*Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control.* James C. Spall

ISBN: 0-471-33052-3

# CHAPTER 12

# STATISTICAL METHODS FOR OPTIMIZATION IN DISCRETE PROBLEMS[1]

This chapter is unique in this book in focusing exclusively on problems where $\theta$ takes on only discrete values. We discuss some statistical methods of comparing the loss values for the possible outcomes for $\theta$, the aim being to find the values of $\theta$ yielding the lowest values of the loss function. The use of statistical testing here differs from most of the rest of the book in that testing is being used to *solve* the optimization problem rather than to *evaluate* other optimization methods. Further, the methods here differ in not generally relying on the notion of updating a $\theta$ estimate in an iterative manner. Rather, the methods sift through the candidate values of $\theta$ to produce an estimate in a batch manner.

The central challenge is that classical statistical testing is best suited to *pairwise* comparisons (is option 1 better than option 2?), yet the primary interest here is a many-way comparison. Numerous approaches have been developed for extending notions of pairwise comparisons to general multiple comparisons. Some of these will be discussed here. Although there may be uniquely "obvious" approaches to pairwise testing—such as the pairwise $t$-tests in Appendix B—there will rarely be such a unique approach when the test involves a comparison of at least three options (i.e., at least three values in the domain $\Theta$ for $\theta$).

This chapter examines several approaches to multiple comparisons tests as they apply in the problem of stochastic optimization. Section 12.1 provides some general background on the testing framework. Section 12.2 describes a popular test for performing an initial statistical assessment of whether there exist some candidate $\theta$ values that are better than other $\theta$ values. Sections 12.3 and 12.4 assume that prior information is available to suggest that a particular $\theta$ is a candidate for being the optimum. This allows for a potentially dramatic reduction in the number of comparisons that need to be tested, thereby improving the ability to make concrete statistical statements. Section 12.3 assumes that the

[1]Without loss of continuity, this chapter may be skipped or postponed by those whose interests focus on algorithms for search and optimization where the estimate is updated recursively. Although this chapter contains important material for discrete problems, the material here is not a prerequisite to the material in other chapters, with the possible exception of Section 14.5 on selection methods for optimization via Monte Carlo simulations.

measurement noise variance is known, whereas Section 12.4 does not. Section 12.5 summarizes some other statistical tests for aiding in the discrete optimization task. These tests are sometimes referred to as ranking and selection methods in the statistics and other literature. Section 12.6 offers some concluding remarks, including some comments about sequential analysis methods—such as multi-armed bandit problems—for optimizing over a discrete set $\Theta$.

Roughly speaking, the multiple comparisons tests that dominate this chapter provide a means for combining *pairwise* comparisons of options in a meaningful way to perform general *multiple* comparisons. The ranking and selection methods, in contrast, are not based directly on the combination of pairwise comparisons. Section 14.5 covers multiple comparisons tests in the special case where Monte Carlo simulations are generating the measurements of $L$. In particular, the property of common random numbers discussed in that chapter is used to advantage in the simulation-based setting.

## 12.1 INTRODUCTION TO MULTIPLE COMPARISONS OVER A FINITE SET

While much of this book concentrates on search and optimization when the elements of $\boldsymbol{\theta}$ can, in principle, take on any real-numbered value satisfying possible constraints in $\Theta$, there are many important problems where there are a finite number of discrete options. (The "in principle" qualifier pertains to the fact that a digital computer implementation is necessarily discrete, although for all practical purposes may be treated as continuous.) Problems having a finite number of options can arise in many ways, including as a natural part of the problem definition or as a simplification of a large-scale continuous or discrete problem with an unbounded number of possibilities. Suppose that the interest is in finding $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$, where $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K\}$ represents the $K$ possible values for $\boldsymbol{\theta}$. It is possible that more than one value of $\boldsymbol{\theta}_i$ can serve as $\boldsymbol{\theta}^*$ (i.e., the set $\Theta^*$ from expression (1.1) contains more than one point), but for convenience in discussion, we generally make singular (versus plural) references to $\boldsymbol{\theta}^*$. If at any value of $\boldsymbol{\theta}$, one can measure $L$ exactly (i.e., without noise), then the obvious means by which one can determine $\boldsymbol{\theta}^*$ is to calculate $L(\boldsymbol{\theta}_i)$ for all $i = 1, 2, \ldots, K$, taking $\boldsymbol{\theta}^*$ as the $\boldsymbol{\theta}_i$ yielding the lowest value of $L$. (Of course, if $K$ is a very large number, even this simple approach may not be feasible.)

We are interested in the more challenging case where $L$ can be observed only in the presence of noise (i.e., Property A in Section 1.1). Hence, simply collecting measurements of $L$ and comparing values—without further analysis—is generally inadequate for making rigorous statements about the optimality of a point. The job of the statistical comparison procedures here is to sort through potentially confusing or ambiguous measurements in an attempt to provide a formal means of estimating $\boldsymbol{\theta}^*$. Because the number of tests required in the statistical procedures grows rapidly with the number of options to be considered,

statistical procedures such as these are appropriate when the number of options, $K$, is not too large. Goldsman and Nelson (1998), for example, suggest $2 \leq K \leq 20$ as a reasonable range.

In problems with a larger $K$, it is possible to combine several methods for screening as a way of narrowing the options (e.g., Nelson et al., 2001). Alternatively, methods such as the direct search techniques of Chapter 2, simulated annealing and related methods (Chapter 8), or evolutionary computation (Chapters 9 and 10) *may* be useful. However, the application of such methods with noisy loss measurements (as we have here) is largely ad hoc. Further, there is a loss of statistical interpretation for the solution.

There also exist sequential methods for optimizing over a discrete set $\Theta$ based on principles of decision theory (see the survey paper by Lai, 2001). These sequential methods involve an adaptive choice of the sampling strategy for determining the optimal $\boldsymbol{\theta}$. As with the methods of this chapter, the sequential methods provide statistical guarantees in the form of hypothesis tests. Important examples of such alternative methods are multiarmed bandit problems, so named by statistical analogies to slot machines. While the sequential methods are powerful and possibly more efficient in terms of the sample sizes required to guarantee certain levels of statistical accuracy, they tend also to be more cumbersome to implement than the relatively simple methods of this chapter.

The conceptual example below illustrates one type of application for the statistical comparisons methods of this chapter. There are countless other types as well. Although the example below is based on Monte Carlo simulations, the comparisons methods can also apply when physical experiments are used to produce the measurements of $L$. Section 14.5 demonstrates how certain properties unique to simulations can be exploited to enhance some types of statistical comparisons procedures.

**Example 12.1—Decision making via Monte Carlo simulation.** Let us summarize a conceptual example where multiple comparisons of the type here might be used. Suppose that a decision maker for a firm is attempting to choose the best alternative from among $K$ possible options. Each of these options represents a significant investment for which it is not possible to run actual field experiments. For example, an option may involve the acquisition of another firm or the construction of a new manufacturing plant. Suppose that the firm has developed a Monte Carlo simulation that provides a realistic evaluation of alternatives. Each run of the simulation produces an estimate of the firm's annual profit under a particular option. To run the simulation, values for system parameters $\boldsymbol{\theta}$ must be specified. The $i$th option is represented in the simulation by setting $\boldsymbol{\theta} = \boldsymbol{\theta}_i$. As a reflection of randomness in the real world, multiple runs of the Monte Carlo simulation at a *fixed* scenario (a fixed $\boldsymbol{\theta}$) produce *different* estimates for the profit.

To help decide which option is best—that is, to find $\boldsymbol{\theta}^*$—the decision maker runs the simulation multiple times at each $\boldsymbol{\theta}_i$. By averaging the simulation

outputs under each scenario, the decision maker has information to judge whether one or more scenarios are clearly superior or inferior in a statistical sense given the inherent variability at each $\boldsymbol{\theta}_i$. This comparison can be carried out using statistical tests of the type in this chapter (and Section 14.5). ❑

One will find in the methods below that determining a likely $\boldsymbol{\theta}^*$ with the comparison-based statistical methods does not yield a "cut and dried" approach (just like most other methods of stochastic search and optimization!). Although there are some rigorous methods that can point to some values of $L(\boldsymbol{\theta}_i)$ as probably being better (lower) than other values, it is difficult to establish a single rigorous probabilistic statement about the likelihood of a specific $\boldsymbol{\theta}_i$ being the optimal value. Rather, one generally has to be satisfied with statistical *indicators* of an optimum. These indicators—though imperfect—provide more quantification of the likelihood of reaching $\boldsymbol{\theta}^*$ than the information that is available to the analyst in the other discrete search methods mentioned above.

The fundamental difficulty arises from the ambiguity inherent in making more than one comparison: If statistical indications suggest that one specific hypothesis is not likely, then there may be many plausible alternatives that need to be considered, only one of which corresponds to the (unknown) truth when $\boldsymbol{\theta}^*$ is unique. This contrasts with pairwise statistical testing, where if the data collected are strongly inconsistent with a null hypothesis of, say, $L(\boldsymbol{\theta}_1) \geq L(\boldsymbol{\theta}_2)$, then there are indications that the *unique* alternative hypothesis $L(\boldsymbol{\theta}_1) < L(\boldsymbol{\theta}_2)$ is true (i.e., $\boldsymbol{\theta}_1$ is optimal). Furthermore, the error in this conclusion is explicitly quantifiable at the false alarm rate of the test. In some cases, it is possible to introduce additional assumptions—such as knowledge that the optimal point produces a loss function value at least $\delta$ units better than all other values—as a way of circumventing the difficulties with a lack of unique alternatives. With such assumptions, there is sometimes an explicit quantifiable error as in the pairwise case. An example is given in Section 12.5. The broad philosophy below is to test a general (null) hypothesis that all points in $\Theta$ are equivalent, in the hopes that this hypothesis can be *rejected*. One must then perform detailed analysis to attempt to isolate the element in $\Theta$ most likely to be $\boldsymbol{\theta}^*$.

To determine $\boldsymbol{\theta}^*$, the statistical approaches here are based on the principle of multiple comparisons. Multiple comparisons is fundamental when one needs to compare more than two quantities. The methods below are based on two broad steps: (i) a test where we simultaneously determine if there is evidence that not all candidate points $\boldsymbol{\theta}_i$ produce the same loss value and (ii) given that there is such evidence, refined analysis on a subset of points as necessary to resolve ambiguities that may remain after the full simultaneous test in (i). The ultimate aim, of course, is to determine clear winners, that is, particular $\boldsymbol{\theta}_i$ that are statistically better than all other $\boldsymbol{\theta}_j$, $j \neq i$. In practice, the analyst may have to settle for less information than clearly identified winners, perhaps instead a strong indication of a likely optimal point or small set of likely optimal points. The two-sample tests discussed in Appendix B are fundamental

in conducting the refined analysis. The treatment of multiple comparisons here is relatively brief. The reader with a serious interest in this branch of search and optimization is directed to more comprehensive references such as Gupta (1965), Miller (1981), Hochberg and Tamhane (1987), Hsu (1996), and Goldsman and Nelson (1998). Although some of these references are of an older vintage, the points raised are still largely relevant.

Statistical hypothesis testing plays a prominent role in executing the broad steps mentioned above. Such testing involves a *null* hypothesis and an *alternative* hypothesis. The null hypothesis in all the tests being considered here is that all candidate $\boldsymbol{\theta}_i$ produce loss values that are effectively equal. The aim in trying to find an optimal $\boldsymbol{\theta}_i$ is, of course, to *reject* this null hypothesis.

