5

Testing Several Contrasts
The Multiple-Comparisons Problem

OVERVIEW OF CHAPTER: RESEARCH QUESTIONS ADDRESSED

In Chapter 4, you learned how to test individual comparisons among means and assess their practical significance through such procedures as confidence intervals and effect size measures. You were also introduced to the concepts of linear independence and orthogonality. As these two concepts demonstrate, several contrasts are often tested on the same set of data. Linear independence and orthogonality concern the degree of overlap in information obtained from testing several contrasts among a groups. This chapter considers another issue that arises in testing several contrasts. Namely, any time multiple tests are performed, the probability of committing a Type I error increases for the collection of tests. As a rough analogy, if we toss a coin enough times, the probability we will at some point get heads increases (i.e., we are more likely to get at least one heads in 5 tosses than 2 tosses, and even more likely to get at least one heads in 10 tosses than 5 tosses). This chapter presents methods that control for multiple testing and thus prevent inflation of the Type I error rate. This is important for science because without such control it often becomes difficult to know whether to trust a report of a statistically significant result. Simmons et al. (2011) provide a compelling example of how flexibility in data collection and analysis can all too easily lead to spurious results. Gelman and Loken (2014) provide a very readable introduction to the larger problem that arises when researchers make many seemingly arbitrary choices during the process of data analysis. They state, "Fisher offered the idea of p-values as a means of protecting researchers from declaring truth based on patterns in noise. In an ironic twist, p-values are now often manipulated to lend credence to noisy claims based on small samples" (2014, p. 460). It is especially important to realize that failing to control for multiple testing may play a major role in contributing to a disappointing failure rate in attempts to replicate published studies in psychology. We will approach this issue primarily from the perspective of hypothesis testing, but we will see that the issue is also relevant for confidence intervals. Here are some examples of types of questions you will learn how to answer in Chapter 5:

- How can the Type I error rate be controlled at .05 if a researcher plans to test more than one contrast?
- How can the Type I error rate be controlled at .05 if a researcher decides to test all possible pairwise comparisons (i.e., comparing the mean of each group with the mean of every other group)?

- How can the Type I error rate be controlled if a researcher wants to be able to consider all possible ways groups might differ from one another, including not only all pairwise comparisons but also complex comparisons?
- How can confidence intervals be formed so that the level of confidence pertains not to a single interval but instead to a collection of intervals?
- Are there tradeoffs between planning a small number of contrasts versus letting the data drive the choice of which groups to compare?

PUBLISHED EXAMPLE

Consider the Kroes et al. (2014) electroconvulsive therapy (ECT) study we introduced in Chapter 4. You may recall that their study involved three groups of participants. In particular, participants were randomly assigned to one of three groups that differed in whether they received ECT (groups A and B) or not (group C), and whether they were tested on the stories immediately (group B) or after a 24-hour delay (groups A and C). Thus, the three groups were:

Group A: ECT, delayed test
Group B: ECT, immediate test
Group C: No ECT, delayed test

Multiple tests are of interest here. For example, comparing group A to group B assesses the effect of timing for individuals who received ECT. On the other hand, comparing group A to group C assesses the effect of ECT for those whose testing was delayed. This interest in multiple research questions leads us to consider how to perform multiple tests and yet control the Type I error rate at the desired level.

MULTIPLE COMPARISONS

Experimentwise and Per-Comparison Error Rates

We begin by considering the example from the beginning of the previous chapter, where there are four treatments for hypertension to be compared. Suppose it was decided to test the following three contrasts:

$$\psi_1 = \mu_1 - \mu_2$$
$$\psi_2 = \tfrac{1}{2}(\mu_1 + \mu_2) - \mu_3$$
$$\psi_3 = \tfrac{1}{3}(\mu_1 + \mu_2 + \mu_3) - \mu_4$$

Assuming equal n, these three contrasts form an orthogonal set, as we verified near the end of Chapter 4. Suppose that each of these contrasts is tested using an alpha level of .05. If the four treatments are in fact equally effective, how likely are we to obtain at least one statistically significant result in our study? In other words, how probable is it that we will make at least one Type I error? The answer is obviously a number greater than .05 because we are performing three different tests at the .05 level. At first glance, the answer might seem to be .05 × 3, or .15. Although .15 is a number we return to momentarily, it is not the answer to this question. Recall from probability theory that probabilities of events cannot be summed unless the events are mutually

exclusive, that is, unless the occurrence of one event rules out the occurrence of another. This is not the case here, because if the H_0 for ψ_1 is mistakenly rejected, the hypotheses for ψ_2 and ψ_3 might or might not be rejected. It turns out that for orthogonal contrasts, the binomial formula provides an answer to our question:

$$\begin{aligned}\Pr(\text{at least one Type I error}) &= 1 - \Pr(\text{no Type I errors}) \\ &= 1-(1-\alpha)^C \end{aligned} \tag{1}$$

where α is the alpha level for a single contrast and C is the number of contrasts tested. For our example, then, $\alpha = .05$ and $C = 3$. Substituting the values for our scenario into Equation 1, we find that the probability of at least one Type I error in our study, with three orthogonal contrasts, is .143.

Before we comment further on this number, several comments on Equation 1 are in order. First, the expression $1 - (1 - \alpha)^C$ is obtained from the binomial formula for the probability of at least one success in C trials, when the probability of a success on a single trial is α. It may be necessary to remind oneself that here a "success" is a Type I error. Second, Equation 1 is only appropriate when the C contrasts to be tested form an orthogonal set, because the binomial requires an assumption that the C trials be statistically independent. This assumption is not met for nonorthogonal contrasts, so Equation 1 is inappropriate unless the contrasts are orthogonal. Third, strictly speaking, Equation 1 holds only for large n, because although the $\hat{\psi}$ values of orthogonal contrasts are uncorrelated, the F tests all use the same denominator term—namely, MS_W, assuming homogeneity of variance. Thus, the F tests are not strictly independent. However, this is a technical point and need not concern us.

Let's return to our value of .143. Remember that this is the probability of committing at least one Type I error if we test three orthogonal contrasts using an alpha level of .05 for each contrast. Is this a problem? After all, it seemed that our alpha level was .05, but now we are saying that our probability of a Type I error is almost three times as large as .05. To clarify this issue, it is helpful to develop some terminology. First, the *error rate per contrast* (α_{PC}) is the probability that a particular contrast will be falsely declared significant. In other words, if a contrast whose true population value is zero were to be tested over and over again in repeated studies, α_{PC} is the proportion of times that the contrast would be found to be statistically significant. Second, the *expected number of errors per experiment* (ENEPE) is the expected number of contrasts that will be falsely declared significant in a single experiment.[1] Notice that ENEPE is not a probability and in fact can exceed one under some circumstances. Third, the *experiment wise error rate* (α_{EW}) is the probability that one or more contrasts will be falsely declared significant in an experiment. In other words, if an experiment were to be conducted repeatedly, α_{EW} is the proportion of those experiments (in the long run) that would contain at least one Type I error. Fourth, in designs with more than one factor, it is necessary to define yet another error rate, called the *familywise error rate* (α_{FW}). As discussed in more detail in Chapter 7, in multifactor designs, significance tests involving different factors are usually regarded as constituting different families. For this reason, a single experiment may contain several families of tests, in which case α_{FW} and α_{EW} are different. However, in single-factor designs, which is all that we have discussed until now, α_{FW} and α_{EW} are identical, so we will wait until Chapter 7 to discuss family wise error rate.

The distinctions among these three types of error rates (i.e., α_{PC}, α_{EW}, and ENEPE) can perhaps best be understood by returning to our example, with four groups and three contrasts to be tested. In this example, α_{PC} is equal to .05, because each comparison was tested at an alpha level of .05. For any single comparison, there is a 5% chance of a Type I error. What is the value of ENEPE? ENEPE will equal .15 because the expected number of Type I errors per contrast is .05 and there are three contrasts tested in the experiment. In general, with C contrasts each tested at

an alpha level of α_{PC}, ENEPE equals $C\alpha_{PC}$. Finally, α_{EW} is the probability of at least one Type I error being made in the experiment. Earlier, we found that this probability equals .143.

That there are three types of error rates and that each has a different numerical value here poses a problem. Even though the value of .05 is somewhat arbitrary, at least it provides an objective standard for making decisions in most disciplines that employ inferential statistics. So, suppose that we can agree that .05 is the standard we wish to use. The problem that immediately confronts us is: which error rate should be .05? In our four-group example, α_{PC} was .05, but ENEPE and α_{EW} exceeded .05. What if we were to have chosen either ENEPE or α_{EW} to be .05? In this case, it turns out that α_{PC} must be less than .05 whenever more than a single contrast is tested in an experiment. Thus, when multiple contrasts are tested, it is impossible to achieve a .05 value for all three types of error. Instead, a decision must be made regarding which type of error is to be controlled at the 5% level.

Although this is an issue about which reasonable people may choose to differ, our preference is to control α_{EW} at .05. The basic argument in favor of this approach is that there must be an explicit control on the number of studies in the literature that contain Type I errors. By keeping α_{EW} at .05, the probability of a Type I error occurring anywhere in a given experiment is at most .05. If instead we were to control α_{PC} at .05, studies with multiple tests would produce Type I errors more than 5% of the time. Many of these studies might then report statistically significant results in the published literature. Even though the error rate for the study as a whole may be inflated, these studies may be accepted for publication unless journals require that α_{EW} instead of α_{PC} be controlled at .05. As Greenwald (1975) pointed out, allowing such studies to be published without controlling α_{EW} is a major problem, because it is difficult in many fields to publish studies claiming to demonstrate the lack of an effect. As a consequence, studies that unbeknownst to us report Type I errors are likely to go uncorrected because subsequent studies of the phenomenon will likely produce non-significant results. Although failing to reject the null hypothesis is obviously appropriate when the null hypothesis is in fact true, there is nevertheless a problem if journals tend not to publish papers claiming to support the null hypothesis. This practice reflects publication bias and historically it has hindered progress toward a cumulative science of psychology. Controlling the alpha level for the study instead of for the individual comparison makes it less likely that studies will report Type I errors in the first place. This is one reason that most methodologists advocate controlling the alpha level for a collection of tests (and in this sense, controlling the alpha level for a study) instead of for each individual comparison. However, in so doing, it does not necessarily follow that 5% of published findings represent Type I errors. In fact, Ioannidis (2005b) has suggested that over half of published findings may be Type I errors even when the α level is set at .05. If, instead, α_{PC} were controlled at .05, studies with multiple contrasts would have a higher Type I-error rate than .05. In this situation, an experimenter could increase his or her chances of obtaining a statistically significant result simply by testing many contrasts. At first glance, it might seem desirable to be able to maximize the opportunity to find a significant result somewhere in the data. However, as Gelman and Loken (2014) along with others have pointed out, failing to adjust the alpha level while conducting multiple tests often has the unfortunate consequence that results deemed to be statistically significant may not be trustworthy. By choosing to set α_{EW} rather than α_{PC} at .05, this problem is avoided. Of course, it might be argued that the structure imposed by a single experiment is rather arbitrary. Miller (1981, pp. 31–32) provides a humorous discussion along these lines:

> Two extremes of behavior are open to anyone involved in statistical inference. A non-multiple comparisonist regards each separate statistical statement as a family, and does not give increased protection to any group of statements through group error rates. At the other extreme is the ultraconservative statistician who has just a single family consisting of

every statistical statement he might make during his lifetime. If all statisticians operated in this latter fashion at the 5 percent level, then 95 percent of the world's statisticians would never falsely reject a null hypothesis, and 5 percent would be guilty of some sin against nullity. There are a few statisticians who would adhere to the first principle, but the author has never met one of the latter variety.

Why do you suppose Miller has never met such an ultraconservative statistician—after all, aren't statisticians stereotypically considered to be rather conservative? Suppose there was such a statistician somewhere. Further suppose he or she figures that the total number of statistical hypotheses he or she might test in a lifetime is 1,000. This set of 1,000 hypotheses then can be thought of as an "experiment" in terms of Type I error. Algebraic manipulation of Equation 1 shows that

$$\alpha_{PC} = 1 - \sqrt[c]{1 - \alpha_{EW}} \tag{2}$$

for unrelated hypotheses. If, for simplicity, we assume that the 1,000 hypothesis tests are independent and that α_{EW} is to be kept at .05, Equation 2 tells us that α_{PC} must be set at .0000513, or, essentially, .05 divided by 1,000. If you remember that there is an inverse relationship between Type I and Type II errors, it should be obvious that in lowering the alpha level from .05 to .00005, we are inevitably increasing the probability of a Type II error. In other words, if we decide to control α_{EW} rather than α_{PC} at .05, we must set α_{PC} at .00005. As a result, the power to detect real effects (differences) in the population is greatly diminished. The same effect occurs anytime we decide to control α_{EW} at .05, although the magnitude of the effect is much weaker when the number of hypotheses in the experiment is not so large. Indeed, in this sense, the decision about controlling α_{EW} or α_{PC} at .05 really involves a trade-off between Type I and Type II errors.

SIMULTANEOUS CONFIDENCE INTERVALS

The concept of multiple comparisons is also relevant for confidence intervals. Consider an example where $a = 3$, and the goal is to form three 95% confidence intervals, one for each pairwise difference between means, that is, $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$. Because more than one interval is being formed, the 95% figure could take on either of two meanings. First, the confidence level might be 95% for each interval considered individually. In other words, for a single pairwise difference, 95% of such intervals would contain the true difference. Second, the 95% figure might pertain to the entire collection of intervals, in which case it is referred to as a 95% simultaneous confidence interval. In other words, (in this case) 95% of the time that three such intervals were constructed, all three would contain their true respective difference. A 95% confidence interval for a single contrast is directly analogous to a hypothesis test where $\alpha_{PC} = .05$, whereas a 95% simultaneous confidence interval is directly analogous to a collection of hypothesis tests where $\alpha_{EW} = .05$.

We will see that simultaneous confidence intervals can be constructed for many but not all multiple-comparison procedures. For the methods we recommend in this book, the formation of simultaneous confidence intervals follows one basic formula. To form a confidence interval for a contrast ψ, under the assumption of homogeneity of variance, the interval is given by

$$\hat{\psi} \pm CV\sqrt{MS_W \sum_{j=1}^{a}\left(c_j^2 / n_j\right)} \tag{3}$$

where (as we will see later in the chapter) CV is a critical value that depends on the specific multiple-comparison procedure employed. Looking back at Chapter 4, you might realize that Equation 3 bears a striking resemblance to Equation 4.40:

$$\hat{\psi} \pm \sqrt{F_{\alpha;1,N-a}}\sqrt{MS_W \sum_{j=1}^{a}\left(c_j^2 / n_j\right)} \qquad \text{(4.40, repeated)}$$

Recall that we developed this equation as a confidence interval for a single contrast. This implies that choosing CV equal to $\sqrt{F_{\alpha;1,N-a}}$ is equivalent to controlling confidence at the per-interval level. In other words, this choice does not produce simultaneous confidence intervals. However, we will see momentarily that by choosing CV differently we can obtain simultaneous confidence intervals. We will also see that similar but slightly more complicated expressions can be used when homogeneity is questionable.

For these recommended procedures, there is a direct correspondence between the confidence interval and the hypothesis test. The null hypothesis is rejected if and only if the simultaneous confidence interval fails to contain the hypothesized value of ψ (which in almost all applications is zero). Thus, proper interpretation of a simultaneous confidence interval conveys the information contained in a hypothesis test for each of these multiple-comparison procedures. However, a simultaneous confidence interval is often more informative than the corresponding hypothesis test because the interval shows both the magnitude of the difference and the precision with which the magnitude is estimated.

Levels of Strength of Inference

Anytime multiple inferences are to be made, researchers face a choice about how to control the error rate. So far we have seen that the error rate can be controlled at either the experiment level (by controlling α_{EW}) or at the per-comparison level (by controlling α_{PC}). In fact, there are yet other choices for level of control that we will now describe. Once a researcher has chosen a desired level of control (also referred to as a level of inference), the next step is to choose a statistical technique to provide the appropriate level of control. Unfortunately, in practice, there is the appearance that all too often the statistical technique is chosen first, and only later (if at all) does the researcher seriously think about level of inference. Thus, our presentation here has two related goals. First, we need to describe the pros and cons of each of the levels of inference, so you can make an informed choice about which level you believe is most appropriate for your situation. Second, we need to describe statistical techniques that allow you to control your Type I error rate at the desired level of inference.

