, but with the variances of all systems either monotonically decreasing or monotonically increasing as in the MDM configuration. When μ_0 is known, σ_0^2 is set to zero.

The example of means and variances configurations is given in Tables 1 and 2 when there are 5 alternatives (6 systems including the standard). The variance of system 0 will be replaced with 0 when μ_0 is known in Table 2.

When CRN was employed we assumed that the assumption of sphericity holds and the correlation between all pairs of systems, ρ, were tested at $\rho = 0.02, 0.25, 0.5, 0.75$.

Thus, we had six configurations when system 1 is the best: SC with equal variances, MDM with equal variances, MDM with increasing variances, MDM with decreasing variances, SC with increasing variances and SC with decreasing variances. There are another set of six configurations when system 0 is the best: EMC with equal variances, MDM with equal variances, MDM with increasing variances, MDM with decreasing variances, EMC with increasing variances and EMC with decreasing variances. We have a total of twelve configurations and each configuration is repeated with known and unknown μ_0 and with independence or with CRN. Note that for known μ_0, we do not employ CRN since it is safer to simulate all alternatives independently.

For each configuration, 1000 macroreplications (complete repetitions) of the entire experiment were performed. In all experiments, the nominal probability of correct selection (PCS) was set at $1 - \alpha = 0.95$. To compare the performance of the procedures we recorded the total number of observations required by each procedure and the estimated PCS.

4.2 Summary of Results

For all configurations with or without CRN, the estimated PCS of FSP was over 0.95. For a few configurations (EMC with increasing variances or SC with increasing variances for $k = 2$), we got an estimated PCS which is slightly lower then 0.95, like 0.944 or 0.946. However, when we made $100,000$ macro-replications for those configurations, we obtained an estimated PCS larger than the nominal PCS (around 0.953). The overall experiments showed that FSP is uniformly superior to NG under any configurations in terms of the total number of observations. Especially under a MDM configuration with increasing variances, FSP's superiority relative to NG was more noticeable as the number of systems increased.

The total number of basic observations consumed by each procedure was increasing much more slowly in FSP than in NG as k increases.

4.3 Some Specific Results

Instead of presenting all results, we present selected results that emphasize the key conclusions through the following four tables:

- Table 3: μ_0 is known and system 1 is the best.
- Table 4: μ_0 is known and system 0 is the best.
- Table 5: μ_0 is unknown and system 1 is the best.
- Table 6: μ_0 is unknown and system 0 is the best.

4.3.1 Effect of Number of Systems

In our experiments the FSP outperformed NG under any configurations; see Tables 3 – 6 for an illustration. NG is very sensitive to the number of systems. For example, Table 3 shows that for MDM with increasing variances NG spent $1,918$ when $k = 5$ and $772,964$ when $k = 100$ while FSP spent 419 when $k = 5$ and $5,726$ when $k = 100$. FSP achieved reductions from 50% up to 97% in the number of basic observations, as compared to NG depending on configurations. When the number of systems is large and the configuration is difficult such as MDM with increasing

Table 3: Sample Average Total Number of Observations for the NG and FSP when μ_0 is Known and System 1 is the Best (Numbers in Parentheses Represent the Amount of Induced Correlation, ρ)

the number of alternatives	procedure type	MDM increasing	MDM decreasing	SC increasing	SC decreasing
$k = 5$	NG (0)	1918	802	1932	795
	FSP (0)	419	307	673	384
$k = 100$	NG (0)	772964	5443	774855	5444
	FSP (0)	5726	1751	187093	2881

Table 4: Sample Average Total Number of Observations for the NG and FSP when μ_0 is Known and System 0 is the Best (Numbers in Parentheses Represent the Amount of Induced Correlation, ρ)

the number of alternatives	procedure type	MDM increasing	MDM decreasing	SC increasing	SC decreasing
$k = 5$	NG (0)	2279	647	2291	644
	FSP (0)	322	124	1283	365
$k = 100$	NG (0)	789504	4984	789973	4976
	FSP (0)	5653	1268	318798	2813

Table 5: Sample Average Total Number of Observations for the NG and FSP when μ_0 is Unknown and System 1 is the Best (Numbers in Parentheses Represent the Amount of Induced Correlation, ρ)

the number of alternatives	procedure type	MDM increasing	MDM decreasing	SC increasing	SC decreasing
$k = 5$	NG (0)	4032	1773	4056	1747
	FSP (0)	991	711	1225	780
	NG (0.02)	3070	1319	3070	1327
	FSP (0.02)	995	716	1245	799
	NG (0.25)	2330	1011	2351	1013
	FSP (0.25)	764	545	941	603
	NG (0.50)	1576	689	1572	686
	FSP (0.50)	517	374	620	412
	NG (0.75)	834	355	835	354
	FSP (0.75)	268	196	331	209
$k = 100$	NG (0)	1257089	9329	1264627	9338
	FSP (0)	7262	2886	191405	3979
	NG (0.02)	739163	5487	736493	5505
	FSP (0.02)	7240	2875	187881	3959
	NG (0.25)	580859	4510	579492	4499
	FSP (0.25)	5884	2427	156862	3209
	NG (0.50)	413682	3450	414250	3445
	FSP (0.50)	4413	1911	120698	2375
	NG (0.75)	245638	2353	243845	2371
	FSP (0.75)	2806	1444	74112	1651