One might wonder why the null hypothesis is not taken to be the primary hypothesis of interest (that a particular $\boldsymbol{\theta}_i$ is the best $\boldsymbol{\theta}$ value). It is customary to make the null hypothesis a nominal state of nature—in this case that $\boldsymbol{\theta}_i$ is no better than the other $\boldsymbol{\theta}$ values—because it is almost always easier to characterize the distribution of the test statistic under the null hypothesis. Then, based on this characterization, one can guarantee that a null hypothesis will only be mistakenly rejected with a small probability (the *type I error* or *false alarm rate* in common vernacular). In other words, rejection of the null hypothesis is strong evidence that the alternative hypothesis is, in fact, true since it is unlikely that the null hypothesis will be rejected when it is true. On the other hand, failing to reject the null hypothesis is usually only weak evidence that the null is true because one usually has only limited (or no!) control over the *type II error*, that of failing to reject the null when the alternative is, in fact, true.

With multiple options, hypothesis testing is complicated in a significant way: the null hypothesis may be false in many ways. For example, $\boldsymbol{\theta}_1$ may yield a loss value different from all other $\boldsymbol{\theta}_i$; or only $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ may yield different loss values; or only $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_3$ may yield different values; and so on. Only one of the many alternative hypotheses corresponds to a specified $\boldsymbol{\theta}_i$ being the best $\boldsymbol{\theta}$ value. Hence the second broad step mentioned above: to attempt to isolate the cause of the rejected null hypothesis. This will often involve an attempt to determine whether the *one* alternative hypothesis demonstrating the superiority of a *specified* element in $\Theta$ is, statistically, the correct alternative hypothesis. That will be the focus of Sections 12.3 and 12.4, following the general test of Section 12.2.

As is typical in statistical practice, acceptance regions are used for testing the null hypothesis. These regions are designed to contain the chosen test statistic with high probability *if* (an important "if") the null hypothesis of equality of the loss values for the various elements of $\Theta$ is true. If the chosen test statistic does not land in the acceptance region, the null hypothesis is rejected. To obtain concrete solutions, to be consistent with the vast majority of multiple comparisons approaches, and to aid in isolating the effects of individual $\boldsymbol{\theta}_i$ in $\Theta$, the acceptance *region* will be restricted to a set of acceptance *intervals*. Each interval corresponds to a pairwise comparison between two of the elements in $\Theta$.

Hence, the acceptance regions of interest here will be multidimensional rectangles.

A fundamental property in multiple comparisons tests is that such intervals must be wider than their single comparison counterpart (as in the pairwise tests of Appendix B). The wider intervals guard against the greater chance of an "extreme" event following from the greater number of opportunities to find significant occurrences. In particular, when there are many possible events, it is not surprising to find at least one oddball! Interestingly, this *multiplicity effect* is sometimes ignored, leading to false conclusions about whether particular events are statistically significant. For example, Benjamini and Yekutieli (2001) discuss a published medical study where clinical trials suggest that a particular treatment for breast cancer provides 6 of 18 possible medical benefits. This result is based on a claim of statistically significant results in 6 of 18 comparisons ($P$-values $\leq 0.05$), but where the multiplicity is ignored. In fact, as noted by Benjamini and Yekutieli (2001), only 2 to 4 (versus 6) of the 18 possible benefits are supported by the data at an overall $P$-value of 0.05 (the range of 2 to 4 benefits depends on the details of the testing procedure).[2]

In carrying out the tests below, at each value of $\boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K\}$, let $\bar{L}_i$ represent a sample mean of measured values of $L$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_i$. In particular, let $y_k(\boldsymbol{\theta}_i) = L(\boldsymbol{\theta}_i) + \varepsilon_{ki}$, represent the $k$th measurement of $L$ at the given $\boldsymbol{\theta}_i$, $k = 1, 2, \ldots, n_i$ (say), and $\varepsilon_{ki}$ represent the mean-zero noise term. Then, $\bar{L}_i = n_i^{-1} \sum_{k=1}^{n_i} y_k(\boldsymbol{\theta}_i)$. A fundamental quantity in the tests below is the difference

$$\delta_{ij} \equiv \bar{L}_i - \bar{L}_j \tag{12.1}$$

for any pair $i, j$. We will make varying assumptions about the distribution of the noise terms, beginning with independent, identically distributed (i.i.d.) normal in Section 12.2 and relaxing this condition in Section 12.3. As discussed in Section 1.1 and illustrated in numerous places throughout this book, many practical problems have noise that is non-i.i.d., often because of a dependence of the noise on the value of $\boldsymbol{\theta}$. In some of the tests below, it is assumed that it is possible to do a table lookup to find critical values of a relevant distribution; these critical values are often based on the noises in the loss measurements (the $\varepsilon$) having desirable properties such as independence and normality. In some applications, such assumptions will clearly not hold, as discussed in Section 1.1. Somerville (1997) provides an automated way of computing the required constants in cases where simple table lookups do not apply. The method is a combination of numerical integration and Monte Carlo sampling.

---

[2]Benjamini and Yekutieli (2001) point out that the *New England Journal of Medicine* (which is not where the study they discussed appeared) requires that researchers account for the multiplicity effect in reporting results. This appears to be a rare requirement in medical journals.

## 12.2 STATISTICAL COMPARISONS TEST WITHOUT PRIOR INFORMATION

A classical—and popular—method for performing simultaneous comparisons was introduced in 1953 by J. W. Tukey in an unpublished technical report from Princeton University entitled "The Problem of Multiple Comparisons." The method was independently introduced in Kramer (1956). This test is a generalization of the familiar two-sample $t$-tests in Appendix B. In fact, Goldsman and Nelson (1998) state: "The most widely used method for forming [simultaneous] intervals is Tukey [–Kramer]'s method, which is implemented in many statistical software packages." The interest here is in constructing acceptance intervals for all $K(K-1)/2$ differences $L(\boldsymbol{\theta}_i) - L(\boldsymbol{\theta}_j)$ (Exercise 12.1). If prior information suggests that a particular $\boldsymbol{\theta}_i$ is a strong candidate for being $\boldsymbol{\theta}^*$, then the Tukey–Kramer procedure may be unnecessary and one should proceed to a more focused test such as in Section 12.3 or 12.4. On the other hand, in the absence of prior information, the construction of the $K(K-1)/2$ simultaneous intervals is a useful step in identifying candidate values $\boldsymbol{\theta}_i$ for consideration as the optimal point $\boldsymbol{\theta}^*$.

The Tukey–Kramer technique is derived under the assumptions that the noises $\varepsilon_{ki}$ in the measurements $y_k(\boldsymbol{\theta}_i) = L(\boldsymbol{\theta}_i) + \varepsilon_{ki}$ are i.i.d. normal with mean zero across $k$ and $i$. In many practical problems, however, these noise assumptions do not hold, motivating the alternative inequality-based approaches in Sections 12.3 and 12.4. Nevertheless, the assumptions are sometimes reasonable. Furthermore, the Tukey–Kramer method is sometimes used irrespective of the distribution of the noise terms, as it is "sensible" in much the way that the $t$-distribution is often used when the parent population is nonnormal as discussed in Appendix B, Mardia et al. (1979, p. 147), and Goldsman and Nelson (1998). If being used in a problem where the conditions are seriously violated, however, one should exercise caution, especially if the inference leads to some "close calls."

For this method, suppose that one is interested in constructing multiple comparison acceptance intervals such that when the null hypothesis, $L(\boldsymbol{\theta}_i) = L(\boldsymbol{\theta}_j)$ for all $i, j$, is true (i.e., all points in $\Theta$ yield the same loss value), then it is known with probability at least $1 - \alpha$ (say) that *all* $K(K-1)/2$ test statistics, each corresponding to one of the differences $L(\boldsymbol{\theta}_i) - L(\boldsymbol{\theta}_j)$, will lie in the joint acceptance region. Hence, under the null hypothesis, the probability is no more than $\alpha$ that at least one of the test statistics will lie outside its acceptance interval. Typical values for $\alpha$—the false alarm rate—from classical hypothesis testing are 0.01, 0.05, or 0.10.

Given that the noises are assumed i.i.d, it is valid to pool all measurements in forming the estimate of the common variance. Let the pooled sample variance from the $n_1 + n_2 + \ldots + n_K$ sample values of the loss function be

$$s^2 \equiv \frac{\sum_{i=1}^{K}\sum_{k=1}^{n_i}[y_k(\boldsymbol{\theta}_i)-\bar{L}_i]^2}{\sum_{i=1}^{K}(n_i-1)}. \tag{12.2}$$

The above is a natural extension of the classical variance estimate for one sample (see Appendix B). Note that $s^2$ is an unbiased estimate of the variance (another unbiased—but inferior—variance estimate is given in Exercise 12.12). The simultaneous acceptance intervals for the differences $\delta_{ij}$ are then

$$\left[-Q_{K,\nu}^{(\alpha)}\frac{s\sqrt{n_i^{-1}+n_j^{-1}}}{\sqrt{2}},\ Q_{K,\nu}^{(\alpha)}\frac{s\sqrt{n_i^{-1}+n_j^{-1}}}{\sqrt{2}}\right], \tag{12.3}$$

where $Q_{K,\nu}^{(\alpha)}$ is the $1-\alpha$ quantile of the studentized range distribution with parameter $K$ and degrees of freedom $\nu = n_1 + n_2 + \ldots + n_K - K$ (the quantile is the point such that the probability of being less than or equal to this point is $1-\alpha$). A table with the critical values of this distribution is given in Miller (1981, pp. 234–237), Hochberg and Tamhane (1987, App. 3, Table 8), and, in condensed form, Goldsman and Nelson (1998). Note that the interval in (12.3) corresponds exactly to the identical variances pairwise $t$-test interval in (B.3b) in Appendix B in the special case where $K = 2$ (i.e., $Q_{2,\nu}^{(\alpha)}/\sqrt{2} = t_\nu^{(\alpha/2)}$).

Hayter (1984) made the important discovery that the intervals in (12.3) are *at least* as wide as necessary to provide the probability $1-\alpha$ coverage. That is, if the null hypothesis of $L(\boldsymbol{\theta}_1) = L(\boldsymbol{\theta}_2) = \ldots = L(\boldsymbol{\theta}_K)$ is true, then the probability of all $\delta_{ij}$ simultaneously lying in their respective intervals from (12.3) is at least $1-\alpha$. So, in conducting an experiment, the probability that at least one $\delta_{ij}$ will not lie in its corresponding interval is less than or equal to $\alpha$. Hence, there is potentially strong evidence that the null hypothesis is false if at least one $\delta_{ij}$ does not lie in its corresponding interval. The formal statement of the result is below.

**Theorem 12.1 (Hayter, 1984).** Suppose that the measurement noises $\varepsilon_{ki}$ are i.i.d., normal, with mean zero across $k$ and $i$. Under the null hypothesis of equal $L(\boldsymbol{\theta}_i)$, the probability of all $\delta_{ij}$ simultaneously lying in the random intervals in (12.3) is greater than or equal to $1-\alpha$. The probability is equal to $1-\alpha$ *only* when $K = 2$ or $n_1 = n_2 = \ldots = n_K$ for $K \geq 3$. If any $n_i \neq n_j$ when $K \geq 3$, then the actual probability is strictly above $1-\alpha$.