Hsu (1996) describes five levels of strength of inference, each of which corresponds to a choice of how best to control the error rate. Table 5.1 lists these levels from weakest to strongest (top to bottom). The weakest level of inference is simply to control the per-comparison alpha level and thus make no adjustment whatsoever for multiple tests. A statistically significant result obtained by a researcher who chooses this level of inference allows a conclusion that this specific contrast is truly non-zero, but provides absolutely no protection against inflating the Type I error rate with regard to the set of contrasts being tested. The next weakest level of inference controls the alpha level for a "test of homogeneity." A statistically significant result at this level allows a researcher to state that a difference exists among a set of population means, but it does not allow a statement about which specific means differ from one another. Hsu's third level is "confident inequalities." This level is stronger than the previous level because it allows a statement about specific means that differ from one another. However, only at the fourth level, "confident directions," is there justification for a statement that one mean is larger or smaller than another. In

TABLE 5.1
HSU'S (1996) LEVELS OF INFERENCE FOR MULTIPLE COMPARISONS

1. Per comparison	Weakest
2. Test of homogeneity	
3. Confident inequalities	
4. Confident directions	
5. Confidence intervals	Strongest

other words, at Levels 2 and 3, a researcher can legitimately claim only that some form of difference exists, but cannot legitimately proclaim the direction of any such differences. The fifth and strongest level, "confidence intervals," not only allows a statement of direction, but also allows a statement about the size of the difference between means.

Thus, one step in choosing how to proceed in the face of multiple inferences is to choose a level of strength of inference from Table 5.1. Our general recommendation, which agrees with Hsu (1996), is to choose techniques (to be described momentarily) that provide inference at either the confidence intervals or confident directions level. Anything less than this simply tells us that means are different, but does not tell us in what direction they differ. Except in rare cases, knowledge of direction is essential, both scientifically and practically.

You might wonder why we are spending so much time on this topic. One reason is that many researchers have failed to appreciate the distinctions among the levels shown in Table 5.1. All too often in the published literature, one finds a journal article where the author has used a statistical technique that provides inference only at level 2 or 3, yet the author has gone right ahead and interpreted his or her results in terms of direction of mean differences. However, such an interpretation demands a technique appropriate for level 4 or 5.

Thus, although we stressed that many of the choices in this chapter are more subjective than simple "right versus wrong," it is also true that some interpretations of results are literally wrong in the sense that the technique chosen to analyze the data does not answer the true question of interest. As Tukey (1991) states:

> Statisticians classically asked the wrong question—and were willing to answer with a lie, and one that was often a downright lie. They asked "Are the effects of *A* and *B* different?" and they were willing to answer "no."
>
> . . . asking "Are the effects different?" is foolish.
>
> What we should be answering first is "Can we tell the direction in which the effects of *A* differ from the effects of *B*?"
>
> (p. 100)

We agree with both Tukey and Hsu, and thus emphasize techniques at the highest level of inference. In any event, the most important point may be that researchers who choose to use techniques corresponding to lower levels must realize the limitations of the statements that they are justified in making based on their choice of technique.

Types of Contrasts

Hsu (1996) classifies techniques for handling multiple inferences in terms of two dimensions. We have just seen that one of these dimensions is level of strength of inference. The second

dimension is the type of contrast to be tested. We will consider four types of contrasts: (1) pairwise, (2) complex, (3) comparisons with a control, and (4) comparisons with the best. We will describe each of these in detail as we progress through the chapter, so for the moment we will provide only a brief preview. The important point for now is that the choice of an appropriate technique to address multiple inference should depend on the type of questions being asked. For example, different techniques are appropriate when all contrasts to be examined are pairwise than when some complex contrasts are also of interest. In some situations, a preselected set of contrasts may be of interest. For example, the specific goal of a study may be to compare each of several treatment groups to a control group, leading to the third type of contrast in Hsu's conceptualization. Or, an example of the fourth type of contrast occurs when the goal is to determine the best treatment based on data obtained in an experiment. As we will see shortly, each of these situations calls for a particular technique to analyze data appropriately.

Overview of Techniques

Not only is the choice of a multiple-comparison technique somewhat arbitrary, but the choice is complicated by the great variety of available options. As of this writing, SAS, SPSS, and R all offer at least a dozen choices. Although the instant availability of these techniques is obviously convenient, it also can lead to researchers reporting results based on a choice that does not really correspond to their situation. The choice may not reflect an appropriate level of strength of inference, or the choice may not correspond to the type of contrast examined by the researcher. Thus, while we will describe how to perform a variety of tests, our primary goal in this chapter is to delineate situations where different types of techniques are appropriate. Table 5.2 is a variation of a table presented by Hsu (1996), showing where the most popular multiple-comparison procedures fit into a two-dimensional framework of (1) level of strength of inference and (2) type of contrast.

TABLE 5.2
TWO-DIMENSIONAL CONCEPTUALIZATION OF MULTIPLE-COMPARISON PROCEDURES†

	Planned	*All Pairwise*	*Some Complex*	*Comparisons With a Control*	*Comparisons With the Best*
Per comparison					
Test of homogeneity		Duncan Newman–Keuls LSD Tukey's-b			
Confident inequalities		REGW-Q REGW-F			
Confident directions				See Hsu (1996)	
Confidence intervals	Bonferroni* Sidak*	Tukey HSD* Dunnett's T3* Games–Howell* Hochberg's GT2 (or SMM) Gabriel	Scheffé*	Dunnett*	Edwards–Hsu* Hsu*

† Rows indicate strength of inference and columns indicate type of contrasts of interest.
* We recommend this technique in certain situations; see the text for details.

As we stated previously, we believe that with rare exception, techniques must provide information about the direction of differences in group means if they are to be useful. Thus, we largely restrict our attention to techniques that satisfy this requirement. Readers who are convinced that their specific situation does not demand this level of inference may want to consult such sources as Hsu (1996) and Toothaker (1991) for descriptions of other techniques.

To provide a structure for reading the rest of the chapter, we present a brief overview of the multiple-comparisons procedures we recommend. First, when a researcher *plans* to test a small number of contrasts based on theoretical hypotheses prior to data collection, a technique known as the Bonferroni adjustment is appropriate. Second, when all pairwise comparisons are of potential interest, Tukey (1953) developed a procedure to maintain α_{EW} at .05. Third, Scheffé's procedure can be used when an investigator decides to test complex comparisons suggested by the data. In other words, Scheffé's method permits "data snooping," so that even after having examined the data, multiple tests can be performed, and α_{EW} will be maintained at .05. Fourth, occasionally, the only questions of interest involve comparing a control group to each of several treatment groups. Dunnett developed a procedure appropriate for this situation. Fifth, especially in applied studies, the major goal may be to determine the best treatment. Hsu has shown how Dunnett's method can be modified in this situation. As of this writing, this method has been greatly underutilized in the behavioral sciences and is not yet widely available in statistical software. However, we believe it deserves much more attention than it has received to date, so we will feature it later in the chapter.

We first consider why it is important whether contrasts to be tested have been selected prior to or after having collected the data. Then we present the Bonferroni, Tukey, Scheffé, and Dunnett procedures. Besides describing the mechanics of how to perform each test, we also develop the logic behind each technique. This discussion of logical underpinnings is especially important because we have seen that the literature is filled with many more multiple-comparisons procedures than just these four. As a result, you need to understand what it is that these four methods accomplish that many competing methods do not. To further attain this goal, after presenting the four techniques we recommend, we also briefly discuss liabilities of some of the more popular competitors. Finally, we also present a flowchart (i.e., a decision tree) to help you decide which technique should be used in a particular situation.

Planned Versus Post Hoc Contrasts

As might be expected, controlling α_{EW} at .05 is considerably more difficult than simply deciding to use an alpha level of .05 for each contrast to be tested. The first step in our task distinguishes between planned and post hoc contrasts. A *planned contrast* is a contrast that the experimenter decided to test prior to any examination of the data. A *post hoc contrast*, on the other hand, is a contrast that the experimenter decided to test only after having observed some or all of the data. For this reason, it is often said that a post hoc contrast is a contrast suggested by the data.

Why is the distinction between planned and post hoc contrasts important? The importance can be illustrated by the following example. Suppose that a researcher obtains the following means in a four-group study: $\bar{Y}_1 = 50$, $\bar{Y}_2 = 44$, $\bar{Y}_3 = 52$, and $\bar{Y}_4 = 60$. Consider the single contrast $\mu_2 - \mu_4$. There is an important difference between deciding in advance to compare Groups 2 and 4 versus deciding after having looked at the data to compare these two groups. The difference can be exemplified most easily by supposing that unbeknownst to the researcher all four population means are equal, that is

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 \tag{4}$$

If the comparison of Groups 2 and 4 has been planned and $\alpha_{PC} = .05$ is used, then (in the long run) 5 out of every 100 times the experiment would be conducted, the contrast would be statistically significant and a Type I error would have been committed. However, suppose that this contrast had not been planned. How would things change? Suppose that the study were repeated, yielding a different set of sample means: $\bar{Y}_1 = 46$, $\bar{Y}_2 = 57$, $\bar{Y}_3 = 49$, and $\bar{Y}_4 = 54$. From inspecting these data, it is doubtful that an experimenter would decide to compare Groups 2 and 4. Instead, this time the data suggest that the comparison of Groups 1 and 2 should be investigated. In other words, if the contrast to be tested is suggested by the data, it is only natural that the largest difference between means be tested because the usual goal of hypothesis testing is to obtain a statistically significant result. Suppose that a procedure was followed where this largest difference is always tested using an alpha level of .05 for the contrast. The result would be that the probability of committing a Type I error would greatly exceed .05, especially for large a. The crucial point is that $\bar{Y}_2 - \bar{Y}_4$ has a very different sampling distribution from $\bar{Y}_{max} - \bar{Y}_{min}$, where $\bar{Y}_{max}$ and $\bar{Y}_{min}$ are the largest and smallest sample means, respectively. The critical value of the F distribution that provides an alpha level of .05 for judging the significance of $\bar{Y}_2 - \bar{Y}_4$ is too small for judging the significance of $\bar{Y}_{max}$ and $\bar{Y}_{min}$. The point of this discussion has simply been to convince you that it matters greatly whether a contrast is planned or has been selected post hoc. We now turn to a consideration of procedures for testing more than one planned comparison in a study. In the following section, we consider how to test post hoc contrasts, where such topics as the sampling distribution of $\bar{Y}_{max} - \bar{Y}_{min}$ become relevant.

MULTIPLE PLANNED COMPARISONS

We illustrate the use of multiple planned comparisons by an example. Consider a four-group study whose purpose is to investigate the effects of strategy training on a memory task for children of two age levels. Independent samples of 6-year-olds and 8-year-olds are obtained. One half of the children in each group are assigned to a strategy-training condition, and the other half receive no training and serve as a control group. The general question of interest concerns the effect of strategy training on mean level of memory task performance for the two age groups.

How should the investigator attempt to answer this question? If the investigator has not planned to compare specific groups prior to collecting data, a test of the omnibus null hypothesis for all four groups could be performed. If the test were non-significant, no further tests would be performed; if the test were significant, contrasts suggested by the data might be further tested. Although this approach is entirely permissible, we defer discussion of it for the moment. Instead, we discuss an alternative approach whereby the investigator plans to test a number of specific hypotheses instead of the general omnibus hypothesis. For example, suppose that an investigator decides prior to obtaining data that he or she is interested in testing the following contrasts in our four-group study:

$$\psi_1 = \mu_{T6} - \mu_{C6}$$
$$\psi_2 = \mu_{T8} - \mu_{C8}$$
$$\psi_3 = \tfrac{1}{2}(\mu_{T6} + \mu_{C6}) - \tfrac{1}{2}(\mu_{T8} + \mu_{C8})$$

In this instance, the first subscript represents treatment (T) or control (C), and the second subscript represents the child's age group (6 or 8). The first contrast equals the effect of training for 6-year-olds, the second equals the effect for 8-year-olds, and the third equals an age effect averaged over condition. We should hasten to point out that we are not claiming these are the

three "correct" contrasts to test. What is "correct" depends on the scientific questions the study is designed to answer. For our purposes, we assume that these three contrasts have been chosen to allow us to address the questions of scientific interest. Keep in mind, however, that a researcher with different goals might formulate a very different set of contrasts. Indeed, even the number of contrasts might be very different from three.

Assuming that these three contrasts have been chosen, how should the investigator proceed? The first step is to compute an F statistic for each contrast. This can be accomplished using any of the approaches described in Chapter 4. For example, if we are willing to assume homogeneity of variance, Equation 4.32 might be used:

$$F = \frac{(\hat{\psi})^2 \Big/ \sum_{j=1}^{a} \left(c_j^2 / n_j\right)}{MS_{\mathrm{W}}} \qquad \text{(4.32, repeated)}$$

Suppose that we have done this calculation for our first contrast, and we obtained an F value equal to 4.23. Let's say there were 11 participants per group, so this F value has 1 and 40 degrees of freedom associated with it. Can we reject the null hypothesis that the population value of the first contrast is zero? If we refer to an F table, the critical F for an alpha of .05 with 1 and 40 degrees of freedom is 4.08. The observed F exceeds the critical F, which would seem to imply a statistically significant result. However, recall our earlier distinction between α_{PC} and α_{EW}. The procedure that was just described used .05 for α_{PC}. However, earlier we demonstrated that if three orthogonal contrasts (with equal n) are each tested with an alpha of .05, then $\alpha_{\mathrm{EW}} = .143$. In other words, if we test ψ_1, ψ_2, and ψ_3 using $\alpha_{\mathrm{PC}} = .05$ for each, there is a 14.3% chance of committing at least one Type I error. This seems to defeat the primary purpose behind inferential statistics, namely, to avoid a declaration of a difference between groups (or a relationship between variables) where in fact none exists in the population.

Bonferroni Adjustment

Instead of letting α_{EW} be at the mercy of α_{PC}, it seems reasonable to work backward. In other words, it would be preferable to control α_{EW} at .05, but to accomplish this, α_{PC} would have to be lowered by some amount. The problem is to determine an appropriate value of α_{PC} to result in $\alpha_{\mathrm{EW}} = .05$. It turns out that there is a remarkably simple and intuitive solution. In the general case of C hypotheses to be tested, set α_{PC} at $.05/C$. It can be proven mathematically that with this procedure, α_{EW} will be .05 or less. To use this approach in our current example, α_{PC} would be set equal to .05/3, or .0167. The critical F for $\alpha = .0167$ with 1 and 40 degrees of freedom is 6.24, which is naturally somewhat larger than the value of 4.08 that we found for α_{PC} equal to .05. In fact, we would now judge our observed F of 4.23 to be non-significant because it fails to exceed the critical value of 6.24. In an actual study, the second and third contrasts would also be tested for significance. The use of α_{PC} values other than .05 can sometimes be awkward in practice, because appropriate tables of the F distribution may be unavailable. There are three possible solutions to this problem. First, Appendix Table 3 can be used to find critical values for an F distribution with 1 numerator degree of freedom and an α_{PC} equal to $.05/C$ (two-tailed). Each row of the table represents a particular value for denominator degrees of freedom, and each column represents a value of C. It should be noted that the table only applies to F distributions with 1 numerator degree of freedom. This limitation poses no problem for testing a contrast because such a test has 1 degree of freedom in the numerator; however, there are other procedures similar to the method we discuss that involve more than 1 numerator degree of freedom. Appendix Table 3 could not be used for this situation. Second, if you have used statistical software to analyze your data and

if the program provides a p value in the output, the Bonferroni adjustment is extremely easy to apply. All that must be done is to compare the p value from the printout with $.05/C$, because $.05/C$ is the per-comparison alpha level. The contrast is statistically significant if and only if the p value is below $.05/C$. Notice that this procedure works as well in the more general case, where the numerator degrees of freedom exceed 1. Third, a single line of R code can provide the appropriate critical F value:

```
qf(1-alpha_PC, df1, df2)
```

For example, qf(1−(.05/3), 1, 40) returns a value of 6.24 (to two decimal places).

