99 percent intervals, using both the paired-t and Welch approaches, with asterisks indicating those intervals missing zero, i.e., those pairs of systems that appear to have different expected operating costs. Once again, note that the two approaches do not always agree in terms of which differences are significant, and neither approach gives intervals with uniformly smaller half-lengths. Furthermore, it is possible to arrive at apparent contradictions in the conclusions. For instance, using the Welch approach we would conclude that neither μ_1 nor μ_2 differs significantly from μ_3, so that we might (crudely) want to say something like "$\mu_1 = \mu_3 = \mu_2$" and thus think logically that "$\mu_1 = \mu_2$." But the confidence interval for $\mu_2 - \mu_1$ misses zero, indicating that we *cannot* regard μ_1 as being equal to μ_2. The problem here is that we dare not interpret the confidence-interval statements as constituting "proof" of equality or inequality; in the above discussion we just could not resolve a difference between either μ_1 or μ_2 in comparison with μ_3, but we could detect a difference between μ_1 and μ_2. Such apparent contradictions become less likely as the intervals become smaller, which could occur by making more replications of the systems or perhaps by using CRN, discussed in Sec. 11.2.

As at the end of Sec. 10.3.1, we note the importance of ensuring the validity of the individual confidence intervals, the possibility of using CRN across the different models, and of taking the above approach for steady-state comparisons by using an appropriate steady-state methodology for the individual intervals.

10.3.3 Multiple Comparisons with the Best

Finally, we mention another kind of comparison goal that forms simultaneous confidence intervals for the differences between the means of each of the k alternatives and that of the best of the other alternatives, even though we do not know which of the others really is the best. This is known as *multiple comparisons with the best* (MCB), and it has as its objective to form k simultaneous confidence intervals on $\mu_i - \min_{l \neq i} \mu_l$ for $i = 1, 2, \ldots, k$, assuming that smaller means are better (if larger is better, then "min" is replaced by "max").

Hsu (1984) gives a technique for addressing the MCB goal, and Hochberg and Tamane (1987) and Hsu (1996) are comprehensive books on multiple-comparison procedures. Nelson (1993) gives a treatment of MCB that allows the use of CRN (see Sec. 11.2) for improved efficiency. Damerdji and Nakayama (1999), Nakayama (1997, 2000), and Yuan and Nelson (1993) address the steady-state MCB problem.

While MCB is useful in its own right, it is also intimately related to the ranking-and-selection procedures discussed in Sec. 10.4.1 [see Nelson and Matejcik (1995)].

10.4 RANKING AND SELECTION

In this section we consider goals that are different—and more ambitious—than simply making a comparison between several alternative systems. In Sec. 10.4.1 we describe procedures whose goal is to select one of the k systems as being the

best one, in some sense, and to control the probability that the selected system really *is* the best one. Section 10.4.2 considers a different goal, picking a subset of m of the k systems so that this selected subset contains the best system, again with a specified probability. (The validity of two of these selection procedures is considered in App. 10A.) Further problems and methods are discussed in Sec. 10.4.3, including the issue of ranking and selection based on steady-state measures of performance.

10.4.1 Selecting the Best of k Systems

As in Secs. 10.2 and 10.3, let X_{ij} be the random variable of interest from the jth replication of the ith system, and let $\mu_i = E(X_{ij})$. For this selection problem, as well as that in Sec. 10.4.2, the X_{ij}'s are assumed to be independent for different replications of the ith system. Except for the Nelson and Matejcik (1995) procedure discussed later in this section, the replications for different systems are also to be made independently. For example X_{ij} could be the average total cost per month for the jth replication of policy i for the inventory model of Examples 10.7 and 10.8.

Let μ_{i_l} be the lth smallest of the μ_i's, so that $\mu_{i_1} \leq \mu_{i_2} \leq \cdots \leq \mu_{i_k}$. Our goal in this section is to select a system with the *smallest* expected response, μ_{i_1}. (If we want the *largest* mean μ_{i_k}, the signs of the X_{ij}'s and μ_i's can simply be reversed.) Let "CS" denote this event of "correct selection."