Section 12.1 mentioned one of the fundamental properties of multiple comparisons tests. That is, the acceptance intervals need to be wider than their pairwise counterparts in order to guard against the greater likelihood of an

extreme event when there is more than one event (i.e., $K \geq 3$ or, equivalently, an evaluation of more than one $\delta_{ij} = \bar{L}_i - \bar{L}_j$). Figure 12.1 illustrates this property for the Tukey–Kramer test. It is assumed that $\alpha = 0.05$, $n_i = 10$ for all $i$, and that $s^2(n_i^{-1} + n_j^{-1}) = 1$ for all $K$. The latter assumption is simply a convenient normalization that follows from $s^2$ being an unbiased estimator of the variance at any $K$. That is, $E[s^2(n_i^{-1} + n_j^{-1})]$ is the same for all $K$. Figure 12.1 shows intervals for four values of $K$. As expected, the intervals grow in width with $K$, although at a slowing rate as $K$ gets larger.

Example 12.2 illustrates the above approach for a simple problem with $K = 4$. This example depicts a realistic experiment where the null hypothesis $L(\boldsymbol{\theta}_1) = L(\boldsymbol{\theta}_2) = L(\boldsymbol{\theta}_3) = L(\boldsymbol{\theta}_4)$ is rejected, but where the optimal point does not unambiguously leap out.

**Example 12.2—Simultaneous acceptance intervals.** Suppose that the true loss values unknown to the analyst are $L(\boldsymbol{\theta}_i) = 2.5$, 3.0, 3.5, and 2.0 for $i = 1$, 2, 3, and 4 respectively, and that the measurement noise is i.i.d. normal with mean zero and unknown variance. The aim is to construct acceptance intervals for the $K(K-1)/2 = 6$ pairs $\delta_{ij}$ when $\alpha = 0.05$. Suppose that each loss measurement represents an expensive evaluation, so that the analyst can afford only 40 measurements for the analysis, allocated as $n_1 = n_2 = n_3 = n_4 = 10$. The following matrix of data is collected, where the four columns, in order, contain the 10 measurements of the loss function at $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, $\boldsymbol{\theta}_3$, and $\boldsymbol{\theta}_4$:

$$\begin{bmatrix} 2.69 & 1.99 & 3.25 & 3.30 \\ 3.11 & 1.23 & 2.97 & 3.25 \\ 2.07 & 5.20 & 4.93 & 0.80 \\ 1.93 & 2.90 & 4.81 & 2.21 \\ 2.42 & 3.83 & 2.99 & 2.77 \\ 2.71 & 3.09 & 3.51 & 0.42 \\ 1.66 & 1.12 & 2.36 & 1.64 \\ 3.33 & 3.15 & 5.17 & 1.43 \\ 3.44 & 3.69 & 2.53 & 1.98 \\ 2.05 & 4.48 & 3.87 & 1.40 \end{bmatrix}.$$

We find that $\bar{L}_i = 2.54$, 3.07, 3.64, and 1.92 for $i = 1$, 2, 3, and 4, respectively. (These sample means—and certain other data-derived values here—may differ slightly from quantities based on the data shown above due to round-off error; the reported means are calculated from original data with more digits of accuracy.) Further, $s = 1.017$ using (12.2) and $Q_{4,36}^{(0.05)} = 3.82$, the latter from Table 8.1 in Goldsman and Nelson (1998). From (12.3), the (common) acceptance interval for use in evaluating the six differences from (12.1) is

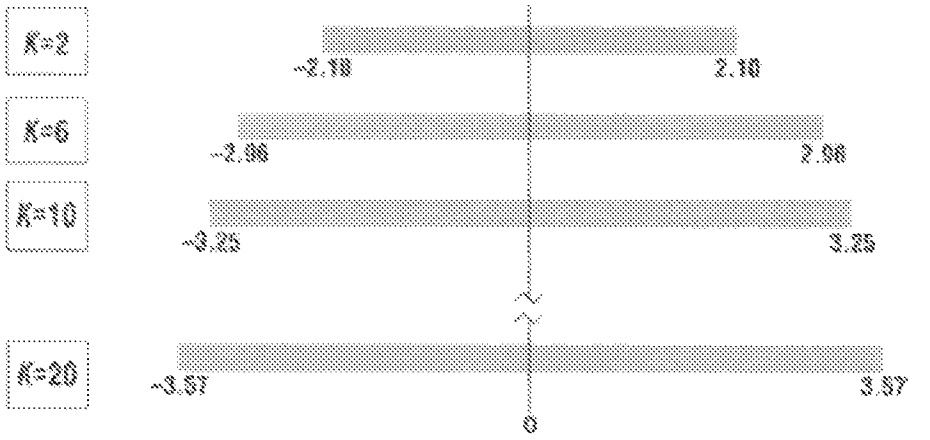


**Figure 12.1.** Relative widths of acceptance intervals (12.3) for varying $K$ in Tukey–Kramer test ($n_1 = n_2 = \ldots = n_K = 10$). Coverage probability is $1 - \alpha = 0.95$.

[–1.23, 1.23]. (For comparison purposes, with only one pairwise test—as in Appendix B—the interval with the same $s$ and sample size $n_i$ is [–0.96, 0.96], illustrating the greater width required to protect against the greater number of possible events. See also Figure 12.1 above.) Table 12.1 shows the values of the differences $\delta_{ij}$ that are to be tested for inclusion in the acceptance interval.

Because $\delta_{34}$ lies outside of the acceptance region (corresponding to the null hypothesis $L(\theta_1) = L(\theta_2) = L(\theta_3) = L(\theta_4)$), one can reject the null hypothesis (at level 0.05) of equality in the loss values. Unfortunately, this test does not provide an "automatic" means of identifying the cause of the rejection, although an objective examination of the table below does correctly suggest that $\theta_4$ is a possible cause of the rejection. The point $\theta_4$ is also associated with another near rejection in comparison to $i = 2$. Note, however, that the test statistic in Table 12.1 for $i = 1$ and $j = 3$ provides some evidence that $\theta_1$ should be evaluated

**Table 12.1.** Difference values for testing hypothesis.

| Indices $i, j$ | $\delta_{ij}$ |
|---|---|
| 1, 2 | –0.53 |
| 1, 3 | –1.10 |
| 1, 4 | 0.62 |
| 2, 3 | –0.57 |
| 2, 4 | 1.15 |
| 3, 4 | 1.72 |

further given that it nearly falls outside of the acceptance interval. Further analysis, possibly of the type discussed in Sections 12.3 and 12.4, would be required to solidify the conclusion that $\theta^* = \theta_4$. ❑

## 12.3 MULTIPLE COMPARISONS AGAINST ONE CANDIDATE WITH KNOWN NOISE VARIANCE(S)

The Tukey–Kramer multiple comparisons method above is useful for obtaining broad information about possible optimal points and likely nonoptimal points. However, because it does not assume any prior information about a likely optimum, the method suffers in needing to produce acceptance intervals that are wide enough to cover the number of expected anomalies in all $K(K-1)/2$ pairs $i, j$. A reduction in the number of comparisons can lead to a sharper analysis by producing tighter acceptance intervals. That is one of the goals of the general approach in this section. The other main goal is to consider some more general noise conditions under which multiple comparisons can be carried out. The aims and general approach of this section and the next are identical, the only difference being that this section assumes that the variances of certain differences $\delta_{ij}$ are known, whereas the next section does not.

The testing approach described below rests on the premise that one has prior information suggesting that one of the $K$ points, say $\theta_m$, is a candidate for being the optimal point. This may have resulted from preliminary testing based on an initial data set different from the data to be used in the formal testing below[3] or it may be a consequence of a "hunch" or physical understanding of the process. The Tukey–Kramer method in Section 12.2 is one of the methods that may be useful for the initial culling to identify a candidate $\theta_m$. In contrast to having no prior insight (where any one of the $K$ values is a candidate to be tested), testing that revolves around one candidate point dramatically reduces the number of comparisons that need to be tested from all $K(K-1)/2$ pairs $i, j$ to just the $K-1$ pairs of $\theta_m$ against each of the other $\theta_j$. (A further reduction may be possible if some of the original $K$ candidates can be clearly eliminated from consideration; in that case the $K$ that is discussed below will be less than the $K$ in the sections above. For ease of discussion, however, we will not distinguish between these "before and after" values of $K$, proceeding under the assumption that all of the original $K$ candidates are viable candidates for consideration.)

Other aspects being equal, this reduction to only $K-1$ pairs to be tested sharpens the inference considerably. In addition, further sharpening is possible since we now consider only *one-sided* acceptance regions. That is, since there is

[3]It is statistically invalid to reuse the data employed in such preliminary analysis in the formal hypothesis tests to follow. The same data should not be used to both construct a hypothesis for testing and to test the hypothesis. Formally, such reuse invalidates the null distribution forming the basis for the hypothesis test.

an identified candidate optimal point, it is only necessary to consider deviations in one direction for each of the differences $\delta_{mj}$. (In the absence of a candidate point, one must consider wider *two-sided* intervals, as in the Tukey–Kramer method.) The "other aspects being equal" qualifier above is relevant here since we include some more general noise conditions than the i.i.d. normal assumption of the Tukey–Kramer method, in which case one may not always see a tightening of the confidence intervals. Methods involving a comparison against one candidate point are called *many-to-one* in Tong (1980, p. 187) and *multiple comparisons with control* or *multiple comparisons with best* in Hsu (1996) and Goldsman and Nelson (1998). We use the term *many-to-one*.

Let us assume that the noise terms have a common (zero or nonzero) mean for all $i, j$; the common mean is needed to ensure that the differences $\delta_{mj}$ have zero mean. Except as noted below, it is not assumed that the noises are mutually independent (mutual independence is stronger than pairwise independence, which in turn is stronger than uncorrelatedness). Further, the methods here are based on knowing the variances of the differences $\delta_{mj}$. From (12.1), knowing the variance—and, if relevant, the correlations—of the noises provides the variance of $\delta_{mj}$ via the relationship

$$\text{var}(\bar{L}_m - \bar{L}_j) = \text{var}\left[\frac{\varepsilon_{m1} + \varepsilon_{m2} + \ldots + \varepsilon_{mn_m}}{n_m} - \frac{\varepsilon_{j1} + \varepsilon_{j2} + \ldots + \varepsilon_{jn_j}}{n_j}\right].$$

The next section extends the methods here to cases where these variances are not known. If, for some $m$, the $\delta_{mj}$ are sufficiently negative for all $j \neq m$, then there is evidence that $\boldsymbol{\theta}_m$ is the optimal point. Unfortunately, it is usually difficult in practice to make a precise statistical statement about the likelihood of $\boldsymbol{\theta}_m$ being the optimal. The fundamental difficulty is that

$$\boldsymbol{\delta}_{m|K} \equiv [\delta_{m1}, \delta_{m2}, \ldots, \delta_{m,m-1}, \delta_{m,m+1}, \ldots, \delta_{mK}]^T$$

represents a vector of $K-1$ *dependent* random variables through the common presence of $\bar{L}_m$ in all of the $\delta_{mj}$ (and possibly through dependence introduced if the noise terms $\varepsilon_{jk}$ are not mutually independent).