At this point, more detail for the rationale behind the $.05/C$ adjustment must be provided. The procedure was first applied to the problem of multiple contrasts by Dunn (1961), so the Bonferroni adjustment is also known as Dunn's procedure. She based the procedure on an inequality derived by the Italian mathematician Bonferroni, who proved mathematically that

$$1 - (1 - \alpha)^C \leq C\alpha \qquad (5)$$

for any value of C whenever $0 \leq \alpha \leq 1$. The practical importance of this inequality for us can be seen by realizing that the left-hand side of Equation 5 is identical to the expression in Equation 1. Thus, it is true that

$$\Pr(\text{at least one Type I error}) \leq C\alpha$$

whenever C orthogonal contrasts are each tested at the same alpha level (indicated simply by α). By setting $\alpha = .05/C$, it follows from Equation 5 that

$$\Pr(\text{at least one Type I error}) \leq .05 \qquad (6)$$

Indeed, this is precisely what is done in the Bonferroni approach. Several comments are pertinent here. First, because Equation 5 is an inequality, it might happen that the actual probability of a Type I error is much less than .05 when the Bonferroni adjustment is used. However, for orthogonal contrasts and small values of alpha, the inequality is for all practical purposes an equality, as Table 5.3 shows. Thus, the adjustment does not result in a conservative test. Second, so far we have only considered orthogonal contrasts. Remember that $1 - (1 - \alpha)^C$ equals the probability of at least one Type I error only for orthogonal contrasts. It turns out that if the set of contrasts is nonorthogonal, the probability of at least one Type I error will always be less than $1 - (1 - \alpha)^C$. Thus, the Bonferroni procedure maintains α_{EW} at .05 for nonorthogonal and for orthogonal contrasts. However, the procedure is somewhat conservative for nonorthogonal contrasts.

A second way of viewing the rationale for the Bonferroni adjustment is in many ways simpler than the first perspective. Recall that in our discussion of error rates we defined the error rate per experiment ENEPE to be the expected number of Type I errors in an experiment. If we perform C tests of significance, each at an alpha value of α_{PC}, then the expected number of Type I errors is simply

$$\text{ENEPE} = C\alpha_{\mathrm{PC}} \qquad (7)$$

If we choose α_{PC} to equal $.05/C$, then obviously ENEPE will equal .05. As a result, the expected number of Type I errors in an experiment will equal .05, regardless of the number of tests that are performed. What is the relationship between α_{EW} and ENEPE? The former equals the proportion

TABLE 5.3
COMPARISON OF $1 - [1 - (.05/C)]^C$ AND .05
FOR ORTHOGONAL CONTRASTS

C	*Actual Probability of at Least One Type I Error* $1 - [1 - (.05/C)]^C$
1	.050000
2	.049375
3	.049171
4	.049070
5	.049010
⋮	⋮
10	.048889
⋮	⋮
20	.048830
⋮	⋮
50	.048794

of experiments that have Type I errors, whereas the latter equals the number of Type I errors per experiment. In symbols,

$$\alpha_{EW} = \frac{\text{number of experiments with errors}}{\text{number of experiments}} \tag{8}$$

$$\text{ENEPE} = \frac{\text{number of errors}}{\text{number of experiments}} \tag{9}$$

Obviously, α_{EW} and ENEPE share the same denominator. However, the numerator for α_{EW} is less than or equal to the numerator for ENEPE, for the same set of data, because the numerator of Equation 8 is at most 1 per experiment, whereas the numerator of Equation 9 is incremented by 1 or more whenever the numerator of Equation 8 is 1. Thus, it is true that $\alpha_{EW} \leq \text{ENEPE}$. We showed a moment ago that the Bonferroni approach yields a value of ENEPE equal to .05. Because $\alpha_{EW} \leq \text{ENEPE}$, the Bonferroni procedure guarantees that $\alpha_{EW} \leq .05$.

The Bonferroni approach has a straightforward counterpart for simultaneous confidence intervals. We saw earlier in the chapter that the general expression for simultaneous confidence intervals is

$$\hat{\psi} \pm CV\sqrt{MS_W \sum_{j=1}^{a}\left(c_j^2 / n_j\right)} \tag{3, repeated}$$

For the Bonferroni method, CV is given by

$$CV = \sqrt{F_{.05/C;1,df_{error}}}$$

Notice that if C were equal to 1, the Bonferroni interval would simply duplicate the interval given by Equation 4.40, which controls confidence at the per-interval level. However, whenever C is

greater than 1, CV for the Bonferroni interval will be wider than CV in Equation 4.40. Thus, the Bonferroni interval will be wider than the per-interval confidence level. This decrease in precision is the price to be paid for obtaining confidence at the simultaneous level.

Two other points are worth mentioning. First, in theory, the α_{EW} of .05 need not be divided into C equal pieces of $.05/C$ in the Bonferroni method. Instead, it is only necessary that the C α_{PC} values sum to .05. For example, an experimenter testing three contrasts might use α_{PC} values of .03, .01, and .01. This could be done if the first contrast were considered most important. Notice that using a larger alpha value for the first contrast would increase the power for detecting an effect, if one exists. However, there is a catch that limits the value of such unequal splitting of alpha in practice—the choice about how to divide .05 must be made prior to any examination of data. Otherwise, the experimenter could capitalize on chance and obtain statistically significant findings too often. Second, you might wonder why we did not use Equation 2 to find the value of α_{PC} that would keep α_{EW} at .05:

$$\alpha_{PC} = 1 - \sqrt[C]{1 - \alpha_{EW}} \tag{2, repeated}$$

Although we derived this equation for C orthogonal contrasts, Sidak (1967) proved that an inequality similar to Bonferroni's holds in the general case of nonorthogonal or orthogonal contrasts. Specifically, if α_{PC} is set equal to $1 - \sqrt[C]{.95}$, then α_{EW} will be .05 or less. It turns out that Sidak's value of α_{PC} is always slightly higher than the Bonferroni value (for $C > 1$), so the Sidak modification is more powerful than the Bonferroni approach. However, the difference in power is very small as long as α_{EW} is low. In addition, the Bonferroni α_{PC} is much easier to calculate and more intuitive. For these reasons, in practice, the Bonferroni approach is usually used instead of Sidak's method (for more detail on the Sidak approach, see Kirk, 2013, or Holland & Copenhaver, 1988).

There is a final point regarding planned multiple comparisons that must be mentioned. The procedure we have described guarantees that α_{EW} will be .05 or less regardless of how many contrasts an experimenter plans. Thus, the overall probability of a Type I error being made somewhere in the experiment is the same as it would be if the researcher were to perform a test of the omnibus null hypothesis instead of planned comparisons. In this way, the chances of obtaining statistical significance in a study are not increased simply by performing multiple tests. At the same time, however, there is a penalty imposed on the investigator who plans a large number of contrasts, because α_{PC} is set at $.05/C$. As C increases, it becomes more difficult to detect each individual true effect, all other things being equal. Although the experiment as a whole has an alpha level of .05, each individual hypothesis is tested at $.05/C$. It could be argued that this puts each hypothesis test at an unfair disadvantage. Indeed, some behavioral statisticians (e.g., Keppel, 1982; Kirk, 1982) used this line of reasoning for planned contrasts. With their approaches, α_{EW} is allowed to exceed .05, because they allow up to $a - 1$ contrasts to be tested with an α_{PC} level of .05. There is disagreement within this camp about whether the $a - 1$ contrasts must form an orthogonal set in order to set α_{PC} at .05. Although this general approach has some appeal, it nevertheless fails to control α_{EW} at .05. We prefer the Bonferroni approach, because it accomplishes this goal.

Modification of the Bonferroni Approach With Unequal Variances

As we discussed in Chapter 4, using MS_W as an error term for testing contrasts is problematic when population variances are unequal. Just as heterogeneous variances affect α_{PC}, they also may affect α_{EW}. However, a rather straightforward solution is available. The Bonferroni procedure is

easily modified by using Equations 4.33 and 4.34 when population variances are unequal. As we showed in Chapter 4, the resultant F statistic in this case is given by

$$F = \frac{(\hat{\psi})^2}{\sum_{j=1}^{a}(c_j^2/n_j)s_j^2} \tag{10}$$

As we mentioned in Chapter 4, the corresponding denominator degrees of freedom are tedious, so a computer program will invariably be invoked for their calculation. Later in this chapter, Tables 5.16 and 5.17 will show the formula for these degrees of freedom in the context of summarizing our recommended multiple-comparison procedures.

The F value shown in Equation 10 will usually be computed by statistical software, which will also yield the p value for this F and its corresponding degrees of freedom. As usual with the Bonferroni adjustment, the obtained p value is then compared to α_{EW}/C (usually $.05/C$) to assess the statistical significance of the contrast. Thus, access to a statistical program that computes F values of the form shown in Equation 10 makes it straightforward to test contrasts without having to assume homogeneity of variance.

We saw that a simultaneous confidence interval can be obtained under the homogeneity assumption by generalizing Equation 40 from Chapter 4. When homogeneity is questionable, a simultaneous confidence interval can be formed by generalizing Equation 41 from Chapter 4. Recall that Equation 41 in Chapter 4 was given by

$$\hat{\psi} \pm \sqrt{F_{\alpha;1,df}}\sqrt{\sum_{j=1}^{a}\left[\left(c_j^2/n_j\right)s_j^2\right]} \qquad \text{(4.41, repeated)}$$

The Bonferroni approach simply uses α/C instead of α to find the critical F value in Equation 41, with the value of the degrees of freedom for each interval being based on the heterogeneous formula from Equation 34 in Chapter 4 (the same as the SPSS output in Table 5.5). Thus, when homogeneity is questionable, a Bonferroni 95% simultaneous confidence interval has the form

$$\hat{\psi} \pm \sqrt{F_{.05/C;1,df}}\sqrt{\sum_{j=1}^{a}\left[\left(c_j^2/n_j\right)s_j^2\right]} \tag{11}$$

Numerical Example

We will illustrate the Bonferroni procedure by considering the data shown in Table 5.4, which are a slight modification of the blood pressure data we originally encountered in Chapter 4. We will now assume that the experiment successfully obtained blood pressure readings from all intended

TABLE 5.4
BLOOD PRESSURE DATA FOR FOUR TREATMENT GROUPS

Group			
1	*2*	*3*	*4*
84	81	98	91
95	84	95	78
93	92	86	85
104	101	87	80
99	80	94	81
106	108	101	76

individuals, so we no longer have to worry about missing data. Although the Bonferroni method easily accommodates unequal sample sizes, we want to use this same data set to illustrate a variety of multiple-comparison procedures, and we will see later in the chapter that some of the procedures yet to come are much simpler with equal sample sizes. We should also emphasize from the outset that we will use the Table 5.4 data to illustrate several different procedures, but in an actual study, a single procedure should be chosen based on an appropriate intersection of the rows and columns of Table 5.2. We will say more at the end of the chapter about choosing an appropriate procedure.

Suppose that before collecting the data shown in Table 5.4, the investigator had decided to examine four planned contrasts: (1) drug versus biofeedback, (2) drug versus diet, (3) biofeedback versus diet, and (4) the average of each of the first three treatments versus the combined treatment. As usual, we are not claiming that these four contrasts are necessarily the only correct contrasts to examine in this data set. On the other hand, we do believe they constitute a plausible set of contrasts that might in fact be examined as planned comparisons in such a study. How should the investigator proceed? To answer this question, look at Table 5.2.

Because all comparisons are planned, a procedure can be chosen from the first column of the table. If the investigator follows our recommendation to obtain strength of inference at the level of confidence intervals, either Bonferroni or Sidak is appropriate.

We will see that using SAS, SPSS, or R to test contrasts with the Bonferroni approach is very simple. However, using these programs to form confidence intervals is sometimes less straightforward. As of this writing, proper Bonferroni (or Sidak) confidence intervals are straightforward in SAS and SPSS only in a very special case, where three conditions are met: (1) all pairwise contrasts are of interest; (2) no complex contrasts are of interest; and (3) homogeneity is assumed. To form confidence intervals for a subset of pairwise contrasts, or for complex contrasts (with or without homogeneity), or for pairwise contrasts without assuming homogeneity, it may be necessary to obtain intermediate values that can then be used as input to appropriate formulas for final hand calculations. For example, Table 5.5 shows output obtained from SPSS for the four contrasts in question (at the top) and hand calculations. It is important to stress that the results in this table were obtained without specifying the Bonferroni option in SPSS, but rather by specifying the contrast coefficients directly. Our situation does not conform to the three conditions previously listed. If we had specified the Bonferroni option, SPSS would have provided tests and confidence intervals for pairwise contrasts under a mistaken assumption that C (the number of contrasts of interest) equals 6, because that is the total number of pairwise contrasts with 4 groups. In reality, $C = 4$ in our situation, so asking for the Bonferroni option in either SAS or SPSS would not be appropriate here. Instead, the results shown here implicitly assume a per-comparison alpha level of .05—contrary to our real goal. Thus, we will have to adjust information in the output ourselves instead of relying entirely on SPSS.

The information in Table 5.5 makes it easy to test the null hypothesis that each contrast equals zero using the Bonferroni method. All we need to do is to compare each p value on the printout with an adjusted per-comparison alpha level of $.05/C$. In our case, $C = 4$, so $\alpha_{PC} = .0125$. Table 5.5 shows that regardless of whether we assume homogeneity, the only contrast with a p value less than .0125 is the fourth contrast. Thus, the only statistically significant difference is between the combined treatment and the average of the three individual treatments.

Forming simultaneous confidence intervals for the contrasts is slightly more complicated.[2] However, we can use our computer output by realizing that Equation 3 (assuming homogeneity) and the comparable variation based on Equation 4.41 (without assuming homogeneity) for forming confidence intervals can be written in words as

$$\text{estimate} \pm (\text{critical value})\,(\text{estimated standard error}) \tag{12}$$

TABLE 5.5
SPSS OUTPUT AND HAND COMPUTATIONS FOR PLANNED COMPARISONS

Selected Portions of Output from SPSS ONEWAY

Contrast Coefficients

Contrast	Group 1	2	3	4
1	1	−1	0	0
2	1	0	−1	0
3	0	1	−1	0
4	1	1	1	−3

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
sbp	Assume equal variances	1	5.83	4.668	1.250	20	.226
		2	3.33	4.668	.714	20	.483
		3	−2.50	4.668	−.536	20	.598
		4	35.83	11.433	3.134	20	.005
	Does not assume equal variances	1	5.83	5.724	1.019	8.947	.335
		2	3.33	4.084	.816	9.222	.435
		3	−2.50	5.284	−.473	7.508	.650
		4	35.83	9.095	3.940	13.288	.002

Hand Computations For Simultaneous Confidence Intervals

Assuming Homogeneity of Variance

Contrast	*Computation*	*Interval*
1	$5.8333 \pm (\sqrt{7.53})(4.6676)$	−6.97, 18.64
2	$3.3333 \pm (\sqrt{7.53})(4.6676)$	−9.47, 16.14
3	$-2.500 \pm (\sqrt{7.53})(4.6676)$	−15.31, 10.31
4	$(35.83/3) \pm (\sqrt{7.53})\,(11.4331/3)$	1.49, 22.40

Without Assuming Homogeneity of Variance

Contrast	*Computation*	*Interval*
1	$5.8333 \pm (\sqrt{10.28})(5.7237)$	−12.52, 24.18
2	$3.3333 \pm (\sqrt{9.68})(4.0838)$	−9.37, 16.04
3	$-2.500 \pm (\sqrt{11.12})(5.2836)$	−20.12, 15.12
4	$(35.83/3) \pm (\sqrt{8.39})(9.0955/3)$	3.16, 20.73

With one possible minor exception to be noted in a moment, the estimate and the estimated standard error can be read directly from the output. Appendix Table 3 provides the necessary critical value (notice we will usually have to round down for the degrees of freedom to use this table based on computer output, but with computer programs the exact critical values can be used with the non-integer degrees of freedom). The "minor exception" is that, as we discussed in Chapter 4, we may need to be aware of the scaling (or metric) of the contrast coefficients. As long as the sum of the absolute value of the coefficients equals 2, we can simply proceed. In our example, this is true for the first three contrasts. However, the coefficients that produced the output for the fourth contrast were 1, 1, 1, and −3. It may seem strange that we did not tell SPSS to use coefficients of 1/3, 1/3, 1/3, and −1. The reason we did not ask SPSS to use these coefficients is because SPSS allows non-integers to be written in decimal form, but it does not allow fractional coefficients. The fraction 1/3 cannot be expressed as an exact decimal (e.g., 0.3333 is close to 1/3, but to be exactly equal to 1/3, we would need to continue adding more and more 3's to our expression until they go on forever), which is why coefficients of 1, 1, 1, and −3 were used here. To obtain the correct result, we need to divide both the estimate and the estimated standard error shown on the output by 3. Incidentally, SAS allows the user to specify a "divisor" as part of the definition of the contrast, in which case the SAS output already reflects the division we have to do by hand with SPSS.