The inherent randomness of the observed X_{ij}'s implies that we can never be *absolutely* sure that we shall make the CS, but we would like to be able to prespecify the *probability* of CS. Further, if μ_{i_1} and μ_{i_2} are actually very close together, we might not care if we erroneously choose system i_2 (the one with mean μ_{i_2}), so that we want a method that avoids making a large number of replications to resolve this unimportant difference. The exact problem formulation, then, is that we want $P(\text{CS}) \geq P^*$ provided that $\mu_{i_2} - \mu_{i_1} \geq d^*$, where the minimal CS probability $P^* > 1/k$ and the "indifference" amount $d^* > 0$ are both specified by the analyst. It is natural to ask what happens if $\mu_{i_2} - \mu_{i_1} < d^*$. (The value d^* is the smallest actual difference that we care about detecting.) The procedure stated below has the nice property that, with probability at least P^*, the expected response of the *selected* system will be no larger than $\mu_{i_1} + d^*$ [see sec. 18.2.3 in Kim and Nelson (2006a)]. Thus, we are protected (with probability at least P^*) against selecting a system with mean that is more than d^* worse than that of the best system (see Fig. 10.4).

The statistical procedure for solving this problem, developed by Dudewicz and Dalal (1975), involves "two-stage" sampling from each of the k systems. In the first stage we make a fixed number of replications of each system, then use the resulting variance estimates to determine how many more replications from each system are necessary in a second stage of sampling in order to reach a decision. It must be assumed that the X_{ij}'s are normally distributed, but (importantly) we need *not* assume that the values of $\sigma_i^2 = \text{Var}(X_{ij})$ are known; nor do we have to assume that the σ_i^2's are the same for different i's. [Assuming known or equal variances (see Table 10.5)

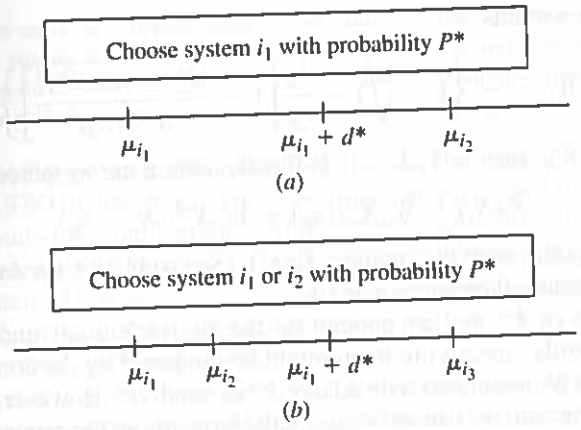

FIGURE 10.4
Selected system(s): (*a*) $\mu_{i_2} > \mu_{i_1} + d^*$; (*b*) $\mu_{i_2} < \mu_{i_1} + d^*$ and $\mu_{i_3} > \mu_{i_1} + d^*$.

is very unrealistic when simulating real systems.] The procedure's performance should be robust to departures from the normality assumption, especially if the X_{ij}'s are averages. (We have verified this robustness when X_{ij} is the average of a fixed number of delays in queue for an *M/M/*1 queueing system.)

In the first-stage sampling, we make $n_0 \geq 2$ replications of each of the k systems and define the first-stage sample means and variances

$$\overline{X}_i^{(1)}(n_0) = \frac{\sum_{j=1}^{n_0} X_{ij}}{n_0}$$

and

$$S_i^2(n_0) = \frac{\sum_{j=1}^{n_0} [X_{ij} - \overline{X}_i^{(1)}(n_0)]^2}{n_0 - 1}$$

for $i = 1, 2, \ldots, k$. Then we compute the total sample size N_i needed for system i as

$$N_i = \max\left\{n_0 + 1, \left\lceil \frac{h_1^2 S_i^2(n_0)}{(d^*)^2} \right\rceil\right\} \tag{10.3}$$

where $\lceil x \rceil$ is the smallest integer that is greater than or equal to the real number x, and h_1 (which depends on k, P^*, and n_0) is a constant that can be obtained from Table 10.11 in App. 10B. Next, we make $N_i - n_0$ *more* replications of system i ($i = 1, 2, \ldots, k$) and obtain the second-stage sample means

$$\overline{X}_i^{(2)}(N_i - n_0) = \frac{\sum_{j=n_0+1}^{N_i} X_{ij}}{N_i - n_0}$$

