Chapter 17: Selecting the Best System

Seong-Hee Kim

School of Industrial & Systems Engineering, Georgia Institute of Technology

Barry L. Nelson

Department of Industrial Engineering & Management Sciences, Northwestern University

Abstract

We describe the basic principles of ranking and selection, a collection of experiment-design techniques for comparing "populations" with the goal of finding the best among them. We then describe the challenges and opportunities encountered in adapting ranking-and-selection techniques to stochastic simulation problems, along with key theorems, results and analysis tools that have proven useful in extending them to this setting. Some specific procedures are presented along with a numerical illustration.

1 Introduction

Over the last twenty years there has been considerable effort expended to develop statistically valid ranking-and-selection (R&S) procedures to compare a finite number of simulated alternatives. There exist at least four classes of comparison problems that arise in simulation studies: selecting the system with the largest or smallest expected performance measure (selection of the best), comparing all alternatives against a standard (comparison with a standard), selecting the system with the largest probability of actually being the best performer (multinomial selection), and selecting the system with the largest probability of success (Bernoulli selection). For all of these problems, a constraint is imposed either on the probability of correct selection (PCS) or on the simulation budget. Some procedures find a desirable system with a guarantee on the PCS, while other procedures maximize the PCS under the budget

Email addresses: `shkim@isye.gatech.edu` (Seong-Hee Kim), `nelsonb@northwestern.edu` (Barry L. Nelson).

constraint. Our focus in this chapter is on selection-of-the-best problems with a PCS constraint. A good procedure is one that delivers the desired PCS efficiently (with minimal simulated data) and is robust to modest violations of its underlying assumptions. Other types of comparison problems and procedures will be discussed briefly in Section 7. In this chapter "best" means maximum expected value of performance, such as expected throughput or profit.

Traditional roles for R&S are selecting the best system from among a (typically small) number of simulated alternatives and screening a relatively large number of simulated alternatives to quickly discard those whose performance is clearly inferior to the best. More recently, R&S procedures are playing an important role in optimization via simulation. Many algorithms for optimization via simulation search the feasible solution space by some combination of randomly sampling solutions and exploring the neighborhood of good solutions (see Chapters 18–21). R&S procedures can be embedded within these algorithms to help them make improving moves correctly and efficiently. In addition, at the end of an optimization-via-simulation search, R&S procedures can be applied to those solutions that were visited by the search to provide a statistical guarantee that the solution returned as best is at least the best of all the solutions actually simulated. See, for instance, Boesel et al. (2003) and Pichitlamken and Nelson (2001) for more on the application of R&S in this context.

Rather than present a comprehensive survey of R&S procedures, or provide a guide for applying them, our goal is to explain how such procedures are constructed, emphasizing issues that are central to designing procedures for computer simulation, and reviewing some key theorems that have proven useful in deriving procedures. We do, however, present three specific R&S procedures as illustrations. See Goldsman and Nelson (1998) and Law and Kelton (2000) for detailed "how to" guides, Bechhofer et al. (1995) for a comprehensive survey of R&S procedures, and Hochberg and Tamhane (1987) or Hsu (1996) for closely related multiple comparison procedures (MCPs).

The chapter is organized as follows: In Section 2 we show how R&S procedures are derived in an easy, but unrealistic, setting. Section 3 lists the challenges and opportunities encountered in simulation problems, along with key theorems and results that have proven useful in extending R&S procedures to this setting. Three specific procedures are presented in Section 4, followed by a numerical illustration in Section 5. Section 6 reviews asymptotic analysis regimes for R&S. Section 7 describes other formulations of the R&S problem and gives appropriate references. Section 8 closes the chapter by speculating on future research directions in this area.

2 Basics of Ranking and Selection

In this section we employ the simplest possible setting to illustrate how R&S procedures address comparison problems. This setting (i.i.d. normal data with known, common variance) allows us to focus on key techniques before moving on to the technical difficulties that arise in designing procedures for realistic simulation problems.

R&S traces its origins to two papers: Bechhofer (1954) established the *indifference-zone formulation*, while Gupta (1956, 1965) is credited with the *subset selection formulation* of the problem. Both approaches are reviewed in this section, and both were developed to compensate for the limited inference provided by hypothesis tests for the homogeniety of k population parameters (usually means). In many experiments, rejecting the hypothesis $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, where μ_i is the parameter associated with the ith population, leads naturally to questions about which one has the largest or smallest parameter. R&S tries to answer such questions. MCPs also provide inference beyond rejection of homogeniety; there is a close connection between R&S and MCPs, as we demonstrate later.

Suppose that there are k systems. Let X_{ij} represent the jth output from system i and let $\mathbf{X}_i = \{X_{ij}\ ; j = 1, 2, \ldots\}$ denote the output sequence from system i. In this section, we assume that the X_{ij} are i.i.d. normal with means $\mu_i = E[X_{ij}]$ and variances $\sigma_i^2 = \mathrm{Var}[X_{ij}]$. Further, we assume that the processes $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k$ are mutually independent, and the variances are known and equal; that is, $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$. These are unrealistic assumptions that will be relaxed later, but we adopt them here because we can derive R&S procedures in a way that illustrates the key issues. Throughout the chapter we assume that a larger mean is better, and we let $\mu_k \geq \mu_{k-1} \geq \cdots \geq \mu_1$, so that (unknown to us) system k is the best system.

2.1 Subset-Selection Formulation

Suppose that we have n outputs from each of the systems. Our goal is to use these data to obtain a subset $I \subseteq \{1, 2, \ldots, k\}$ such that

$$\Pr\{k \in I\} \geq 1 - \alpha \tag{1}$$

where $1/k < 1 - \alpha < 1$. Ideally $|I|$ is small, the best case being $|I| = 1$. Gupta's solution was to include in the set I all systems ℓ such that

$$\bar{X}_\ell(n) \geq \max_{i \neq \ell} \bar{X}_i(n) - h\sigma\sqrt{\frac{2}{n}} \tag{2}$$

where $\bar{X}_i(n)$ is the sample mean of the (first) n outputs from system i, and h is a constant whose value will depend on k and $1-\alpha$. The proof that rule (2) provides guarantee (1) is instructive and shows what the value of h should be:

$$\begin{aligned}
\Pr\{k \in I\} &= \Pr\left\{\bar{X}_k(n) \geq \max_{i \neq k} \bar{X}_i(n) - h\sigma\sqrt{\frac{2}{n}}\right\} \\
&= \Pr\left\{\bar{X}_k(n) \geq \bar{X}_i(n) - h\sigma\sqrt{\frac{2}{n}}, \forall i \neq k\right\} \\
&= \Pr\left\{\frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\sigma\sqrt{2/n}} \leq h - \frac{(\mu_i - \mu_k)}{\sigma\sqrt{2/n}}, \forall i \neq k\right\} \\
&\geq \Pr\left\{Z_i \leq h, i = 1, 2, \ldots, k-1\right\} = 1 - \alpha
\end{aligned}$$

where $(Z_1, Z_2, \ldots, Z_{k-1})$ have a multivariate normal distribution with means 0, variances 1, and common pairwise correlations 1/2. Therefore, to provide the guarantee (1), h needs to be the $1-\alpha$ quantile of the maximum of such a multivariate normal random vector, a quantile that turns out to be relatively easy to evaluate numerically. Notice the inequality in the final step where we make use of the fact that $\mu_k \geq \mu_i$.

A theme that runs throughout much of R&S is first using appropriate standardization of estimators and then bounding the resulting probability statements in such a way that a difficult multivariate probability statement becomes one that is readily solvable.

2.2 Indifference-Zone Formulation

A disadvantage of the subset-selection procedure in Section 2.1 is that the retained set I may, and likely will, contain more than one system. However, there is no procedure that can guarantee a subset of size 1 and satisfy (1) for arbitrary n. Even when n is under our control, as it is in computer simulation, the appropriate value will depend on the true differences $\mu_k - \mu_i, \forall i \neq k$. To address this problem, Bechhofer (1954) suggested the following compromise: guarantee to select the single best system, k, whenever $\mu_k - \mu_{k-1} \geq \delta$, where $\delta > 0$ is the smallest difference the experimenter feels is worth detecting. Specifically, the procedure should guarantee

$$\Pr\{\text{select } k | \mu_k - \mu_{k-1} \geq \delta\} \geq 1 - \alpha \tag{3}$$

where $1/k < 1-\alpha < 1$. If there are systems whose means are within δ of the best, then the experimenter is "indifferent" to which of these is selected,

leading to the term indifference-zone (IZ) formulation.

The procedure is as follows: From each system, take

$$n = \left\lceil \frac{2h^2\sigma^2}{\delta^2} \right\rceil \tag{4}$$

outputs, where h is an appropriate constant (determined below) and $\lceil x \rceil$ means to round x up; then select the system with the largest sample mean as the best. Assuming $\mu_k - \mu_{k-1} \geq \delta$,

$$\begin{aligned}
\Pr\{\text{select } k\} &= \Pr\left\{\bar{X}_k(n) > \bar{X}_i(n), \forall i \neq k\right\} \\
&= \Pr\left\{\frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\sigma\sqrt{2/n}} < -\frac{(\mu_i - \mu_k)}{\sigma\sqrt{2/n}}, \forall i \neq k\right\} \\
&\geq \Pr\left\{\frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\sigma\sqrt{2/n}} < \frac{\delta}{\sigma\sqrt{2/n}}, \forall i \neq k\right\} \\
&\geq \Pr\left\{\frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\sigma\sqrt{2/n}} < h, \forall i \neq k\right\} \\
&= \Pr\left\{Z_i < h, i = 1, 2, \ldots, k-1\right\} = 1 - \alpha
\end{aligned}$$

where again $(Z_1, Z_2, \ldots, Z_{k-1})$ has a multivariate normal distribution with means 0, variances 1, and common pairwise correlations 1/2, implying h needs to be the $1-\alpha$ quantile of the maximum of such a multivariate normal random vector.

Notice that the first inequality results from the assumption that $\mu_k - \mu_{k-1} \geq \delta$, while the second occurs because $\sqrt{n} \geq \sqrt{2}h\sigma/\delta$. Both of these tricks are standard: the first frees the probability statement of dependence on the true means, while the second frees it of dependence on the value of the variance.

It is worth noting that, over all configurations of the true means such that $\mu_k - \mu_{k-1} \geq \delta$, the configuration $\mu_i = \mu_k - \delta, \forall i \neq k$ minimizes the PCS; it is therefore known as the *least-favorable configuration* (LFC). In this chapter we break from the statistics literature in that we will not be concerned with identifying the LFC; our only interest is insuring that (3) is met.

