Proceedings of the 1994 Winter Simulation Conference
ed. J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila

RANKING, SELECTION AND MULTIPLE COMPARISONS IN COMPUTER SIMULATION

David Goldsman

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

Barry L. Nelson

Department of Industrial, Welding & Systems Engineering
The Ohio State University
Columbus, Ohio 43210, U.S.A.

ABSTRACT

We present a state-of-the-art review of ranking, selection and multiple-comparison procedures that are used to compare system designs via computer simulation. We describe methods for four classes of problems: screening a large number of system designs, selecting the best system, comparing all systems to a standard and comparing alternatives to a default. Rather than give a comprehensive review, we present the methods we would be likely to use in practice and emphasize recent results. Where possible, we unify the ranking-and-selection and multiple-comparison perspectives.

1 INTRODUCTION

Simulation experiments are typically performed to compare, in some fashion, two or more system designs. The statistical methods of *ranking and selection* and *multiple comparisons* are applicable when comparisons among a finite and typically small number of systems (say 2 to 20) are required. The particular method that is appropriate depends on the type of comparison desired and properties of the simulation output data. In this state-of-the-art review we describe methods for four classes of problems: screening a large number of system designs, selecting the best system, comparing all systems to a standard and comparing alternatives to a default. Rather than give a comprehensive review, we present the methods we would be likely to use in practice. And where possible, we unify the ranking-and-selection and multiple-comparison perspectives.

Ranking and selection procedures (R&S) are statistical methods specifically developed to select the best system, or a subset of systems that includes the best system, from among a collection of competing alternatives. Provided certain assumptions are met, these methods usually guarantee that the probability of a correct selection will be at least some user-specified value. Multiple-comparison procedures (MCPs) treat the comparison problem as an inference problem on the performance parameters of interest. MCPs account for the error that arises when making simultaneous inferences about differences in performance among the systems. Both types of procedures are relevant in the context of computer simulation because the assumptions behind the procedures can frequently be satisfied: The assumption of normally distributed data can often be secured by batching large numbers of (cheaply generated) outputs. Independence can be obtained by controlling random-number assignments. And multiple-stage sampling—which is required by some methods—is feasible in computer simulation because a subsequent stage can be initialized simply by retaining the final random-number seeds from the preceding stage.

To facilitate the discussion that follows we define the following notation: Let Y_{ij} represent the jth simulation output from system design i, for $i = 1, 2, \ldots, k$ alternatives and $j = 1, 2, \ldots$. For fixed i, we will always assume that the outputs from system i, $Y_{i1}, Y_{i2}, \ldots$, are independent and identically distributed (i.i.d.). These assumptions are plausible if $Y_{i1}, Y_{i2}, \ldots$ are outputs across independent replications, or if they are appropriately defined batch means from a single replication after accounting for initialization effects. Let $\mu_i = \mathrm{E}[Y_{ij}]$ denote the expected value of an output from the ith system, and let $\sigma_i^2 = \mathrm{Var}[Y_{ij}]$ denote its variance. Further, let

$$p_i = \Pr\left\{Y_{ij} > \max_{\ell \neq i} Y_{\ell j}\right\}$$

be the probability that Y_{ij} is the largest of the jth outputs across all systems *when* $Y_{1j}, Y_{2j}, \ldots, Y_{kj}$ *are mutually independent.*

The methods we describe make comparisons based on either μ_i or p_i. Although not a restriction on either R&S or MCPs, we will only consider situations in which there is no known functional relationship among the μ_i or p_i (other than $\sum_{i=1}^{k} p_i = 1$). Therefore, there are is no potential information to be gained about one system from simulating the others—such as might occur if the μ_i were a function of some explanatory variables—and no potential efficiency to be gained from fractional-factorial experiment designs, group screening designs, etc.

The paper is organized into four additional sections, one for each type of comparison problem.

2 SCREENING PROBLEMS

Example 1 *A brain-storming session produces 15 potential designs for the architecture of a new computer system. Response time is the performance measure of interest, but there are so many designs that a careful simulation study will be deferred until a pilot simulation study determines which designs are worth further scrutiny. Smaller response time is preferred.*

If the expected response time is the performance measure of interest, then the goal of the pilot study is to determine which designs are the better performers, which have similar performance, and which can be eliminated as clearly inferior.