Then define the weights

$$W_{i1} = \frac{n_0}{N_i}\left\{1 + \sqrt{1 - \frac{N_i}{n_0}\left[1 - \frac{(N_i - n_0)(d^*)^2}{h_1^2 S_i^2(n_0)}\right]}\right\}$$

and $W_{i2} = 1 - W_{i1}$, for $i = 1, 2, \ldots, k$. Finally, define the weighted sample means

$$\tilde{X}_i(N_i) = W_{i1}\overline{X}_i^{(1)}(n_0) + W_{i2}\overline{X}_i^{(2)}(N_i - n_0)$$

and select the system with the smallest $\tilde{X}_i(N_i)$. (See App. 10A for an explanation of the seemingly bizarre definition of W_{i1}.)

The choices of P^* and d^* depend on the analyst's goals and the particular systems under study; specifying them might be tempered by the computing cost of obtaining a large N_i associated with a large P^* or small d^*. However, choosing n_0 is more troublesome, and we can only say, on the basis of our experiments and various statements in the literature, that n_0 be at least 20. If n_0 is too small, we might get a poor estimate $S_i^2(n_0)$ of σ_i^2; in particular, it could be that $S_i^2(n_0)$ is much greater than σ_i^2, leading to an unnecessarily large value of N_i. On the other hand, if n_0 is too large, we could "overshoot" the necessary numbers of replications for some of the systems, which is wasteful. Table 10.11, in App. 10B, gives values of h_1 for $P^* = 0.90$ and 0.95, $n_0 = 20$ and 40, and for $k = 2, 3, \ldots, 10$. If values of h_1 are needed for other P^*, n_0, or k values, we refer the reader to Dudewicz and Dalal (1975) or Koenig and Law (1985).

EXAMPLE 10.9. For the inventory model of Sec. 1.5 (and Examples 10.7 and 10.8), suppose that we want to compare the $k = 5$ different (s, S) policies, as given in Table 10.4, on the basis of their corresponding expected average total costs per month for the first 120 months of operation, which we denote by μ_i for the ith policy. Our goal is to select a system with the smallest μ_i and to be $100P^* = 90$ percent sure that we have made the correct selection provided that $\mu_{i_2} - \mu_{i_1} \geq d^* = 1$. We made $n_0 = 20$ initial independent replications of each system, so that $h_1 = 2.747$ from Table 10.11. The results of the first-stage sampling are given in the $\overline{X}_i^{(1)}(20)$ and $S_i^2(20)$ columns of Table 10.8. From the $S_i^2(20)$'s, h_1, and d^*, we next computed the total sample size N_i for each system, as shown in Table 10.8. Then we made $N_i - 20$ additional replications for each policy, i.e., 90 more replications for policy 1, 41 more for policy 2, etc., and computed the second-stage sample means $\overline{X}_i^{(2)}(N_i - 20)$, as shown. Finally, we calculated the weights W_{i1} and W_{i2} for each system and the weighted sample means $\tilde{X}_i(N_i)$. Since $\tilde{X}_2(N_2)$ is the smallest weighted sample mean, we select policy 2 ($s = 20$ and $S = 80$) as being the lowest-cost configuration. Note from the $S_i^2(20)$ and N_i columns of Table 10.8 that

TABLE 10.8
Selecting the best of the five inventory policies

i	$\overline{X}_i^{(1)}(20)$	$S_i^2(20)$	N_i	$\overline{X}_i^{(2)}(N_i - 20)$	W_{i1}	W_{i2}	$\tilde{X}_i(N_i)$
1	126.48	14.52	110	124.45	0.21	0.79	124.87
2	121.92	7.96	61	121.63	0.39	0.61	121.74
3	127.16	9.45	72	126.11	0.32	0.68	126.44
4	130.71	8.25	63	132.03	0.37	0.63	131.54
5	144.07	6.20	47	144.83	0.46	0.54	144.48

the procedure calls for a higher value of the final N_i if the variance estimate $S_i^2(20)$ is high; this is simply reflecting the fact that we need more data on the more variable systems. Note that if $d^* = 2$, then 28, 21, 21, 21, and 21 total replications are required for policies 1 through 5, respectively.