Bechhofer's procedure is essentially a power calculation: how large a sample is required to detect differences of at least δ? When true differences are greater than δ, Bechhofer's n may be much larger than needed. By taking observations and making decisions sequentially, it is often possible to reach an earlier decision. Sequential selection procedures can be traced back at least to Wald

(1947), but here we present a procedure due to Paulson (1964) that better illustrates the approach that has had the most impact in computer simulation. Paulson's procedure takes observations *fully sequentially*—meaning one at a time—and *eliminates* systems from continued sampling when it is statistically clear that they are inferior. Thus, simulation for a problem with a single dominant alternative may terminate very quickly.

Using the same notation as above, let $\bar{X}_i(r)$ be the sample mean of the first r outputs of system i. At each stage $r = 1, 2, \ldots, n$, one output is taken from each system whose index is in I, where initially $I = \{1, 2, \ldots, k\}$. At stage r, system ℓ is retained in I only if

$$\bar{X}_\ell(r) \geq \max_{i \in I} \bar{X}_i(r) - \max\{0, a/r - \lambda\} \tag{5}$$

where $a > 0$ and $0 < \lambda < \delta$ are constants to be determined, and $n = \lfloor a/\lambda \rfloor$, with $\lfloor \cdot \rfloor$ meaning round down. The procedure ends when $|I| = 1$, which requires no more than $n + 1$ stages. Parallels with Gupta's subset selection and Bechhofer's IZ ranking are obvious: At each stage a subset selection is performed, with the hedging factor $(a/r - \lambda)$ decreasing as more data are obtained. In the end, if the procedure makes it that far, the system with the largest sample mean is selected.

The following result is used to establish the PCS: Suppose $Z_1, Z_2, \ldots$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$ with $\mu < 0$. Then it can be shown that

$$\Pr\left\{\bar{Z}(r) > \frac{a}{r}, \text{ for some } r < \infty\right\} \leq \exp\left(\frac{2\mu}{\sigma^2} a\right). \tag{6}$$

This result is a consequence of Wald's lemma (Wald, 1947, p. 146). *Large deviation results, frequently based on the analysis of approximating Brownian motion processes, are central to the design of fully sequential procedures that involve frequent looks at the data.*

The approach in this case is to bound the probability of an *incorrect selection* (ICS). An ICS event occurs if system k is eliminated at some point during the procedure. Let $\Pr\{\mathrm{ICS}_i\}$ be the probability of an incorrect selection if only systems i and k are included in the competition.

The first key inequality is

$$\Pr\{\mathrm{ICS}\} \leq \sum_{i=1}^{k-1} \Pr\{\mathrm{ICS}_i\}. \tag{7}$$

Decomposition into some form of paired comparisons is a key step in many sequential procedures.

This decomposition allows us to focus only on $\Pr\{\text{ICS}_i\}$. Notice that

$$\begin{aligned}
&\Pr\{\text{ICS}_i\} \\
&\leq \Pr\left\{\bar{X}_k(r) < \bar{X}_i(r) + \lambda - a/r, \text{ for some } r \leq n+1\right\} \\
&= \Pr\left\{\bar{X}_i(r) - \bar{X}_k(r) + \lambda > a/r, \text{ for some } r \leq n+1\right\} \\
&\leq \Pr\left\{\bar{X}_i(r) - \bar{X}_k(r) + \lambda > a/r, \text{ for some } r < \infty\right\} \\
&\leq \exp\left(\frac{(\mu_i - \mu_k + \lambda)}{\sigma^2} a\right) \\
&\leq \exp\left(\frac{(\lambda - \delta)}{\sigma^2} a\right).
\end{aligned}$$

The third inequality comes from the large deviation result (6), while the fourth inequality exploits the indifference-zone assumption. If we set

$$a = \ln\left(\frac{k-1}{\alpha}\right)\frac{\sigma^2}{\delta - \lambda} \tag{8}$$

then $\Pr\{\text{ICS}_i\} \leq \alpha/(k-1)$ and

$$\Pr\{\text{ICS}\} \leq (k-1)\frac{\alpha}{(k-1)} = \alpha.$$

2.3 Connection to Multiple Comparisons

MCPs approach the comparison problem by providing simultaneous confidence intervals on selected differences among the systems' parameters. Hochberg and Tamhane (1987) and Hsu (1996) are good comprehensive references. As noted by Hsu (1996, pp. 100-102), the connection between R&S and MCPs comes through multiple comparisons with the best (MCB). MCB forms simultaneous confidence intervals for $\mu_i - \max_{\ell \neq i} \mu_\ell, i = 1, 2, \ldots, k$, the difference between each system and the best of the rest. Specialized to the known-variance case, the intervals take the form

$$\mu_i - \max_{\ell \neq i} \mu_\ell \in \left[-\left(\bar{X}_i(n) - \max_{\ell \neq i} \bar{X}_\ell(n) - h\sigma\sqrt{\frac{2}{n}}\right)^-\right.,$$

$$\left(\bar{X}_i(n) - \max_{\ell \neq i} \bar{X}_\ell(n) + h\sigma\sqrt{\frac{2}{n}}\right)^+\Bigg] \tag{9}$$

where h is the same critical value used in Bechhofer's and Gupta's procedures, $-x^- = \min\{0, x\}$ and $x^+ = \max\{0, x\}$. Under our assumptions these k confidence intervals are simultaneously correct with probability $\geq 1 - \alpha$.

Consider the set I containing the indices of all systems whose MCB upper confidence bound is greater than 0. Thus, for $i \in I$,

$$\bar{X}_i(n) > \max_{\ell \neq i} \bar{X}_\ell(n) - h\sigma\sqrt{\frac{2}{n}}$$

meaning these are the same systems that would be retained by Gupta's subset-selection procedure. Since $\mu_k - \max_{\ell \neq k} \mu_\ell > 0$, and these intervals are simultaneously correct with probability $\geq 1 - \alpha$, system k will be in the subset identified by the MCB upper bounds with the required probability.

Now suppose that n has been selected such that $n \geq 2h^2\sigma^2/\delta^2$, implying that

$$h\sigma\sqrt{\frac{2}{n}} \leq \delta$$

as in Bechhofer's procedure. Let B be the index of the system with the largest sample mean. Then the MCB lower bounds guarantee with probability $\geq 1-\alpha$ that

$$\begin{aligned}\mu_B - \max_{\ell \neq B} \mu_\ell &\geq -\left(\bar{X}_B(n) - \max_{\ell \neq B} \bar{X}_\ell(n) - h\sigma\sqrt{\frac{2}{n}}\right)^- \\ &\geq -\delta.\end{aligned}$$

The final inequality follows because $\bar{X}_B(n) - \max_{\ell \neq B} \bar{X}_\ell(n) \geq 0$ by the definition of B, and $h\sigma\sqrt{2/n} \leq \delta$ because of our choice of n. This establishes that the system selected by Bechhofer's procedure is guaranteed to be within δ of the true best, with probability $\geq 1-\alpha$, *under any configuration of the means.* Further, if $\mu_k - \mu_{k-1} > \delta$, then $\Pr\{B = k\} \geq 1 - \alpha$ as required.

As a consequence of this analysis both Bechhofer's and Gupta's procedures can be augmented with MCB confidence intervals "for free," and Bechhofer's procedure is guaranteed to select a system within δ of the best. Nelson and Matejcik (1995) establish very mild conditions under which these results hold for far more general R&S procedures.

3 Simulation Issues and Key Results

In the previous section we illustrated different approaches to the R&S problem under assumptions such as independence, normality, and known and equal variances. Unfortunately, such assumptions rarely hold in simulation experiments. There are also opportunities available in simulation experiments that are not present in physical experiments. In the following subsections we describe these issues and opportunities, and present key theorems and results that have been useful in deriving R&S procedures that overcome or exploit them.

3.1 Unknown and Unequal Variances

Unknown and unequal variances across alternatives is a fact of life in system simulation problems, and the variances can differ dramatically. In the simple inventory model presented in Section 5 the ratio of the largest to smallest variance is almost 4.

There are many subset-selection procedures that permit an unknown, common variance (see Goldsman and Nelson 1998 for one). When variances are unknown and unequal, however, the subset-selection problem is essentially equivalent to the famous Behrens-Fisher problem. One approach is to work with the standardized random variables

$$\frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\left(\frac{S_i^2}{n} + \frac{S_k^2}{n}\right)^{1/2}}, i = 1, 2, \ldots, k-1 \tag{10}$$

where S_i^2 is the sample variance of the outputs from system i. Neither the joint nor marginal distributions of these quantities are conveniently characterized. If we break the required joint probability statement up into statements about the individual terms, using techniques described below, then there are at least two solutions available. Welch (1938) suggested approximating each term in (10) as having a $t_{\hat{\nu}}$ distribution, where the degrees of freedom $\hat{\nu}$ is an approximation based on the values of S_i^2 and S_k^2. Banerjee (1961) proposed a probability bound that we specialize to our case:

Theorem 1 (Banerjee 1961) *Suppose Z is $\mathrm{N}(0,1)$ and independent of Y_i and Y_k, which are themselves independent χ^2_ν random variables. Then for arbitrary but fixed $0 \le \gamma \le 1$,*

$$\Pr\left\{\frac{Z^2}{\gamma\frac{Y_i}{\nu} + (1-\gamma)\frac{Y_k}{\nu}} \le t^2_{1-\alpha/2,\nu}\right\} \ge 1 - \alpha \tag{11}$$

where $t_{1-\alpha/2,\nu}$ is the $1-\alpha/2$ quantile of the t distribution with ν degrees of freedom.

To employ Banerjee's inequality in our context, identify

$$Z = \frac{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k)}{\left(\frac{\sigma_i^2}{n} + \frac{\sigma_k^2}{n}\right)^{1/2}}$$

and

$$\begin{aligned}\gamma\frac{Y_i}{\nu} + (1-\gamma)\frac{Y_k}{\nu} &= \frac{\frac{S_i^2}{n} + \frac{S_k^2}{n}}{\frac{\sigma_i^2}{n} + \frac{\sigma_k^2}{n}} \\ &= \left(\frac{\sigma_i^2}{\sigma_i^2 + \sigma_k^2}\right)\frac{S_i^2}{\sigma_i^2} + \left(\frac{\sigma_k^2}{\sigma_i^2 + \sigma_k^2}\right)\frac{S_k^2}{\sigma_k^2}.\end{aligned}$$

This inequality is used in Procedure NSGS presented in Section 4.

For some time it has been known that it is not possible to provide a guaranteed PCS, in the IZ sense, with a single stage of sampling when variances are unknown (see Dudewicz 1995 for a comprehensive discussion of this result). Thus, practically useful IZ procedures work sequentially—meaning two or more stages of sampling—with the first stage providing variance estimates that help determine how much, if any, additional sampling is needed in the succeeding stages. However, one cannot simply substitute variance estimators into Bechhofer's or Paulson's procedures and hope to achieve a guaranteed PCS. Instead, the uncertainty in the variance estimators enters into the determination of the sample sizes, invariably leading to more sampling than would take place if the variances were known.