Table 6: Sample Average Total Number of Observations for the NG and FSP when μ_0 is Unknown and System 0 is the Best (Numbers in Parentheses Represent the Amount of Induced Correlation, ρ)

the number of alternatives	procedure type	MDM increasing	MDM decreasing	SC increasing	SC decreasing
$k = 5$	NG (0)	4615	1623	4618	1603
	FSP (0)	599	360	2239	867
	NG (0.02)	3439	1192	3483	1189
	FSP (0.02)	595	347	2229	844
	NG (0.25)	2650	919	2673	912
	FSP (0.25)	461	270	1696	654
	NG (0.50)	1809	631	1805	624
	FSP (0.50)	311	184	1153	441
	NG (0.75)	959	336	958	337
	FSP (0.75)	162	104	633	244
$k = 100$	NG (0)	1279121	8800	1281718	8788
	FSP (0)	6370	1903	346924	4081
	NG (0.02)	753097	5176	749379	5177
	FSP (0.02)	6346	1884	341809	4004
	NG (0.25)	594238	4229	591112	4230
	FSP (0.25)	5239	1682	267538	3272
	NG (0.50)	418327	3204	420295	3205
	FSP (0.50)	3931	1425	188194	2481
	NG (0.75)	246252	2166	245564	2183
	FSP (0.75)	2574	1191	104430	1786

variances or SC with increasing variances, the benefit of FSP becomes larger.

It is interesting to notice that NG spent almost the same number of observations for MDM and SC cases with increasing or decreasing variances while FSP clearly spent fewer observations for MDM cases than for SC cases. This is because NG takes account for variances only when it determines N_i. FSP also takes account for variances only when it determines N_i, but the large difference in means between alternatives makes it easy to detect inferior systems. Therefore, the difference in means is considered in FSP indirectly and this makes the procedure efficient.

4.3.2 Effect of Correlation

Kim and Nelson (2001) suggest that positive correlation larger than 0.02 is sufficient for the FSP with CRN to outperform the FSP assuming independence for the procedures developed to find the best alternatives. As shown in the empirical results in Tables 5 and 6, the FSP under independence shows similar performance as the FSP under CRN when $\rho = 0.02$ in terms of the number of observations. This implies that we can gain the benefit of CRN even with a small amount of positive correlation (say, larger than 0.02). A larger positive correlation makes the FSP even more efficient, and this holds across all of the configurations that were used in our experiments.

5 CONCLUSION

We proposed efficient procedures for comparison with a standard. Even though it is clear that the proposed procedures can be much more efficient than NG in terms of number of required observations to find the best, they are more computationally intensive and more complicated to perform due to switching between systems, stopping, and restarting simulation of each system. However, with parallel computing environments, this problem is now less restrictive.

REFERENCES

Boesel, J., B. L. Nelson, and S.-H. Kim. 2002. Using ranking and selection to clean up after a simulation search. *Operations Research*, forthcoming.

Chen, H.-C., C.-H. Chen, L. Dai, and E. Yücesan. 1997. New development of optimal computing budget allocation for discrete event simulation, In *Proc. 1997 Winter Simulation Conference*, ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B.L. Nelson, 334-341. Piscataway, New Jersey: IEEE.

Chen, H.-C., C.-H. Chen, and E. Yücesan. 2000. Computing efforts allocation for ordinal optimization and discrete event simulation. *IEEE Transactions on Automatic Control*, 45: 960–964.

Chick, S. E. 1997. Selecting the best system: A decision-theoretic approach. In *Proc. 1997 Winter Simulation Conference*, ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B.L. Nelson, 326-333. Piscataway, New Jersey: IEEE.

Chick, S., and K. Inoue. 2001a. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research*, 49:1609–1624.

Chick, S., and K. Inoue. 2001b. New procedures for identifying the best simulated system using common random numbers. *Management Science*, 47: 1133–1149.

Goldsman, D., and B. L. Nelson. 1998. Comparing Systems via Simulation. In *Handbook of simulation: Principles, Methodology, Advances, Applications, and Practice*, ed. J. Banks, 273–306. New York: John Wiley & Sons.

Kim, S.-H. 2002. Comparison with a standard via fully sequential procedures. Technical Reports. School of Industrial and Systems Engineering, Georgia Tech.

Kim, S.-H., and B. L. Nelson. 2001. A fully sequential procedure for indifference-zone selection in simulation. *ACM TOMACS*, 11:251–273.

Law, A. M., and D. Kelton. 2000. *Simulation modeling and analysis*, 3rd ed. New York: McGraw-Hill.

Nelson, B. L., and D. Goldsman. 2001. Comparisons with a standard in simulation experiments, *Management Science*, 47:449–463.

Nelson, B. L., J. Swann, D. Goldsman, and W. Song. 2002. Simple procedures for selecting the best simulated system when the number of alternatives is large. *Operations Research*, 49:950–963.

AUTHOR BIOGRAPHY

SEONG-HEE KIM is an Assistant Professor in the School of Industrial Systems and Engineering at Georgia Tech. Her research interests include simulation output analysis and ranking and selection. Her e-mail and web addresses are <skim@isye.gatech.edu> and <www.isye.gatech.edu/~skim/>.