2.1 Multiple Comparisons Approach

Let μ_i denote the expected response time for architecture i. Multiple comparisons attacks the screening problem by forming simultaneous confidence intervals on the parameters $\mu_i - \mu_j$ for all $i \neq j$. These $k(k-1)/2$ confidence intervals indicate the magnitude and direction of the difference between each pair of alternatives. The most widely used method for forming the intervals is Tukey's procedure, which is implemented in many statistical software packages. We review the procedure here and cite some recent advances. General references include Hochberg and Tamhane (1987) and Miller (1981).

Suppose that the systems are simulated independently, and we obtain i.i.d. outputs $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ from system i. Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}/n_i$ be the sample mean from system i, and let

$$S^2 = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

be the pooled sample variance. Tukey's simultaneous confidence intervals are

$$\mu_i - \mu_j \in \bar{Y}_i - \bar{Y}_j \pm \frac{Q_{k,\nu}^{(\alpha)}}{\sqrt{2}} S\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

for all $i \neq j$, where $Q_{k,\nu}^{(\alpha)}$ is the $1-\alpha$ quantile of the Studentized range distribution with parameter k and $\nu = \sum_{i=1}^{k}(n_i - 1)$ degrees of freedom (see Hochberg and Tamhane 1987, Appendix 3, Table 8, for instance).

When the Y_{ij} are normally distributed with common (unknown) variance, and $n_1 = n_2 = \cdots = n_k$, these intervals achieve simultaneous coverage probability $1-\alpha$. Hayter (1984) showed that the coverage probability is strictly greater than $1-\alpha$ when the sample sizes are not equal.

When there are a large number of comparisons, as in the example above, then the confidence intervals should be displayed a manner that allows the analyst to easily perceive the magnitude and direction of significant differences, and to recognize systems that are practically equivalent. Hsu and Peruggia (1994) recently proposed the *mean-mean scatter plot* for this purpose. The plot is constructed by letting each pair $(\bar{Y}_i, \bar{Y}_j)$ with $\bar{Y}_i > \bar{Y}_j$ be a point in two-dimensional Euclidean space, then drawing a line segment of length $Q_{k,\nu}^{(\alpha)} S\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$ with $-45°$ slope centered at the point. Since it is easier to perceive vertical and horizontal alignments, the entire plot may be rotated 45° counterclockwise.

An example is shown in Figure 1 for $k = 4$ systems. Notice that line segments that do not cross the diagonal "0" line indicate differences that are statistically significant.

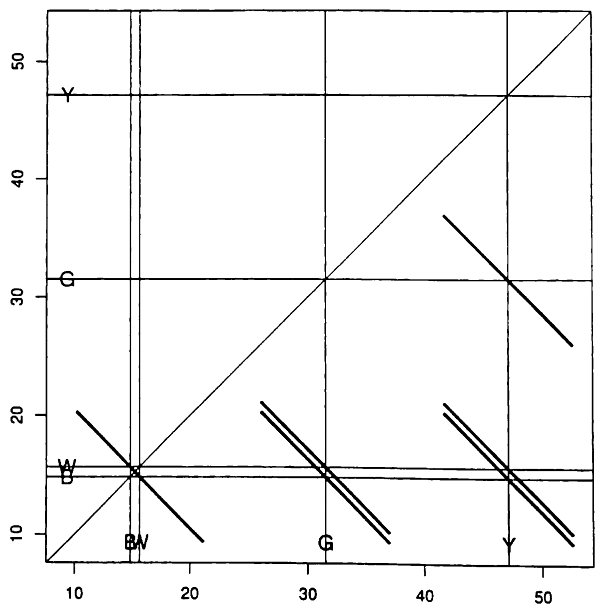

Figure 1: Mean-Mean Scatter Plot of Tukey's All-Pairwise Comparisons

2.2 Subset Selection Approach

The subset selection approach is a screening device that attempts to select a (random-size) *subset* of the $k = 15$ competing designs of Example 1 that contains the design with the smallest expected response time. Gupta (1956, 1965) proposed a single-stage procedure for this problem that is applicable in cases when the data from the competing designs are balanced (i.e., $n_1 = \cdots = n_k = n$) and are normal with common (unknown) variance σ^2.

