Objectives

To extend the analysis of variance by examining ways of making comparisons within a set of means.

Contents

12.1 Error Rates
12.2 Multiple Comparisons in a Simple Experiment on Morphine Tolerance
12.3 A Priori Comparisons
12.4 Confidence Intervals and Effect Sizes for Contrasts
12.5 Reporting Results
12.6 Post Hoc Comparisons
12.7 Tukey’s Test
12.8 Which Test?
12.9 Computer Solutions
12.10 Trend Analysis
A significant $F$ in an analysis of variance is simply an indication that not all the population means are equal. It does not tell us which means are different from other means. As a result, the overall analysis of variance often raises more questions than it answers. We now face the problem of examining differences among individual means, or sets of means, for the purpose of isolating significant differences or testing specific hypotheses. We want to make statements of the form $\mu_1 = \mu_2 = \mu_3$, and $\mu_4 = \mu_5$, but the first three means are different from the last two, and all of them are different from $\mu_6$.

Many different techniques for making comparisons among means are available, and the list grows each year. Here we will limit coverage to only the most common and useful ones. A thorough discussion of this topic can be found in Hochberg and Tamhane (1987) and Toothaker (1991). Keselman, Holland, and Cribbie (2005) offer an excellent review of some of the newer methods. The papers by Games (1978a, 1978b) are also helpful, as is the paper by Games and Howell (1976) on the treatment of unequal sample sizes.

It may be helpful to the reader to understand how this chapter has changed through various editions. The changes largely reflect the way people look at experimental results and focus on the most useful procedures. Originally this chapter covered a few of the most common test procedures and left it at that. Then as time went on I kept adding to the number of procedures and focused at length on ways to make many individual comparisons among means. But in this edition I am returning to covering only a few test procedures, which are the ones that almost everyone now uses. I am also emphasizing the fact that we should direct our attention to those differences we really care about and not fill our results section with all of the other differences that we can test but don’t actually care about. This philosophy carries over to calculating effect sizes and selecting appropriate error terms.

If you are interested in a few specific comparisons, then taking a standard multiple comparison test such as Tukey’s (which is an excellent test for the purpose for which it was designed) and testing every conceivable pairwise null hypothesis is a very poor idea. It wastes power, it often leads to the use of inappropriate error terms, it gives poor measures of effect size, and generally confuses what is often a clear and simple set of results. The fact that you are able to do something is rarely a sufficient reason for actually doing it. I will cover Tukey’s test because it is such a commonly used approach, but I think that it is more in line with Tukey’s general approach to statistics to lay the main emphasis elsewhere.

### 12.1 Error Rates

The major issue in any discussion of multiple-comparison procedures is the question of the probability of Type I errors. Most differences among alternative techniques result from different approaches to the question of how to control these errors. The problem is in part technical, but it is really much more a subjective question of how you want to define the error rate and how large you are willing to let the maximum possible error rate be.

Here we will distinguish two basic ways of specifying error rates, or the probability of Type I errors.¹ (Later we will discuss an alternative view of error rates called the False Discovery Rate, which has received a lot of attention in the last few years.) In doing so, we shall use the terminology that has become more or less standard since an extremely important unpublished paper by Tukey in 1953. (See also Ryan, 1959; O’Neil and Wetherill, 1971.)

---

¹ There is another error rate called the error rate per experiment (PE), which is the expected number of Type I errors in a set of comparisons. The error rate per experiment is not a probability, and we typically do not attempt to control it directly. We can easily calculate it, however, as $PE = ca$, where $c$ is the number of comparisons and $\alpha$ is the per comparison error rate.
Section 12.1  Error Rates

**Error Rate Per Comparison (PC)**

We have used the error rate per comparison (PC) in the past and it requires little elaboration. It is the probability of making a Type I error on any given comparison. If, for example, we make a comparison by running a $t$ test between two groups and we reject the null hypothesis because our $t$ exceeds $t_{0.05}$, then we are working at a per comparison error rate of .05.

**Familywise Error Rate (FW)**

When we have completed running a set of comparisons among our group means, we will arrive at a set (often called a family) of conclusions. For example, the family might consist of the statements

\[ \mu_1 < \mu_2 \]
\[ \mu_3 < \mu_4 \]
\[ \mu_1 < (\mu_3 + \mu_4)/2 \]

The probability that this family of conclusions will contain at least one Type I error is called the familywise error rate (FW). Many of the procedures we will examine are specifically directed at controlling the FW error rate, and even those procedures that are not intended to control FW are still evaluated with respect to what the level of FW is likely to be.