The bottom of Table 5.5 shows the simultaneous confidence intervals using the Bonferroni procedure, first when we are willing to assume homogeneity of variance and then when we are not. With or without assuming homogeneity, only the fourth interval does not contain zero, so only this contrast is statistically significantly different from zero. Notice that this result is consistent with comparing the p value in the output with a critical p value of $.05/C$, as it must be. Also notice that in these particular data, it makes little practical difference whether we assume homogeneity. Two of the intervals are appreciably wider when we relax the assumption, but the other two intervals are slightly narrower without the assumption. It is important to keep in mind that in other situations the assumption of homogeneity can make more of a difference, and in particular can change the conclusions, as we will see in some of the exercises at the end of the chapter.

PAIRWISE COMPARISONS

Frequently, a researcher decides to consider only pairwise differences between groups. In other words, no complex comparisons will be tested. How can α_{EW} be controlled at .05 in this situation? One possible approach would be to use a Bonferroni adjustment. However, care must be taken in using the proper value of C. Most often, $C = a(a-1)/2$ for testing pairwise comparisons. The reason is that with a levels of the factor, there are $a(a-1)/2$ pairs of means that can be compared. Thus, when all pairwise comparisons might be tested, using the Bonferroni adjustment the α_{EW} value of .05 must be divided by $a(a-1)/2$.

It is important to understand the connection between this section on pairwise comparisons and the previous discussion of planned comparisons. To solidify this connection, suppose that a researcher is conducting a four-group study and is interested for theoretical reasons in comparing the following pairs of means: μ_1 versus μ_2, μ_2 versus μ_3, and μ_3 versus μ_4. As long as these comparisons have been selected prior to collecting data, α_{EW} can be maintained at .05 by using an α_{PC} equal to .05/3. Thus, $C = 3$ in this situation, even though there are a total of six pairs of means; using $C = 3$ restricts the investigator to ignore the other three pairs of means, no matter how interesting such differences might appear after having collected data. For example, it would not be permissible to decide after examining data that μ_1 versus μ_4 should also be tested, and then redefine C to equal 4. Doing so leads to an exploratory and not confirmatory framework. This post hoc change in choosing which groups to compare makes the p value and confidence interval incorrect when

C is simply increased by 1. Similarly, suppose that the investigator originally planned to test all pairwise comparisons, but after looking at data, decided not to test μ_1 versus μ_3 or μ_2 versus μ_4. Again, it would not be legitimate to define $C = 4$; instead, the value of C must be set at 6.

Thus, when a specific subset of mean differences is chosen in advance of collecting data, C equals the number of comparisons in the subset. However, C must be set equal to $a(a-1)/2$ if any of the following conditions apply:

1. All pairwise comparisons are to be tested.
2. The original intent was to test all pairwise comparisons, but after looking at the data, fewer comparisons are actually tested.
3. The original intent was to test a subset of all possible pairwise comparisons, but after looking at the data, one or more additional pairwise comparisons are also to be tested.

In any case, the Bonferroni adjustment can be used to control α_{EW} when performing pairwise comparisons. However, when one of these three conditions applies, so that C must be set at $a(a-1)/2$, the Bonferroni approach is usually not as powerful as other special-purpose techniques that have been developed specifically for testing all pairwise comparisons. The technique we generally recommend for testing pairwise comparisons in between-subjects designs was developed by Tukey (1953) and is referred to as Tukey's HSD (or, interchangeably, Tukey's WSD).[3] This technique generally is more powerful than the Bonferroni approach when $C = a(a-1)/2$, and yet it allows a researcher to test all possible pairwise comparisons and still maintain the α_{EW} level at .05 (or any other desired level). That is to say, Tukey's HSD accomplishes the same goal of controlling the Type I error rate, but it does so in a more powerful way than the Bonferroni approach.

Before describing the details of Tukey's HSD, we want to mention that Tukey developed several different multiple-comparison procedures, so it is important to be certain that "Tukey" as operationalized in statistical software is in fact HSD. For example, as of this writing, SPSS provides two "Tukey" options. The option labeled simply as "Tukey" is indeed HSD, but an alternative labeled "Tukey-b" is a variation that does not necessarily control α_{EW} at the intended level, and thus we do not recommend it.

Tukey's HSD Procedure

Tukey's HSD procedure allows a researcher to perform tests of all possible pairwise comparisons in an experiment and still maintain the α_{EW} level at .05 regardless of whether or not the comparisons have been planned in advance.[4] This control of α_{EW} is accomplished by adopting a critical value appropriate for testing the significance of that pair of means that is found post hoc to yield a larger F value than any other pair of means. To make things concrete, suppose that $a = 3$. In this situation, there are three pairwise comparisons that can be tested:

$$\psi_1 = \mu_1 - \mu_2$$
$$\psi_2 = \mu_1 - \mu_3$$
$$\psi_3 = \mu_2 - \mu_3$$

For the moment, we restrict ourselves to the case of equal n and homogeneity of variance. (Tukey made both of these assumptions in deriving his procedure; in a later section, we consider modifications when either condition is not satisfied.) To test the significance of the three contrasts, Equation 4.19 can be applied three times. With equal n, this yields

$$F_{\psi_1} = \frac{n(\bar{Y}_1 - \bar{Y}_2)^2}{2MS_W}$$
$$F_{\psi_2} = \frac{n(\bar{Y}_1 - \bar{Y}_3)^2}{2MS_W}$$
$$F_{\psi_3} = \frac{n(\bar{Y}_2 - \bar{Y}_3)^2}{2MS_W}$$

It is obvious from these three equations that the largest F value will be obtained for the pair of sample means whose values are most different from one another. In symbols,

$$F_{\text{pairwise maximum}} = \frac{n(\bar{Y}_{\max} - \bar{Y}_{\min})^2}{2MS_W}$$

(Notice that because the difference between means is squared, we could just as well subtract $\bar{Y}_{\max}$ from $\bar{Y}_{\min}$.) How can we achieve our goal of maintaining α_{EW} at .05? If we were to use a single critical value (which we will abbreviate as CV) against which to judge each contrast, there would be a statistically significant result in the experiment if and only if $F_{\text{pairwise maximum}} > CV$. Our goal is that the α_{EW} should be .05, so we need to determine how large CV must be so that $F_{\text{pairwise maximum}}$ will exceed it only 5% of the time when the null hypothesis is true. The appropriate value of CV can be found from the sampling distribution of $F_{\text{pairwise maximum}}$, which has been derived mathematically. Specifically, it can be shown that the expression $\sqrt{2F_{\text{pairwise maximum}}}$ has a "studentized range" distribution if all ANOVA assumptions are met. It is traditional to represent the studentized range with the letter q, so we can write $q = \sqrt{2F_{\text{pairwise maximum}}}$.

To obtain an alpha level of .05, the critical value CV is chosen to be that value in the right tail of the q distribution beyond which lies 5% of the area. Appendix Table 4 presents critical values of the studentized range distribution for both $\alpha = .05$ and $\alpha = .01$, which we can convert into critical values for F (or t). Before we examine this table, we summarize the mechanics of Tukey's procedure. To employ Tukey's method, an observed F is calculated in the usual way for each pairwise contrast. However, instead of comparing this observed F to a critical F value, we take the square root of $2F$ and compare this number to a critical q value, or equivalently the observed F value can be compared to $q^2/2$, where q is the critical q value. This procedure is repeated for each contrast to be tested. It follows that yet another option is to calculate an observed t value for the contrast and compare it to $q/\sqrt{2}$.

TABLE 5.6
COMPARISON OF CORRESPONDING PER-COMPARISON, TUKEY, AND BONFERRONI CRITICAL F VALUES FOR TESTING ALL PAIRWISE COMPARISONS WITH $df_{\text{error}} = 12$

Number of Groups	*Number of Comparisons*	*Critical F Value*		
		Per Comparison	*Tukey*	*Bonferroni*
2	1	4.75	4.75	4.75
3	3	4.75	7.12	7.73
4	6	4.75	8.81	9.92
5	10	4.75	10.16	11.76
6	15	4.75	11.28	13.32

What makes Tukey's method different from the previously encountered methods for testing contrasts is the use of a different critical value. Instead of comparing an observed F to a critical F with an alpha level of .05 or .05/C, the observed F is compared to $q^2/2$. Notice, then, that the observed test statistic itself is unchanged; what has changed is the critical value for assessing significance. As mentioned earlier, this critical value is chosen to maintain α_{EW} at .05 specifically in the context of all pairwise comparisons.

Table 5.6 illustrates how the use of $q^2/2$ controls α_{EW} at the desired value. The specific values in the table are for $df_{error} = 12$, but the general pattern would hold for other values as well. For the moment, we concentrate on the first two columns of critical values, which show that whenever $a > 3$, the critical value for Tukey's method is larger than the critical value that would be used if α_{PC} were set at .05. The table also shows that the Tukey critical value increases dramatically as a increases. This is not surprising, because the rationale for Tukey's approach is that $F_{\text{pairwise maximum}}$ exceeds the Tukey critical value only 5% of the time. As a increases, there are more pairs of groups to be contrasted, so that $F_{\text{pairwise maximum}}$ tends to be larger in the long run even if in reality all groups have the same population mean. This is why the critical value for Tukey's method increases as a increases. (Of course, it is also true that in Table 5.6, the degrees of freedom for error is 12 regardless of a; as a increases, there are necessarily fewer participants per group.) For this reason, the Tukey critical value is larger for higher values of a. In this way, α_{EW} is maintained at .05. A necessary consequence is that implicitly α_{PC} is less than .05 whenever $a > 2$ for Tukey's method.

This leads us to a comparison of the second and third columns of critical values. Suppose that an investigator plans to test all $a(a-1)/2$ pairwise contrasts in an a-group study. From our earlier discussion, it would seem that the Bonferroni adjustment could be applied, in which case the third column would display the appropriate critical values. In the four cases where $a > 2$, the Bonferroni critical value is larger than the Tukey critical value. The smaller critical value for Tukey's method illustrates the point made earlier, that *Tukey's HSD is more powerful than the Bonferroni procedure for testing all pairwise comparisons.* Although both procedures are guaranteed to control α_{EW} at the desired level (as long as necessary statistical assumptions are met), Tukey's technique is preferable in between-subjects designs because it is more powerful. However, in Chapter 13, we will see that the Bonferroni approach may be preferable to Tukey's HSD in within-subjects designs because of the restrictive assumptions required by the HSD approach in such designs. We should also add that when only a subset of pairwise comparisons have been planned in advance, the Bonferroni approach can sometimes be more powerful than Tukey's method.

So far, the presentation of Tukey's method has been restricted to the equal sample size, equal variance condition. We now discuss modifications that can be employed when either or both of these conditions fail to hold.

Modifications of Tukey's HSD

Consider an experiment where the sample sizes of the various groups are unequal, but homogeneity of variance is assumed. The recommended procedure to employ here was developed by Kramer (1956). Recall that we developed Tukey's procedure using Equation 4.19, which is a special case of Equation 4.18 to be used only when $n_1 = n_2$. Kramer's approach for unequal n is simply to compute the observed F using the general form of Equation 4.18 that allows for unequal n. In other words, the F is calculated in exactly the same way as it was calculated for planned contrasts with unequal n. As with Tukey's approach, the observed F for each contrast is compared to a critical value given by $q^2/2$, where q is found in Appendix Table 4, using the appropriate α, a, and degrees of freedom for error.

When population variances are unequal, the situation is considerably more complicated. As we discussed in Chapter 4, procedures using Equation 4.18 are not robust to violations of

homogeneity of variance, so neither Tukey's procedure nor the Kramer modification is appropriate when variances are heterogeneous. A number of modifications of these procedures, which involve different formulas for calculating an observed F value and for calculating a critical value, have been suggested. Our recommendation is based on a synthesis of findings reported in Games, Keselman, and Rogan (1981), Hochberg and Tamhane (1987), and Wilcox (1987b). When a researcher is unwilling to assume homogeneity of variance, the observed F statistic for comparing groups g and h should be calculated as

$$F = \frac{(\bar{Y}_g - \bar{Y}_h)^2}{\frac{s_g^2}{n_g} + \frac{s_h^2}{n_h}} \tag{13}$$

where the g and h subscripts refer to the two groups involved in the specific comparison. This expression for the F statistic is simply a special case of the more general formula developed in Chapter 4 for dealing with heterogeneity (also see Equation 5.10). Similarly, the appropriate degrees of freedom in this special case can be written as

$$df = \frac{(s_g^2/n_g + s_h^2/n_h)^2}{s_g^4/n_g^2(n_g - 1) + s_h^4/n_h^2(n_h - 1)} \tag{14}$$

Fortunately, current versions of widely distributed statistical software calculate a t-statistic analog to Equations 13 and 14. In other words, they calculate the square root of the F in Equation 13; the degrees of freedom are the same for the F and the t statistics. For example, as of this writing, these values are obtained from the separate variance estimate calculations in an SPSS independent samples t test and an SPSS one-way ANOVA and from the unequal variances calculation in SAS PROC TTEST. These equations form the basis for the default independent-samples t test in R (using the t.test() function).

Once the observed t (or F) has been obtained, it must be compared to a critical value to assess statistical significance. Statisticians have proposed numerous critical values as possibilities in this context. Current evidence suggests that when sample sizes are small (e.g., fewer than 50 per group), a critical value suggested by Dunnett is most appropriate. For larger samples, a different critical value suggested by Games and Howell is better. Dunnett's procedure, which is called Dunnett's T3 (the T comes from a statistician named Tamhane, who developed the predecessor to T3), is based on the studentized maximum modulus distribution.[5] The observed t statistic is compared to a critical value V, obtained from Appendix Table 5. (Alternatively, F can be compared to V squared.) The columns of Appendix Table 5 correspond to the number of groups,[6] and the rows correspond to degrees of freedom calculated from Equation 14. When the observed t exceeds the critical V, the contrast is statistically significant.

For larger samples, we recommend a procedure suggested by Games and Howell (1976). To use their procedure, the observed t statistic is compared to $q/\sqrt{2}$ (or, equivalently, F is compared to $q^2/2$), where the degrees of freedom for the studentized range again come from Equation 14. If the observed t exceeds $q/\sqrt{2}$, the contrast is statistically significant. Notice that although at first glance the only difference between this method and Tukey's HSD might seem to be basing the critical value on Equation 14, there is a second difference because the Games and Howell method uses a separate error term based on Equation 13 for the observed F (or t), whereas the standard Tukey method assumes homogeneity of variance and thus uses MS_W as the error term. The reason Dunnett's T3 is recommended instead of the Games–Howell procedure for smaller sample sizes is that Dunnett (1980) found that the Games–Howell approach becomes slightly liberal (i.e., α_{EW} is slightly above .05) when sample sizes are small. Fortunately, both Dunnett's T3 and the

Games–Howell procedure are currently available as options in an SPSS one-way ANOVA, so these tests (as well as simultaneous confidence intervals) can easily be obtained. As of this writing, SAS does not offer these options, which implies that although PROC TTEST can be used to find the relevant observed F value and its degrees of freedom, the critical value must be calculated by hand using either Appendix Table 5 (for T3) or Appendix Table 4 (for Games–Howell).