A fundamental result in parametric statistics is the following: If $X_1, X_2, \ldots, X_n$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$, then $\bar{X}$ and S^2 are independent random variables. The result extends in the natural way to random vectors $\mathbf{X}_j$ that are multivariate normal. An extension of a different sort, due to Stein (1945), is fundamental to R&S procedures with unknown variances:

Theorem 2 (Stein 1945) *Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, and S^2 is $\sigma^2\chi^2_\nu/\nu$ and independent of $\sum_{i=1}^n X_j$ and of $X_{n+1}, X_{n+2}, \ldots$.*

(1) If $N \geq n$ is a function only of S^2 then

$$\frac{\bar{X}(N) - \mu}{S/\sqrt{N}} \sim t_\nu. \tag{12}$$

(2) If $\xi > 0$ and

$$N = \max\left\{\left\lceil \frac{S^2}{\xi^2} \right\rceil, n+1\right\}$$

then for any weights $w_1, w_2, \ldots, w_N$ satisfying $\sum_{j=1}^N w_j = 1$, $w_1 = w_2 = \cdots = w_n$, and $S^2 \sum_{j=1}^N w_j^2 = \xi^2$ we have

$$\frac{\sum_{j=1}^N w_j X_j - \mu}{\xi} \sim t_\nu. \tag{13}$$

In the usual case where S^2 is the sample variance of the first n observations, $\nu = n-1$. The importance of this result in R&S is that it allows determination of a sample size large enough to attain the desired power against differences of at least δ without requiring knowledge of the process variance.

If comparisons of only $k = 2$ systems were necessary, then Stein's result would be enough (at least in the i.i.d. normal case). But our problem is multivariate and requires joint probability statements about

$$\frac{\bar{X}_i(N_i) - \bar{X}_k(N_k) - (\mu_i - \mu_k)}{\mathcal{S}_{ik}}, i = 1, 2, \ldots, k-1 \tag{14}$$

where $\mathcal{S}_{ik}^2$ is a variance estimate of the difference between systems i and k based on an initial sample of size (say) n, and N_i and N_k are the final sample sizes from systems i and k. The joint distribution of these random variables is quite messy in general, even if all systems are simulated independently (as we assume in this section). One approach is to condition on $\mathcal{S}_{ik}$ and $\bar{X}_k(N_k)$ and apply inequalities such as the following to bound the joint probability:

Theorem 3 (Kimball 1951) *Let $V_1, V_2, \ldots, V_k$ be independent random variables, and let $g_j(v_1, v_2, \ldots, v_k), j = 1, 2, \ldots, p$, be nonnegative, real-valued functions, each one nondecreasing in each of its arguments. Then*

$$\mathrm{E}\left[\prod_{j=1}^p g_j(V_1, V_2, \ldots, V_k)\right] \geq \prod_{j=1}^p \mathrm{E}\left[g_j(V_1, V_2, \ldots, V_k)\right].$$

Kimball's theorem is actually only for the case $k = 1$; see Hochberg and Tamhane (1987) for the extension.

Theorem 4 (Slepian 1962) *Let $(Z_1, Z_2, \ldots, Z_k)$ have a k-variate normal distribution with zero mean vector, unit variances, and correlation matrix*

$\mathbf{R} = \{\rho_{ij}\}$. *Let* $\xi_1, \xi_2, \ldots, \xi_k$ *be some constants. If all the* $\rho_{ij} \geq 0$, *then*

$$\Pr\left\{\bigcap_{i=1}^{k}(Z_i \leq \xi_i)\right\} \geq \prod_{i=1}^{k} \Pr\{Z_i \leq \xi_i\}.$$

Notice that, conditional on the $\mathcal{S}_{ik}^2$, the terms in (14) are positively correlated (due to the common $\bar{X}_k(N_k)$ term), providing the opening to apply Slepian's inequality. Kimball's inequality then can be applied to simplify the unconditioning on $\mathcal{S}_{ik}^2$. Both of these ideas are employed in the design of Procedure NSGS in Section 4.

3.2 Initial Sample Size Problem

When variances are unknown, then at least two stages of sampling are required to deliver a guaranteed PCS. In a typical two-stage R&S procedure, such as Rinott's (1978) procedure, the total sample size required of, say, system i is:

$$N_i = \max\left\{n_0, \left\lceil\left(\frac{hS_i}{\delta}\right)^2\right\rceil\right\} \tag{15}$$

where $h = h(k, 1-\alpha, n_0)$ is a constant determined by k, the number of systems being compared; $1-\alpha$, the desired confidence level; and n_0, the number of first-stage observations used to produce the variance estimator, S_i^2. The constant h increases in k, and decreases in α and n_0. The experiment design factor that is under our control is n_0.

Figure 1 presents the typical form of $\mathrm{E}[N_i]$ as a function of n_0. The figure shows that increasing n_0, up to a point, decreases $\mathrm{E}[N_i]$, but if n_0 is too large then more data are obtained in the first stage than required to deliver the PCS guarantee. Unfortunately, the location of the minimizing value of n_0 depends on the unknown variance. Nevertheless, it is clear that there is a huge penalty for selecting n_0 too small, which forces an excessive second-stage sample to compensate for the highly unstable variance estimator. Taking $n_0 \geq 10$ is a common recommendation.

3.3 Non-normality of Output Data

Raw output data from industrial and service simulations are rarely normally distributed. Surprisingly, non-normality is usually not a concern in simulation experiments that (a) are designed to make multiple independent replications,

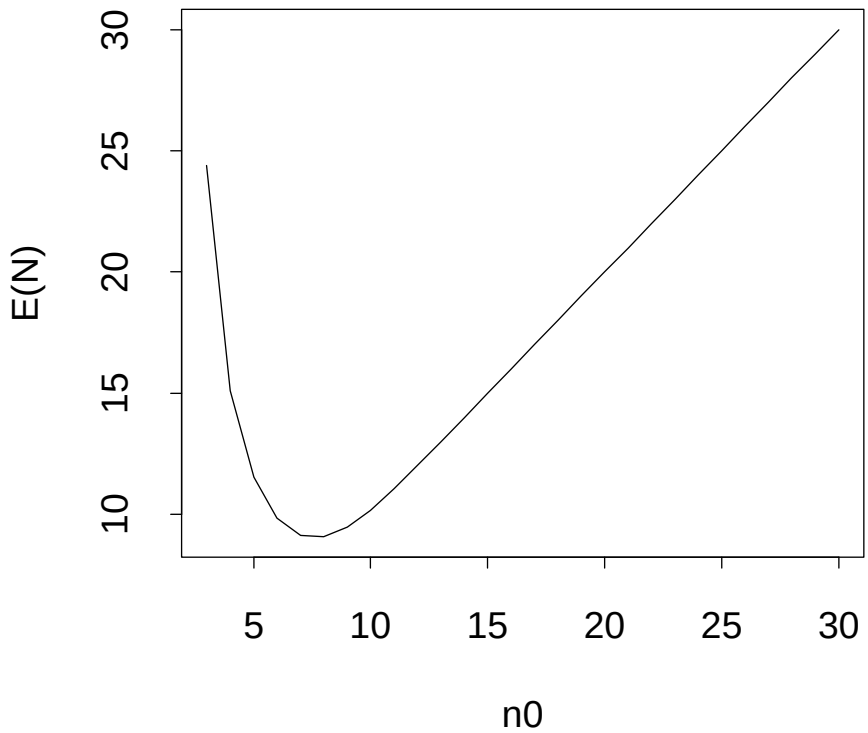

Fig. 1. Illustration of the impact of n_0 on E[N].

and (b) use a within-replication average of a large numbers of raw simulation outputs as the basic summary measure. This is frequently the situation for so-called "terminating simulations" in which the initial conditions and stopping time for each replication are an inherent part of the definition of the system. A standard example is a store that opens empty at 6 AM, then closes when the last customer to arrive before 9 PM leaves the store. If the output of interest is the average customer delay in the checkout line over the course of the day, and comparisons will be based on the expected value of this average, and the average is over many individual customer delays, then the Central Limit Theorem suggests that the replication averages will be approximately normally distributed.

Difficulties arise in so-called "steady-state simulations" where the parameter of interest is defined by a limit as the time index of a stochastic process approaches infinity (and therefore forgets its initial conditions). Some steady-state simulations are amenable to multiple replications of each alternative and within-replication averages as summary statistics, in which case the preceding

discussion applies. Unfortunately, severe estimator bias due to residual effects of the initial conditions sometimes force an experiment design consisting of a single, long replication from each alternative. The raw outputs within each replication are typically neither normally distributed nor independent. For example, waiting times of individual customers in a queueing system are usually dependent because a long delay for one customer tends to increase the delays of the customers who follow. The best we can hope for is an approximately stationary output process from each system, but neither normality nor independence.

The most common approach for dealing with this problem is to transform the raw data into *batch means*, which are averages of large number of raw outputs. The batch means are often far less dependent and non-normal than the raw output data. There are problems with the batching approach for R&S, however. If a "stage" is defined by batch means rather than raw output, then the simulation effort consumed by a stage is a multiple of the batch size. When a large batch size is required to achieve approximate independence—and batch sizes of several thousand are common—then the selection procedure is forced to make decisions at long intervals, wasting outputs and time. This inefficiency becomes serious when fully sequential procedures are employed because the elimination decisions for clearly inferior systems must wait for an entire batch to be formed. Therefore, for steady-state simulations, selection procedures that use individual raw outputs as basic observations are desirable.

Although no known procedures provide a guaranteed PCS for single-replication designs, some procedures have shown good empirical performance (e.g., Sullivan and Wilson 1989), while others have been shown to be asymptotically valid (e.g., Procedure $\mathcal{KN}$++ in Section 4). See Law and Kelton (2000) or Chapter 15 for a general discussion of replications versus batching, Glynn and Iglehart (1990) for conditions under which the batch means method is asymptotically valid for confidence intervals, and Section 6 for a review of asymptotic analysis of R&S procedures.

3.4 Common Random Numbers

The procedures described in Section 2 assume that data across the k alternative systems are independent. In simulation experiments this assumption can be made valid by using different sequences of random numbers to drive the simulation of each system (see Chapter 3). However, since we are making comparisons, there is a potential advantage of using common random numbers (CRN) to drive the simulation of each system because

$$\mathrm{Var}[X_{ij} - X_{\ell j}] = \mathrm{Var}[X_{ij}] + \mathrm{Var}[X_{\ell j}] - 2\mathrm{Cov}[X_{ij}, X_{\ell j}].$$

If implemented correctly (see Banks, et al. 2005), CRN tends to make $\mathrm{Cov}[X_{ij}, X_{\ell j}] > 0$ thereby reducing the variance of the difference.