As mentioned in Section 12.1, multiple comparisons generally involve two broad steps. If the Tukey–Kramer approach of Section 12.2 is used for general determination of nonequivalence of the $\boldsymbol{\theta}_i$, the many-to-one approaches of this section may broadly be considered as addressing the isolation problem of the second step. Even within this isolation problem, however, there is a similar pattern of steps: (i) a test simultaneously determining if the suspected $\boldsymbol{\theta}_m$ is significantly different from the other points $\boldsymbol{\theta}_j, j \neq m$, and (ii) refined analysis on a subset of points, as necessary, to resolve ambiguities that may remain after the full simultaneous test in (i). The first step relies on the construction of a multiple comparison test for simultaneously testing the null hypothesis that $L(\boldsymbol{\theta}_m) \geq L(\boldsymbol{\theta}_j)$

for all $j \neq m$. In other words, the null hypothesis is that all $\boldsymbol{\theta}_j$ produce loss values at least as low as $\boldsymbol{\theta}_m$. The aim in showing the superiority of $\boldsymbol{\theta}_m$ is to *reject* this null hypothesis.

There are, however, many ways in which the null hypothesis that $\boldsymbol{\theta}_m$ is no better than $\boldsymbol{\theta}_j$ may not be true. Only one of the ways corresponds to $\boldsymbol{\theta}_m$ being the best $\boldsymbol{\theta}$ value. That is the role of the second part of the test: to attempt to isolate the cause of the rejected null hypothesis as a means of determining whether the one alternative hypothesis demonstrating the superiority of $\boldsymbol{\theta}_m$, $L(\boldsymbol{\theta}_m) < L(\boldsymbol{\theta}_j)$ for all $j \neq m$, is, statistically, the correct alternative hypothesis. For the first part, suppose that the interest is in constructing a multiple comparison region such that when $L(\boldsymbol{\theta}_m) \geq L(\boldsymbol{\theta}_j)$ for all $j$ (i.e., the conjectured $\boldsymbol{\theta}_m$ is no better than any other point), then the probability is at least $1 - \alpha$ (say) that $\boldsymbol{\delta}_{m|K}$ will lie in this $(K-1)$-dimensional acceptance region.

Let $\overline{\delta}_{mj} < 0$ represent a critical value for the random variable $\delta_{mj}$. This critical value defines the lower limit in the acceptance region for $\delta_{mj}$. That is, $[\overline{\delta}_{mj}, \infty)$ is the acceptance region. Associated with this region is the acceptance event

$$E_{mj} \equiv \left\{\delta_{mj} \geq \overline{\delta}_{mj}\right\}.$$

As in the Tukey–Kramer test, the critical values in a multiple comparison test must be chosen more conservatively (more negative here) than in a conventional one-comparison test to guard against the greater chance of an extreme event. If the events $E_{mj}$ are true for all $j \neq m$, then there is insufficient evidence to reject the null hypothesis that $L(\boldsymbol{\theta}_j) \leq L(\boldsymbol{\theta}_m)$ for all $j \neq m$. That is, there is insufficient evidence that $\boldsymbol{\theta}_m$ is a possible optimal point at the confidence level $1 - \alpha$. The fact that $\overline{\delta}_{mj}$ is *strictly* negative suggests that the evidence for rejecting the null hypothesis is not sufficient until $\overline{L}_m$ is "considerably" less than $\overline{L}_j$.

The test philosophy outlined above is in the spirit of the *union–intersection principle* of multivariate testing (e.g., Mardia et al., 1979, pp. 127–131). In this principle, the null hypothesis is composed of the *intersection* of events, as in

$$E_{m|K} \equiv E_{m1} \cap E_{m2} \cap \ldots \cap E_{m,m-1} \cap E_{m,m+1} \cap \ldots \cap E_{mK}.$$

For the null hypothesis to be true, all of the events $E_{mj}$, $j = 1, 2, \ldots, m-1, m+1, \ldots, K$, must be true. Hence, the null hypothesis is rejected if any one of the events $E_{mj}$ is not true. This leads to a rejection region (corresponding to the alternative hypothesis) that is the *union* of the events $E_{mj}^c = \{\delta_{mj} < \overline{\delta}_{mj}\}$ (i.e., the rejection region is $\bigcup_{j \neq m} E_{mj}^c$, where the superscript $c$ denotes set complement).

An essential aspect of this union–intersection principle is that when the null hypothesis is rejected, one can try to determine the cause of the rejection by determining which of the events $E_{mj}^c$ occurred. This contrasts with some other

well-known hypothesis testing approaches (e.g., the likelihood ratio test, as in Mardia et al., 1979, pp. 123–127; not to be confused with the likelihood ratio gradient estimation method of Chapter 15). In these other approaches, the rejection of the null hypothesis provides little or no guidance about the cause of the rejection. In this section, we are especially interested in the cause of the rejection since one particular combination of events, $\bigcap_{j \neq m} E_{mj}^c \subseteq \bigcup_{j \neq m} E_{mj}^c$, corresponds to the outcome most consistent with the hypothesis of principal interest: that $\boldsymbol{\theta}_m$ is the best value of $\boldsymbol{\theta}$. The union–intersection principle is ideal for this because we can simply determine whether each of the events $E_{mj}$ occurred or did not occur. A further advantage of the union–intersection approach is that there is no need to know the full (joint) distribution of the $K-1$ elements in $\boldsymbol{\delta}_{m|K}$ forming the basis of the test. Through the use of probability inequalities discussed below, less than complete distributional information is sufficient for conducting a test.

The following joint probability is related to conducting a hypothesis test of $\boldsymbol{\theta}_m$ possibly being the optimal $\boldsymbol{\theta}$ value:

$$P(E_{m|K}) = P(E_{m1} \cap E_{m2} \cap \ldots \cap E_{m,m-1} \cap E_{m,m+1} \cap \ldots \cap E_{mK}). \quad (12.4)$$

From (12.4) it is possible to determine the above-mentioned critical values $\overline{\delta}_{mj}$. In an ideal case where the *joint* distribution of the elements in $\boldsymbol{\delta}_{m|K}$ is known to be normal and the noises are independent, one can, in principle, compute the probability and determine the critical values using the fact that

$$\text{cov}(\delta_{mi}, \delta_{mj}) = \text{var}(\overline{L}_m) \text{ for all } i \neq j.$$

(See Exercise 12.4; see also Appendix C for a discussion of joint normality. Recall that the elements of a random vector $\boldsymbol{X}$ are jointly normally distributed if, and only if, $\boldsymbol{a}^T\boldsymbol{X}$ has a univariate normal distribution for all conformable nonrandom vectors $\boldsymbol{a} \neq \boldsymbol{0}$. Each element of $\boldsymbol{X}$ having a normal distribution is not alone sufficient for joint normality.)

Because of the dependence among the elements in $\boldsymbol{\delta}_{m|K}$, no table lookup is available for readily determining the critical values. That is, since $P(E_{m|K})$ is not a product of the $P(E_{mj})$, $j \neq m$, it is not possible to simply compute critical values from a normal table for the marginal probabilities $P(E_{mj})$ and multiply the marginal probabilities to obtain the chosen probability $1-\alpha$ (although see the Slepian inequality below). Hence, in practice, a numerical search procedure is typically needed to solve for the $\overline{\delta}_{mj}$. The numerical search would work by setting the expression in (12.4) to the chosen $1-\alpha$ level and solving for the $\overline{\delta}_{mj}$. If $\boldsymbol{\delta}_{m|K}$ is jointly normally distributed, then deterministic search methods may be used to find the $\overline{\delta}_{mj}$.

If the full joint normality assumption above does not apply or if one seeks an analytical solution for analysis purposes, there are two inequalities that

provide partial insight into the probability in (12.4), and, in the course of this, allow for a relatively easy computation of critical values $\overline{\delta}_{mj}$. These are the Bonferroni and Slepian inequalities (Tong, 1980, pp. 143–144 for Bonferroni; pp. 8–12 for Slepian). The Bonferroni and Slepian inequalities are both means of converting this difficult-to-compute joint probability into a much-easier-to-compute expression based only on the marginal probabilities, $P(E_{mj})$, or bounds to these probabilities. These inequalities yield only partial insight because, as inequalities, they provide conservative (lower) bounds to the probability of interest $P(E_{m|K})$, yielding critical values that are more negative than strictly required (i.e., wider-than-necessary acceptance intervals for the $\delta_{mj}$). Hence, rejection of the null hypothesis under these critical values is an even stronger indication of the validity of the alternative hypothesis than would be suggested by the false alarm rate $\alpha$. On the other hand, some rejections may be missed due to the conservatism.

The *Bonferroni inequality* for converting the joint probability $P(E_{m|K})$ into a function of the marginal probabilities is

$$P(E_{m|K}) \geq 1 - \sum_{\substack{j=1 \\ j \neq m}}^{K} [1 - P(E_{mj})]. \tag{12.5}$$

Prior information, if available, should be used to determine $P(E_{mj})$. Ideally, the marginal distribution (typically, normal) of each $\delta_{mj}$ will provide the values of $P(E_{mj})$. If a precise calculation is not available, a one-sided Chebyshev inequality can be used to produce a conservative bound for $P(E_{mj})$ using only information about the variance of the $\delta_{mj}$. The Chebyshev bound is $P(X \leq c) \geq 1 - \text{var}(X)/[\text{var}(X) + c^2]$, where $X$ is a mean-zero random variable and $c > 0$ (see Tong, 1980, p. 155 or Subsection 8.2.2 here). This bound yields the following lower bound to the probability of interest:

$$P(E_{mj}) = P(\delta_{mj} \geq \overline{\delta}_{mj}) = P(-\delta_{mj} \leq -\overline{\delta}_{mj}) \geq 1 - \frac{\text{var}(\delta_{mj})}{\text{var}(\delta_{mj}) + \overline{\delta}_{mj}^2}. \tag{12.6}$$

Note that the second equality in (12.6) is employed because the Chebyshev bound pertains to probabilities of the form $P(X \leq c)$, where $c$, like $-\overline{\delta}_{mj}$, is *positive*. In the absence of an exact value for $P(E_{mj})$, we seek a *lower* bound to $P(E_{mj})$ to retain the validity of the lower bound in (12.5). Hence, the Chebyshev bound of (12.6) is in the correct direction to maintain the validity of the lower bound in (12.5).

The *Slepian inequality* applies when the elements of $\boldsymbol{\delta}_{m|K}$ are *jointly* normally distributed with mean zero:

$$P(E_{m|K}) \geq \prod_{\substack{j=1 \\ j \neq m}}^{K} P(E_{mj}). \tag{12.7}$$

(Recall the comment above that it is necessary—but *not sufficient*—for joint normality that the marginal distributions of all components be normal.) The mean-zero condition holds under the above-mentioned assumption that the noise terms $\varepsilon_{km}$ and $\varepsilon_{kj}$ have a common mean since this common mean will subtract out in forming the $\delta_{mj}$. The joint normality for $\boldsymbol{\delta}_{m|K}$ holds if the noise terms are jointly normally distributed. More generally, it approximately applies in other cases, such as when the noise terms in all $K$ loss measurements are mutually independent (not necessarily normal) and satisfy the modest conditions for one of the central limit theorems (e.g., the variances are uniformly bounded above and uniformly bounded away from zero over $k = 1, 2, \ldots, \infty$; see Exercise 12.5). So, the Slepian inequality applies (at least approximately) more broadly than the narrow case where the noise terms are jointly normally distributed. Note, however, that when $\alpha$ is very small, the Slepian bound offers only a negligible improvement over the Bonferroni bound (see Exercise 12.7).