First specify the desired probability $1 - \alpha$ of actually including the best design in the selected subset. Calculate the k sample means $\bar{Y}_i = \sum_{j=1}^{n} Y_{ij}/n$, for $i = 1, 2, \ldots, k$, and an unbiased pooled estimate of σ^2,

$$S^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_i)^2}{k(n-1)}.$$

Include the ith design in the selected subset if

$$\bar{Y}_i \leq \min_{1 \leq j \leq k} \left\{ \bar{Y}_j + gS\sqrt{2/n} \right\}$$

where $g = T^{(\alpha)}_{k-1,k(n-1),0.5}$ is an equicoordinate critical point of the equicorrelated multivariate central t-distribution; this constant can be found in Hochberg and Tamhane (1987), Appendix 3, Table 4; Bechhofer, Santner and Goldsman (BSG) (1995); or by using the FORTRAN program AS251 of Dunnett (1989). Gupta and Huang (1976) proposed a similar procedure (requiring more obscure tables) for the unbalanced case.

3 SELECTING THE BEST

Example 2 (Goldsman, et al. 1991) *For the purpose of evaluation prior to purchase, simulation models of four different airline-reservation systems have been developed. The single measure of system performance is the time to failure (TTF), so that larger TTF is better. A reservation system works if either of two computers works. The four systems arise from variations in parameters affecting the TTF and time-to-repair distributions. Differences of less than about two days are considered practically equivalent.*

3.1 Indifference-Zone Selection Approach

If expected TTF is taken as the performance measure of interest, then the goal in this example is to select the system with the largest expected TTF. In a stochastic simulation such a "correct selection" can never be guaranteed with certainty. A compromise solution offered by *indifference-zone selection* is to guarantee to select the best system with high probability whenever it is at least a user-specified amount better than the others; this practically-significant difference is called the indifference zone. In the example the indifference zone is $\delta = 2$ days.

Law and Kelton (1991) present indifference-zone procedures that have proven useful in simulation, while BSG (1995) provide a comprehensive review of R&S procedures. We present two procedures, one due to Rinott (1978) that is applicable when the output data are normally distributed and all systems are simulated independently of each other, and the other due to Matejcik and Nelson (1993) that works in conjunction with common random numbers.

Multiple comparisons attacks the problem of determining the best system by forming simultaneous confidence intervals on the parameters $\mu_i - \max_{j \neq i} \mu_j$ for $i = 1, 2, \ldots, k$, where μ_i denotes the expected TTF for the ith reservation system. These confidence intervals are known as *multiple comparisons with the best (MCB)*, and they bound the difference between the expected performance of each system and the best of the others. The first MCB procedures were developed by Hsu (1984); a thorough review is found in Hochberg and Tamhane (1987).

Matejcik and Nelson (1992) and Nelson and Matejcik (1993) established a fundamental connection between indifference-zone selection and MCB by showing that *most indifference-zone procedures can simultaneously provide MCB confidence intervals with the width of the intervals corresponding to the indifference zone.* The procedures we display below are combined indifference-zone selection and MCB procedures. The advantage of a combined procedure is that we not only select a system as best, we also gain information about how close each of the inferior systems is to being the best. This information is useful if secondary criteria that are not reflected in the performance measure (such as ease of installation, cost to maintain, etc.) may tempt us to choose an inferior system if it is not deficient by much.

In the combined procedures below we use the convention that a "." subscript indicates averaging with respect to that subscript. For example, $\bar{Y}_{i\cdot}$ is the sample average of $Y_{i1}, Y_{i2}, \ldots, Y_{in_0}$.

Rinott + MCB (independent sampling)

1. Specify δ, α and n_0. Let h solve Rinott's integral for n_0, k and α (see the tables in Wilcox 1984 or BSG 1995).

2. Take an i.i.d. sample $Y_{i1}, Y_{i2}, \ldots, Y_{in_0}$ from each of the k systems *simulated independently*.

3. Compute the marginal sample variances
$$S_i^2 = \frac{\sum_{j=1}^{n_0}(Y_{ij} - \bar{Y}_{i\cdot})^2}{n_0 - 1}$$
for $i = 1, 2, \ldots, k$.

4. Compute the final sample sizes
$$N_i = \max\{n_0, \lceil (hS_i/\delta)^2 \rceil\}$$
for $i = 1, 2, \ldots, k$, where $\lceil\cdot\rceil$ is the integer "round-up" function.

5. Take $N_i - n_0$ additional i.i.d. observations from system i, independently of the first-stage sample and the other systems, for $i = 1, 2, \ldots, k$.