R&S procedures often need to make probability statements about the collection of random variables

$$\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k), i = 1, 2, \ldots, k-1. \tag{16}$$

The appearance of the common term $\bar{X}_k(n)$ causes dependence among these random variables, but it is often easy to model or tightly bound. The introduction of CRN induces dependence between $\bar{X}_i(n)$ and $\bar{X}_k(n)$ as well. Even though the sign of the induced covariance is believed known, its value is not, making it difficult to say anything about the dependence among the differences (16).

Two approaches are frequently used. The first is to replace the basic data $\{X_{ij}; i = 1, 2, \ldots, k; j = 1, 2, \ldots, n\}$ with pairwise differences $\{X_{ij} - X_{\ell j}; i \neq \ell; j = 1, 2, \ldots, n\}$ because the variance of the sample mean of the difference includes the effect of the CRN-induced covariance. The second is to apply the Bonferroni inequality to break up joint statements about (16) into statements about the individual terms. Recall that for events $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_{k-1}$, the Bonferroni inequality states that

$$\Pr\left\{\bigcap_{i=1}^{k-1} \mathcal{E}_i\right\} \geq 1 - \sum_{i=1}^{k-1} \Pr\left\{\mathcal{E}_i^c\right\}. \tag{17}$$

In the R&S context $\mathcal{E}_i$ corresponds to an event like $\{\bar{X}_i(n) - \bar{X}_k(n) - (\mu_i - \mu_k) \leq h\}$.

Approaches based on the Bonferroni inequality make no assumption about the induced dependence, and therefore are very conservative. A more aggressive approach is to assume some structure for the dependence induced by CRN. One standard assumption is that all pairwise correlations $\rho = \mathrm{Corr}[X_{ij}, X_{\ell j}]$ are positive, and identical, and all variances are equal; this is known as *compound symmetry*. Nelson and Matejcik (1995) extended Rinott's procedure (1978)—one of the simplest and most popular IZ procedures—in conjunction with CRN under a more general structure called *sphericity*. The specific assumption is

$$\mathrm{Cov}[X_{ij}, X_{\ell j}] = \begin{cases} 2\beta_i + \tau^2, & i = \ell \\ \beta_i + \beta_\ell, & i \neq \ell \end{cases} \tag{18}$$

with $\tau^2 > 0$, which is equivalent to assuming that $\mathrm{Var}[X_{ij} - X_{\ell j}] = 2\tau^2$ for all $i \neq \ell$, a type of variance balance. This particular structure is useful because there exists an estimator $\widehat{\tau}^2$ of τ^2 that is independent of the sample means and has a χ^2 distribution (allowing a pivotal quantity to be formed and Stein's theorem to be applied). Nelson and Matejcik (1995) showed that procedures based on this assumption are robust to departures from sphericity, at least in part because assuming sphericity is like assuming that all pairwise correlations equal the average pairwise correlation.

3.5 The Sequential Nature of Simulation

Suppose an IZ ranking procedure is applied in the study of k new blood pressure medications. Then "replications" correspond to patients, and the idea of using a fully sequential procedure (assign one patient at a time to each drug, then wait for the results before recruiting the next patient) seems absurd. In simulation experiments, however, data are naturally generated sequentially, at least within each simulated alternative, making multi-stage procedures much more attractive. However, there are some issues:

- In multiple-replication designs, sequential sampling is particularly attractive. All that needs to be retained to start the next stage of sampling is the ending random number seeds from the previous stage. In single-replication designs it can be more difficult to resume sampling from a previous stage, since the entire state of the system must be retained and restored.
- A hidden cost of using multi-stage procedures is the computational overhead in switching among the simulations of the k alternatives. On a single-processor computer, switching can involve saving output, state and seed information from the current system; swapping the program for the current system out of, and for the next system into, active memory; and restoring previous state and seed information for the next system. Thus, the overall computation effort includes both the cost of generating simulated data and the cost of switching. Hong and Nelson (2005) look at sequential IZ procedures that attempt to minimize the total computational cost.
- If k processors are available, then an attractive option is to assign each system to a processor and simulate in parallel. This is highly effective in conjunction with R&S procedures that require little or no coordination between the simulations of each system, such as subset-selection procedures or IZ-ranking procedures that use only variance information (and not differences among the sample means). Unfortunately, a fully sequential procedure with elimination would defeat much of the benefit of parallel processing because communication among the processors is required after generating each output.

Many sequential procedures are based on results for Brownian motion processes. Let $\mathcal{B}(t;\Delta)$ be a standard Brownian motion process with drift Δ. Consider the partial sum of the pairwise difference $D_i(r) = \sum_{j=1}^{r}(X_{kj} - X_{ij})$, $r = 1, 2, \ldots$. If the X_{ij} are i.i.d. normal, and $\mu_k - \mu_i = \delta$, then $\{D_i(r), r = 1, 2, \ldots\} \stackrel{\mathcal{D}}{=} \{\sigma\mathcal{B}(t;\delta/\sigma), t = 1, 2, \ldots\}$, where $\sigma^2 = \mathrm{Var}[X_{kj} - X_{ij}]$ (with or without CRN). In other words, $D_i(r)$ is a Brownian motion process with drift observed only at discrete (integer) points in time. A great deal is known about the probability of Brownian motion processes crossing boundaries in various ways (see, for instance, Siegmund 1985 or Jennison and Turnbull 2000); we display one specific result below. Thus, it seems natural to design R&S procedures for $\sigma\mathcal{B}(t;\delta/\sigma)$ and apply them to $D_i(r)$.

Let $c(t)$ be a symmetric (about 0) continuation region for $\sigma\mathcal{B}(t;\delta/\sigma)$, and let an incorrect selection correspond to the process exiting the region in the wrong direction (down, when the drift is positive). If $T = \inf\{t \geq 0 : |\sigma\mathcal{B}(t;\delta/\sigma)| > c(t)\}$, then

$$\Pr\{\mathrm{ICS}_i\} = \Pr\{\sigma\mathcal{B}(T;\delta/\sigma) < 0\}.$$

Of course $\sigma\mathcal{B}(t;\delta/\sigma)$ is only an approximation for $D_i(r)$. However, Jennison, et al. (1980) show that under very general conditions, $\Pr\{\mathrm{ICS}_i\}$ is no greater if the Brownian motion process is observed at discrete times; thus, procedures designed for $\sigma\mathcal{B}(t;\delta/\sigma)$ provide an upper bound on the probability of incorrect selection for $D_i(r)$. In conjunction with a decomposition into pairwise comparisons, as in (7), this result can be used to derive R&S procedures for $k \geq 2$.

Fabian (1974) tightened the triangular continuation region used by Paulson, and this was exploited by Hartmann (1988, 1991), Kim and Nelson (2001, 2005) and Hong and Nelson (2005).

Theorem 5 (Fabian 1974) *Let $\{\mathcal{B}(t,\Delta), t \geq 0\}$ be a standard Brownian motion with drift $\Delta > 0$. Let*

$$\begin{aligned} l(t) &= -a + \lambda t \\ u(t) &= a - \lambda t \end{aligned}$$

for some $a > 0$ and $\lambda = \Delta/(2b)$ for some positive integer b. Let $c(t)$ denote the continuation region $(l(t), u(t))$ and let T be the first time that $\mathcal{B}(t,\Delta) \notin c(t)$. Then

$$\Pr\{\mathcal{B}(T,\Delta) < 0\} \leq \sum_{j=1}^{b}(-1)^{j+1}\left(1 - \frac{1}{2}\mathcal{I}(j = b)\right)\exp\{-2a\lambda(2b - j)j\}.$$

Fabian's bound on Pr{ICS} is particularly useful because a is the term that depends on the sample variance (see Paulson's a in Equation (8) for intuition). Thus, appropriately standardized, $\exp(-a)$ is related to the moment generating function of a chi-squared random variable, which simplifies unconditioning on the sample variance.

3.6 Large Number of Alternatives

The number of alternatives of interest in simulation problems can be quite large, with 100 or more being relatively common. However, Bechhofer-like IZ procedures were developed for relatively small numbers of alternatives, say no more than 20. They can be inefficient when the number of alternatives is large because they were developed to protect against the LFC—the configuration of system means under which it is most difficult to correctly select the best—to free the procedure from dependence on the true differences among the means. The Slippage Configuration (SC), $\mu_i = \mu_k - \delta$ for $i = 1, 2, \ldots, k-1$, is known to be the LFC for many procedures.

When the number of systems is large we rarely encounter anything remotely like the SC configuration, because large numbers of alternatives typically result from taking all feasible combinations of some controllable decision variables. Thus, the performance measures of the systems are likely to be spread out, rather than all clustered near the best. Paulson-like procedures with elimination might seem to be a cure for this ill, but the inequalities used to decompose the problem of k systems into paired comparisons with system k are typically quite conservative and become much more so with increasing k (although Kim and Nelson's (2001) fully sequential procedure $\mathcal{KN}$, described in the next section, has been shown to work well for up to $k = 500$ systems).

To overcome the inefficiency of IZ approaches for large numbers of alternatives, one idea is to try to gain the benefits of screening, as in Paulson-like procedures, but avoid the conservatism required to compensate for so many looks at the data. Nelson, et al. (2001) proposed spending some of the α for incorrect selection on an initial screening stage (using a Gupta-like subset-selection procedure), and spending the remainder on a second ranking stage (using a Bechhofer-like IZ procedure). Additive and multiplicative α spending is possible, depending on the situation (see Nelson, et al. 2001 and Wilson 2001). The resulting procedure, named NSGS, is presented in the next section.

This so-called "α-spending" approach—spreading the probability of incorrect selection across multiple stages—is a general-purpose tool, and there is no inherent reason to use only a single split. See Jennison and Turnbull (2000) for a thorough discussion.

4 Example Procedures

In this section we present three specific procedures to illustrate the concepts described in earlier sections. The NSGS procedure, due to Nelson, et al. (2001), and the $\mathcal{KN}$ procedure, due to Kim and Nelson (2001), are appropriate for terminating simulations or for steady-state simulations when multiple replications are employed. Procedure $\mathcal{KN}$++, due to Kim and Nelson (2005), is specifically designed for steady-state simulations employing a single replication from each alternative. All of the procedures employ the IZ approach and utilize elimination to gain efficiency in the case of many systems. In all three procedures variances are considered unknown and unequal.

The NSGS procedure requires that the output data from each system be i.i.d. normal, and that outputs across systems be independent, which leaves out CRN. NSGS is the combination of a Gupta-like subset-selection procedure, to reduce the number of alternatives still in play after the first stage of sampling, and a Bechhofer-like ranking procedure applied to the systems in the subset. The procedure uses α-spending between the subset selection and ranking to control the overall PCS. Banerjee's inequality allows the subset-selection procedure to handle unequal variances.