Numerical Example

We will return to the blood pressure data of Table 5.4 to illustrate Tukey's HSD and its modifications. Table 5.7 shows SPSS output for the six pairwise comparisons we can examine in our

TABLE 5.7
SPSS OUTPUT FOR PAIRWISE COMPARISONS

Multiple Comparisons

Dependent Variable: sbp

	(I) group	(J) group	Mean Difference (I − J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Tukey HSD	1	2	5.833	4.668	.604	−7.23	18.90
		3	3.333	4.668	.890	−9.73	16.40
		4	15.000*	4.668	.021	1.94	28.06
	2	1	−5.833	4.668	.604	−18.90	7.23
		3	−2.500	4.668	.949	−15.56	10.56
		4	9.167	4.668	.235	−3.90	22.23
	3	1	−3.333	4.668	.890	−16.40	9.73
		2	2.500	4.668	.949	−10.56	15.56
		4	11.667	4.668	.091	−1.40	24.73
	4	1	−15.000*	4.668	.021	−28.06	−1.94
		2	−9.167	4.668	.235	−22.23	3.90
		3	−11.667	4.668	.091	−24.73	1.40
Tamhane	1	2	5.833	5.724	.913	−13.38	25.04
		3	3.333	4.084	.967	−10.26	16.93
		4	15.000*	3.957	.027	1.65	28.35
	2	1	−5.833	5.724	.913	−25.04	13.38
		3	−2.500	5.284	.998	−21.17	16.17
		4	9.167	5.186	.535	−9.49	27.83
	3	1	−3.333	4.084	.967	−16.93	10.26
		2	2.500	5.284	.998	−16.17	21.17
		4	11.667*	3.288	.032	.91	22.42
	4	1	−15.000*	3.957	.027	−28.35	−1.65
		2	−9.167	5.186	.535	−27.83	9.49
		3	−11.667*	3.288	.032	−22.42	−.91

Based on observed means.
Table error term is Mean Square(Error) = 65.358.
* The mean difference is significant at the .05 level.

four-group study. The table actually shows 12 comparisons under the assumption of homogeneity, as well as 12 additional comparisons without the assumption, because the table shows not only $\mu_1 - \mu_2$ but also $\mu_2 - \mu_1$ and so forth.

Table 5.7 shows that under the homogeneity assumption, the only statistically significant pairwise difference is between the drug and combination therapies. We know that 95% of similarly constructed intervals in exact replications of this study would include the true value of the contrast [or, put differently, this interval suggests that the population mean of the combination therapy is at least 2 units (rounding off) and at most 28 units lower than the population mean of the drug therapy]. Notice that although this difference is statistically significant, the small sample size has produced a very wide confidence interval, leaving considerable uncertainty about the extent to which the combination therapy is truly superior to the drug therapy. In other words, we can be reasonably confident that the combination therapy is truly better than the drug therapy, but the data provide limited evidence about the magnitude of this advantage. Further, it is plausible that all five other differences may be either positive or negative, so we cannot establish which of these therapies is better than the other.[7]

If we decide not to assume homogeneity of variance, an additional pairwise comparison becomes statistically significant. For these data, we can still assert that combination therapy is superior to drug therapy, but we can now also claim that combination therapy is superior to diet. In general, the number of statistically signficant differences can either be larger or smaller when one relaxes the homogeneity assumption. In either case, the decision as to whether to assume homogeneity should clearly be made on grounds other than the number of significant results obtained.

POST HOC COMPLEX COMPARISONS

The previous section provides a method for maintaining α_{EW} at .05 when all pairwise contrasts are tested. Now the Scheffé method is introduced to maintain α_{EW} at .05 when at least some of the contrasts to be tested are complex and suggested by the data. Although in many situations the data may suggest that a researcher compare all pairs of groups, there are times when other comparisons may also be of interest, as we saw in the hypertension example of Chapter 4. To consider another such example, suppose that the effects of different dosage levels of a drug on some aspect of behavior are being investigated. A researcher might conduct a three-group study, where the groups are defined by the dosage level they receive, such as 1 ml, 2 ml, or 3 ml. Assume that on examination of the data, the intermediate dosage seems to be most effective. Then, one contrast of interest might be

$$\psi_1 = \frac{1}{2}(\mu_1 + \mu_3) - \mu_2$$

to see whether the average of the effects of 1 and 3 ml equals the effect of 2 ml. Suppose that the researcher also wants to test the three pairwise contrasts:

$$\psi_2 = \mu_1 - \mu_2$$
$$\psi_3 = \mu_1 - \mu_3$$
$$\psi_4 = \mu_2 - \mu_3$$

Although it would be possible to use the Bonferroni approach if these are planned contrasts, we assume for the moment that they have instead been formed post hoc. After developing an appropriate technique for testing these contrasts post hoc, we return to the planned versus post hoc distinction.

If these four contrasts are to be tested post hoc, neither the Bonferroni nor the Tukey method is appropriate. The Bonferroni method is not applicable, because these particular contrasts were

not selected prior to examining the data. Thus, it would be incorrect to set C equal to 4 and use the Bonferroni adjustment with $\alpha_{PC} = .05/4$. Tukey's method is not applicable either: not all the contrasts are pairwise because ψ_1 involves three groups. We now turn to Scheffé's approach for a method that allows all four contrasts to be tested post hoc and yet keep α_{EW} at .05.

Our presentation of the logic underlying Scheffé's method is similar to the presentation of the rationale for Tukey's method. Recall that for Tukey's approach, we considered the sampling distribution of $F_{\text{pairwise maximum}}$. Now, however, we do not want to restrict ourselves to only pairwise contrasts. The logic of Scheffé's method is to consider the sampling distribution of F_{maximum}, which represents the largest possible F value for any contrast in the data, either pairwise or complex. Although finding this distribution would seem to be an extremely difficult task, it actually becomes rather easy with a few additional facts at our disposal.

Proof That $SS_{\max} = SS_B$

We detour momentarily to develop these facts. Recall that we are interested in finding the sampling distribution of F_{maximum}. Notice that the contrast that produces the largest F value is whatever contrast yields the largest sum of squares, because the F value is simply the sum of squares divided by mean square within, that is

$$F_\psi = SS_\psi / MS_W$$

It then follows that

$$F_{\text{maximum}} = SS_{\max}/MS_W \tag{15}$$

where $SS_{\max}$ is the sum of squares of the contrast with the largest sum of squares. We now show that for any set of data, $SS_{\max}$ equals the between groups sum of squares SS_B.

First, we must convince you that the sum of squares for a contrast must always be less than or equal to the between group sum of squares, that is

$$SS_\psi \le SS_B \tag{16}$$

This must be true, because we learned in Chapter 4 that one measure of effect size for comparisons is R^2_{alerting}, which was defined as

$$R^2_{\text{alerting}} = SS(\psi)/SS_{\text{Between}} \tag{4.46}$$

However, as a squared multiple correlation, R^2_{alerting} can be no larger than 1.0, in which case it immediately follows that $SS(\psi)$ can be no larger than SS_{Between}. In other words,

$$SS_\psi \le SS_B \tag{16, repeated}$$

which is what we were seeking to prove. Thus, it follows that no contrast can have a sum of squares larger than the between-group sum of squares, implying that

$$SS_{\max} \le SS_B \tag{17}$$

The final step in this argument is to show that it is always possible (after obtaining the data) to find a contrast whose sum of squares will equal SS_B. This is accomplished by defining contrast coefficients to be equal to (or proportional to) the weighted deviations of each group mean from the grand mean, where the weights for the deviations are given by the sample sizes of the groups. In other words, the contrast whose sum of squares equals SS_B has coefficients of the form

$$c_j = n_j(\bar{Y}_j - \bar{Y}) \tag{18}$$

For any sample data, the contrast whose coefficients are defined as in Equation 18 will have an R^2_{alerting} value of 1.0 and a sum of squares equal to SS_B.[8] This contrast is then necessarily the contrast with the largest possible sum of squares, because we saw earlier from Equation 17 that

$$SS_{\max} \le SS_B \tag{17, repeated}$$

However, as proved in Footnote 8, there is always a contrast whose sum of squares equals SS_B. Combining these two facts allows us to amend Equation 17. We can now say that

$$SS_{\max} = SS_B \tag{19}$$

Earlier, we argued that

$$F_{\text{maximum}} = SS_{\max}/MS_W \tag{15, repeated}$$

Substituting Equation 19 into Equation 15 yields

$$F_{\text{maximum}} = SS_B/MS_W \tag{20}$$

Thus, for a given set of data, the largest F value for a contrast always equals SS_B/MS_W.

Remember that the task at hand was to find the sampling distribution of F_{maximum}. This is made simple now that we know $F_{\text{maximum}} = SS_B/MS_W$, because we can rewrite this as

$$F_{\text{maximum}} = (a-1)MS_B/MS_W \tag{21}$$

because $SS_B = (a-1)MS_B$. However, if all necessary assumptions are met, MS_B/MS_W is distributed as an F variable with $a-1$ and $N-a$ degrees of freedom under the null hypothesis. It follows that F_{maximum} is simply distributed as $(a-1)$ times such an F variable. Therefore, even if the omnibus null hypothesis is true, so every contrast has a population value of zero, F_{maximum} exceeds

$$(a-1)F_{.05;a-1,N-a} \tag{22}$$

only 5% of the time. By using $(a-1)F_{.05;a-1,N-a}$ as a critical value against which to judge the significance of a contrast, we guarantee ourselves of maintaining α_{EW} at .05, regardless of how many contrasts we test, *even after having looked at the data*. (Of course, as always, the necessary assumptions must be met in order for the actual alpha level to equal the nominal value.) Notice

that, once again, in order to use the Scheffé method, the observed F value is calculated from Equation 4.32 (or one of its equivalent forms). What distinguishes this method from the other multiple-comparison procedures is the use of Equation 22 for the critical value.

A nice feature of Scheffé's method is that it has a direct correspondence to the test of the omnibus null hypothesis. Remember that the omnibus null hypothesis will be rejected if and only if

$$MS_B/MS_W > F_{.05;a-1,N-a} \tag{23}$$

Suppose that we were to test the contrast corresponding to F_{maxmium} with Scheffé's approach. Recall that $F_{\text{maximum}} = (a-1)MS_B/MS_W$. The critical value for Scheffé is $(a-1)F_{.05;a-1,N-a}$.

This contrast is judged to be statistically significant if and only if its observed F value exceeds the Scheffé critical value, that is, if and only if

$$F_{\text{maximum}} > (a-1)F_{.05;a-1,N-a} \tag{24}$$

However, from Equation 21

$$F_{\text{maximum}} = (a-1)MS_B/MS_W \qquad \text{(21, repeated)}$$

Substituting this result into Equation 24, we see that the contrast is significant if and only if

$$(a-1)MS_B/MS_W > (a-1)F_{.05;a-1,N-a}$$

However, we can cancel the $(a-1)$ terms, implying that the contrast is significant if and only if

$$MS_B/MS_W > F_{.05;a-1,N-a}$$

However, this repeats Equation 23, which is the condition under which the omnibus null hypothesis is rejected. Thus, the maximum contrast is statistically significant by Scheffé's method if and only if the omnibus null hypothesis is rejected.

Thus, if the omnibus null hypothesis is rejected, at least one contrast exists that is significant by Scheffé's method (namely the contrast corresponding to F_{maximum}). Conversely, if the omnibus null hypothesis is not rejected, it is impossible to find a significant contrast using Scheffé's method. All this should seem reasonable. After all, if we declare the means to be different from one another, we should be able to specify how they are different. On the other hand, if we cannot declare them to be different, it makes no sense to think we should be able to say how they are different. Although this is indeed reasonable, not all multiple-comparison procedures share this property. For example, with Tukey's method, inconsistencies can occur. It is possible to reject the omnibus null hypothesis and yet reject none of the pairwise differences. The opposite can also occur: it is possible to reject one or more of the pairwise contrasts, although the omnibus null hypothesis cannot be rejected.

Comparison of Scheffé to Bonferroni and Tukey

Scheffé's method is very useful in that it allows the researcher to test literally any contrast that may be suggested by the data (or by theory). Because the critical value is based on the sampling

TABLE 5.8
COMPARISON OF BONFERRONI AND SCHEFFÉ CRITICAL VALUES FOR $a = 4$ AND $df_{\text{error}} = 30$

C	*Bonferroni*	*Scheffé*
	$F_{.05/C;1,30}$	$3F_{.05;3,30}$
1	4.17	8.76
2	5.57	8.76
3	6.45	8.76
4	7.08	8.76
5	7.56	8.76
6	8.01	8.76
7	8.35	8.76
8	8.64	8.76
9	8.94	8.76
10	9.18	8.76

distribution of F_{maximum}, all possible contrasts could be tested for an experiment, and α_{EW} would still be maintained at .05. As we have noted before, the number of contrasts that may be tested is infinite. Although many of these may have little or no scientific meaning, they can all be tested for significance with Scheffé's method. On the other hand, what if we are really interested in testing just a few of these contrasts? In this situation, the Scheffé method is typically quite conservative, in that the actual α_{EW} for the few contrasts we actually test may be considerably less than the .05 that would result from testing all possible contrasts. Indeed, this points out the advantage of planned contrasts. If the experimenter plans the contrasts prior to the study and if the number of contrasts to be tested is relatively small, the Bonferroni critical value will be less than the Scheffé critical value, so the Bonferroni approach will be more powerful. Table 5.8 illustrates this point, where $a = 4$ and $df_{\text{error}} = 30$. From the table, we can see that as many as eight planned comparisons could be tested and still use a lower critical value with the Bonferroni than with the Scheffé. Only an investigator who might be interested in more than eight contrasts among the four groups or who is unable to specify all possible contrasts of interest in advance would find the Scheffé method superior.

Table 5.9 provides a more complete view of the choice between Bonferroni and Scheffé. Each entry in the table is the maximum number of contrasts that could be planned and still have the Bonferroni critical value less than the Scheffé. The entries are a function of a and df_{error}. Notice that the entry for $a = 4$ and $df_{\text{error}} = 30$ is 8, agreeing with Table 5.8. This table is useful for helping decide whether you should perform planned contrasts or use the Scheffé method for testing your contrasts post hoc. If the number of contrasts you might conceivably test is less than or equal to the number in Table 5.9 for your values of a and df_{error}, the Bonferroni approach is better. On the other hand, if you might test more contrasts than the number in the table, Scheffé's method is better, even if all the contrasts are planned. In the face of this discussion of how the Bonferroni and Scheffé techniques compare, do not forget Tukey's method for pairwise comparisons, either planned or post hoc.

TABLE 5.9
MAXIMUM NUMBER OF PLANNED CONTRASTS THAT SHOULD BE TESTED IN A STUDY WITH THE BONFERRONI APPROACH

	Number of Groups							
df_{error}	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
5	2	4	8	12	17	24	31	40
6	2	5	9	14	21	30	41	55
7	2	5	10	16	25	37	52	71
8	2	6	11	18	29	44	64	89
9	2	6	12	20	33	51	75	107
10	2	6	12	22	37	58	87	127
12	3	7	13	25	43	70	110	166
14	3	7	14	28	49	82	132	205
16	3	7	15	30	54	93	153	243
18	3	7	16	32	58	103	173	281
20	3	7	17	33	63	112	191	316
30	3	8	18	39	78	147	267	470
40	3	8	20	43	87	170	320	586
50	3	8	20	45	94	187	360	674
60	3	8	21	47	98	199	390	743
70	3	9	21	48	102	209	414	799
80	3	9	21	49	105	217	433	844
90	3	9	22	50	107	223	449	882
100	3	9	22	50	109	228	462	913
110	3	9	22	51	111	232	473	941
120	3	9	22	51	112	236	483	964

We saw earlier that Tukey's method is generally superior to Bonferroni's for testing all pairwise contrasts. The Scheffé is even less appropriate. Notice that almost all values in Table 5.9 exceed $a(a-1)/2$ (the number of pairwise contrasts), indicating that the Bonferroni is almost always better than the Scheffé for this number of contrasts. But we have already seen that Tukey is superior to Bonferroni here; Tukey is also superior to Scheffé for this purpose. Thus, using Scheffé's method to test pairwise comparisons sacrifices power. Scheffé's method should not be used unless at least one of the comparisons to be tested is complex.

Modifications of Scheffé's Method

When population variances are unequal, it may be desirable to use a separate variances modification of Scheffé's method for testing comparisons. Such a modification was proposed by Brown and Forsythe (1974). Several simulation studies have suggested that their modification successfully controls α_{EW} when variances are heterogeneous.[9]

The Brown-Forsythe procedure is based on the same F statistic we have repeatedly seen previously when a separate variances approach is taken. Specifically, an observed F is calculated as

$$F = \frac{(\hat{\psi})^2}{\sum_{j=1}^{a}(c_j^2/n_j)s_j^2} \qquad \text{(10, repeated)}$$

As we mentioned in Chapter 4, the corresponding denominator degrees of freedom are arduous, so a computer program will invariably be invoked for their calculation. Later in this chapter, Tables 5.16 and 5.17 will show the formula for these degrees of freedom in the context of summarizing our recommended multiple-comparison procedures. The observed F from Equation 10 is compared to a critical F equal to $(a-1)F_{.05;a-1,df}$.