In an experiment in which only one comparison is made, both error rates will be the same. As the number of comparisons increases, however, the two rates diverge. If we let $\alpha'$ represent the error rate for any one comparison and $c$ represent the number of comparisons, then

- Error rate per comparison (PC): $\alpha = \alpha'$
- Familywise error rate (FW): $\alpha = 1 - (1 - \alpha')^c$
  (if comparisons are independent)

If the comparisons are not independent, the per comparison error rate remains unchanged, but the familywise rate is affected. In most situations, however, $1 - (1 - \alpha')^c$ still represents a reasonable approximation to FW. It is worth noting that the limits on FW are $PC \leq FW \leq ca$; in most reasonable cases FW is in the general vicinity of $ca$. This fact becomes important when we consider the Bonferroni tests.

**The Null Hypothesis and Error Rates**

Until now we have been speaking as if the null hypothesis in question is what is usually called the complete, or omnibus, null hypothesis ($\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_6$). This is the null hypothesis tested by the overall analysis of variance. In many, if not most, experiments, however, nobody is seriously interested in the complete null hypothesis; rather, people are concerned about a few more restricted null hypotheses, such as ($\mu_1 = \mu_2 = \mu_3$, $\mu_4 = \mu_5$, $\mu_6 = \mu_7$), with differences among the various subsets. If this is the case, the problem becomes more complex, and it is not always possible to specify FW without knowing the pattern of population means. We will need to take this into account in designating the error rates for the different tests we shall discuss.

---

2 This error rate is frequently referred to, especially in older sources, as the “experimentwise” error rate. However, Tukey’s term “familywise” has become more common. In more complex analyses of variance, the experiment often may be thought of as comprising several different families of comparisons.
A Priori versus Post Hoc Comparisons

In the earlier editions of this book I carefully followed tradition and distinguished between a priori comparisons, which are chosen before the data are collected, and post hoc comparisons, which are planned after the experimenter has collected the data, looked at the means, and noted which of the latter are far apart and which are close together. This is a traditional distinction, but one that seems to be less and less important to people who run such comparisons. In practice the real distinction seems to come down to the difference between deliberately making a few comparisons that are chosen because of their theoretical or practical nature (and not just because the means looked different) and making comparisons among all possible pairs of means. I am going to continue to make the a priori/post hoc distinction because it organizes the material nicely and is referred to frequently, but keep in mind that the distinction is a rather fuzzy one.

To take an example that you have already seen, we will return to the study by Foa, Rothbaum, Riggs, and Murdock (1991), which formed Exercise 11.8 in the preceding chapter. Foa et al. (1991) conducted a study evaluating four different types of therapy for rape victims. The Stress Inoculation Therapy (SIT) group received instructions on coping with stress. The Prolonged Exposure (PE) group went over the events in their minds repeatedly. The Supportive Counseling (SC) group was taught a general problem-solving technique. Finally, the Waiting List (WL) control group received no therapy.

Suppose you ran that study and your real goal was to compare Stress Inoculation Therapy (SIT) with Prolonged Exposure (PE), and you also wanted to check that these two therapies considered as a set were more effective than the two control conditions (SC and WL). But then suppose that you looked at your results and found that SC, which you expected to be quite ineffective, appeared to do much better than the Waiting List control. That is a comparison that you never planned to make, but it looks as if you should. So what do you do? That is no longer a simple question. The first two comparisons are clearly a priori—that’s why you ran the study. But the third comparison is post hoc—you weren’t intending to make it until you saw the results. The traditional answer is to use a post hoc test to make all pairwise comparisons. Otherwise you risk making a Type I error just because of the unusual difference between the control groups. An alternative approach, not as well grounded in theory, would be to treat this last comparison as a priori as well but to be a bit cautious if it comes out to be significant, especially if the probability is near your critical cutoff. We will consider both approaches in what follows.

It is important to realize that when we speak of a priori tests, we commonly mean a relatively small set of comparisons. If you are making all possible pairwise comparisons among several means, for example, it won’t make any difference whether that was planned in advance or not. (I would wonder, however, if you really wanted to make all possible comparisons.)