Procedure NSGS

(1) *Setup.* Select the overall desired PCS $1-\alpha$, IZ parameter δ, and common first-stage size $n_0 \geq 2$. Set

$$t = t_{n_0-1, 1-(1-\alpha/2)^{\frac{1}{k-1}}}$$

and obtain Rinott's constant $h = h(n_0, k, 1-\alpha/2)$ from the tables in Wilcox (1984) or Bechhofer et al. (1995). See also Table 8.3 in Goldsman and Nelson (1998).

(2) *Initialization.* Obtain n_0 outputs X_{ij} $(j = 1, 2, \ldots, n_0)$ from each system i $(i = 1, 2, \ldots, k)$ and let $\bar{X}_i(n_0) = n_0^{-1} \sum_{j=1}^{n_0} X_{ij}$ denote the sample mean of the first n_0 outputs from system i. Calculate the marginal sample variances

$$S_i^2 = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} \left(X_{ij} - \bar{X}_i(n_0) \right)^2,$$

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

(3) *Subset Selection.* Calculate the quantity

$$W_{i\ell} = t \left(\frac{S_i^2 + S_\ell^2}{n_0} \right)^{1/2}$$

for all $i \neq \ell$. Form the screening subset I, containing every alternative i such that $1 \leq i \leq k$ and

$$\bar{X}_i(n_0) \geq \bar{X}_\ell(n_0) - (W_{i\ell} - \delta)^+ \quad \text{for all } \ell \neq i.$$

(4) *Ranking.* If $|I| = 1$, then stop and return the system in I as the best. Otherwise, for all $i \in I$, calculate the second-stage sample sizes

$$N_i = \max\left\{n_0, \lceil (hS_i/\delta)^2 \rceil\right\},$$

where $\lceil \cdot \rceil$ is the ceiling function.

(5) Take $N_i - n_0$ additional outputs from all systems $i \in I$.

(6) Calculate the overall sample means $\bar{X}_i(N_i)$ for all $i \in I$. Select the system with the largest $\bar{X}_i(N_i)$ as best.

Nelson et al. (2001) showed that any subset-selection procedure and any two-stage IZ ranking procedure that satisfy certain mild conditions can be combined in this way while guaranteeing the overall probability of correct selection. The NGSG procedure can handle a relatively large number of systems because the first-stage screening is pretty tight. Nelson et al. (2001) provide a revised version of the NGSG procedure, the Group-Screening procedure, in which one can avoid simulating all the systems simultaneously. Boesel et al. (2003) extended the Group-Screening procedure for "clean up" after optimization via simulation.

The $\mathcal{KN}$ procedure is *fully sequential* because it takes only a single basic output from each alternative still in contention at each stage. Also, if there exists clear evidence that a system is inferior, then it will be eliminated from consideration immediately—unlike the NSGS procedure, where elimination occurs only after the first stage. $\mathcal{KN}$ also requires i.i.d. normal data, but does allow CRN. $\mathcal{KN}$ exploits the ideas of using paired differences, and controlling the Pr{ICS} on pairs to control it overall. Fabian's result is used to bound the error of a Brownian motion process that approximates each pair.

Procedure $\mathcal{KN}$

(1) *Setup.* Select the overall desired PCS $1-\alpha$, IZ parameter δ and common first-stage sample size $n_0 \geq 2$. Set

$$\eta = \frac{1}{2}\left[\left(\frac{2\alpha}{k-1}\right)^{-2/(n_0-1)} - 1\right].$$

(2) *Initialization.* Let $I = \{1, 2, \ldots, k\}$ be the set of systems still in contention, and let $h^2 = 2\eta(n_0 - 1)$.

Obtain n_0 outputs X_{ij} ($j = 1, 2, \ldots, n_0$) from each system i ($i = 1, 2, \ldots, k$) and let $\bar{X}_i(n_0) = n_0^{-1} \sum_{j=1}^{n_0} X_{ij}$ denote the sample mean of

the first n_0 outputs from system i.

For all $i \neq \ell$ calculate

$$S_{i\ell}^2 = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} \left(X_{ij} - X_{\ell j} - \left[\bar{X}_i(n_0) - \bar{X}_\ell(n_0) \right] \right)^2,$$

the sample variance of the difference between systems i and ℓ. Set $r = n_0$.

(3) *Screening.* Set $I^{\text{old}} = I$. Let

$$I = \Big\{ i : i \in I^{\text{old}} \text{ and } \bar{X}_i(r) \geq \bar{X}_\ell(r) - W_{i\ell}(r), \forall \ell \in I^{\text{old}}, \ell \neq i \Big\},$$

where

$$W_{i\ell}(r) = \max \left\{ 0, \frac{\delta}{2r} \left(\frac{h^2 S_{i\ell}^2}{\delta^2} - r \right) \right\}.$$

(4) *Stopping Rule.* If $|I| = 1$, then stop and select the system whose index is in I as the best.

Otherwise, take one additional output $X_{i,r+1}$ from each system $i \in I$, set $r = r + 1$ and go to *Screening*.

The $\mathcal{KN}$ procedure requires simulation of all systems simultaneously and a lot of switching among them. As discussed in Section 3, the switching cost can overwhelm the sampling cost, but this has become less of an issue in modern computing environments.

Both NSGS and $\mathcal{KN}$ can be applied to steady-state simulations if one is willing to use within-replication averages or batch means as the basic observations. However, as discussed in Section 3, employing within-replication averages or batch means as basic observations may be inefficient, so it is desirable to use individual outputs from within a single replication of each system if possible. Damerdji and Nakayama (1999) developed two-stage multiple-comparison procedures to select the best system for steady-state simulation that use batch means in the first stage of sampling, but can use individual outputs thereafter. Similarly, Goldsman et al. (2001) and Kim and Nelson (2005) proposed three R&S procedures that make a single replication from each system and use individual output as basic observations. One is a two-stage procedure based on Rinott's procedure, and the others are extensions of $\mathcal{KN}$ to steady-state simulation. One extension of $\mathcal{KN}$, called $\mathcal{KN}$++, updates the variance estimators as more outputs are available and has been shown to be highly efficient. We present the procedure below.

In $\mathcal{KN}$++, we assume that the output from each system i, X_{ij}, $j = 1, 2, \ldots$, is a stationary stochastic process that satisfies a Functional Central Limit Theorem condition (see Kim and Nelson 2005 for detailed conditions), and further

that the systems are simulated independently. Variance estimation centers on the *asymptotic variance constant* $v_i^2 = \lim_{r\to\infty} r\mathrm{Var}[\bar{X}_i(r)]$. See Goldsman et al. (2001) and Chapter 15 for reviews of different methods for the estimation of v_i^2. $\mathcal{KN}$++ extends $\mathcal{KN}$ to steady-state simulation by replacing its first-stage variance estimator with an estimator of the appropriate asymptotic variance constant. Moreover, $\mathcal{KN}$++ updates the variance estimator as more data are obtained based on a batching sequence m_r which is an integer-valued and nondecreasing function of r. The batching sequence needs to be carefully chosen to guarantee the strong consistency of the variance estimator in use; Goldsman et al. give three examples of such batching sequences. In general, m_r satisfies the property that $m_r \to \infty$ as $r \to \infty$.

Procedure $\mathcal{KN}$++

(1) *Setup.* Select the overall desired PCS $1-\alpha$, indifference-zone parameter δ, common first-stage sample size $n_0 \geq 2$ and initial batch size $m_{n_0} < n_0$. Set $r = n_0$. Calculate

$$\eta = \frac{1}{2}\left\{\left[2\left(1-(1-\alpha)^{1/k-1}\right)\right]^{-2/f} - 1\right\}$$

(2) *Initialization.* Let $I = \{1, 2, \ldots, k\}$ be the set of systems still in contention, and let $h^2 = 2\eta f$, where f is a function of the number of batches, b_r that depends on the variance estimator in use.

Obtain n_0 outputs $X_{ij}, j = 1, 2, \ldots, n_0$, from each system $i = 1, 2, \ldots, k$.

(3) *Update.* If m_r has changed since the last update, then for all $i \neq \ell$, calculate $m_r V_{i\ell}^2(r)$, the sample asymptotic variance of the difference between systems i and ℓ based on b_r batches of size m_r. Update f, η, and h^2.

(4) *Screening.* Set $I^{\text{old}} = I$. Let

$$I = \Big\{i : i \in I^{\text{old}} \text{ and } \bar{X}_i(r) \geq \bar{X}_\ell(r) - W_{i\ell}(r), \forall \ell \in I^{\text{old}}, \ell \neq i\Big\}$$

where

$$W_{i\ell}(r) = \max\left\{0, \frac{\delta}{2cr}\left(\frac{h^2 m_r V_{i\ell}^2(r)}{\delta^2} - r\right)\right\}.$$

(5) *Stopping Rule.* If $|I| = 1$, then stop and select the system whose index is in I as the best.

Otherwise, take one additional output $X_{i,r+1}$ from each system $i \in I$ and set $r = r + 1$ and go to *Update*.

Even if the output data fed to $\mathcal{KN}$++ were i.i.d. normal, the procedure does not provide a guaranteed PCS in finite samples. However, using techniques

Table 1
The Five Alternative Inventory Policies

Policy i	s	S	Expected Cost
1	20	40	114.176
2	20	80	112.742
3	40	60	130.550
4	40	100	130.699
5	60	100	147.382

described in Section 6.1, $\mathcal{KN}$++ can be shown to guarantee PCS $\geq 1 - \alpha$ asymptotically.

5 Application

This section illustrates NSGS and $\mathcal{KN}$ using an (s, S) inventory system with the five inventory policies as described in Koenig and Law (1985). The goal of this study is to compare the five policies given in Table 1 and find the one with the smallest expected average cost per month for the first 30 months of operation. Table 1 also contains the expected cost (in thousands of dollars) of each policy, which can be analytically computed in this case. We set $\delta = \$1$ thousand, $n_0 = 10$ initial replications, and $1 - \alpha = 0.95$.

Table 2 shows the results of the simulation study for each procedure, including the total number of outputs taken and the sample average cost per month for each policy. In NSGS, policies 3, 4, and 5 were eliminated after the first stage of sampling, so only policies 1 and 2 received second-stage samples. In $\mathcal{KN}$, only policies 4 and 5 were eliminated after the first stage, but the elimination of policies 3 and 1 occurred after they received 16 and 98 observations, respectively. This illustrates the value of the tighter initial screen in NSGS, which takes only one look at the data, and the potential savings from taking many looks, as $\mathcal{KN}$ does. Both procedures chose policy 2 as the best (which is in fact correct). Since the true difference is larger than δ, NSGS and $\mathcal{KN}$ will choose the true best with 95% confidence. However, in general we do not have any information about the true differences; therefore, the most we can conclude without prior knowledge is that policy 2 is either the true best, or has expected cost per month within \$1 thousand of the true best policy, with 95% confidence.