6. Compute the overall sample means
$$\bar{\bar{Y}}_{i\cdot} = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij}$$
for $i = 1, 2, \ldots, k$.

7. Select the system with the largest $\bar{\bar{Y}}_{i\cdot}$ as best.

8. Simultaneously form the MCB confidence intervals
$$\mu_i - \max_{j\neq i}\mu_j \in$$
$$\left[-\left(\bar{\bar{Y}}_{i\cdot} - \max_{j\neq i}\bar{\bar{Y}}_{j\cdot} - \delta\right)^-, \left(\bar{\bar{Y}}_{i\cdot} - \max_{j\neq i}\bar{\bar{Y}}_{j\cdot} + \delta\right)^+\right]$$
for $i = 1, 2, \ldots, k$, where $(a)^+ = \max\{0, a\}$ and $-(b)^- = \min\{0, b\}$.

Rinott's procedure, and the accompanying MCB intervals, simultaneously guarantee a probability of correct selection and confidence-interval coverage probability greater than or equal to $1 - \alpha$ under the stated assumptions.

A fundamental assumption of the Rinott+MCB procedure is that the k systems are simulated independently (see Step 2 above). In practice this means that different streams of (pseudo)random numbers are assigned to the simulation of each system. However, under fairly general conditions assigning common random numbers (CRN) to the simulation of each system decreases the variances of estimates of the pairwise differences in performance. Unfortunately, CRN also complicates the statistical analysis when $k > 2$ systems are involved. The following new procedure provides the same guarantees as Rinott+MCB under a more complex set of conditions, but has been shown to be quite robust to departures from those conditions. And unlike Rinott+MCB, it is designed to exploit the use of CRN to reduce the total number of observations required to make a correct selection.

NM + MCB (common random numbers)

1. Specify δ, α and n_0. Let $g = T^{(\alpha)}_{k-1,(k-1)(n_0-1),0.5}$ (see Hochberg and Tamhane 1987, Appendix 3, Table 4; and BSG 1995).

2. Take an i.i.d. sample $Y_{i1}, Y_{i2}, \ldots, Y_{in_0}$ from each of the k systems *using CRN across systems.*

3. Compute the approximate sample variance of the difference of the sample means
$$S^2 = \frac{2\sum_{i=1}^{k}\sum_{j=1}^{n_0}\left(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot}\right)^2}{(k-1)(n_0-1)}.$$

4. Compute the final sample size
$$N = \max\{n_0, \lceil (gS/\delta)^2 \rceil\}.$$

5. Take $N - n_0$ additional i.i.d. observations from each system, using CRN across systems.

6. Compute the overall sample means
$$\bar{\bar{Y}}_{i\cdot} = \frac{1}{N}\sum_{j=1}^{N} Y_{ij}$$
for $i = 1, 2, \ldots, k$.

7. Select the system with the largest $\bar{\bar{Y}}_{i\cdot}$ as best.

8. Simultaneously form the MCB confidence intervals as in Rinott+MCB.

3.2 Multinomial Selection Approach

Another approach to the airline-reservation problem is to select the system that is most likely to have the largest *actual* TTF. To this end, one can define p_i as the probability that design i will produce the largest TTF from a given vector-observation $(Y_{1j}, Y_{2j}, \ldots, Y_{kj})$. The goal now is to select the design associated with the largest p_i-value. This goal is equivalent to that of finding the multinomial category having the largest probability of occurrence; and there is a rich body of literature concerning such problems.

More specifically, suppose that we wish to select the correct category with probability $1 - \alpha$ whenever the ratio of the largest to second largest p_i is greater than some user-specified constant, say $\theta > 1$. The indifference constant θ can be regarded as the smallest ratio "worth detecting."

The following *single-stage* procedure was proposed by Bechhofer, Elmaghraby and Morse (BEM) (1959) to guarantee the above probability requirement.

BEM

1. For the given k, and (α, θ) specified prior to the start of sampling, find n from the Tables in BEM (1959), Gibbons, Olkin and Sobel (1977) or BSG (1995).

2. Take a random sample of n multinomial observations $\boldsymbol{X}_j = (X_{1j}, X_{2j}, \ldots, X_{kj})$, for $j = 1, 2, \ldots, n$, in a *single* stage, where

$$X_{ij} = \begin{cases} 1, & \text{if } Y_{ij} > \max_{\ell \neq i}\{Y_{\ell j}\} \\ 0, & \text{otherwise.} \end{cases}$$

3. Let $W_i = \sum_{j=1}^{n} X_{ij}$ for $i = 1, 2, \ldots, k$. Select the design that yielded the largest W_i as the one associated with the largest p_i (randomize in the case of ties).