Notice that this is the same critical F as used in Scheffé's method (see Equation 22), except that the denominator degrees of freedom are no longer simply equal to $N-a$.

Numerical Example

We will return once again to our hypertension data to illustrate Scheffé's procedure. Suppose that four contrasts are of interest: (1) drug versus biofeedback, (2) drug versus diet, (3) biofeedback versus diet, and (4) the average of each of the first three treatments versus the combined treatment. You may realize that these are exactly the same contrasts we examined with the Bonferroni method, as shown in Table 5.5. Now, however, we will suppose that we have chosen these specific contrasts after having seen the data, in which case the Bonferroni method is no longer applicable. Because the contrasts are now post hoc and because they include a complex comparison, Scheffé is the method of choice.The current version of SPSS provides significance tests and confidence intervals for complex and pairwise comparisons using Scheffé's method under an assumption of homogeneity. Ironically, however, as of this writing PROC GLM and PROC ANOVA in SAS provide the capability to apply Scheffé only to pairwise comparisons, despite the fact that the situation that calls for Scheffé involves complex comparisons.

The top portion of Table 5.10 shows SPSS output for all pairwise comparisons. We can obviously use this output to find tests and confidence intervals for our pairwise contrasts of interest. Below the SPSS output are hand calculations needed for forming confidence intervals for the complex contrast as well as for all contrasts when we decide not to assume homogeneity. As in our discussion of the Bonferroni method, we can conceptualize each confidence interval as

$$\text{estimate} \pm (\text{critical value})(\text{estimated standard error}) \qquad \text{(12, repeated)}$$

Notice that the estimate and the estimated standard error for a contrast do not depend on which multiple-comparison procedure we are using, so these two values are the same for any given contrast as they were for Bonferroni or Tukey. As usual, what we must take into account when using Scheffé is the appropriate critical value. Thus, the calculations seen in Table 5.10 differ from those we saw earlier in Table 5.5 only in that the critical values differ. In particular, the critical value for Scheffé is larger than the critical value for Bonferroni (notice how this is consistent with Table 5.9), so it necessarily follows that Scheffé intervals will be wider than Bonferroni intervals. Indeed, comparing Table 5.10 with Table 5.5 shows that each Scheffé interval is wider than the corresponding Bonferroni interval. This is the price we must pay for choosing our contrasts of interest post hoc instead of planned. Nevertheless, in these particular data, regardless of whether our four contrasts were post hoc or planned, we find that only in the case of the complex contrast can we make a confident statement of directionality, namely that the combined group is more effective than the average of the other three groups.

TABLE 5.10
POST HOC CONTRASTS USING SCHEFFÉ'S METHOD

Multiple Comparisons

Dependent Variable: sbp

Scheffé

(I) group	(J) group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	5.8333	4.6676	.673	−8.3971	20.0638
	3	3.3333	4.6676	.915	−10.8971	17.5638
	4	15.0000*	4.6676	.036	.7696	29.2304
2	1	−5.8333	4.6676	.673	−20.0638	8.3971
	3	−2.5000	4.6676	.962	−16.7304	11.7304
	4	9.1667	4.6676	.307	−5.0638	23.3971
3	1	−3.3333	4.6676	.915	−17.5638	10.8971
	2	2.5000	4.6676	.962	−11.7304	16.7304
	4	11.6667	4.6676	.135	−2.5638	25.8971
4	1	−15.0000*	4.6676	.036	−29.2304	−.7696
	2	−9.1667	4.6676	.307	−23.3971	5.0638
	3	−11.6667	4.6676	.135	−25.8971	2.5638

* The mean difference is significant at the .05 level.

Simultaneous Confidence Intervals
Intervals for pairwise comparisons are shown directly in the output. Hand computations are required for the complex comparison and when homogeneity of variance is not assumed.

Assuming Homogeneity of Variance			
Contrast[a]	*Computation*[b]	*95% Confidence Interval*	
4	$(35.83/3) \pm (\sqrt{9.30})(11.4331/3)$	0.32	23.57
Without Assuming Homogeneity of Variance			
Contrast[a]	*Computation*[b]	*95% Confidence Interval*	
1	$5.8333 \pm (\sqrt{12.21})(5.7237)$	−14.17	25.83
2	$3.3333 \pm (\sqrt{11.58})(4.0838)$	−10.56	17.23
3	$-2.500 \pm (\sqrt{13.05})(5.2836)$	−21.59	16.59
4	$(35.83/3) \pm (\sqrt{10.23})(9.0955/3)$	2.25	21.64

[a]The first three contrasts are pairwise. In order, they are (1) Group 1 minus Group 2, (2) Group 1 minus Group 3, and (3) Group 2 minus Group 3. The fourth contrast is the average of the first three groups minus the fourth group.

[b]The critical values used in the hand computation of confidence intervals without assuming homogeneity of variance are obtained by multiplying by $a - 1 = 3$ the tabled values for F with 3 numerator degrees of freedom and the adjusted denominator degrees of freedom shown in Table 5.5 where homogeneity of variance is not assumed. For example, for contrast 1 we have $3F(3,8) = 3(4.07) = 12.21$.

OTHER MULTIPLE-COMPARISON PROCEDURES

Although the Bonferroni, Tukey, and Scheffé multiple-comparison procedures are probably the most widely used and most generally appropriate techniques, we have already seen that they are far from the only ones that have been developed. In particular, referring back to Table 5.2, notice that we have now described appropriate procedures for establishing direction and magnitude of mean differences for the first three columns of the table, that is, for examining (1) planned, (2) pairwise, and (3) complex comparisons.

We now proceed to describe procedures for the final two columns of Table 5.2. First, Dunnett's procedure is particularly useful when one of the groups in a study is a control group. Second, a variation of Dunnett's method can be used when the purpose of the study is to identify the best treatment group.

Dunnett's Procedure for Comparisons With a Control

In some studies, the primary tests of interest may involve comparing one of the groups with each of the other $a - 1$ groups individually. For example, a researcher might plan to compare each of $a - 1$ different treatments with a control group. Although the Bonferroni procedure could be used, Dunnett (1955) developed a test that is more powerful in this situation, which is often referred to as "many-one" testing because many groups are each compared to one other group. (Do not confuse this procedure with Dunnett's T3, which is an entirely different procedure we presented earlier for performing pairwise comparisons with unequal variances.) This method is used much less often in behavioral research than Bonferroni, Tukey, or Scheffé, primarily because Dunnett's method does not accommodate tests comparing the $a - 1$ treatment groups to one another.

There is no change in the calculation of the observed F test statistic for Dunnett's procedure. As in the methods previously encountered, however, the critical value is altered to maintain the α_{FW} level at .05. Appendix Tables 6 and 7 provide the necessary critical values for two-tailed and one-tailed tests, respectively. The columns of the table correspond to the number of groups, including the control group. The entry in the table must be squared to establish a critical value for the F statistic. In other words, the entries in the table are critical t values against which to judge an observed t statistic. Confidence intervals are formed as usual, the only difference being the choice of critical value.

Numerical Example

Suppose the only questions of interest in the hypertension example involved comparing the combined treatment to each of the separate treatments. In other words, suppose we have no interest in comparing the separate treatment groups to one another. Such a "many-one" set of comparisons is exactly the situation where Dunnett's method is appropriate. We will illustrate two-tailed tests and two-sided confidence intervals, although a case could be made for one-tailed tests and one-sided intervals here.

Table 5.11 shows SAS output for comparing the combined treatment against each of the three separate treatments. According to Dunnett's method, the 95% confidence interval for the difference between the combined treatment and drug therapy alone (the first treatment) ranges from approximately 3 to 27. We cannot unequivocally establish the direction of difference for the combined therapy versus either diet or biofeedback. Notice that the Dunnett intervals shown in Table 5.11 are all narrower than the corresponding Tukey intervals we saw in Table 5.7. The increased precision of the Dunnett intervals reflects the advantage we have gained in being willing to forgo consideration of how the separate therapy conditions compare to one another. Thus, in practice,

TABLE 5.11
CONTRASTS OF EACH SEPARATE TREATMENT VERSUS COMBINED TREATMENT

The GLM Procedure

Dunnett's t *Tests for SBP*

Note. This test controls the Type 1 experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error degrees of freedom	20
Error mean square	65.35833
Critical value of Dunnett's *t*	2.54043
Minimum significant difference	11.858

**** Comparisons significant at the 0.05 level.*

Group Comparison	*Difference Between Means*	*Simultaneous 95% Confidence Interval*	
1–4	15.000	3.142	26.858***
3–4	11.667	−0.191	23.524
2–4	9.167	−2.691	21.024

there is a trade-off here between precision and the number of contrasts we choose to consider. As Exercise 12 at the end of the chapter shows, the Dunnett intervals are also narrower than Bonferroni intervals for this same set of contrasts.

Procedures for Comparisons With the Best

Especially in applied settings, the fundamental question of interest may be to establish which treatment is best, that is, either the treatment with the largest or the smallest population mean, depending on which end of the scale is most desirable. Hsu (1996) describes a modification of Dunnett's procedure that can be used to answer this question. We will see that we can continue to use the same Dunnett critical value we developed in the immediately preceding discussion. The primary difference is that in the previous use of Dunnett, we knew in advance which group was the control group, whereas in the application we are about to develop here, we will not know in advance of collecting the data which single group we should focus on. Thus, the precise details of the procedure must be modified, even though we will continue to use much of what we developed in the previous section.

From another perspective, the goal here is similar to and yet not quite the same as that of examining pairwise comparisons. Tukey's HSD and similar methods in the second column of Table 5.2 implicitly assume that we are interested in comparing every possible pair of groups. However, when the goal is to identify the best group, we are not interested in pairwise differences that do not include the best group. Thus, although we could use Tukey's HSD in an attempt to identify the best group, we would implicitly be including comparisons of no real interest for our specific purpose and therefore considering a larger family of contrasts than necessary. The practical implication is that the method we will present for identifying the best treatment will yield greater power and precision than Tukey's HSD. Thus, Tukey's method is inefficient for identifying the best treatment. Indeed, its inefficiency for this purpose may explain why some researchers have resorted to using such alternatives as Newman–Keuls and Duncan in an effort

to increase power and precision. However, if the goal is to identify the best treatment, we should use the modification of Dunnett's method that we will now present.

Our remaining presentation of this method will consist of three sections. First, we will present the rationale for the method in more detail. Second, because most computer software does not include an explicit option for identifying the best treatment, we will spend more time than usual on the mechanics of performing the test. Third, we will conclude with a numerical example based on our hypertension data.

We will now consider the logic behind the method Hsu (1996) presents for identifying the best treatment group. For example, suppose we wonder whether Group 1 might be the best group in a study. For the moment, let's assume that higher scores are better on the dependent variable. (The same basic logic applies when lower scores are better, but as we will see later, we then have to reverse our comparisons.)

Our question is whether μ_1 is the largest population mean. Equivalently, we could ask whether μ_1 is larger than all other population means in the study. Formally, the question centers around an expression of the form

$$\mu_1 - \max_{j \neq 1} \mu_j$$

The term after the minus sign is to be understood as the largest population mean among the remaining $a - 1$ groups. In theory, it is trivial with this formulation to identify the best treatment. If

$$\mu_1 - \max_{j \neq 1} \mu_j > 0$$

it immediately follows that the first group is in fact the best. On the other hand, if

$$\mu_1 - \max_{j \neq 1} \mu_j < 0$$

it similarly follows that some other group has a larger mean than μ_1, in which case the first group is not the best. Notice that we said that this formulation makes our answer clear—in theory. The complication here is that we do not know the population means, but instead have access only to sample data.

Fortunately, we can solve this problem by forming a confidence interval for the difference between the population mean of Group 1 and the population mean of the best group other than Group 1. As usual, we will be especially interested in discovering whether this interval contains zero. For example, if the interval lies entirely above zero, we can be 95% confident that the population mean of Group 1 is larger than the population mean of any other group, implying that Group 1 is indeed the best treatment. On the other hand, if the interval lies entirely below zero, we can be 95% confident that the population mean of Group 1 is smaller than that of some other group, in which case we can conclude that Group 1 is not the best treatment. Finally, if the interval straddles zero, the mean of Group 1 might be either larger or smaller than the largest mean of the remaining groups, in which case Group 1 may or may not be the best treatment. In this final scenario, we cannot rule out Group 1 as a possibility, but we also cannot be 95% confident that it is truly best.

This same approach is then applied to each group in the study. The end result will be a set of intervals that show plausible values for the difference between each individual group mean and the best of the remaining group means. As a consequence, in the simplest case, we will be able to identify one group as the best with some desired level of confidence (typically 95%). However,

in some situations, one group will not stand out from the others, in which case all we can do is to identify a subset of groups, any one of which may be best in the population. In other words, in this scenario, we can eliminate some groups as possibly being best, but we cannot reduce the number of contenders to a single group. Of course, when group differences are all small in the sample, it may happen that we cannot be certain that any groups are truly different from one another, in which case we are left with the unfortunate conclusion that no groups can be eliminated from contention as possibly the best.

Hsu (1996) describes two slightly different techniques for identifying the best treatment, one based on unconstrained intervals and another based on constrained intervals. Unconstrained intervals provide slightly more information but are somewhat wider than constrained intervals. In particular, unconstrained intervals include information about how much better the best group is than other groups, whereas constrained intervals only allow a conclusion that a group is best without estimating the extent to which it is best. Thus, as often happens, there is a trade-off between how much we can say and how precisely we can say it. In our judgment, both variations have their place, so we will present both.

As of this writing, procedures for comparisons with the best are available in only a small number of software programs, including JMP, Minitab, and SAS/INSIGHT. Largely for this reason, we will describe the mechanics of these procedures in more detail than has been typical of other procedures. Specifically, we will present four scenarios: (1) unconstrained intervals where high scores on the dependent variable are best, (2) unconstrained intervals where low scores are best, (3) constrained intervals where high scores are best, and (4) constrained intervals where low scores are best. Ideally, our descriptions will allow you not only to use these methods but also to understand them more fully.

Table 5.12 describes the necessary steps for forming unconstrained intervals. The top half of the table pertains to situations where high scores are best, while the bottom half is relevant when low scores are best. These steps follow the same basic logic we developed for using Dunnett's method to compare a set of treatment groups to a control group. For example, suppose that high scores on the dependent variable are best in a three-group study and that the second group has the largest sample mean. Then the second step consists of using Dunnett's method to find confidence intervals based on the sample mean differences $\bar{Y}_1 - \bar{Y}_2$ and $\bar{Y}_3 - \bar{Y}_2$. Step 3 then involves a similar interval, but this time the interval involves subtracting the next largest mean from $\bar{Y}_2$. Each of the resulting intervals then provides lower and upper limits for the difference between the population mean in question and the largest of all the other $a - 1$ population means.

Table 5.13 describes the necessary steps for forming constrained intervals. As in Table 5.12, the top half of the table pertains to situations where high scores are best, whereas the bottom half is relevant when low scores are best. Once again, the essential logic follows Dunnett's method, although the precise steps are somewhat more complicated. Notice that the fundamental difference between the unconstrained intervals of Table 5.12 and the constrained intervals of Table 5.13 is that the former use the two-tailed version of Dunnett's method, whereas the latter use the one-tailed version. As a result, constrained intervals will always be narrower than unconstrained intervals. However, only unconstrained intervals allow you to estimate how much better the best treatment is than all other treatments, so when you want to be able to make this type of statement, the increased width of intervals may be a small price to pay for being able to say more about the best treatment.

The steps shown in Tables 5.12 and 5.13 are appropriate when two conditions hold: equal sample sizes and homogeneity of variance. When population variances are equal but sample sizes are unequal, complications emerge in calculating appropriate critical values for the procedure, necessitating statistical software. Appropriate methods when variances are unequal await further research.

TABLE 5.12

COMPARISONS WITH THE BEST: STEPS FOR FORMING UNCONSTRAINED INTERVALS

When High Scores on the Dependent Variable Are Best

1. Identify the group with the largest sample mean.
2. Use Dunnett's method to form two-sided confidence intervals, comparing the group with the largest sample mean to each of the other $a - 1$ groups. Be certain to calculate the difference so that the largest mean is subtracted from each other mean, in which case the center of each interval should be below zero. This is the current default in the way both SAS and SPSS implement Dunnett when the group with the largest mean is designated to be the "control" group.
3. The first two steps provide an interval for every group except the group with the largest mean. An interval for this group can be obtained from the results in Step 2. Specifically, identify the group with the second largest mean. Then reverse the signs of the lower and upper endpoints of the confidence interval, comparing this group to the group with the largest mean, to obtain the upper and lower endpoints, respectively, for the group with the largest mean.
4. If the interval for the group with the largest mean does not contain zero, this group is deemed to be the best treatment. If this interval does contain zero, all groups with intervals containing zero are included in a set of groups, any one of which is possibly the best treatment. Any group whose interval is entirely below zero is ruled out as the best treatment.