Significance of the Overall F

Some controversy surrounds the question of whether one should insist that the overall F on treatments be significant before conducting multiple comparisons between individual group means. In the past, the general advice was that without a significant group effect, individual comparisons were inappropriate. In fact, the rationale underlying the error rates for Fisher’s least significant different test, to be discussed in Section 12.4, required overall significance.

However, this is a case where the general advice is wrong. The logic behind most of our multiple comparison procedures does not require overall significance before making specific comparisons. First of all, the hypotheses tested by the overall test and a multiple-comparison
Multiple Comparisons in a Simple Experiment on Morphine Tolerance

In discussing the various procedures, it will be helpful to have a data set to which each of the approaches can be applied. We will take as an example a study similar to an important experiment on morphine tolerance by Siegel (1975). Although the data are fictitious and a good deal of liberty has been taken in describing the conditions, the means (and the significance of the differences among the means) are the same as those in Siegel’s paper. It will be necessary to describe this study in some detail, but the example is worth the space required. It will be to your advantage to take the time to understand the hypotheses and the treatment labels.

Morphine is a drug that is frequently used to alleviate pain. Repeated administrations of morphine, however, lead to morphine tolerance, in which morphine has less and less of an effect (pain reduction) over time. (You may have experienced the same thing if you eat spicy food very often. You will find that the more you eat it, the hotter you have to make it to taste the way it did when you started.) A common experimental task that demonstrates morphine tolerance involves placing a rat on an uncomfortably warm surface. When the heat becomes too uncomfortable, the rat will lick its paws, and the latency of the paw-lick is used as a measure of the rat’s sensitivity to pain. A rat that has received a single morphine injection typically shows a longer paw-lick latency, indicating a reduced pain sensitivity. The development of morphine tolerance is indicated by a progressive shortening of paw-lick latencies (indicating increased sensitivity, or decreased insensitivity) with repeated morphine injections.

Siegel noted that there are a number of situations involving drugs other than morphine in which conditioned (learned) drug responses are opposite in direction to the unconditioned (natural) effects of the drug. For example, an animal injected with atropine will usually show a marked decrease in salivation. However if physiological saline (which should have no effect whatsoever) is suddenly injected (in the same physical setting) after repeated injections of atropine, the animal will show an increase in salivation. It is as if the animal were compensating for the anticipated effect of atropine. In such studies, it appears that a learned compensatory mechanism develops over trials and counterbalances the effect of the drug. (You experience the same thing if you leave the seasoning out of...
food that you normally add seasoning to. It will taste unusually bland, though the Grape Nuts you eat for breakfast does not taste bland—and I hope that you don’t put seasoning on Grape Nuts.)

Siegel theorized that such a process might help to explain morphine tolerance. He reasoned that if you administered a series of pretrials in which the animal was injected with morphine and placed on a warm surface, morphine tolerance would develop. Thus, if you again injected the subject with morphine on a subsequent test trial, the animal would be only as sensitive to pain as would a naive animal (one who had never received morphine) because of the tolerance that has fully developed. Siegel further reasoned that if on the test trial you instead injected the animal with physiological saline in the same test setting as the normal morphine injections, the conditioned hypersensitivity that results from the repeated administration of morphine would not be counterbalanced by the presence of morphine, and the animal would show very short paw-lick latencies and heightened sensitivity. Siegel also reasoned that if you gave the animal repeated morphine injections in one setting but then tested it with morphine in a new setting, the new setting would not elicit the conditioned compensatory hypersensitivity to counterbalance the morphine. As a result, the animal would respond as would an animal that was being injected for the first time. Heroin is a morphine derivative. Imagine a heroin addict who is taking large doses of heroin because he has built up tolerance to it. If his response to this now large dose were suddenly that of a first-time (instead of a tolerant) user, because of a change of setting, the result could be, and often is, lethal. We’re talking about a serious issue here, and drug overdoses often occur in novel settings.

You may think that an experiment conducted 30 years ago, which is before most of the readers of this book were born, is too old to be interesting. But a quick search of Google will reveal a great many recent studies that have derived directly from Siegel’s early work. A particularly interesting one by Mann-Jones, Ettinger, Baisden, and Baisden has shown that a drug named Dextromethorphan can counteract morphine tolerance. That becomes interesting when you learn that Dextromethorphan is an important ingredient in cough syrup. This suggests that heroin addicts don’t want to take cough syrup any more than they want to administer heroin in novel environments. The study can be found at www.eou.edu/psych/re/morphinetolerance.doc.