A more efficient procedure, due to Bechhofer and Goldsman (1986), uses *closed, sequential* sampling; that is, the procedure stops when one design is "sufficiently ahead" of the others.

BG

1. For the given k, and (α, θ) specified prior to the start of sampling, find the *truncation number* (i.e., an upper bound on the number of vector-observations) n_0 from the tables in Bechhofer and Goldsman (1986) or BSG (1995).

2. At the mth stage of experimentation ($m \geq 1$), take the random multinomial observation $\boldsymbol{X}_m = (X_{1m}, X_{2m}, \ldots, X_{km})$ (defined above) and calculate the *ordered* category totals $W_{[1]m} \leq W_{[2]m} \leq \cdots \leq W_{[k]m}$; also calculate

$$Z_m = \sum_{i=1}^{k-1} (1/\theta)^{(W_{[k]m} - W_{[i]m})}.$$

3. Stop sampling at the first stage when *either*

$$Z_m \leq \alpha/(1-\alpha) \quad \textit{or} \quad m = n_0$$

$$\textit{or} \quad W_{[k]m} - W_{[k-1]m} \geq n_0 - m,$$

whichever occurs first.

4. Let N (a random variable) denote the stage at which the procedure terminates. Select the design that yielded the largest W_{iN} as the one associated with the largest p_i (randomize in the case of ties).

4 COMPARISONS WITH A STANDARD

Example 3 *Several different investment strategies will be simulated to evaluate their expected rate of return. The strategy ultimately chosen may not be the one with the largest expected return—since factors such as risk could be considered—but none of the strategies will be chosen unless its expected return is larger than a zero-coupon bond that offers a known, fixed return.*

Here the goal is to select the best investment strategy *only if it is better than the standard*; if no strategy is better than the standard, we continue with the standard. More precisely, we have the following probability requirement: Denote the standard by μ_0 and the ordered means by $\mu_{[1]} \leq \mu_{[2]} \leq \cdots \leq \mu_{[k]}$. For constants $\{\delta, P_0, P_1\}$ with $0 < \delta < \infty$, $2^{-k} < P_0 < 1$ and $(1 - 2^{-k})/k < P_1 < 1$, specified prior to the start of experimentation, we require

$$P\{\text{Select the standard}\} \geq P_0 \quad \text{whenever} \quad \mu_{[k]} \leq \mu_0$$

and

$$P\{\text{Select best strategy}\} \geq P_1 \quad \text{whenever}$$

$$\mu_{[k]} \geq \max\{\mu_0, \mu_{[k-1]}\} + \delta.$$

We present a procedure due to Bechhofer and Turnbull (1978) for selecting the best system relative to a given standard when the responses are normal with common unknown variance σ^2. It requires that an initial sample of $n_0 \geq 2$ observations be taken from each system in order to estimate σ^2 in the first stage.

BT

1. For the given (k, μ_0) and specified (δ, P_0, P_1), fix a number of observations $n_0 \geq 2$ to be taken in Stage 1.

2. Choose constants (g, h) from Bechhofer and Turnbull (1978) corresponding to the k, n_0, P_0 and P_1 of interest.

3. In Stage 1, take a random sample of n_0 observations Y_{ij} ($j = 1, 2, \ldots, n_0$) from the k strategies. Calculate the first-stage sample means, $\bar{Y}_i = \sum_{j=1}^{n_0} Y_{ij}/n_0$ ($i = 1, 2, \ldots, k$) and the unbiased pooled estimate of σ^2,

$$S^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_0} (Y_{ij} - \bar{Y}_i)^2 / k(n_0 - 1).$$

4. In Stage 2, take a random sample of $N - n_0$ *additional* observations from each of the strategies, where
$$N = \max\left\{n_0, \lceil (gS/\delta)^2 \rceil\right\}.$$

5. Calculate the cumulative sample means
$$\bar{\bar{Y}}_i = \sum_{j=1}^{N} Y_{ij}/N$$
for $i = 1, 2, \ldots, k$.

6. If the largest sample mean $\bar{\bar{Y}}_{[k]} > \mu_0 + h\delta/g$, select the strategy that yielded it as the one associated with $\mu_{[k]}$; otherwise, select no system, i.e., select the standard as best.