When Low Scores on the Dependent Variable Are Best

1. Identify the group with the smallest sample mean.
2. Use Dunnett's method to form two-sided confidence intervals, comparing the group with the smallest sample mean to each of the other $a - 1$ groups. Be certain to calculate the difference so that the smallest mean is subtracted from each other mean, in which case the center of each interval should be above zero. This is the current default in the way both SAS and SPSS implement Dunnett when the group with the smallest mean is designated to be the "control" group.
3. The first two steps provide an interval for every group except the group with the smallest mean. An interval for this group can be obtained from the results in Step 2. Specifically, identify the group with the second smallest mean. Then reverse the signs of the lower and upper endpoints of the confidence interval, comparing this group to the group with the smallest mean, to obtain the upper and lower endpoints, respectively, for the group with the smallest mean.
4. If the interval for the group with the smallest mean does not contain zero, this group is deemed to be the best treatment. If this interval does contain zero, all groups with intervals containing zero are included in a set of groups, any one of which is possibly the best treatment. Any group whose interval is entirely above zero is ruled out as the best treatment.

TABLE 5.13

COMPARISONS WITH THE BEST: STEPS FOR FORMING CONSTRAINED INTERVALS

When High Scores on the Dependent Variable Are Best

1. Identify the group with the largest sample mean.
2. Form the lower limit of each confidence interval by using Dunnett's method to form one-sided confidence intervals, comparing the group with the largest sample mean to each of the other $a - 1$ groups. Choose DUNNETTU in SAS or DUNNETTR in SPSS syntax (shown as "> control" in menu options). Also, be certain to calculate the difference so that the largest mean is subtracted from each other mean, in which case the lower limit of each interval should be below zero. This is the current default in the way both SAS and SPSS implement Dunnett when the group with the largest mean is designated to be the "control" group.
3. The next step is to form the upper limit of each interval. To do so, use Dunnett's method once again to compare the group with the largest sample mean to each of the other $a - 1$ groups. However, this time choose DUNNETTL in SAS as well as SPSS syntax (also shown as "< control" in SPSS menu options). As before, be certain to calculate the difference so that the largest mean is subtracted from each other mean. Again, this is the current default in both SAS and SPSS when the group with the largest mean is designated to be the control group.

(*Continued*)

TABLE 5.13
(Continued)

4. Consider the value calculated for each upper limit in Step 3. Any interval with a negative upper limit is assigned a new upper limit of zero. Intervals with positive upper limits retain their original upper limit.
5. The preceding steps provide an interval for every group except the group with the largest mean. An interval for this group can be obtained from the results in Step 4. Specifically, identify the group with the second largest mean. Then reverse the signs of the lower and upper endpoints (after having assigned the new upper limit of zero in Step 4, if relevant) of the confidence interval comparing this group to the group with the largest mean, to obtain the upper and lower endpoints, respectively, for the group with the largest mean.
6. If the interval for the group with the largest mean has a lower limit of zero, this group is deemed to be the best treatment. If the interval contains zero, all groups with intervals containing zero are included in a set of groups, any one of which is possibly the best treatment. Any group whose interval has an upper limit of zero is ruled out as the best treatment.

When Low Scores on the Dependent Variable Are Best

1. Identify the group with the smallest sample mean.
2. Form the upper limit of each confidence interval by using Dunnett's method to form one-sided confidence intervals comparing the group with the smallest sample mean to each of the other $a - 1$ groups. Choose DUNNETTL in SAS as well as SPSS (also shown as "< control" in SPSS menu options). Also, be certain to calculate the difference so that the smallest mean is subtracted from each other mean, in which case the upper limit of each interval should be above zero. This is the current default in the way both SAS and SPSS implement Dunnett when the group with the smallest mean is designated to be the control group.
3. The next step is to form the lower limit of each interval. To do so, use Dunnett's method once again to compare the group with the smallest sample mean to each of the other $a - 1$ groups. However, this time choose DUNNETTU in SAS or DUNNETTR in SPSS syntax (shown as "> control" in menu options). As before, be certain to calculate the difference so that the smallest mean is subtracted from each other mean. Again this is the current default in both SAS and SPSS when the group with the smallest mean is designated to be the control group.
4. Consider the value calculated for each lower limit in Step 3. Any interval with a positive lower limit is assigned a new lower limit of zero. Intervals with negative lower limits retain their original lower limit.
5. The preceding steps provide an interval for every group except the group with the smallest mean. An interval for this group can be obtained from the results in Step 4. Specifically, identify the group with the second smallest mean. Then reverse the signs of the lower and upper endpoints (after having assigned the new lower limit of zero in Step 4, if relevant) of the confidence interval, comparing this group to the group with the smallest mean, to obtain the upper and lower endpoints, respectively, for the group with the smallest mean.
6. If the interval for the group with the smallest mean has an upper limit of zero, this group is deemed to be the best treatment. If the interval contains zero, all groups with intervals containing zero are included in a set of groups, any one of which is possibly the best treatment. Any group whose interval has a lower limit of zero is ruled out as the best treatment.

Numerical Example

We will once again rely on our blood pressure data to illustrate how to perform comparisons with the best. We will first show how to form unconstrained intervals, and then we will show how to form constrained intervals. Remember that low blood pressure scores are best, so we will be referring to the bottom portions of Tables 5.12 and 5.13. Readers desiring an opportunity to practice the procedures when high scores are best should know that Exercise 17 at the end of the chapter provides this practice. In addition, we will use SAS to obtain Dunnett intervals for these

data, but the same values should be available from any statistical software that will form two-sided and one-sided Dunnett intervals.

We will begin by assuming that we want to form unconstrained confidence intervals to find which treatment produces the lowest mean blood pressure readings. The bottom half of Table 5.12 shows that the first step is to identify the group with the smallest sample mean. In our case, that is the combined treatment group, which has a mean of 81.83. Step 2 requires that we use Dunnett's method to compare this group with each of the other groups. Table 5.14 shows SAS output providing two-sided Dunnett's intervals. Notice that the values shown in Table 5.14 are exactly the same as those we saw earlier in Table 5.11, because in both cases we are comparing the combined group to each of the other groups. However, the remaining Steps and ultimate interpretation of the intervals is somewhat different in the two cases. Step 3 requires that we find the group with the second smallest sample mean, which is the biofeedback group, with a mean of 91.00. Reversing the sign of the limits of this interval in the SAS output yields an additional interval for the combined group, with a lower limit of −21.024 and an upper limit of 2.691. Notice that this interval contains zero, so according to Step 4, all groups with intervals containing zero should be regarded as possibly the best treatment. Given our data, it is plausible that either biofeedback, diet, or the combined treatment is the best treatment. In other words, with 95% confidence, we can only rule out drug therapy as possibly the best treatment. Further, notice we can be 95% confident that the combined therapy is at most 2.69 units less effective than the best treatment, whereas the biofeedback and diet treatments may be as much as 21.02 and 23.52 units less effective, respectively. Even though our single best guess is that the combined treatment is best, the data are not strong enough for us to eliminate biofeedback and diet from consideration. A larger sample size would have given us more power to detect the best treatment.

Suppose we are not interested in estimating the extent to which the best treatment is truly best. In this case, we should form constrained intervals. The bottom half of Table 5.13 shows that the first step is to find the group with the smallest sample mean, which is the combined therapy

TABLE 5.14
TWO-SIDED DUNNETT INTERVALS TO IDENTIFY THE BEST TREATMENT FOR HYPERTENSION

The GLM Procedure

Dunnett's t *Tests for SBP*

Note. This test controls the Type 1 experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error degrees of freedom	20
Error mean square	65.35833
Critical value of Dunnett's t	2.54043
Minimum significant difference	11.858

*** Comparisons significant at the 0.05 level.

Group Comparison	*Difference Between Means*	*Simultaneous 95% Confidence Interval*	
1–4	15.000	3.142	26.858***
3–4	11.667	−0.191	23.524
2–4	9.167	−2.691	21.024

TABLE 5.15
ONE-SIDED DUNNETT INTERVALS TO IDENTIFY THE BEST TREATMENT FOR HYPERTENSION

Upper Limits			
Group Comparison	*Difference Between Means*	*Simultaneous 95% Confidence Limits*	
1–4	15.000	−Infinity	25.233
3–4	11.667	−Infinity	21.899
2–4	9.167	−Infinity	19.399
Lower Limits			
Group Comparison	*Difference Between Means*	*Simultaneous 95% Confidence Limits*	
1–4	15.000	4.767	Infinity
3–4	11.667	1.434	Infinity
2–4	9.167	−1.066	Infinity
Final Confidence Limits			
		Lower Limit	*Upper Limit*
Group 1	Drug	0.00	25.23
Group 2	Biofeedback	−1.07	19.40
Group 3	Diet	0.00	21.90
Group 4	Combined	−19.40	1.07

group in our data. The top half of Table 5.15 shows one-sided Dunnett upper limits obtained from SAS in accordance with Step 2 of Table 5.13. The next section of Table 5.15 shows lower limits obtained as described in Step 3. The final section of the table shows the results of applying Steps 4 and 5, where values of zero are assigned and an interval is formed for the combined treatment group. The intervals for both the combined and the biofeedback groups contain zero, so we can be 95% confident that one of them is the best treatment. In other words, we can rule out drug and diet as possibly being the best treatment. Notice that for these data, forming constrained intervals allowed us to rule out more treatments than we were able to do with unconstrained intervals. However, if the data had allowed us to identify a single group as best, constrained intervals would not have allowed us to estimate how much better this group is than the remaining groups, reflecting the trade-off we have previously mentioned.

Fisher's LSD (Protected *t*)

We finish this section with a description of yet one other method, Fisher's least significant difference (LSD). Notice that this method appears in the second row of Table 5.2. Because it does not allow us to establish confident directions, much less confidence intervals, we generally recommend against its use. It may seem strange, therefore, that we have decided to include it in this chapter. Our primary purpose in including it is to illustrate why we believe it suffers from a serious shortcoming. However, another reason for including it is that Levin, Serlin, and Seaman (1994) have shown that this general shortcoming is not a problem in the very special case of three groups. For reasons we will explain momentarily, in this specific situation, Fisher's LSD may be useful.

Fisher (1935) developed a procedure known as Fisher's LSD method or, equivalently, as the protected *t*-test method. Although this technique was developed nearly 20 years earlier than the other multiple-comparison methods described in this chapter, it is still in use today, partly because of its simplicity. The test proceeds in two stages. First, the omnibus null hypothesis is tested. If it is not rejected, no further tests are performed; if it is rejected, the process continues to the second stage. At this stage, individual contrasts among the groups are tested using an α_{PC} level of .05 for each contrast. Traditionally, only pairwise contrasts are tested in the second step (Fisher developed the procedure for this purpose), but as Keppel (1982) and Levin et al. (1994) point out, the logic of the procedure does not rule out complex comparisons in the second stage. The LSD has another advantage besides its simplicity—the critical values at the second stage are less than those for the Bonferroni, Tukey, or Scheffé methods. This is true because the LSD uses an α_{PC} of .05, whereas the others use an α_{EW} of .05. The implication is that the LSD has more power to detect true differences. You may be thinking that this is an unfair comparison, however, because the objective of a multiple-comparisons procedure is to control α_{EW} at .05.

We now consider whether the LSD succeeds in maintaining α_{EW} at .05. The basic logic behind the LSD is that, because it requires a statistically significant omnibus F value, in only 5 of every 100 studies (in the long run) will the process mistakenly lead to Stage 2, when in fact all the population means are equal. Even if the second stage were to always produce a statistically significant result, only 5% of the time would a Type I error be committed, because the omnibus test of Stage 1 protects tests performed in the second stage. It seems that by requiring the omnibus test to be significant before testing individual contrasts, the goal of maintaining α_{EW} at .05 is accomplished. Indeed, the reasoning to this point is valid. As long as all a population means are equal (i.e., the complete null hypothesis is true), α_{EW} is held at .05 by the LSD. However, suppose that some but not all of the null hypothesis is true. For example, with $a = 11$, it might happen that the first 10 groups all have identical population means. The 11th treatment, however, has been included in the study because prior evidence suggests it to be very different from the first 10. If the 11th group is different enough from the first 10, the omnibus null hypothesis will be rejected with a probability approaching 1.0 for a large enough sample size. Conceivably, then, the second stage of the LSD will be reached with a high probability. Now, however, the LSD offers no further protection for contrasts among the first 10 groups. In other words, there is no protection for that part of the complete null hypothesis that is true. If all pairwise contrasts among the 10 truly identical groups are performed, the probability of at least one significant result using $\alpha_{PC} = .05$ is approximately .60. Thus, in 60 of every 100 such experiments (in the long run), a Type I error would be committed. Thus, the LSD fails to maintain α_{EW} at .05, except in the special case where the entire null hypothesis is true. None of the other approaches (Bonferroni, Tukey, Scheffé, or Dunnett) suffer from this limitation—they all maintain α_{EW} at .05 under all circumstances as long as the basic ANOVA assumptions are satisfied. Indeed, as mentioned earlier, Tukey's method is referred to as the wholly significant difference (WSD) precisely because the whole set of pairwise contrasts is protected at .05 with his approach.

Levin et al. (1994) realized that the flaw disclosed in the previous paragraph is no longer applicable in the special case of three groups. In particular, they point out that there are three relevant patterns of possible population means with three groups:

a. All population means are equal.
b. Two of the population means are equal to one another (but the third mean is different).
c. All three population means are different from one another.

Levin et al. (1994) then explain why the LSD performs satisfactorily in any of these three situations. First, if all population means are equal, the omnibus test will reject only 5% of the time,

so the procedure moves to the second stage only 5% of the time, which means that the experimentwise error rate will be at most 5%. Second, if one of the means is different from the other two (i.e., case *b*), the second stage will generally be reached more than 5% of the time. However, once it is reached, there is only one true null hypothesis, so testing it at the 5% level is perfectly appropriate and succeeds in controlling α_{EW} at .05. Third, if all three population means are different from one another, there are no true null hypotheses to test, so no Type I errors can be committed. Thus, Fisher's LSD successfully controls α_{EW} at .05 in the special case of three groups, and can be recommended in this specific situation. However, even in this favorable situation, it does not allow the formation of appropriate simultaneous confidence intervals, so its use is restricted to hypothesis testing.

Several other methods have been developed that in many ways represent a compromise between the LSD and Tukey's approach for testing pairwise comparisons. Two of the most widely used compromises are the Newman–Keuls and the Duncan procedures. Both are referred to as either multiple-range tests or layered methods, because they involve testing ranges between groups in a layered fashion. Unfortunately, both of these methods suffer from the same liability as Fisher's LSD (with more than three groups) in that they do not necessarily control α_{EW} at .05. Thus, we recommend against their use when the goal is to control α_{EW} at .05. More recently, other alternatives, such as REGW-Q and REGW-F, have been developed and included in statistical software. These methods succeed in properly controlling α_{EW} at .05. Their major drawback is that although they succeed in properly identifying true differences in population means, they do not allow a statement of confident directions or the formation of confidence intervals. In this sense, they are appropriate only in a rather special case where the only goal is to establish differences without regard to the direction and magnitude of those differences. Nevertheless, some researchers favor them for pairwise comparisons because they offer more power than Tukey's HSD.

FALSE DISCOVERY RATE

A recurring theme throughout this chapter has been the importance of the distinction between the per-comparison error rate and the experimentwise error rate. In 1995, Benjamini and Hochberg introduced yet another way of conceptualizing error rates in the presence of more than one test per experiment. They called their new conceptualization the "false discovery rate" (FDR). Although as of this writing it is still a newcomer in the long history of multiple comparisons, it has nevertheless begun to attract quite a bit of attention, because in a practical sense it represents a compromise between the more traditional ideas of the per-comparison error rate and the experimentwise error rate.