Table 2
Simulation Results of the (s, S) Inventory Example

	NSGS		$\mathcal{KN}$	
Policy i	# Obs.	Average	# Obs.	Average
1	209	114.243	98	114.274
2	349	112.761	98	113.612
3	10	130.257	16	130.331
4	10	128.990	10	128.990
5	10	147.133	10	147.133
Total	588		232	

6 Asymptotic Analysis

In order of importance, the key performance measures for R&S procedures are the ability to deliver the nominal PCS and the ability to deliver it efficiently. Although many procedures provide a guaranteed PCS under ideal conditions (e.g., i.i.d. normal outputs), and the expected sample size of simple procedures can be explicitly calculated, when conditions are not ideal, or when the procedure is more complex (e.g., it includes early elimination), small-sample performance may be difficult to derive. Fortunately, asymptotic analysis—driving the sample sizes to infinity—can sometimes provide meaningful insights. The power of asymptotic analysis is that many of the problem-specific details that thwart small-sample analysis wash out in the limit. Appropriate asymptotic analysis can establish conditions under which procedures work (at least approximately), and the superiority of one procedure over another. In the R&S literature there are at least three asymptotic regimes:

PCS as $\delta \to 0$: To evaluate the ability of a procedure to provide a PCS guarantee under a range of conditions, the indifference-zone parameter δ may be driven to zero. Done naively, this drives the sample sizes from all systems to infinity and the PCS to 1. Therefore, to make the analysis useful, the selection problem must become more difficult as $\delta \to 0$. We describe this approach in Section 6.1.

Efficiency as $\delta \to 0$: The indifference-zone parameter δ may also be driven to zero to evaluate the efficiency of a procedure that estimates unknown variances, relative to a corresponding known-variance procedure. To date this type of analysis has only been applied to procedures whose sample sizes are independent of the true means (that is, the procedure does not take advantage of a favorable configuration of the means, e.g., Bechhofer 1954), so there is no need to change the selection problem as $\delta \to 0$. We briefly describe this approach in Section 6.2.

Efficiency as $(1-\alpha) \to 1$**:** To compare the efficiency of competing procedures, the nominal PCS may be driven to 1. This, too, will drive the sample sizes to infinity, but if the rate at which they grow can be determined then that rate can be compared to the rate achieved by other procedures. We describe this approach in Section 6.3.

See also Damerdji and Nakayama (1999) for a related asymptotic analysis of multiple-comparison procedures.

6.1 Asymptotic Probability of Correct Selection

There is a close connection between the PCS in R&S and the power in statistical hypothesis testing. Consider a hypothesis testing problem of the form

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta > \theta_0$$

Suppose that the power of the test cannot be calculated explicitly. As the sample size n goes to infinity, any reasonable test has asymptotic power 1 against any fixed alternative (say, $\theta = \theta_0 + \delta$). As noted by Lehmann (1999, Section 3.3), the trick is to embed the actual situation into a suitable sequence (n, θ_n) that makes the discrimination problem more difficult as the sample size increases in such a way that a meaningful limit < 1 is reached. A sequence that frequently works is

$$\theta_n = \theta_0 + \frac{\delta}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right)$$

with $\delta > 0$.

IZ R&S procedures are essentially power calculations, since their goal is to detect the best with given probability (power) when the best is at least a significant amount $\delta > 0$ better than the rest (think $H_1 : \mu_k > \mu_{k-1} + \delta$). However, instead of n driving the parameter θ_n, as in the hypothesis test, it makes more sense to have $\delta \to 0$ drive N_i, the number of observations to be taken from system i; frequently $\sqrt{N_i} \propto 1/\delta$.

Mukhopadhyay and Solanky (1994) say that an IZ procedure is *asymptotically consistent* if

$$\liminf_{\delta \to 0} \mathrm{PCS} \geq 1 - \alpha$$

for all $\mu_1, \mu_2, \ldots, \mu_k$ such that $\mu_k - \mu_{k-1} \geq \delta$. For a procedure that assumes normally distributed output data, Dalal and Hall (1977) declare an IZ procedure to be *asymptotically robust* if

$$\liminf_{\delta \to 0} \inf_{F \in \mathcal{F}} \text{PCS} \geq 1 - \alpha$$

where $\mathcal{F}$ is a suitable family of location parameter distributions $F(x - \mu_i)$ (containing the normal distribution) with $\mu_1, \mu_2, \ldots, \mu_k$ such that $\mu_k - \mu_{k-1} \geq \delta$.

An example of a procedure that does not provide a guaranteed PCS for finite samples, but can be shown to be asymptotically consistent, is due to Robbins, Sobel and Starr (1968). Their procedure generalizes Bechhofer's (1954) known, common-variance procedure in Section 2.2 to the unknown common variance case (still assuming normally distributed data). They suggest taking N observations from each system, where

$$N = N(\delta) = \inf\{n \geq n_0 : n \geq h^2 S^2(n)/\delta^2\}$$

with $n_0 \geq 2$ and $S^2(n)$ the usual pooled estimator of σ^2 based on n observations. Notice that the variance estimator is updated as more data are collected, which makes it impossible to establish the finite-sample PCS.

The proof of asymptotic consistency illustrates a key idea in this approach. After some manipulation one can show that

$$\text{PCS} \geq \text{E}\left[\int_{-\infty}^{\infty} \left\{\Phi\left(y + \sqrt{N}\delta/\sigma\right)\right\}^{k-1} \phi(y)dy\right] \tag{19}$$

where Φ and ϕ are the cdf and pdf, respectively, of the standard normal distribution. Now since $\sqrt{N}\delta/\sigma \to h$ with probability 1 as $\delta \to 0$, the right-hand side of (19) converges to

$$\int_{-\infty}^{\infty} \{\Phi(y + h)\}^{k-1} \phi(y)dy = 1 - \alpha.$$

Notice that in the limit the unknown-variance procedure behaves like Bechhofer's known-variance procedure. The asymptotic validity of $\mathcal{KN}$++ (see Section 4) is based on an analogous argument showing that as $\delta \to 0$ the (appropriately standardized) output processes behave like (known variance and drift) Brownian motion processes (Kim and Nelson 2005). See also Damerdji, Glynn, Nakayama and Wilson (1996).

6.2 Asymptotic Efficiency

Let n be the sample size (per system) of a Bechhofer-like, known-variance R&S procedure, and let N be the sample size of a corresponding unknown-variance procedure where an initial n_0 observations from each system are used to estimate the unknown variance. Typically N takes the form

$$N = \max\left\{n_0, \left\lceil\left(\frac{hS}{\delta}\right)^2\right\rceil\right\}$$

where S^2 is a pooled estimator of the unknown variance, and h is an appropriately adjusted constant.

Mukhopadhyay and Solanky (1994) say that a procedure is *asymptotically first-order efficient* if

$$\lim_{\delta\to 0} \mathrm{E}\left(\frac{N}{n}\right) = 1$$

and *asymptotically second-order efficient* if $\lim_{\delta\to 0} \mathrm{E}\,(N-n)$ is bounded. They show that the typical procedure for which n_0 is a fixed value, the variance is estimated only once, and N grows as $1/\delta^2$ is neither asymptotically first- nor second-order efficient. However, if N grows somewhat more slowly than $1/\delta^2$ then an asymptotically first-order efficient procedure can be obtained, while asymptotic second-order efficiency typically requires that the variance estimator be updated as more data are obtained.

6.3 Asymptotic Sample Size

Suppose that we want to know the expected sample size of Paulson's procedure in Section 2.2. The fact that systems can be eliminated before the terminal stage implies that the expected sample size depends on the differences between the true means, and that we must account for the complication that any system has a chance to eliminate any other. However, consider what happens as we drive $(1-\alpha) \to 1$ (the following heuristic argument is made precise by Perng 1969):

- As $(1-\alpha) \to 1$, the procedure stops making mistakes; the best system survives and all of the inferior systems are eliminated by the best one.
- As the sample sizes are driven to infinity, $\bar{X}_i(r)$ behaves more and more like μ_i. Thus, the stage at which system $i \neq k$ is eliminated is the first r for

which

$$\mu_i \leq \mu_k - (a/r - \lambda).$$

This occurs (approximately) when $r_i = a/(\mu_k - \mu_i + \lambda)$.

- Recall that

$$a = \ln\left(\frac{k-1}{\alpha}\right)\frac{\sigma^2}{\delta - \lambda}.$$

Therefore, as $(1-\alpha) \to 1$, the expected sample size from system $i \neq k$ is approximately

$$r_i \approx \ln\left(\frac{k-1}{\alpha}\right)\frac{\sigma^2}{(\delta-\lambda)(\mu_k - \mu_i + \lambda)}$$

while for $i = k$ it is

$$r_k \approx \ln\left(\frac{k-1}{\alpha}\right)\frac{\sigma^2}{(\delta-\lambda)(\mu_k - \mu_{k-1} + \lambda)}.$$

Thus, the expected total sample size as, $(1-\alpha) \to 1$, is $\approx \sum_{i=1}^{k} r_i$.

Notice that the impact of the true differences $\mu_k - \mu_i$ and the choice of λ become apparent from this analysis. The growth rate of $\ln((k-1)/\alpha)$ is common to many procedures (see Dudewicz 1969), so the differences in their asymptotic efficiency is the term that multiplies $\ln((k-1)/\alpha)$. For an example of this type of analysis for a more complex procedure see Jennison, Johnstone and Turnbull (1982).

7 Other Formulations

Throughout this chapter we have focused on the problem of finding the best when the best is defined as the system with the largest or smallest mean performance measure. As discussed in Section 1, there exist other types of comparison problems. Here we briefly visit each type of comparison problem and provide useful references.

7.1 Comparisons with a Standard

The goal of comparison with a standard is to find systems whose expected performance measures are larger (or smaller) than a standard and, if there are any, to find the one with the largest (or smallest) expected performance.

For this type of problem, each alternative needs to be compared to the standard as well as the other alternative systems, and the standard may be a known value or the expected value of a designated system (simulated or real). Such procedures first appeared in Paulson (1952) and Bechhofer and Turnbull (1978).

Clearly, the standard could be treated as just another system and the problem formulated as selection of the best. Specially tailored procedures are required when the standard is to be given special status, specifically a guarantee that no alternative will be selected unless it beats the standard substantially.