5 COMPARISONS WITH A DEFAULT

Example 4 *A manufacturing company will replace an existing storage-and-retrieval system if one can be found that is superior to the system currently in place. Five vendors have proposed hardware-software systems, and simulation models have been developed for each. Systems will be evaluated in terms of their retrieval times, but the system ultimately chosen might not be the one with the smallest retrieval time because of differences in cost, ease of installation, etc.*

If the expected retrieval time is the performance measure of interest, then the goal of the simulation study is to determine which designs are better than the system currently in place, which we term the "default" (or "control"). Data on the performance of the default system may be obtained either from the system itself or from a simulation model of it.

5.1 Multiple Comparisons Approach

Let μ_i denote the expected retrieval time for system i, where $i = k$ corresponds to the default system. A fundamental principle of multiple comparisons is that we should limit ourselves to the comparisons that are necessary for the decision at hand, because the fewer confidence statements that are required to be true simultaneously the sharper the inference. Therefore, when comparing to a default we find simultaneous confidence intervals for $\mu_i - \mu_k$ for all $i \neq k$, rather than $\mu_i - \mu_j$ for all $i \neq j$. Such comparisons are called *multiple comparisons with a control (MCC)*. Further, if we know that only differences in *a specified direction* are of interest—in the example we are only interested in systems with smaller expected retrieval time than the default—then we should form one-sided confidence intervals.

Suppose that the data are acquired independently from each system (a necessity if data are collected from the default system itself), and we obtain i.i.d. outputs $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ from system i. Let
$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$
be the sample variance from system i. MCC procedures are well known for the case when the variances across systems are equal (see for instance Hochberg and Tamhane 1987 and Miller 1981). Here we present a simple procedure due to Tamhane (1977) that is valid when variances may not be equal.

The simultaneous, upper one-sided confidence intervals are
$$\mu_i - \mu_k \le \bar{Y}_i - \bar{Y}_k + \sqrt{\frac{t^2_{n_i-1,\beta}S_i^2}{n_i} + \frac{t^2_{n_k-1,\beta}S_k^2}{n_k}}$$
for all $i \neq k$, where $t_{\nu,\beta}$ is the $\beta = 1 - (1-\alpha)^{1/(k-1)}$ quantile of the t distribution with ν degrees of freedom. When the output data are normally distributed, these intervals are guaranteed to have coverage at least $1 - \alpha$ regardless of the system variances. If the upper bound for $\mu_i - \mu_k$ is less than or equal to 0 then we can conclude that system i has lower expected retrieval time than the default.

5.2 A Selection Procedure

The problem considered here differs from that in §4 since we now consider selection with respect to a default system. The following single-stage procedure was proposed by Paulson (1952) for the case in which σ^2 is *known*. For the given k and specified (δ, P_0, P_1), let $h = Z^{(1-P_0)}_{k-1,1/2}$ be determined from, e.g., Table A.1 of BSG (1995). (These Z constants used to implement the procedure are upper equicoordinate points of a certain multivariate normal distribution.) Further, let n be the solution of
$$\int_{-\infty}^{\infty} \Phi\left(x + \frac{\delta\sqrt{n}}{\sigma} - h\sqrt{2}\right)\Phi^{k-2}\left(x + \frac{\delta\sqrt{n}}{\sigma}\right)d\Phi(x) \ge P_1$$
where $\Phi(\cdot)$ is the standard normal c.d.f. This equation can be solved for n using the FORTRAN program MVNPRD in Dunnett (1989).

P

1. Take a random sample of n observations $Y_{i1}, Y_{i2}, \ldots, Y_{in}$ in a single stage from each system, including the default, k.

2. Calculate the k sample means

$$\bar{Y}_i = \sum_{j=1}^{n} Y_{ij}/n$$

for $i = 1, 2, \ldots, k$, and let $\bar{Y}_{[k-1]} = \max\{\bar{Y}_1, \ldots, \bar{Y}_{k-1}\}$ denote the largest (non-default) sample mean.

3. If $\bar{Y}_{[k-1]} > \bar{Y}_k + h\sigma\sqrt{2/n}$, select the treatment associated with $\bar{Y}_{[k-1]}$; otherwise, select the default.