The motivation behind the FDR stems from a desire to control the proportion of false discoveries that appear in the literature. By "false discoveries," Benjamini and Hochberg (1995) mean false rejections of the null hypothesis (i.e., Type I errors). What is different about the FDR is that Benjamini and Hochberg (1995) define it in terms of the total number of discoveries, that is, in terms of the total number of rejected hypotheses, some of which may have been falsely rejected, but others of which may have been correctly rejected. The idea here is to make certain that the proportion of false discoveries relative to total discoveries is kept appropriately small, such as at the 5% level. It turns out that defining error rates in this way leads to methods that are less prone to Type I errors than are methods that simply control the per-comparison alpha level, but are often more powerful than methods that control the experimentwise alpha level. In a moment we will look at FDR in more detail, but its essence can be thought of in terms of the following fraction:

$$\frac{\text{number of false rejections}}{\text{number of total rejections}}$$

The goal of the FDR is to control the average value of this fraction. In other words, we want to use a method that ensures that only a small fraction of the discoveries (i.e., statistically significant results) we report are unbeknownst to us false discoveries.

We need to consider three additional aspects of the FDR:

1. Exactly how is controlling FDR different from controlling α_{EW}?
2. What methods provide appropriate control of the FDR?
3. Under what types of circumstances might the FDR be most appropriate?

In order to understand the FDR, it is first helpful to revisit exactly what it means to control α_{EW}. Suppose an experiment involves testing m null hypotheses. Define a variable P whose value depends on the outcome of the experiment in the following manner:

$P = 0$ if no true null hypotheses are rejected.
$P = 1$ if one or more true null hypotheses are rejected.

Thus, in any single experiment, P will have a value of either 0 or 1. As a thought experiment, we could contemplate replicating any given study over and over again many times. Each time we will record the value of P. (Of course, in actual research, we never know with certainty which if any null hypotheses are true, but remember that this is a thought experiment.) Across replications, our values of P will be a string of zeros and ones, where we would hope that the zeros predominate. What it means from this perspective to control α_{EW} at .05 is that the average value of P is at most .05. In other words, in the long run, only 1 time out of 20 will an experiment contain a false rejection of the null hypothesis. Thus, from this perspective, if our goal is to control α_{EW} at .05, we need to use a method that leads to an average value of P no larger than .05.

How does this compare to the FDR? Once again, consider an experiment consisting of m hypothesis tests. Now, however, define a variable Q whose value depends on the outcome of the experiment in the following manner:

$Q = 0$ if no true null hypotheses are rejected.
$Q = V/(V + S)$ if V true null hypotheses are incorrectly rejected and S false null hypotheses are correctly rejected.

Notice the relationship between Q and the fraction we defined a few paragraphs ago. Whenever $V + S$ is greater than zero, Q is literally the fraction of falsely rejected null hypotheses relative to the total number of rejections. In cases where $V + S$ equals 0, the fraction is itself considered to be zero, because no Type I errors have been committed. In general, then, we can regard Q as an index of the number of false rejections relative to the number of total rejections. Whereas α_{EW} is defined as the average value of P, FDR is defined as the average value of Q.

Thus, to better understand the FDR, we need to better understand Q. First, let's consider a situation where (unbeknownst to us) all m null hypotheses we are testing are true. Notice then that $S = 0$ in every replication, because we have no false null hypotheses available to be rejected. Because $S = 0$ in every replication, Q takes on either of two values. If we do not reject any true null hypotheses in a specific replication, $Q = 0$. However, $Q = 1$ anytime we reject one or more true null hypotheses. As a result, in this situation, we will get a string of Q values consisting of

zeros and ones. The crucial point to realize is that this string of Q values will be identical to the string of P values we would obtain for the same replications. Thus, FDR and α_{EW} will be the same here. In general, when all null hypotheses being tested are true, FDR and α_{EW} are equal to one another.

What if some of the null hypotheses we are testing are false? In particular, suppose we test 10 null hypotheses, where 7 of them are true and 3 are false. Suppose we decided to use the Bonferroni approach, and thus set the per-comparison alpha level at .005 in order to control α_{EW} at .05. What would our string of values for P look like? The string necessarily consists of zeros and ones, and in this case the proportion of ones would be something less than .05. What about our string of values for Q? Assuming we still use a per-comparison alpha level of .005, Q will equal zero for any specific replication in the string if and only if $P = 0$ for that same replication. But what about when $P = 1$? In *this* instance, Q will tend to be less than 1, because S will tend to be greater than 0 to the extent that we have reasonable power for testing our three false null hypotheses. For example, as the power of each of these tests approaches 1.0, S will approach a value of 3 for every replication, which means that the ratio of V to $V + S$ is likely to be 0.25 (except when we falsely reject more than one of the seven true null hypotheses in the same study). The practical point is that the average value of the string of Q values will be noticeably less than the average value of the comparable string of P values. In other words, when some of the hypotheses we are testing are false, FDR will be less than α_{EW}. Thus, we can allow α_{EW} to exceed .05 and yet succeed in controlling FDR at .05 when some of our hypotheses should be rejected because they are false. In such a situation, deciding to control FDR instead of α_{EW} provides more power to detect those hypotheses that should in fact be rejected.

To pursue the difference between FDR and α_{EW} one step further, we will continue to consider our hypothetical study involving a total of 10 hypotheses, 7 of which are true and 3 of which are false. Suppose we used a critical value for each comparison so that 80% of the time none of the true null hypotheses were rejected, but 20% of the time at least one false rejection occurs. In this instance, 80% of our P values would be zero, and 20% would be 1. Thus, α_{EW} would equal 0.20, a rate likely to be deemed too large. Now consider our values of Q using the same critical value. As with P, 80% of our Q values will equal zero. But what about the other 20%? For simplicity, suppose we have power close to 1.00 for rejecting the three false hypotheses. Then the remaining 20% of Q values will tend to equal 0.25 (i.e., $1/(1 + 3)$, because only rarely will we reject more than one of the true nulls, and we will almost always reject all three of the false nulls). What will the overall average of Q equal here? If 80% of values are zero and 20% of values are 0.25, the overall average will equal 0.05. Thus, the same critical value that produces an experimentwise error rate of 0.20 produces an FDR of 0.05. We would almost certainly decide that our decision rule is unacceptable in terms of α_{EW} and thus be forced to increase our critical value, inevitably lowering power to detect true differences. However, the original critical value would presumably be judged as acceptable in terms of FDR, so we would proceed with the smaller critical value and hence enjoy the benefits of greater power.

Suppose we decide to use FDR as the basis for controlling our error rate in the face of multiple hypotheses. What method can we use to ensure proper control of the FDR? Before answering this question, remember that we have seen throughout this chapter that the choice of method depends on two dimensions: the definition of error control and the type of contrast being tested. Recall that these two dimensions are reflected in Table 5.2. The situation is no different when controlling the FDR in the sense that different methods are appropriate for different types of contrasts. However, the situation for the FDR is further complicated by the fact that this is a newly emerging area of research, so new suggestions about ideal methods for controlling the FDR continue to appear. We cannot possibly do justice to the entire topic here, so instead we have chosen to

focus on a single method that is applicable for either of two situations: testing independent (i.e., orthogonal) hypotheses and testing pairwise comparisons.

The name of the specific method we will present is Benjamini and Hochberg's Linear Step Up procedure. Benjamini and Hochberg (1995) have shown that this method controls the FDR for independent tests, and Keselman, Cribbie, and Holland (1999) have shown that it controls FDR for pairwise comparisons. However, other methods are needed for testing other patterns of hypotheses where test statistics may be correlated with one another. Benjamini, Drai, Elmer, Kafkafi, and Golani (2001) describe a slightly more complicated procedure that controls FDR for any arbitrary pattern of correlated test statistics.

The Linear Step Up (LSU) procedure begins by rank ordering the p values obtained from testing each of the m hypotheses. Specifically, the p values are ordered from smallest to largest. The LSU begins with the largest p value. If this p value is less than .05, all null hypotheses are rejected. If this p value is greater than .05, the hypothesis corresponding to this p value is not rejected, and the procedure moves on to consider the next p value in the ordering. If this second largest p value is less than $((m - 1)/m).05$, this hypothesis and all remaining hypotheses are rejected. However, if the p value is greater than $((m - 1)/m).05$, then this hypothesis is not rejected, and the procedure moves on once again to consider the next p value in the ordering. No hypotheses at all are rejected if the procedure reaches the smallest p value and it exceeds $.05/m$.

Notice that the alpha levels of the LSU procedure reveal how it represents a compromise between per-comparison control and experimentwise control. In the LSU, the largest p value is compared against .05, which corresponds to the per-comparison alpha level. The smallest p value is compared against $.05/m$, which corresponds to the experimentwise alpha level. All other p values are compared against intermediate values. The LSU procedure can be described more formally as consisting of the following steps, for testing a total of m hypotheses:

1. Rank order the p values from smallest to largest.
2. Assign an index value i to each p value to represent its rank. The smallest p value is assigned an index value i of 1, the next smallest is assigned an index value i of 2, and so forth, up to the largest p value, which receives an index value of m.
3. Compare each p value to $(i/m)\alpha$, where α is the desired value of the FDR (usually .05).
4. Define k to be the largest value of i for which the observed p value is less than $(i/m)\alpha$.
5. All hypotheses with index values less than or equal to k are rejected, whereas all hypotheses with index values larger than k are not rejected.
6. If all observed p values are larger than $(i/m)\alpha$, no hypotheses are rejected.

To see how the LSU procedure works, let's consider an example where $m = 5$ and we want to control FDR at .05. Suppose we obtained the following five p values: .029, .026, .167, .048, and .001. How can we assess the statistical significance of each corresponding hypothesis? We begin by ranking the p values from smallest to largest, which yields .001, .026, .029, .048, and .167. We then compare the largest p value of .167 to (5/5).05. The observed p value is larger than .05, so we do not reject this hypothesis and instead move on to the next p value. We compare this p value of .048 to (4/5).05. Once again, the observed p value of .048 is larger than .040, so we again move on to the next p value. This time we compare the p value of .029 to (3/5).05. Now .029 is less than .030, so we reject the hypothesis corresponding to this p value. Formally, we have determined that $k = 3$ for these data, so according to Step 5, we reject the hypotheses associated with the three smallest p values. Notice that we do so even though the second smallest p value of .026 would not have been less than (2/5).05. As soon as we identify a p value smaller than $(i/m).05$, testing stops.

Under what types of circumstances might we prefer to define error control in terms of FDR instead of α_{EW}? Any answer to this question is necessarily somewhat subjective. Before providing any type of answer, it is essential to realize that the FDR is a different way of thinking about error control. Only when all null hypotheses are true is FDR identical to α_{EW}. Whenever at least one null hypothesis is false, choosing to control FDR at .05 necessarily means that α_{EW} is being allowed to exceed .05. Whether this is good or bad depends on one's perspective. At any rate, such authors as Benjamini and Hochberg (1995); Keselman et al. (1999); and Williams, Jones, and Tukey (1999) have suggested a variety of situations where FDR might be preferred to α_{EW}. Examples include exploratory studies where discovering unexplored results may be more important than in later stages of confirmatory research. Similarly, in neuroscience, hypotheses involving brain activity may be tested in a very large number of areas of the brain, in which case controlling α_{EW} may wreak havoc with power. Instead of ignoring the problem altogether and using a per-comparison alpha level, FDR may be an appropriate compromise, especially when statistically significant results obtained for regions for a single individual can then be compared across individuals. Yet another example is genetics (including behavioral genetics), where a genome wide association study (GWAS) may investigate the possible importance of thousands of SNPs in a single study. It is also worth noting that Shaffer, Kowalchuk, and Keselman (2013) show that the Newman-Keuls method of multiple comparisons may be preferable to the LSU procedure for controlling FDR although it has largely been regarded as undesirable because it does not properly control α_{EW}.

CHOOSING AN APPROPRIATE PROCEDURE

The practical implication of this chapter is that we have recommended five multiple-comparison procedures for general use: Bonferroni, Tukey, Scheffé, Dunnett, and Hsu. Figure 5.1 is a flowchart (i.e., a decision tree) that is intended to provide a general guideline for choosing from among these five procedures in a particular situation. We should stress the phrase "general guideline" here; it is important that you understand the principles we have presented in the chapter so that you can use this flowchart as an aid in choosing a technique without being at the complete mercy of a set of mechanical rules to follow blindly. For example, you may want to consider the FDR as an appropriate measure of error control, but we have not attempted to incorporate this conceptualization into Figure 5.1. Similarly, keep in mind that Fisher's LSD properly controls α_{EW} in the special case of three groups, but even there it does not allow for simultaneous confidence intervals. For this reason, we have excluded it from the figure, but if you are content with hypothesis tests unaccompanied by confidence intervals, you may want to consider the LSD with three groups. For these and other reasons, the flowchart is meant to help you organize some of the issues of this chapter, but it cannot possibly reflect all of the nuances we have discussed.

We would be remiss if we did not take a moment to explain where the omnibus test of Chapter 3 fits into this framework. In particular, it is important that you understand that the omnibus test is not a prerequisite for the Bonferroni, Tukey, Dunnett, or Hsu procedures. Instead, these methods should be viewed as substitutes for the omnibus test because they control α_{EW} at the desired level all by themselves. Requiring a significant omnibus test before proceeding to perform any of these analyses, as is sometimes done, only serves to lower α_{EW} below the desired level (Bernhardson, 1975) and hence inappropriately decreases power, although it could be said that they are used for somewhat different but often complementary purposes.

The proper role of the omnibus test is that it addresses a global question of whether any differences exist among the groups. At times, this is an important question in its own right, independently of the likely desire to proceed by investigating which specific groups or combinations of

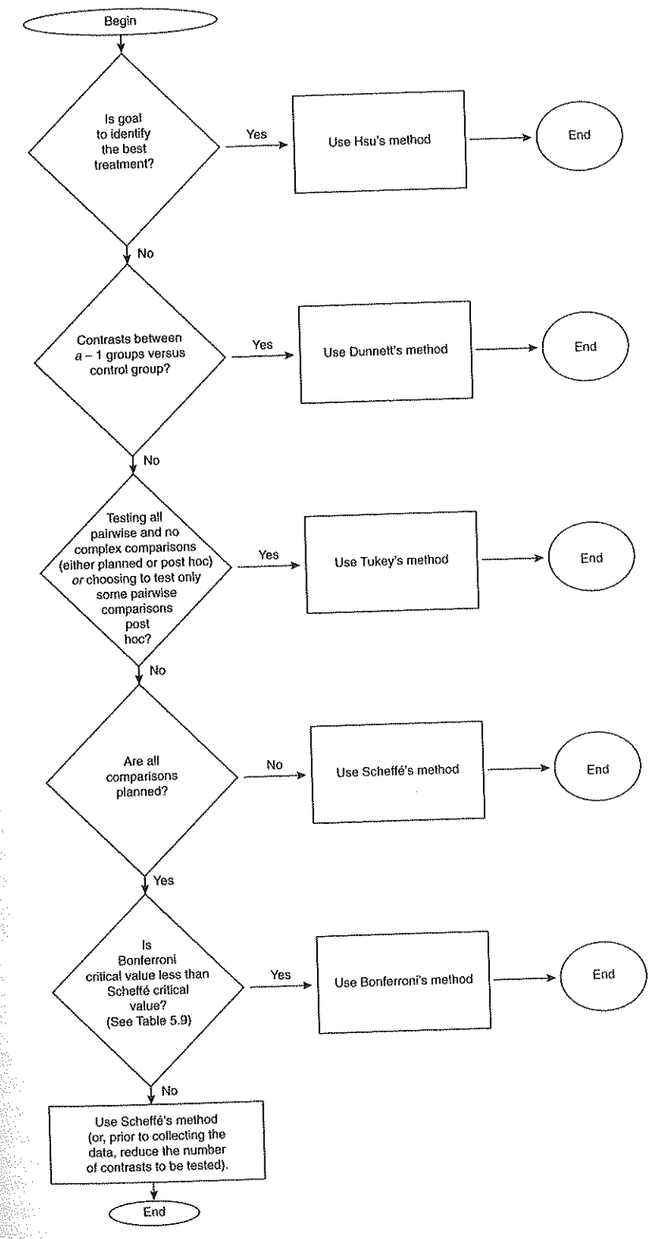

FIG. 5.1 General guideline for choosing a multiple-comparison procedure.