There is another popular procedure for selecting the best of k systems that is due to Rinott (1978) (denoted $\mathcal{R}$). It uses the usual sample means (based on all first-stage and second-stage replications) from the k systems to make its selection, whereas the Dudewicz and Dalal ($\mathcal{DD}$) procedure uses the weighted sample means from the k systems. However, the $\mathcal{DD}$ procedure generally requires fewer replications than the computationally simpler $\mathcal{R}$ procedure, since h_1 is smaller than the comparable "h value" for the $\mathcal{R}$ procedure.

*The $\mathcal{DD}$ and $\mathcal{R}$ procedures discussed above for selecting the best system assume that the k systems are simulated independently. However, in some cases it might be advantageous (in terms of the sample sizes required to make the correct selection) to use CRN in simulating the k systems. In this regard, Nelson and Matejcik (1995) introduced a two-stage procedure (denoted $\mathcal{NM}$) for selecting the best system that explicitly allows for the use of CRN. Let Σ denote the covariance matrix (see Sec. 6.10.1) of the random variables $X_{1j}, X_{2j}, \ldots, X_{kj}$. The $\mathcal{NM}$ procedure assumes that Σ has a particular structure called *sphericity*, which is defined by

$$\Sigma = \begin{bmatrix} 2\psi_1 + \tau^2 & \psi_1 + \psi_2 & \cdots & \psi_1 + \psi_k \\ \psi_2 + \psi_1 & 2\psi_2 + \tau^2 & \cdots & \psi_2 + \psi_k \\ \vdots & \vdots & & \vdots \\ \psi_k + \psi_1 & \psi_k + \psi_2 & \cdots & 2\psi_k + \tau^2 \end{bmatrix}$$

where the ψ_i's and τ^2 are constants, and $\tau^2 > \sqrt{k \sum_{i=1}^k \psi_i^2} - \sum_{i=1}^k \psi_i$ is required to make Σ positive definite. Sphericity implies that $\text{Var}(X_{ij} - X_{lj}) = 2\tau^2$ for $i \neq l$ (see Prob. 10.11). This means that the variances of all pairwise differences across the systems are equal, even though the marginal variances and covariances may be unequal.

As for the $\mathcal{DD}$ procedure, let n_0 be the first-stage sample size, d^* the indifference amount, and $P^* = 1 - \alpha$ the probability of correct selection. Also, let $g = T^{1-\alpha}_{k-1,(k-1)(n_0-1),0.5}$ be the $(1 - \alpha)$-quantile of the maximum of a $(k - 1)$-dimensional multivariate t distribution with $(k - 1)(n_0 - 1)$ degrees of freedom and a common correlation of 0.5. Values of g are given in table B.3 of Bechhofer et al. (1995) and in table 4 of Hochberg and Tamhane (1987). Then the following is a statement of the $\mathcal{NM}$ procedure:

1. In the first-stage sampling, make $n_0 \geq 2$ independent replications of the ith system using CRN across the k systems (for $i = 1, 2, \ldots, k$).

*The remainder of this section may be skipped on a first reading.

2. Based on the assumption of sphericity, compute the sample variance of the pairwise differences as

$$S^2 = \frac{2\sum_{i=1}^{k}\sum_{j=1}^{n_0}(X_{ij} - \overline{X}_{i\cdot} - \overline{X}_{\cdot j} + \overline{X}_{\cdot\cdot})^2}{(k-1)(n_0-1)}$$

where $\overline{X}_{i\cdot}$ denotes the sample mean of $X_{i1}, X_{i2}, \ldots, X_{in_0}$; $\overline{X}_{\cdot j}$ denotes the sample mean of $X_{1j}, X_{2j}, \ldots, X_{kj}$, etc. (see Prob. 10.12).