To determine a bound for the probability $P(E_{m|K})$, one can fix $\alpha$ and then use (12.5) or (12.7) as appropriate to determine the critical values $\overline{\delta}_{mj}$. Note that when the noise terms $\varepsilon_{kj}$ are i.i.d., the critical values may be chosen identically (i.e., $\overline{\delta} \equiv \overline{\delta}_{m1} = \overline{\delta}_{m2} = \ldots$) as a way of forcing equality of either the marginal probabilities $P(E_{mj})$ or the Chebyshev bounds to the marginal probabilities according to (12.6). In other cases, it may be necessary to vary the critical values to preserve equality of the marginal probabilities $P(E_{mj})$ (or their bounds). For all $j \neq m$ and a level $\alpha$ test, (12.5) implies that $P(E_{mj}) = 1 - \alpha/(K-1)$ in the Bonferroni case and (12.7) implies that $P(E_{mj}) = (1-\alpha)^{1/(K-1)}$ in the Slepian case. After collecting the sample means $\overline{L}_1, \overline{L}_2, \ldots, \overline{L}_K$, the vector of differences $\boldsymbol{\delta}_{m|K}$ is formed. We then determine if the null hypothesis is rejected by determining if $\boldsymbol{\delta}_{m|K}$ lies in the acceptance region $E_{m|K}$. If the vector does not lie in $E_{m|K}$, the null hypothesis is rejected. If the vector lies in $E_{m|K}$, there is insufficient evidence to consider $\boldsymbol{\theta}_m$ as a candidate optimal point.

Alternatively, if the prior information suggesting the optimality of $\boldsymbol{\theta}_m$ is strong enough, we may increase the sample sizes ($n_j, j = 1, 2, \ldots, K$) and repeat the experiment. One should be aware, however, that repeating the experiment increases the effective false alarm rate from that implied by $\alpha$ since by random chance we are sure to reject the null hypothesis if the experiment is repeated often enough. If the null hypothesis has been rejected and $\boldsymbol{\delta}_{m|K} \in \bigcap_{j \neq m} E_{mj}^{c}$, then we are in the ideal unambiguous situation providing strong evidence that $\boldsymbol{\theta}_m$ is, in fact, the optimal point.

Even more information is available by evaluating the $P$-values (see Appendix B) associated with each $\delta_{mj}$. Very small values provide strong

evidence in support of the hypothesis that $\boldsymbol{\theta}_m$ is the optimal point. More likely than this ideal situation, perhaps, is the ambiguous case where the null hypothesis has been rejected but not all $\delta_{mj}$ lie in their respective rejection regions $E_{mj}^c$. In such cases, one can subject the subset of $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K\}$ corresponding to the nonrejections to more detailed analysis, perhaps involving an increase in sampling (the $n_i$) and a repeat of the experimental approach above.

Unfortunately, even in the ideal situation, $\boldsymbol{\delta}_{m|K} \in \bigcap_{j \neq m} E_{mj}^c$, it is difficult to assign an easy probabilistic interpretation to the rejection. The reason again comes back to the fundamental problem of lack of uniqueness: The "strength" of one's conclusions comes largely from the $P$-value of some test statistic under a unique null distribution. However, if $\bigcap_{j \neq m} E_{mj}^c$ is the only alternative hypothesis of interest, there is no unique complementary null hypothesis under which a distribution for the test statistic can be derived. (This is the dual relationship to the nonuniqueness of alternative hypotheses for a fixed null hypothesis, as discussed in Section 12.1.) Hence, no simple $P$-value is possible. Note also that the Bonferroni inequality applied directly to $P(\boldsymbol{\delta}_{m|K} \in \bigcap_{j \neq m} E_{mj}^c)$ is not helpful in characterizing the $P$-value since it provides a *lower* bound to the probability, which is the wrong direction for an event intended to have a small probability.

One way to cope with the above is to eliminate all competitors to $\boldsymbol{\theta}_m$ except the one that appears to be the strongest challenger (this information should be available based on the marginal $P$-values for each $\delta_{mj}$ or, equivalently, the degree to which each $\delta_{mj}$ lies outside of its acceptance interval $[\overline{\delta}_{mj}, \infty)$). Then, one can perform a final pairwise test between $\boldsymbol{\theta}_m$ and this competitor. With methods such as in Appendix B, a unique null and alternative hypothesis can be determined, leading to a $P$-value. This approach is illustrated in Example 12.7 for the case where the variances are estimated. The equivalent approach would be used here when the variances are assumed known.

Below we present three related examples. Example 12.3 shows how the Bonferroni and Slepian inequalities compare in the calculation of critical values $\overline{\delta}_{ij}$. Examples 12.4 and 12.5 depict testing outcomes of the "unambiguous" and "ambiguous" types. These examples illustrate the relative ease with which the critical values can be obtained via the inequalities.

**Example 12.3—Calculation of critical values.** Suppose that $K = 4$, $n = 15$ (identical sample sizes for the four options), and that, based on prior information, we have reason to believe that $\boldsymbol{\theta}_2$ may be the optimal point. Further, suppose that the measurement noise variance has been reliably estimated as $\text{var}(\varepsilon_{kj}) = 2.0$ for all $k$, $j$, and that the noise terms are mutually independent. The aim is to determine the critical values $\overline{\delta}_{21}$, $\overline{\delta}_{23}$, and $\overline{\delta}_{24}$ for testing the alternative hypothesis that $\boldsymbol{\theta}_2$ is, in fact, the best point. Note that with $n = 15$, the standard

deviation of $\delta_{21}$, $\delta_{23}$, and $\delta_{24}$ is $\sqrt{(2.0+2.0)/15} = 0.516$. Let us compare the critical values and results from a hypothesis test under three different options:

1. The Bonferroni test with Chebyshev inequality-based calculation for $P(E_{ij})$.
2. The Bonferroni test with a normal distribution-based calculation for $P(E_{ij})$.
3. The Slepian inequality test.

These options are listed in increasing order of the prior information required. Given the equal variances for all of $\delta_{21}$, $\delta_{23}$, and $\delta_{24}$, we seek a common critical value $\bar{\delta} \equiv \bar{\delta}_{21} = \bar{\delta}_{23} = \bar{\delta}_{24}$. (In some applications—not this one—the noise terms will depend on $\boldsymbol{\theta}$, suggesting that the critical values should vary along the lines of the discussion above.)

Consider a test at level $\alpha = 0.05$. Under option 1, which makes no assumptions about the forms of the distributions for $\delta_{21}$, $\delta_{23}$, and $\delta_{24}$,

$$P(E_{2|4}) \geq 1 - \sum_{\substack{j=1 \\ j\neq 2}}^{4} \frac{1}{1+\bar{\delta}_{2j}^2/\mathrm{var}(\delta_{2j})} = 1 - \frac{3}{1+\bar{\delta}^2/0.516^2} = 1-0.05. \quad (12.8)$$

From the last equality in (12.8), $\bar{\delta} = -3.96$.

Option 2 is useful when the marginal probabilities $P(E_{2j})$ can be computed (e.g., the marginal distributions of $\delta_{21}$, $\delta_{23}$, and $\delta_{24}$ are normal), but the *joint* distribution of $\delta_{21}$, $\delta_{23}$, and $\delta_{24}$ is not known to be normal. One way this might occur is if the noise terms $\{\varepsilon_{k2}\}$ are independent (with conventional central limit theory implying the approximate normality of $\bar{L}_2$) while the noise terms $\{\varepsilon_{k1}, \varepsilon_{k3}, \varepsilon_{k4}\}$ are *not* mutually independent but are independent of $\{\varepsilon_{k2}\}$. Suppose a central limit theorem for dependent random variables (e.g., Laha and Rohatgi, 1979, p. 355) implies that $\bar{L}_j$, $j \neq 2$, is approximately normal. The independence of $\{\varepsilon_{k2}\}$ and $\{\varepsilon_{kj}\}$, $j \neq 2$, implies that each of $\delta_{21}$, $\delta_{23}$, and $\delta_{24}$ is approximately normal but the dependence of $\{\varepsilon_{k1}, \varepsilon_{k3}, \varepsilon_{k4}\}$ destroys the *joint* normality needed in the Slepian inequality. By the identical marginal distributions of $\delta_{21}$, $\delta_{23}$, and $\delta_{24}$, (12.5) implies that

$$P(E_{2|4}) \geq 1 - 3P\left(Z < \frac{\bar{\delta}}{0.516}\right) \approx 1 - 0.05, \quad (12.9)$$

where $Z \sim N(0, 1)$. From the equality part of (12.9), it is found that $\bar{\delta} = -1.10$.

Option 3 has the strongest requirements, assuming all of the noise terms $\{\varepsilon_{k1}, \varepsilon_{k2}, \varepsilon_{k3}, \varepsilon_{k4}\}$ to be mutually independent, with the noises either being all normally distributed or with the sample size $n = 15$ being large enough so that central limit theorem effects have fully taken hold. Then from the Slepian inequality (12.7),

$$P(E_{2|4}) \geq P\left(Z \geq \frac{\overline{\delta}}{0.516}\right)^3 = 1-0.05, \tag{12.10}$$

where $Z$ is the standard normal variable as above. Solving for $\overline{\delta}$ in the right-hand equality in (12.10) yields $\overline{\delta} = -1.09$. There is little change in $\overline{\delta}$ from the value under option 2; this small change is consistent with the relatively small change in the regularity conditions (see also Exercise 12.7). (See Exercise 12.8 for a numerical computation of $\overline{\delta}$ in the setting of option 3 where the noises are i.i.d. normally distributed.) ❑

**Example 12.4—Results of unambiguous hypothesis test.** Based on the setting in Example 12.3, suppose that $n = 15$ values of the loss function are measured at each of the four candidate points. The sample mean values are $\overline{L}_i = 2.3$, 1.0, 3.7, and 3.2 for $i = 1$, 2, 3, and 4, respectively. This yields $\delta_{21} = -1.3$, $\delta_{23} = -2.7$, and $\delta_{24} = -2.2$. If it is not possible to make assumptions about the distributional form of the random variables $\delta_{21}$, $\delta_{23}$, and $\delta_{24}$, then (12.8) applies. Since $\delta_{21}$, $\delta_{23}$, and $\delta_{24}$ are all less negative than the value $\overline{\delta} = -3.96$, we cannot reject the null hypothesis that $\theta_2$ is indistinguishable from the other three points.