Let μ_0 denote the expected performance of the standard (which may be known or unknown), and let $\mu_1, \mu_2, \ldots, \mu_k$ be the unknown means of the alternatives, as in selection of the best. In comparisons with a standard we require

$$\Pr\{\text{select } 0 | \mu_0 \geq \mu_k\} \geq 1 - \alpha \tag{20}$$

$$\Pr\{\text{select } k | \mu_k - \mu_0 \geq \delta, \mu_k - \mu_{k-1} \geq \delta\} \geq 1 - \alpha. \tag{21}$$

Thus, we try to protect the standard, but if the best system is substantially better then we want to select it.

Nelson and Goldsman (2001) proposed two-stage procedures for this problem that are specifically designed for computer simulation. Similar to Paulson (1952) and Bechhofer and Turnbull (1978), at the end of their procedures the standard is retained if $\bar{X}_0 + c > \bar{X}_i$ for $i = 1, 2, \ldots, k$, otherwise the system with the largest sample mean is selected. The following result provides guidance for designing the algorithm and specifying the value of $c > 0$:

Theorem 6 (Nelson and Goldsman 2001) *If the distribution of $\bar{X}_{ij} - \mu_i$ is independent of μ_i, for $i = 0, 1, 2, \ldots, k$, and if*

$$\begin{aligned}
&\Pr\{(\bar{X}_i - \bar{X}_0) - (\mu_i - \mu_0) \leq c, i = 1, 2, \ldots, k\} \geq 1 - \alpha \\
&\Pr\{(\bar{X}_k - \bar{X}_0) - (\mu_k - \mu_0) > c - \delta, \\
&\qquad (\bar{X}_k - \bar{X}_i) - (\mu_k - \mu_i) > -\delta, i = 1, 2, \ldots, k-1\} \geq 1 - \alpha
\end{aligned}$$

then (20) and (21) hold.

The two conditions are intuitive: The first insures that, when the standard is best, no inferior system's sample mean beats it by too much. The second condition guarantees that when system k is best by δ or more, then its sample mean is enough bigger than the standard's sample mean, and is bigger than the sample mean of every other system, so that it is selected.

Kim (2002) proposed fully sequential procedures for comparison with a standard. A procedure such as Paulson (1964) or $\mathcal{KN}$ is not directly applicable

because it would require $\mu_0 \geq \mu_k + \delta$, not just $\mu_0 \geq \mu_k$, for the standard to be retained with the desired probability. But since $\mathcal{KN}$ and other procedures that are similar to Paulson (1964) focus on all pairwise comparisons, and the identity of the standard is known, the following reformulation in Kim (2005) works: In any comparison with the standard, revise (5) from

$$\bar{X}_0(r) \geq \max_{i \in I} \bar{X}_i(r) - \max\{0, a/r - \lambda\}$$

to

$$\bar{X}_0(r) + \delta/2 \geq \max_{i \in I} \bar{X}_i(r) - \max\{0, a/r - \lambda\} \tag{22}$$

and further, select a and λ to detect differences of size $\delta/2$ instead of δ.

Why does this work? Suppose that $\mu_0 = \mu_k$ so that the standard should be retained. Then $\bar{X}_0(r) + \delta/2$ has expected value at least $\delta/2$ better than all the other systems and will be retained with the desired probability. On the other hand, if $\mu_k = \mu_0 + \delta$, so that system k should be selected, then $\bar{X}_0(r) + \delta/2$ has expected value that is $\delta/2$ inferior to the best and will be eliminated with the appropriate probability. The procedure is set up for, and detects, differences of size δ for comparisons among the alternatives, but whenever the standard is involved in a comparison, the procedure is adjusted to detect $\delta/2$.

7.2 Selecting the System Most Likely to be the Best

In multinomial selection problems, the definition of "best" is the system that is *mostly likely* to be the best in a single trial. Historically, these procedures were designed for experiments that have a categorical response (e.g., which among 5 soft drinks a subject will say that they prefer). If there are k categories, p_i is the probability that the ith category is selected in a single trial, and the trials are independent, then the number of times each category is selected has a multinomial distribution. More precisely, let N_i be the number of times that category i is chosen in n independent trials. Then

$$\Pr\{N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\} = \frac{n!}{\prod_{i=1}^k n_i!} \prod_{i=1}^k p_i^{n_i}$$

where $\sum_{i=1}^k n_i = n$.

For convenience of notation (but unknown to us), assume that $p_k \geq p_{k-1} \geq \cdots \geq p_1$. Therefore, a correct selection in this context is selecting category k. Multinomial selection procedures seek to provide a guaranteed PCS.

The compromise that makes this possible is to guarantee the PCS whenever $p_k/p_{k-1} \geq \theta$, where $\theta > 1$ is interpreted as the smallest p_k/p_{k-1} ratio worth detecting (and therefore defines another form of indifference zone). Bechhofer, Elmaghraby, and Morse (BEM) (1959) proposed a single-stage procedure that satisfies this requirement. Other work on this problem includes Bechhofer and Goldsman (1986), who proposed a procedure that uses closed, sequential sampling. See Bechhofer et al. (1995) for a review.

Goldsman (1984ab) first suggested the more general use of this type of procedure to find the simulated system most likely to produce the "most desirable" observation on a given trial, where "most desirable" can be almost any criterion of goodness. This often means identifying the system i with the largest value of p_i, where $p_i = \Pr\{X_{ij} > X_{\ell j}, \forall \ell \neq i\}$ for a problem in which a larger simulation output response is better. For instance, in a reliability setting the simulation output X might be the time to system failure and the goal is to select the system that is most likely to survive the longest. The key difference from the categorical data context is that *each trial involves obtaining a response value from each simulated system and the winner is determined by comparing these values*. Stated differently, a trial or replication produces a vector response $(X_{1j}, X_{2j}, \ldots, X_{kj})$ that is transformed into a categorical response $(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ with the 1 indicating the system with the largest output on the jth replication.

With this in mind, Miller et al. (1998) devised a single-stage procedure that achieves a higher probability of correct selection than does BEM in the case where both the replications (vector observations) and the systems themselves are simulated independently (no common random numbers). The key insight is that the formation of vector observations by replication number—$(X_{1j}, X_{2j}, \ldots, X_{kj})$—is arbitrary; any vector formed with one output from each system has the same distribution. Thus, n replications from each system can form n^k vector observations. Of course, these vectors are no longer independent, since they share observations, so this is not the same as having n^k independent replications. However, Miller et al. (1998) showed that forming all vector comparisons increases the effective sample size by at least one third, and their procedure exploits this additional information to achieve the desired PCS with fewer total replications.

The role of CRN in multinomial selection is interesting and worthy of discussion. In the means-based procedures that are the focus of this chapter, CRN was introduced as an experiment design technique to increase efficiency but it has no effect on the problem parameters, specifically $\mu_1, \mu_2, \ldots, \mu_k$. However, in multinomial selection the value of $p_i = \Pr\{X_{ij} > X_{\ell j}, \forall \ell \neq i\}$ will, in general, be different if the systems are simulated with CRN as opposed to independently, as noted by Mata (1993). Miller and Bauer (1997) observed that the identity of the best is typically the same with or without CRN, although

this is not guaranteed, but the relative dominance of the best can increase or decrease even if the identity is unchanged. Thus, in multinomial selection the decision as to whether or not to use CRN should be based on whether the actual performance of the real systems would be affected by like or common factors, or whether their actual performance would be independent.

7.3 Selecting the Largest Probability of Success

In Bernoulli selection problems, the basic output from each system on each independent replication, denoted X_{ij}, takes either the value 1 ("success") or 0 ("failure"), and the best system is the one with the largest probability of success, $p_i = \Pr\{X_{ij} = 1\}$. Simulation applications include comparing systems in terms of their ability to survive a mission or to meet a goal such as on-time performance. To our knowledge there has been little research on, or application of, Bernoulli selection in simulation despite the obvious relevance.

Assume that (unknown to us) $p_k \geq p_{k-1} \geq \cdots \geq p_1$ so that a correct selection is choosing system k. At least three types of indifference-zone parameters have been considered in Bernoulli selection:

Difference: $p_k - p_{k-1} \geq \delta$

Odds Ratio: $\dfrac{p_k/(1-p_k)}{p_{k-1}/(1-p_{k-1})} \geq \theta$

Relative Risk: $p_k/p_{k-1} \geq \theta$

where $\delta > 0$ and $\theta > 1$ are user-specified parameters. A PCS $\geq 1 - \alpha$ is desired in any case. Clearly the Difference formulation is analogous to the IZ formulation for normal-theory procedures described throughout this chapter. A concern about the Difference formulation is that it seems unnatural for a significant difference not to be tied to the sizes of the success probabilities themselves. The other two formulations attempt to incorporate this feature. See Chapter 7 of Bechhofer, et al. (1995) for a discussion of this issue and a list of procedures.

To obtain a sense of the analysis involved in developing Bernoulli selection procedures, suppose that the IZ is of the odds-ratio form, there are only $k = 2$ systems and the two systems are simulated independently (no CRN). We want to develop a procedure that terminates when $\sum_j (X_{2j} - X_{1j}) = \pm a$, where a is a nonnegative integer. Thus, the procedure terminates whenever system 2 has a more successes than system 1, or vice versa. In this case a correct selection will occur if $\sum_j (X_{2j} - X_{1j}) = a$.

Let $S_n = \sum_{j=1}^{n}(X_{2j} - X_{1j}), n = 0, 1, \ldots$, a random walk on $\{-a, -a+1, \ldots, a-1, a\}$ with initial state $S_0 = 0$. Although we could work with this process, it will be more useful to consider a related process

$$Y_m = \text{value of } S_n \text{ after its } m\text{th change in state.}$$

Stated differently, $\{Y_m; m = 0, 1, 2, \ldots\}$ is the process that results from ignoring the transitions of $\{S_n\}$ that do not change its state. Assume that

$$\frac{p_k/(1-p_k)}{p_{k-1}/(1-p_{k-1})} = \theta$$

so that the IZ condition is an equality. It is easy to show that $\{Y_m; m = 0, 1, 2, \ldots\}$ is a time-homogeneous discrete-time Markov chain with one-step transition probabilities

$$q_{ij} = \Pr\{Y_{m+1} = j | Y_m = i\} = \begin{cases} 1, & i = j = a, -a \\ \frac{\theta}{1+\theta}, & j = i+1, i < a \\ \frac{1}{1+\theta}, & j = i-1, i > -a \\ 0, & \text{otherwise} \end{cases}$$

Notice that the IZ assumption leads to transition probabilities that are independent of the actual values of p_1 and p_2.

Using standard Gambler's ruin results (e.g., Ross 2000), the probability that the process is absorbed in state a—the state that would cause us to declare system 2 as best—is $\theta^a/(1+\theta^a)$. Therefore, to obtain PCS $\geq 1 - \alpha$ we set

$$a = \left\lfloor \frac{\ln\left(\frac{1-\alpha}{\alpha}\right)}{\ln(\theta)} \right\rfloor.$$

Random-walk analysis is at the heart of many sequential procedures for Bernoulli selection, and Smith (1995) shows that it is often useful for evaluating the efficiency of such procedures.