REFERENCES

Bechhofer, R. E., S. Elmaghraby and N. Morse. 1959. A single-sample multiple decision procedure for selecting the multinomial event which has the highest probability. *Annals of Mathematical Statistics* 30:102–119.

Bechhofer, R. E., and D. M. Goldsman. 1986. Truncation of the Bechhofer-Kiefer-Sobel sequential procedure for selecting the multinomial event which has the largest probability (II): extended tables and an improved procedure. *Communications in Statistics—Simulation and Computation* B15:829–851.

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

Bechhofer, R. E., and B. W. Turnbull. 1978. Two $(k+1)$-decision selection procedures for comparing k normal means with a specified standard. *Journal of the American Statistical Association* 73:385–392.

Dunnett, C. W. 1989. Multivariate normal probability integrals with product correlation structure. *Applied Statistics* 38:564–579. Correction: 42:709.

Gibbons, J. D., I. Olkin and M. Sobel. 1977. *Selecting and ordering populations: A new statistical methodology.* New York: John Wiley and Sons.

Goldsman, D., B. L. Nelson and B. Schmeiser. 1991. Methods for selecting the best system. In *Proceedings of the 1991 Winter Simulation Conference*, ed. B. L. Nelson, W. D. Kelton, and G. M. Clark, 177–186. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Gupta, S. S. 1956. *On a decision rule for a problem in ranking means.* Ph.D. Dissertation (Mimeo. Ser. No. 150). Inst. of Statist., Univ. of North Carolina, Chapel Hill, NC.

Gupta, S. S. 1965. On some multiple decision (selection and ranking) rules. *Technometrics* 7:225–245.

Gupta, S. S., and D.-Y. Huang. 1976. Subset selection procedures for the means and variances of normal populations: unequal sample sizes case. *Sankhyā* B38:112–128.

Hayter, A. J. 1984. A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Statistics* 12:61–75.

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

Hsu, J. C. 1984. Constrained simultaneous confidence intervals for multiple comparisons with the best. *Annals of Statistics* 12:1136–1144.

Hsu, J. C., and M. Peruggia. 1994. Graphical representations of Tukey's multiple comparison method. *Journal of Computational and Graphical Statistics*, forthcoming.

Law, A. M., and W. D. Kelton. 1991. *Simulation modeling & analysis.* 2nd ed. New York: McGraw-Hill, Inc.

Matejcik, F. J., and B. L. Nelson. 1992. Two-stage multiple comparisons with the best for computer simulation. *Operations Research*, forthcoming.

Miller, R. G. 1981. *Simultaneous statistical inference.* 2nd ed. New York: Springer-Verlag.

Matejcik, F. J., and B. L. Nelson. 1993. Simultaneous ranking, selection and multiple comparisons for simulation. In *Proceedings of the 1993 Winter Simulation Conference*, ed. G. W. Evans, M. Mollaghasemi, E. C. Russell and W. E. Biles, 386–392. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Nelson, B. L., and F. J. Matejcik. 1993. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. Submitted to *Management Science.*

Paulson, E. 1952. On the comparison of several experimental categories with a control. *Annals of Mathematical Statistics* 23:239–246.

Rinott, Y. 1978. On two-stage selection procedures and related probability-inequalities. *Communications in Statistics—Theory and Methods* A7:799–811.

Tamhane, A. C. 1977. Multiple comparisons in model I: One-way ANOVA with unequal variances. *Communications in Statistics—Theory and Methods* A6:15–32.

Wilcox, R. R. 1984. A table for Rinott's selection procedure. *Journal of Quality Technology* 16:97–100.

AUTHOR BIOGRAPHIES

DAVID GOLDSMAN is an Associate Professor in the School of Industrial and Systems Engineering at the Georgia Institute of Technology. His research in-

terests include simulation output analysis and ranking and selection. He is the Area Editor for Simulation of *IIE Transactions*, and the President of the TIMS College on Simulation.

BARRY L. NELSON is an Associate Professor in the Department of Industrial, Welding and Systems Engineering at The Ohio State University, with a joint appointment in the Department of Statistics. His research interests are experiment design and analysis of stochastic simulations, specifically variance reduction and multivariate output analysis. He teaches courses in simulation and stochastic processes, and is the author of *Stochastic Modeling: Analysis and Simulation* published by McGraw-Hill. He is Past-President of the TIMS College on Simulation, the Area Editor for Simulation of *Operations Research*, and an Associate Editor for *ACM Transactions on Modeling and Computer Simulation*.