On the other hand, stronger results follow when the above partial or full independence conditions for the noises hold. This allows the use of (12.9) or (12.10) given the further assumptions that the central limit theorem holds or that the noises are jointly normally distributed. Then, it is possible to reject the null hypothesis in favor of the alternative hypothesis that $\theta_2$ is better than at least one other $\theta_j$ since the values of $\delta_{21}$, $\delta_{23}$, and $\delta_{24}$ are all less than $\overline{\delta} = -1.10$. A more precise conclusion than this nebulous alternative hypothesis is available by looking at the marginal $P$-values. That is, each observed $\delta_{2j}$, $j = 1$, 3, and 4, provides information about the strength of the rejection. For instance, based on the normal distribution and the fact that the standard deviation of $\delta_{2j}$ is 0.516 for $j = 1$, we have $P(\delta_{21} \leq -1.3) = P(\delta_{21}/0.516 \leq -1.3/0.516) = P(Z \leq -2.52) = 0.006$. For $j = 3$, 4, the $P$-values are both less than $10^{-5}$. The small marginal $P$-values provide strong evidence that $\theta_2$ is, in fact, the best point. (If desired, a standard pairwise statistical test can be used to compare $\theta_2$ and $\theta_4$ in a final confirmation of this conclusion.) ❑

**Example 12.5—Results of ambiguous hypothesis test.** Continuing with the setting of Examples 12.3 and 12.4, now assume that the data yield $\overline{L}_i = 5.4$, 1.0, 0.8, and 3.2 for $i = 1$, 2, 3, and 4, respectively. Based on the formulation of Example 12.3, we would reject the null hypothesis that $\theta_2$ is indistinguishable from the other three points under any of the distributional assumptions behind (12.8), (12.9), or (12.10) because $\delta_{21} = -4.4$ is more negative than any one of the three values of $\overline{\delta}$. This is even stronger than the conclusion of Example 12.4, where it was not possible to reject the null hypothesis under the weak

assumptions of (12.8). However, since $\delta_{23} = 1.0 - 0.8 > 0$, there is clearly insufficient evidence to claim that $\boldsymbol{\theta}_2$ is optimal, although under the normality assumption of (12.9) and (12.10), the marginal $P$-value of $P(\delta_{23} \leq 0.2) = P(Z \leq 0.39) = 0.65$ is not close enough to 1.0 to suggest that $\boldsymbol{\theta}_3$ is clearly better than $\boldsymbol{\theta}_2$. The other $P$-values, being less than $10^{-5}$, provide strong evidence of $\boldsymbol{\theta}_2$ being better than $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_4$. This suggests that further analysis is required comparing $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$, perhaps by increasing the sample size from the current $n = 15$. ❑

## 12.4 MULTIPLE COMPARISONS AGAINST ONE CANDIDATE WITH UNKNOWN NOISE VARIANCE(S)

The many-to-one approaches of Section 12.3 assume that knowledge of the variances of the $\delta_{mj}$ is available for $j \neq m$. Let us now consider the case where the variances are estimated. The general philosophical and solution strategy issues raised in Section 12.3 do not change in this setting. Consistent with standard practice when variances need to be estimated, the well-known $t$-distribution is used here. As in the Tukey–Kramer method of Section 12.2, the noise distributions must be normal to guarantee the validity of the technique, although, as noted in that section, $t$-distribution-based techniques have a long history of success in practical nonnormal settings. As in the known variance case, the events $E_{mj} \equiv \{\delta_{mj} \geq \overline{\delta}_{mj}\}$ and joint probability in (12.4) are central to determining the acceptance region of a test for the vector of differences $\boldsymbol{\delta}_{m|K}$. The main distinction here is that the values $\overline{\delta}_{mj}$ are *random* because of their dependence on the appropriate variance estimate (versus the exact variance above).

The Bonferroni inequality of (12.5) applies here as well with the marginal probabilities determined from one of the two-sample $t$-tests described in Appendix B. In particular, suppose that the measurement noise sequences $\{\varepsilon_{km}\}$ and $\{\varepsilon_{kj}\}$, $j \neq m$, are mutually independent with $\{\varepsilon_{km}\}$ being identically distributed and $\{\varepsilon_{kj}\}$ being identically distributed; note that $\text{var}(\varepsilon_{km})$ is not necessarily equal to $\text{var}(\varepsilon_{kj})$. The two tests described in (B.3b) and (B.3c) are relevant here, (B.3b) when the measurement noises in the two sequences have identical variances and (B.3c) when they do not. The advantage in the former case is that the estimated variance needed in the $t$-test is an estimate from pooling both data sets together to form a set of size $n_m + n_j$. In the latter case where the variances are not equal, the two required estimates are formed from either the set of size $n_m$ or the set of size $n_j$. Let the individual variance estimates $s_i^2$, $i = 1, 2, \ldots, K$ be computed in the standard way:

$$s_i^2 \equiv \frac{1}{n_i - 1} \sum_{k=1}^{n_i} \left[ y_k(\boldsymbol{\theta}_i) - \overline{L}_i \right]^2 .$$

When the noise sequences have identical variances, then, as in (B.2), the pooled variance estimate for the samples at $\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}_j$, say $s_{mj}^2$, is given by

$$s_{mj}^2 = \frac{n_m - 1}{n_m + n_j - 2} s_m^2 + \frac{n_j - 1}{n_m + n_j - 2} s_j^2, \tag{12.11}$$

which represents a weighted sum of the two constituent sample variances.

In forming the acceptance regions for the tests under the identical or nonidentical variance case, the above formulas are used to construct the acceptance regions $E_{mj}$. In the identical variance case,

$$\bar{\delta}_{mj} = -t_{n_m+n_j-2}^{(\alpha')} s_{mj} \sqrt{n_m^{-1} + n_j^{-1}}, \tag{12.12}$$

where $t_{n_m+n_j-2}^{(\alpha')}$ is the $1-\alpha'$ quantile of the $t$-distribution with degrees of freedom $n_m + n_j - 2$ and probability level $\alpha' = \alpha/(K-1)$. Note that with this probability level, the sum of the terms, $1 - P(E_{mj})$, in the Bonferroni inequality will equal $\alpha$ as required (i.e., $\sum_{j \neq m}[1 - P(E_{mj})] = (K-1)\alpha' = \alpha$). A $P$-value can be computed if desired by appealing to the test statistic in (B.4). In the nonidentical variance case, there is no known exact test statistic, as discussed in Appendix B. However, the formula in (B.3c) can be used to construct an event $E_{mj}$ with associated test statistic that has an approximate $t$-distribution.

Relative to the simultaneous comparison of all $K(K-1)/2$ pairs $\delta_{ij}$ in the Tukey–Kramer test, Figure 12.2 shows the tightening in the acceptance intervals

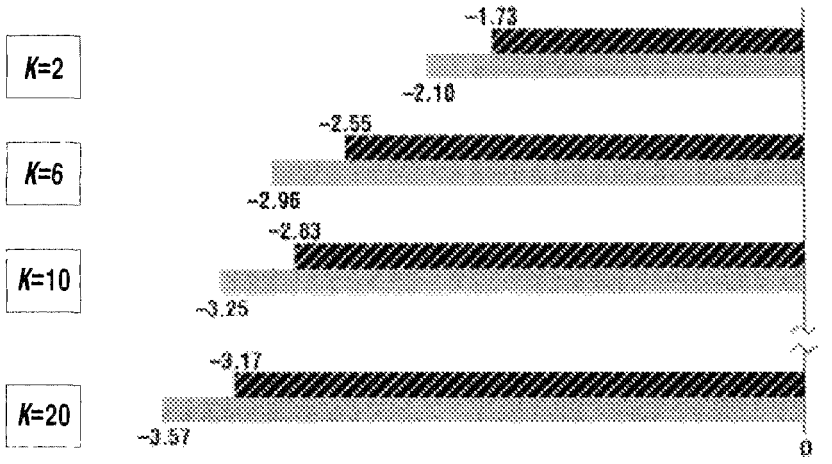


**Figure 12.2.** Relative widths of acceptance intervals (< 0) for Tukey–Kramer and many-to-one tests ($n_1 = n_2 = \ldots = n_K = 10$ for all $K$).

possible by considering only the $K-1$ comparisons in the Bonferroni-based many-to-one approach. The figure shows only the portions of the intervals to the left of zero, which are the relevant parts of the intervals for the many-to-one test since that comparison is based on testing whether $\delta_{mj}$ is sufficiently negative to reject the hypothesis $L(\boldsymbol{\theta}_m) \geq L(\boldsymbol{\theta}_j)$. We consider the same setting as Figure 12.1 ($\alpha = 0.05$, $n_i = 10$ for all $i$, and $s\sqrt{n_i^{-1}+n_j^{-1}} = s_{mj}\sqrt{n_m^{-1}+n_j^{-1}} = 1$ for all $K$ and $i, j, m$, where $s^2$ is the pooled variance estimate from all samples in the Tukey–Kramer test). The four intervals for the Tukey–Kramer test are the same as the intervals in Figure 12.1. The figure illustrates the smaller intervals at each $K$ for the many-to-one approach and the fact that the intervals grow with $K$ at approximately the same rate as the intervals in the Tukey–Kramer approach. Nontabled values associated with the $t$-distribution are calculated via the TINV and TDIST functions in MS EXCEL.

Examples 12.6 and 12.7 illustrate the Bonferroni approach outlined above in the context of a problem where the variances must be estimated.

**Example 12.6—Multiple comparisons against one candidate with estimated variances.** Let us return to the setting of Example 12.2, where it was found that sufficient evidence exists to know at level $\alpha = 0.05$ that all four values in $\Theta$ do not provide the same loss value. Further, the data suggested that $\boldsymbol{\theta}_4$ is possibly better than the other values of $\boldsymbol{\theta}$, but ambiguities remained, suggesting a more refined test is needed. Let us again consider a level $\alpha = 0.05$ test, indicating that the critical values from the $t$-distribution will be chosen such that $P(E_{4j}) = \alpha/(K-1) = 0.05/3 = 0.0167$ for $j = 1, 2$, and 3 since the three probabilities are being combined in the Bonferroni inequality.

Suppose that we collect a *new* data set of size $n_1 = n_2 = n_3 = n_4 = 30$ (recalling the need to have data independent of the data forming the basis for the hypothesis). From these data we find sample means $\bar{L}_i = 2.52$, 3.44, 3.83, and 1.91 and estimated standard deviations $s_i = 0.96$, 0.89, 0.92, and 1.04 for $i = 1$, 2, 3, and 4, respectively. With $m = 4$, from (12.11), we find $s_{4j} = 1.00$, 0.97, and 0.98 and $\delta_{4j} = -0.61$, $-1.53$, and $-1.92$ for $j = 1$, 2, and 3, respectively. Given the identical sample sizes for each of the candidate points, the common critical value from the $t$-distribution at level 0.0167 for use in (12.12) is $t_{58}^{(0.0167)} = 2.18$ (with the degrees of freedom equaling $30 + 30 - 2$). From (12.12), it is found that $\bar{\delta}_{4j} = -0.56$, $-0.55$, and $-0.55$ for $j = 1$, 2, and 3, respectively.