3. Compute the total required sample size N (constant for all k systems) as

$$N = \max\left\{n_0, \left\lceil \frac{g^2S^2}{(d^*)^2} \right\rceil\right\} \tag{10.4}$$

4. In the second-stage sampling, make $N - n_0$ independent replications of the ith system using CRN across the k systems (for $i = 1, 2, \ldots, k$).
5. Compute the overall sample mean for the ith system as

$$\overline{X}_i(N) = \frac{\sum_{j=1}^{N} X_{ij}}{N} \qquad \text{for } i = 1, 2, \ldots, k$$

6. Select the system with the smallest $\overline{X}_i(N)$ as being the best alternative.

Nelson and Matejcik show that the probability of correct selection for the $\mathcal{NM}$ procedure is at least P^* when $\mu_{i_2} - \mu_{i_1} \geq d^*$, provided that Σ satisfies the property of sphericity. If $\mu_{i_2} - \mu_{i_1} < d^*$, then a system is returned whose mean is within d^* of the best mean. They also show that their procedure is robust to departures from sphericity when the covariances σ_{ij} are nonnegative, which is the assumed effect of CRN.

EXAMPLE 10.10. Consider the problem of selecting the best of the five inventory policies in Example 10.9, where $100P^* = 90$ percent and $d^* = 1$. We made $n_0 = 20$ first-stage independent replications for each of the five policies using partial CRN as described in Example 11.7; specifically, we made the interdemand times and demand sizes the same across the policies, but generated the delivery lags independently. From the resulting X_{ij}'s, we found that $S^2 = 3.71$ and computed the required total sample size N as

$$N = \max\left\{n_0, \left\lceil \frac{g^2S^2}{(d^*)^2} \right\rceil\right\} = \max\left\{20, \frac{(1.86)^2(3.71)}{(1)^2}\right\} = 20$$

where the value of $g = 1.86$ was taken from table B.3 of Bechhofer et al. (1995). Since $N = 20 = n_0$, it was not necessary to make any second-stage replications. Furthermore, since the five first-stage sample means (i.e., the $\overline{X}_i$.'s) were 125.64, 121.48, 126.16, 131.61, and 144.52, respectively, we once again selected policy 2 as being the best. Note that the $\mathcal{NM}$ procedure using CRN required 100 total replications to select policy

2 as being the best, whereas the $\mathcal{DD}$ procedure required a total of 353 independent replications in Example 10.9. Thus, the $\mathcal{NM}$ procedure reduced the computational effort by approximately 72 percent.

In Sec. 10.3.3 we briefly discussed multiple comparisons with the best (MCB), which has the objective of forming k simultaneous confidence intervals on $\mu_i - \min_{l \neq i} \mu_l$ for $i = 1, 2, \ldots, k$. Nelson and Matejcik (1995) showed that the output of most indifference-zone procedures (e.g., $\mathcal{DD}$ and $\mathcal{NM}$) can be used to construct MCB confidence intervals, and simultaneously guarantee both the correct selection and the coverage of the MCB differences with overall confidence level P^*. This approach allows one to pick the system with the smallest mean and to draw inferences about the differences between the means of the systems, which may facilitate decision making based on a secondary criterion. For example, if the mean of the second-best system does not differ much from the mean of the best system, then it may be desirable to choose the second-best system because of political or economic reasons.

To make these ideas more concrete, consider once again the $\mathcal{NM}$ procedure. Then the following seventh step can be appended to their procedure:

7. For $i = 1, 2, \ldots, k$, construct the MCB confidence interval for $\mu_i - \min_{l \neq i} \mu_l$ as

$$[-(\bar{X}_{i\cdot} - \min_{l \neq i} \bar{X}_{l\cdot} - d^*)^-, (\bar{X}_{i\cdot} - \min_{l \neq i} \bar{X}_{l\cdot} + d^*)^+]$$

where $-x^- = \min(0, x)$ and $x^+ = \max(0, x)$.

EXAMPLE 10.11. For the five inventory policies of Example 10.10, the calculations for the MCB confidence intervals are given in Table 10.9. Overall we are at least 90 percent confident that policy 2 is the best *and* that the five confidence intervals contain their respective MCB differences. From the second confidence interval, we conclude that policy 2 is no worse than the other policies (the upper endpoint is 0), and it may be as much as \$5.16 less expensive than the others (the lower endpoint is −5.16). The other confidence intervals tell us that policies 1, 3, 4, and 5 are no better than policy 2 (the lower endpoints of their intervals are 0) and may be as much as \$5.16, \$5.68, \$11.13, and \$24.04 more expensive, respectively.