The role of CRN in Bernoulli selection is largely unexplored. Continuing the previous example, suppose now that the data are generated as follows:

$$X_{ij} = \begin{cases} 0, \; U_j \leq 1 - p_i \\ 1, \text{ otherwise} \end{cases}$$

for $i = 1, 2$, where $U_1, U_2, \ldots$ are i.i.d. $U(0,1)$ random variables. This set up induces the largest possible correlation between two Bernoulli random variables and has a profound effect on our procedure because now the outcome $(X_{2j} = 0, X_{1j} = 1)$ *cannot occur*. Therefore, the one-step transition probabilities of $\{Y_m\}$ become

$$q_{ij} = \Pr\{Y_{m+1} = j | Y_m = i\} = \begin{cases} 1, \; i = j = a, -a \\ 1, \; j = i+1, i < a \\ \\ 0, \text{ otherwise} \end{cases}$$

and the PCS of the selection procedure is 1. This might seem like a desirable outcome until you consider the efficiency of the procedure. Remember that the cost of running the procedure is not the number of Y_m's that are required, but rather the number of X_{ij}'s. In the independent case, the expected number of X_{ij}'s required for each transition of Y_m is

$$\frac{2}{p_2 - p_1 + 2p_1(1 - p_2)}$$

but under CRN it is greater, specifically

$$\frac{2}{p_2 - p_1}.$$

In many cases this is enough to make the procedure *less efficient* when CRN is employed (assuming the value of a is not altered). For instance, if $p_2 = 4/5, p_1 = 3/4, 1 - \alpha = 0.95$ and $\theta = 4/3$, then we can show that the expected number of outputs that must be generated under independent sampling is about 357, while under CRN it is 400. Obviously a should be altered when CRN is employed—in fact $a = 1$ is adequate in this illustration—but to date no procedure has been developed. Tamhane (1980, 1985) does provide a procedure for $k = 2$ systems, but it requires being able to provide an upper bound on $\Pr\{X_{1j} \neq X_{2j}\}$.

7.4 Bayesian Procedures

Instead of providing a PCS guarantee, Bayesian procedures attempt to allocate a finite computation budget to maximize the posterior PCS of the selected system. Chen, et al. (2000) and Chick and Inoue (2001) are two recent references; see Chapter 9 for a thorough review of this approach

8 Future Directions

The following are some directions in which future breakthroughs are most needed:

- Procedures specifically designed for very large numbers of alternatives, particularly when the alternatives are not all available at the same time (such as occurs during the search phase of an optimization-via-simulation algorithm).
- Procedures that exploit common random numbers for very large numbers of alternatives without employing such conservative inequalities that the impact of CRN is overwhelmed.
- Development of constrained selection-of-the-best procedures; for instance, procedures that select the best based on one performance measure, subject to a constraint or condition on a different measure.

Acknowledgments

Portions of this chapter were published previously in the *Proceedings of the 2003 Winter Simulation Conference*. This work was partially supported by NSF Grant No. DMI-0217690. The authors gratefully acknowledge the many helpful comments of Ajit Tamhane.

References

Banerjee, S. 1961. On confidence interval for two-means problem based on separate estimates of variances and tabulated values of t-table. *Sankhyā* A23:359–378.

Banks, J., J. S. Carson, B. L. Nelson, and D. Nicol. 2005. *Discrete-Event System Simulation*. Fourth Edition. Upper Saddle River, NJ: Prentice Hall.

Bechhofer, R. E. 1954. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics* 25:16–39.

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. 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 B* 15:829–851.

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

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.

Boesel, J., B. L. Nelson, and S.-H. Kim. 2003. Using ranking and selection to "clean up" after simulation optimization. *Operations Research* 51:814–825.

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

Dalal, S. R. and W. J. Hall. 1977. Most economical robust selection procedures for location parameters. *Annals of Statistics* 7:1321–1328.

Damerdji, H., P. W. Glynn, M. K. Nakayama and J. R. Wilson. 1996. Selecting the best system in transient simulations with variances known. In *Proceedings of the 1996 Winter Simulation Conference* (ed. J. M. Charnes, D. J. Morrice, D. T. Brunner and J. J. Swain), 281–286, The Institute of Electrical and Electronics Engineers, Piscataway, N.J.

Damerdji, H., and M. K. Nakayama. 1999. Two-stage multiple-comparison procedures for steady-state simulation. *ACM TOMACS* 9:1–30.

Dudewicz, E. J. 1969. An approximation to the sample size in selection problems. *Annals of Mathematical Statistics* 40:492–497.

Dudewicz, E. J. 1995. The heteroscedastic method: Fifty+ years of progress 1945–2000, and professor Minoru Siotani's award-winning contributions. *American Journal of Mathematical and Management Sciences* 15:179–197.

Fabian, V. 1974. Note on Anderson's sequential procedures with triangular boundary. *Annals of Statistics* 2:170–176.

Glynn, P. W., and D. L. Iglehart. 1990. Simulation output analysis using standardized time series. *Mathematics of Operations Research* 15:1–16.

Goldsman, D. 1984a. A multinomial ranking and selection procedure: Simulation and applications. In *Proceedings of the 1984 Winter Simulation Conference* (ed. S. Shepard, U. W. Pooch, and C. D. Pegden), 259–264, The Institute of Electrical and Electronics Engineers, Piscataway, N.J.

Goldsman, D. 1984b On selecting the best of K systems: An expository survey of indifference-zone multinomial procedures. In *Proceedings of the 1984 Winter Simulation Conference* (ed. S. Shepard, U. W. Pooch, and C. D. Pegden), 107–112, The Institute of Electrical and Electronics Engineers, Piscataway, N.J.

Goldsman, D., S.-H. Kim, W. Marshall, and B. L. Nelson. 2001. Ranking and selection procedures for steady-state simulation: Perspectives and procedures. *INFORMS Journals on Computing* 14:2–19.

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

Gupta, S. S. 1956. On a decision rule for a problem in ranking means. Doctoral dissertation, Institute of Statistics, Univ. of North Carolina, Chapel Hill, NC.

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

Hartmann, M. 1988. An improvement on Paulson's sequential ranking procedure. *Sequential Analysis* 7:363–372.

Hartmann, M. 1991. An improvement on Paulson's procedure for selecting the population with the largest mean from k normal populations with a common unknown variance. *Sequential Analysis* 10:1–16.

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

Hong, L. J., and B. L. Nelson. 2005. The tradeoff between sampling and switching: New sequential procedures for indifference-zone selection. *IIE Transactions* 37:623–634.

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

Jennison, C., I. M. Johnstone, and B. W. Turnbull. 1980. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. Technical Report, Dept. of ORIE, Cornell Univ., Ithaca, NY.

Jennison, C., I. M. Johnstone, and B. W. Turnbull. 1982. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. In *Statistical Decision Theory and Related Topics III, Vol. 2*, ed. S. S. Gupta and J. O. Berger, 55–86. New York: Academic Press.

Jennison, C., and B. W. Turnbull. 2000. *Group Sequential Methods with Applications to Clinical Trials*. New York: Chapman & Hall.

Koenig, L. W., and A. M. Law. 1985. A procedure for selecting a subset of size m containing the ℓ best of k independent normal populations, with applications to simulation. *Communications in Statistics—Simulation and Computation* B14:719–734.

Kim, S.-H. 2005. Comparison with a standard via fully sequential procedures. *ACM TOMACS* 15:155–174.

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

Kim, S.-H., and B. L. Nelson. 2005. On the asymptotic validity of fully sequential selection procedures for steady-state simulation. *Operations Research*, forthcoming.

Kimball, A. W. 1951. On dependent tests of significance in the analysis of variance. *Annals of Mathematical Statistics* 22:600–602.

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

Lehmann, E. L. 1999. *Elements of Large-Sample Theory*. New York: Springer.

Mata, F. 1993. Common random numbers and Multinomial selection. *Computers and Industrial Engineering* 25:33–36.

Miller, J. O. and K. W. Bauer. 1997. How common random numbers affect

multinomial selection. In *Proceedings of the 1997 Winter Simulation Conference* (ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson), 342-347, The Institute of Electrical and Electronics Engineers, Piscataway, N.J.

Miller, J. O., B. L. Nelson, and C. H. Reilly. 1998. Efficient multinomial selection in simulation. *Naval Research Logistics* 45:459–482.

Mukhopadhyay, N. and T. K. S. Solanky. 1994. *Multistage Selection and Ranking Procedures: Second-Order Asymptotics*. New York: Marcel Dekker.

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

Nelson, B. L., and F. J. Matejcik. 1995. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Management Science* 41:1935–1945

Nelson, B. L., J. Swann, D. Goldsman, and W.-M. T. Song. 2001. Simple procedures for selecting the best system when the number of alternatives is large. *Operations Research* 49:950–963.

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

Paulson, E. 1964. A sequential procedure for selecting the population with the largest mean from k normal populations. *Annals of Mathematical Statistics* 35:174–180.

Perng, S. K. 1969. A comparison of the asymptotic expected sample sizes of two sequential procedures for ranking problem. *Annals of Mathematical Statistics* 40:2198–2202.

Pichitlamken, J. and B. L. Nelson. 2001. "Selection-of-the-best Procedures for Optimization via Simulation," in *Proceedings of the 2001 Winter Simulation Conference*, 401–407.

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

Robbins, H., M. Sobel and N. Starr. 1968. A sequential procedure for selecting the largest of k means. *Annals of Mathematical Statistics* 39:88–92.

Ross, S. M. 2000. *Introduction to Probability Models*. 7th ed. New York: Academic Press.

Siegmund, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. New York: Springer-Verlag.

Slepian, D. 1962. The one-sided barrier problem for Gaussian noise. *Bell Systems Technical Journal* 41:463–501.

Smith, M. J. 1995. Ranking and selection: Open sequential procedures for Bernoulli populations. Unpublished M.S. thesis, School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA.

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.

Sullivan, D. W., and J. R. Wilson. 1989. Restricted subset selection for simulation. *Operations Research* 37:52–71.

Tamhane, A. C. 1980. Selecting the better Bernoulli treatment using a matched samples design. *Journal of the Royal Statistical Society* B42:26–30.

Tamhane, A. C. 1985. Some sequential procedures for selecting the better Bernoulli treatment using a matched samples design. *Journal of the American Statistical Association* 80:455–460.

Wald, A. 1947. *Sequential Analysis*. New York: John Wiley.

Welch, B. L. 1938. The significance of the difference between two means when the population variances are unequal. *Biometrika* 25:350–362.

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

Wilson, J. R. 2001. A multiplicative decomposition property of the screening-and-selection procedures of Nelson et al. *Operations Research* 49:964–966.