Because at least one value of $\delta_{4j}$ is more negative than its corresponding value of $\bar{\delta}_{4j}$, we can reject the null hypothesis of no difference between $\boldsymbol{\theta}_4$ and the other $\boldsymbol{\theta}_j$. Moreover, since *all* values of $\delta_{4j}$ are more negative than the corresponding values of $\bar{\delta}_{4j}$, we have strong evidence that $\boldsymbol{\theta}_4$ is, in fact, the optimal $\boldsymbol{\theta}$. That is, the vector $\boldsymbol{\delta}_{m|K} = \boldsymbol{\delta}_{4|4} \in \bigcap_{j \neq 4} E_{4j}^c$, which is the ideal unambiguous situation providing strong evidence that $\boldsymbol{\theta}_m = \boldsymbol{\theta}_4$ is the optimal point. ❑

**Example 12.7—Pairwise comparison.** Recall from Section 12.3 that in the absence of additional information, it is difficult to attach a precise statistical interpretation to the "strong evidence" in Example 12.6 since there is no unique null hypothesis (and associated null distribution) complementary to this alternative hypothesis. One way around this problem is to perform an appropriate final pairwise test. Here, the rejection of a null hypothesis provides a clearer statistical interpretation. From the analysis above, it is apparent that if $\theta_4$ is not the optimum, then $\theta_1$ is the most likely choice. Therefore, as a final test, suppose that we collect 30 new measurements at each of $\theta_1$ and $\theta_4$ and apply the standard two-sample test (B.3b). ((B.3b) can be used because the underlying true variances are assumed identical via Examples 12.6 and 12.2.) For these new measurements, it is found that $\bar{L}_1 = 2.50$, $\bar{L}_4 = 2.01$, $s_1 = 0.98$, and $s_4 = 1.10$, leading to a pooled standard deviation from (12.11) of 1.04. The value of the test statistic using (B.4) is $-1.79$, which corresponds to a one-sided $P$-value of 0.039 using a $t$-distribution with $30 + 30 - 2 = 58$ degrees of freedom. This provides evidence that $\theta_4$ is the optimum. ❑

Note that the Slepian bound of Section 12.3 will also sometimes apply in the case where the variances are to be estimated, as discussed in Tong (1980, pp. 37–39) and Miller (1981, pp. 254–255). We do not pursue this $t$-distribution analogue of the Slepian inequality here because the improvement to the Bonferroni bound is slight for the typically small values of $\alpha$ used in practice (see Exercise 12.7). However, Exercise 12.13 provides some conditions under which the Slepian bound applies.

## 12.5 EXTENSIONS TO BONFERRONI INEQUALITY; RANKING AND SELECTION METHODS IN OPTIMIZATION OVER A FINITE SET

The approaches described above represent some of the fundamental methods for determining solutions to $\arg\min_{\theta \in \Theta} L(\theta)$ when there are a finite—and relatively small—number of values in the domain $\Theta$ and only noisy measurements of $L$ are available. (Some of the other approaches in this book apply to discrete problems with a larger—possibly unbounded—number of options. The price to be paid is that one loses the statistical guarantees associated with the statistical comparisons methods.) As in the discussion above, we focus on the case where $\theta^*$ is unique. This section is a brief discussion of some other methods related to multiple comparisons, including improvements to the Bonferroni inequality, an indifference zone method for ensuring with high probability that a good (if not best) solution is obtained, and a method for culling the original pool of $K$ candidates to narrow the search to a smaller number of options. These latter two topics are often labeled *ranking and selection* methods and are sometimes treated as distinct entities from the multiple comparisons methods that have dominated this chapter (Fu, 1994, e.g., discusses this dichotomy).

The methods for multiple comparisons against one candidate in Sections 12.3 and 12.4 rest principally on the Bonferroni inequality of probability theory with small improvement possible via the Slepian inequality if its conditions hold. Although the Bonferroni inequality is simple and has been known from the beginnings of formal probability theory, it generally provides a surprisingly tight bound to the joint probability of interest in applications to the types of multiple comparisons considered here. In fact, Miller (1981, p. 254) states: "I knew the Bonferroni inequality was very useful, but over the course of the past ten years, I have become even more impressed with the tightness of the bound....Although special techniques and distribution theory can improve on it, the improvement is very often only minor." Note that this sentiment on the power of the Bonferroni inequality is focused on the multiple comparisons framework here. In *other* applications, the Bonferroni inequality may perform poorly, especially when the underlying random variables are strongly dependent (e.g., Naiman and Priebe, 2001, for problems in genetics and medical image analysis; Hill and Spall, 2000, for a problem in multiple component reliability).

In some critical applications, even slight improvements to the Bonferroni inequality are welcome. Naiman and Wynn (1992) pursue enhancements to the Bonferroni inequality via the introduction of *joint* probability information, à la $P(E_{mi} \cap E_{mj})$, $P(E_{mi} \cap E_{mj} \cap E_{mk})$, and so on. The authors pursue both geometric and combinatoric approaches, with the geometric approach providing interesting insight and computer-implementable algorithms. Monte Carlo sampling methods are used to compute the required joint probabilities, which are demonstrated in the determination of critical values (analogous to $\bar{\delta}_{mj}$ here). Nakayama (1997) exploits structure in multiple comparisons problems where Monte Carlo simulations are being used to generate the noisy loss measurements. Using time series methods, the bounds produced in Nakayama (1997) are guaranteed to have a false alarm rate no higher than $\alpha$ as the simulation run length approaches infinity and to be more precise than the Bonferroni bounds. Hill and Spall (2000) provide a summary of special cases of improvements to the Bonferroni inequality that are based on only pairwise (not higher-order) joint probabilities, such as $P(E_{mi} \cap E_{mj})$ here.

Goldsman and Nelson (1998) provide a summary of several procedures based on the notion of an *indifference zone*. Section 14.5 discusses the use of indifference zone methods when Monte Carlo simulations are used to generate the loss measurements $y_k(\boldsymbol{\theta}_i)$. Based on some results in Rinott (1978), indifference zone approaches are guaranteed with probability at least $1 - \alpha$ to provide a solution equal to $\boldsymbol{\theta}^*$ when the true loss value at $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}^*$ is at least $\delta$ units greater than $L(\boldsymbol{\theta}^*)$. The term *indifference* arises because the user is willing to tolerate any solution $\boldsymbol{\theta}_i$ such that the loss $L(\boldsymbol{\theta}_i)$ is in the range $[L(\boldsymbol{\theta}^*), L(\boldsymbol{\theta}^*) + \delta)$, sometimes called the indifference zone.

The basic steps in indifference zone procedures are as follows: (i) Collect a first stage of data at each element of $\Theta$ and calculate sample means and variances; (ii) use the specified value of $\delta$, the sample variances calculated at the

first stage, and an appropriate distributional table to determine the number of measurements needed at each $\boldsymbol{\theta}_i$ in the second stage (the smaller the value of $\delta$, the larger the required sample sizes); (iii) combine the data collected in the two stages to form sample means at each element of $\Theta$; and (iv) take as an estimate of $\boldsymbol{\theta}^*$ the $\boldsymbol{\theta}_i$ producing the lowest sample mean in step (iii).

In a university–industry collaboration, Goldsman et al. (1999) illustrate the application of indifference zone methods in three nontrivial applications: inventory control, queuing systems, and project management. The indifference zone techniques may sometimes be used to circumvent the general difficulty discussed in Section 12.1 of giving a precise statistical interpretation (e.g., false alarm rate) for the alternative hypothesis of one, and only one, of the $\boldsymbol{\theta}_i$ being optimal. In particular, *if* it is known a priori that the unique unknown $\boldsymbol{\theta}^*$ yields a loss value at least $\delta$ units better than the other $\boldsymbol{\theta}_i$, (i.e., $L(\boldsymbol{\theta}^*) \leq L(\boldsymbol{\theta}_i) - \delta$ for $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}^*$), then the indifference zone techniques provide a "clean," quantifiable error rate. In particular, the unknown optimal point will be correctly identified with probability of at least $1 - \alpha$.

Although the indifference zone methods have the advantage of quantifiable error rate and the advantage of not requiring the noise variances to be known (or even to be equal at the different $\boldsymbol{\theta}$ values), they are based on the assumption of normally distributed data and they require a potentially large number of measurements (proportional to $\delta^{-2}$) at *all* of the $K$ values of $\boldsymbol{\theta}$. (As discussed in Section 14.5, the use of common random numbers in simulation can help mitigate this potentially large number of measurements.) Further, in general, the user must select the size of the indifference zone via picking $\delta$.

Related to the above methods for selecting an optimal solution are methods for screening the $K$ candidates to find a subset likely to contain the optimum. These are sometimes called *ranking and selection* methods. Gupta (1965) describes such a procedure when the sample sizes satisfy $n_1 = n_2 = \ldots = n_K = n$ and the measurement noises are i.i.d. normal across all measurements (see also Goldsman and Nelson, 1998; Nelson et al., 2001). This is a stronger set of assumptions than the Tukey–Kramer procedure, where the sample sizes and measurement noise distributions need not be identical. Under these assumptions, this screening procedure is guaranteed with probability at least $1 - \alpha$ to produce a subset that contains $\boldsymbol{\theta}^*$. The subset is chosen to include all points $\boldsymbol{\theta}_i$ satisfying

$$\bar{L}_i \leq \min_{1 \leq j \leq K} \bar{L}_j + s\, t^{(\alpha)}_{K-1,K(n-1)} \sqrt{\frac{2}{n}}\,, \tag{12.13}$$

where $s$ is as in (12.2) (simplified here due to the identical $n_i$) and $t^{(\alpha)}_{K-1,K(n-1)}$ is a critical value from a multivariate $t$-distribution tabled in Goldsman and Nelson (1998, Table 8.2) and Hochberg and Tamhane (1987, App. 3, Table 4). The following example illustrates (12.13).

**Example 12.8—Subset selection.** Consider an expensive-to-run system where only $n = 6$ measurements are available at each of $K = 5$ options, and where it is believed that the noises are i.i.d. normal. The following data are obtained: $\bar{L}_i$ = 70.6, 76.1, 84.9, 63.0, and 79.5 for $i$ = 1, 2, 3, 4, and 5, respectively, with $s$ = 11.5 using (12.2). We seek a subset of the five options that is guaranteed to contain $\boldsymbol{\theta}^*$ with probability 0.95. From the above-mentioned table in Goldsman and Nelson (1998), $t_{4,25}^{(0.05)} = 2.27$. Therefore, the procedure selects those systems satisfying

$$\bar{L}_i \leq 63.0 + 11.5 \times 2.27 \sqrt{\frac{2}{6}} = 78.1 .$$

Hence, with a probability of at least 0.95, it is known that $\boldsymbol{\theta}^*$ corresponds to one of options 1, 2, and 4. At this point, the test of multiple comparisons against one candidate in Section 12.4 might be used to isolate the optimum from the three candidates. Alternatively, one could carry out additional subset selection from options 1, 2, and 4 using a *new* set of data. This represents a recursive application of subset selection. ❑

## 12.6 CONCLUDING REMARKS

This chapter presented methods for statistically characterizing the relative quality of the multiple options in a discrete set $\Theta = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K\}$. The approaches here differ in character from most of those in the rest of this book. Most of the other methods in this book are based on iteratively updating $\boldsymbol{\theta}$ from an initial guess (or *set* of guesses in the population-based methods of Chapters 9 and 10). In contrast, the statistical approaches here are based on collecting loss measurements at all candidate $\boldsymbol{\theta}$ values and comparing in some formal manner the resulting sample means at each $\boldsymbol{\theta}$. In the other iterative search and optimization methods, all possible values of $\boldsymbol{\theta}$ are typically *not* evaluated en route to finding an optimum. In fact, the very concept is nonsensical when considering continuous search spaces. (In the other search methods, it is not necessary to evaluate all possible $\boldsymbol{\theta}$ values to ensure convergence because of assumptions made about the form of the loss $L = L(\boldsymbol{\theta})$.)