The $\mathcal{DD}$ and $\mathcal{NM}$ procedures are typically used when the number of alternative systems, k, is 20 or fewer. These procedures are designed to produce the desired

TABLE 10.9
MCB confidence intervals for the five inventory policies

i	Lower MCB endpoint	$\bar{X}_{i\cdot} - \min_{l \neq i} \bar{X}_{l\cdot}$	Upper MCB endpoint
1	0	4.16	5.16
2	−5.16	−4.16	0
3	0	4.68	5.68
4	0	10.13	11.13
5	0	23.04	24.04

probability of correct selection, P^*, when $\mu_{i_1} + d^* = \mu_{i_2} = \cdots = \mu_{i_k}$ (see App. 10A), an arrangement of the μ_i's known as the *least-favorable configuration* (LFC). This is the worst-case situation in that it makes the best system as hard to distinguish from the others as possible, given that it's at least d^* better than everything else. {This assumption is made because it makes the calculation of N_i or N [see Eqs. (10.3) and (10.4), respectively] independent of the true and sample means.} Thus, when k is "large" and the μ_i's differ widely, the $\mathcal{DD}$ and $\mathcal{NM}$ procedures may prescribe larger sample sizes than needed to deliver the desired probability of correct selection. As a result of these considerations, Nelson et al. (2001) introduced a screen-and-select procedure (denoted $\mathcal{NSQSP}$) for use when k is large. The screening stage first produces a subset (of random size) that excludes clearly inferior systems. Then in the succeeding selection stage an indifference-zone procedure (e.g., $\mathcal{DD}$ or $\mathcal{NM}$) is applied to the set of remaining systems to choose the "best" system, and the combined procedure guarantees an overall probability of correct selection. Because no selection-stage observations are collected on inferior systems, the combined procedure may require fewer observations than the use of an indifference-zone procedure alone. The screening part of this procedure has been implemented in the Process Analyzer for the Arena simulation package (see Sec. 3.5.1).

10.4.2 Selecting a Subset of Size m Containing the Best of k Systems

Now we consider a different kind of selection problem, that of selecting a subset of exactly m of the k systems (m is prespecified) so that, with probability at least P^*, the selected subset will contain a system with the smallest mean response μ_{i_1}. This could be a useful goal in the initial stages of a simulation study, where there may be a large number (k) of alternative systems and we would like to perform an initial screening to eliminate those that appear to be clearly inferior. Thus, we could avoid expending a large amount of computer time getting precise estimates of the behavior of these inferior systems.

We define X_{ij}, μ_i, μ_{i_j}, and σ_i^2 as in Sec. 10.4.1. Here we assume that all X_{ij}'s are independent and normal (CRN is not allowed), and for fixed i, $X_{i1}, X_{i2}, \ldots$ are IID; the σ_i^2's are unknown and need not be equal. Here, correct selection (CS) is defined to mean that the subset of size m that is selected contains a system with mean μ_{i_1} and we want $P(\text{CS}) \geq P^*$ provided that $\mu_{i_2} - \mu_{i_1} \geq d^*$; here we must have $1 \leq m \leq k - 1$, $P^* > m/k$, and $d^* > 0$. (If $\mu_{i_2} - \mu_{i_1} < d^*$, then with probability at least P^*, the subset selected will contain a system with expected response that is no larger than $\mu_{i_1} + d^*$.)

The procedure is very similar to the $\mathcal{DD}$ procedure of Sec. 10.4.1, and has been derived by Koenig and Law (1985). We take a first-stage sample of $n_0 \geq 2$ replications from each system and define $\overline{X}_i^{(1)}(n_0)$ and $S_i^2(n_0)$ for $i = 1, 2, \ldots, k$ exactly as in Sec. 10.4.1. Next we compute the total number of replications, N_i, needed for the ith system exactly as in Eq. (10.3), except that h_1 is replaced by h_2 (which depends on m as well as on k, P^*, and n_0), as found in Table 10.12 in App. 10B. [For values of h_2 that might be needed for other P^*, n_0, k, or m values, see Koenig and Law