A type of recursive implementation is possible with the statistical methods here. Namely, the comparisons methods may first be used to screen all $K$ options to determine if there is evidence that all options are not of equal value in terms of the loss function (à la the Tukey–Kramer test of Section 12.2). Given evidence that at least some options are better than others, the options can be pared via the selection and screening methods of Sections 12.3–12.5. These paring methods require stronger assumptions, such as prior knowledge of likely optimal $\boldsymbol{\theta}_i$. The above approach of testing and paring can be repeated after reducing the number of options.

Most of the methods here are relatively mature. Commercial software is available that implements most of the techniques (see, e.g., Goldsman et al., 1999). While the methods are presented here in a generic fashion, problem structure can sometimes enhance performance. As discussed in Section 14.5, for instance, common random numbers in simulations can reduce the sample sizes required to attain guaranteed levels of statistical performance when simulation runs provide the measurements of $L$.

The goals of the statistical methods in this chapter are connected to the goals of certain methods in the field of sequential analysis, particularly to the multiarm bandit problem (so named by analogies to slot machines). In such methods, one adaptively determines how the sampling should be done with the aim of optimizing some criterion. As above, the sampling may be assumed to come from a population of $K$ possible generating mechanisms. So, if (as above) one's aim is to identify the $\boldsymbol{\theta}_i$ that minimizes $L$, the sequential sampling methods provide rules governing how many samples should be taken at each $\boldsymbol{\theta}_i$. If evidence is acquired early in the sampling process that certain values of $\boldsymbol{\theta}$ are clearly inferior, then sampling can be focused on the remaining $\boldsymbol{\theta}$ that are legitimate candidates to be $\boldsymbol{\theta}^*$. This contrasts with the comparisons methods in this chapter, where the sampling strategy is generally specified *prior* to collecting any information about the competing alternatives. Such sequential sampling may increase the efficiency in the use of resources.

Lai (2001) provides a *tour de force* of sequential methods, including the multiarmed bandit problem. He notes that one of the major areas of application for such methods is stochastic adaptive control based on Markov chains. There are also close connections to stochastic approximation of the type introduced in Chapter 4 and to experimental design as discussed in Chapter 17.

## EXERCISES

**12.1** Prove that if $\Theta$ contains $K$ elements, there are $K(K-1)/2$ differences $L(\boldsymbol{\theta}_i) - L(\boldsymbol{\theta}_j)$ to be tested in the absence of any prior information eliminating some of the elements in $\Theta$.

**12.2** Suppose that the noise conditions for the Tukey–Kramer method of multiple comparisons apply and $K = 5$. Suppose that $\bar{L}_i = 71, 86, 72, 63$, and 70 for $i = 1, 2, 3, 4$, and 5, respectively. Let $n_1 = n_2 = n_3 = 10$ and $n_4 = n_5 = 6$. Further, suppose that $s = 9.8$ using (12.2).

**(a)** Using the level 0.05 quantile value $Q_{5,37}^{(0.05)} = 4.06$, discuss why the null hypothesis of equality of loss values is rejected at some level *less* than $\alpha = 0.05$.

**(b)** Demonstrate that $\boldsymbol{\theta}_2$ can be effectively ruled out as a candidate for $\boldsymbol{\theta}^*$.

**12.3** For $K = 2$ with $n_1 = 12$ and $n_2 = 8$, suppose that $s^2 = 1.0$. For $\alpha = 0.05$, verify that the Tukey–Kramer interval computed via (12.3) is the same as the

classical two-sample $t$-test interval in (B.3b) in Appendix B. (Note: $Q_{2,18}^{(0.05)} = 2.971$ from Miller, 1981, p. 234.)

**12.4** In addition to having a common mean, assume that all noise terms are mutually uncorrelated. Show that $\text{cov}(\delta_{mi}, \delta_{mj}) = \text{var}(\bar{L}_m)$ for $i \neq m$, $j \neq m$, and $i \neq j$.

**12.5** In the setting of Section 12.3, suppose that the noises $\{\varepsilon_{11}, \varepsilon_{21}, \ldots, \varepsilon_{n1}, \varepsilon_{12}, \ldots, \varepsilon_{nK}\}$ are mutually independent and that a central limit theorem shows that the normalized sample means, $\sqrt{n}[\bar{L}(\boldsymbol{\theta}_i) - E(\bar{L}(\boldsymbol{\theta}_i))]$, have a limiting normal distribution for each $i = 1, 2, \ldots, K$. Given that the common sample size $n$ is large, establish that $\boldsymbol{\delta}_{m|K} = [\delta_{m1}, \delta_{m2}, \ldots, \delta_{m,m-1}, \delta_{m,m+1}, \ldots, \delta_{mK}]^T$ is approximately *jointly* normally distributed.

**12.6** (More difficult.) The basic Slepian inequality states: If $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{R} = [r_{ij}]$ and $\boldsymbol{S} = [s_{ij}]$ are two correlation matrices with $r_{ij} \geq s_{ij}$ for all $i, j$, then $P_{\boldsymbol{\Sigma}=\boldsymbol{R}}\left(\bigcap_j \{X_j \leq a_j\}\right) \geq P_{\boldsymbol{\Sigma}=\boldsymbol{S}}\left(\bigcap_j \{X_j \leq a_j\}\right)$ holds for all $\{a_j\}$, where the subscript on the probability indicates whether the probability is computed under the $N(\boldsymbol{0}, \boldsymbol{R})$ or $N(\boldsymbol{0}, \boldsymbol{S})$ distribution. (Note: A correlation matrix is a covariance matrix scaled such that all diagonal elements are equal to 1 and each off-diagonal element is the correlation between a given pair of random variables.) Show that this inequality implies (12.7) in the nondegenerate case where $\text{cov}(\boldsymbol{\delta}_{m|K}) > \boldsymbol{0}$ (i.e., is positive definite).

**12.7** Suppose that conditions for the Slepian probability bound apply and that $P(E_{mj}) = \beta$ for all $j \neq m$. Show that:

**(a)** The Slepian probability bound is greater than the Bonferroni bound.

**(b)** The two bounds will be close to each other when $\beta \approx 1$.

**12.8** In the setting of Example 12.3, assume that $\{\varepsilon_{k1}, \varepsilon_{k2}, \varepsilon_{k3}, \varepsilon_{k4}\}$ are jointly normally distributed. Calculate the value of $\bar{\delta}$ via one of the direct search methods in Chapter 2 or any other method such as a deterministic search technique. Contrast this value with the three values based on probability inequalities as given in Example 12.3. (Hint: Recall that $\text{cov}(\delta_{mi}, \delta_{mj}) = \text{var}(\bar{L}_m)$ for $i \neq j$. This problem requires the evaluation of a trivariate integral, which can be done numerically via deterministic or Monte Carlo means.)

**12.9** Suppose that measurements $y_k(\boldsymbol{\theta}_i) = L(\boldsymbol{\theta}_i) + \varepsilon_{ki}$, $k = 1, 2, \ldots, n$, are available with the $\varepsilon_{ki}$ being i.i.d. with mean 0 and variance 1 for all $i, k$. Let $n = 10$, $K = 6$, and the false alarm rate be $\alpha = 0.10$. Determine a common critical value $\bar{\delta}$ for using the many-to-one comparisons test in Section 12.3 to help determine whether the $m$th element of $\Theta$ is the optimal point at the chosen $\alpha$ in the following three manners:

**(a)** Use the distribution-free approach based on the Chebyshev and Bonferroni inequalities.

**(b)** Use the Bonferroni inequality together with exact knowledge of $P(E_{ij})$ when it is assumed that the noises are normally distributed.

**(c)** Use the approach based on the Slepian inequality under the same normal distribution assumption.

**12.10** Consider the setting of Exercise 12.9 with $\boldsymbol{\theta} = [t_1, t_2]^T$, $L(\boldsymbol{\theta}) = |t_1 - t_2|$, and domain $\Theta = \{-1, 0, 1\} \times \{1, 2\}$. Using a normal distribution for the noise, generate values of $\bar{L}(\boldsymbol{\theta}_j)$ by Monte Carlo for each of the six possible values of $\boldsymbol{\theta}_j$. Use $n = 10$. Suppose that one correctly picks $\boldsymbol{\theta}_m = \boldsymbol{\theta}^*$. Based on the three critical values computed in Exercise 12.9, test the null hypothesis that $\boldsymbol{\theta}_m$ is no better than any other $\boldsymbol{\theta}_j$. Report $P$-values for the $\delta_{mj}$ generated under the (correct) assumption of normally distributed noise. Comment on whether these tests allow one to statistically conclude that $\boldsymbol{\theta}_m$ is the optimal point.

**12.11** In the setting of Example 12.2, suppose that one had prior knowledge that $\boldsymbol{\theta}_4$ may provide the lowest loss value. Using the data of Example 12.2, carry out an analysis based on multiple comparisons against one candidate when the variances must be estimated. Use $\alpha = 0.10$.

**12.12** (More difficult.) A unbiased variance estimator different from (12.2), which sometimes appears in the literature, is

$$\tilde{s}^2 \equiv \frac{1}{K}\sum_{i=1}^{K}\frac{1}{n_i - 1}\sum_{k=1}^{n_i}[y_k(\boldsymbol{\theta}_i) - \bar{L}_i]^2 .$$

This estimate has a natural interpretation as an average variance because each of the summands in $\sum_{i=1}^{K}(\cdot)$ is the sample variance for the sample of size $n_i$. Show that $s^2$ in (12.2) is a better variance estimator in the sense that $\text{var}(s^2) \leq \text{var}(\tilde{s}^2)$.

**12.13** When the variances must be estimated (versus being known), Tong (1980, pp. 37–38) gives conditions such that the $t$-distribution analogue of the Slepian inequality for a random vector $\boldsymbol{X}$ will hold. Two of the conditions are: (i) $\boldsymbol{X}$ is distributed according to a $N(\boldsymbol{0}, \sigma^2\boldsymbol{R})$ distribution where $\boldsymbol{R}$ is a correlation matrix (see definition in Exercise 12.6), and (ii) there exists an $s$ constructed from the components of $\boldsymbol{X}$ such that $Ms^2/\sigma^2$ has a chi-squared distribution with $M$ degrees of freedom for some $M$. (A special case of the Slepian-like result most useful to us is $P(\bigcap_i \{T_i \geq a_i\}) \geq \prod_i P(T_i \geq a_i)$, where $T_i = X_i/s$, $X_i$ is the $i$th component of the vector $X$, and the $a_i$ are constants.) Prove that conditions (i) and (ii) hold in the special case where $\boldsymbol{X} = \boldsymbol{\delta}_{m|K}$ has a multivariate normal distribution with $\{\bar{L}_1, \bar{L}_2, \ldots, \bar{L}_K\}$ being i.i.d. and having distribution $\bar{L}_i \sim N(0, \sigma^2/2)$.

**12.14** Consider a system where 10 measurements are available at each of $K = 6$ options and where the noises are i.i.d. normal. The following data are obtained: $\bar{L}_i$ = 2.3, 1.1, 0.8, 1.9, 2.0, and 1.7 for $i$ = 1, 2, 3, 4, 5, and 6, respectively. Further, $s^2 = 1.6$ using (12.2). Identify the subset of options that is guaranteed to contain $\boldsymbol{\theta}^*$ with probability 0.95. (Note: $t_{5,54}^{(0.05)} = 2.29$ from Table 8.2 in Goldsman and Nelson, 1998.)