Our version of Siegel’s experiment is based on the predictions just outlined. The experiment involved five groups of rats. Each group received four trials, but the data for the analysis come from only the critical fourth (test) trial. The groups are designated by indicating the treatment on the first three trials and then the treatment on the fourth trial. Group M-M received morphine on the first three trials in the test setting and then again on the fourth trial in the same test setting. This is the standard morphine-tolerant group, and, because morphine tolerance develops very quickly, we would expect to see normal, or at least near-normal, levels of pain sensitivity by that fourth trial. Group M-S received morphine on the first three trials but then received saline on the fourth trial (in the same test setting). These animals would be expected to be hypersensitive to the pain stimulus because the conditioned hypersensitivity would not be balanced by any compensating effects of morphine. Group M(cage)-M (abbreviated Mc-M) received morphine on the first three trials in their home cage but then received morphine on the fourth trial in the standard test setting, which was new to them. For this group, cues originally associated with morphine injection were not present on the test trial, and therefore, according to Siegel’s model, the animals should not exhibit morphine tolerance on that trial. The fourth group (group S-M) received saline on the first three trials (in the test setting) and morphine on the fourth trial. These animals would be expected to show the least sensitivity to pain because there has been no opportunity for morphine tolerance to develop. Finally, group S-S received saline on all four trials.
If Siegel’s model is correct, group S-M should show the longest latencies (indicating least sensitivity), whereas group M-S should show the shortest latency (most sensitivity). Group Mc-M should resemble group S-M, because cues associated with group Mc-M’s first three trials would not be present on the test trial. Groups M-M and S-S should be intermediate. Whether group M-M will be equal to group S-S will depend on the rate at which morphine tolerance develops. The pattern of anticipated results is

\[
\begin{align*}
& S-M \\
& Mc-M \\
& M-M \neq S-S \\
& M-S
\end{align*}
\]

The “?” indicates no prediction. The dependent variable is the latency (in seconds) of paw-licking.

The results of this experiment, which closely follow Siegel’s results, are presented in Table 12.1a, and the overall analysis of variance is presented in Table 12.1b. Notice that the within-group variances are more or less equal (a test for heterogeneity of variance was not significant), and there are no obvious outliers. The overall analysis of variance is clearly significant, indicating differences among the five treatment groups.

### Magnitude of Effect

We can calculate \( \eta^2 \) for these data as \( \frac{SS_{\text{treat}}}{SS_{\text{total}}} = \frac{3497.60}{4617.60} = .76 \), indicating that treatment differences account for 76% of the variation in the study. A nearly unbiased estimate would be \( \omega^2 \), which would be

\[
\omega^2 = \frac{SS_{\text{treat}} - (k - 1)MS_{\text{error}}}{SS_{\text{total}} + MS_{\text{error}}} = \frac{3497.60 - 4(32)}{4617.60 + 32} = \frac{3369.6}{4649.6} = 0.72
\]

Both estimates indicate that group treatment differences account for a very substantial proportion of the variation in this study.
### 12.3 A Priori Comparisons

There are two reasons for starting our discussion with a priori comparisons and \( t \) tests. In the first place, standard \( t \) tests between pairs of means can be a perfectly legitimate method of comparison. Second, the basic formula for \( t \), and minor modifications on it, are applicable to a large number of procedures (a priori and post hoc), and a review at this time is useful.

As we have seen, a priori comparisons (also called contrasts) are planned before the data have been collected. There are several different kinds of a priori comparison procedures, and we will discuss them in turn.

#### Multiple \( t \) Tests

One of the simplest methods of running preplanned comparisons is to use individual \( t \) tests between pairs of groups. In running individual \( t \) tests, if the assumption of homogeneity of variance is tenable, we usually replace the individual variances, or the pooled variance estimate, with \( MS_{\text{error}} \) from the overall analysis of variance and evaluate the \( t \) on \( df_{\text{error}} \) degrees of freedom. When the variances are heterogeneous but the sample sizes are equal, we do not use \( MS_{\text{error}} \), but instead use the individual sample variances and evaluate \( t \) on \( 2(n-1) \) degrees of freedom. Finally, when we have heterogeneity of variance and unequal sample sizes, we use the individual variances and correct the degrees of freedom using the Welch–Satterthwaite approach (see Chapter 7). (In Chapter 7 we saw that Hayes and Cai (2007) argued against the basic idea of pooling variances when running \( t \) tests on independent means. However, when we have an analysis of variance with several groups, we lose a considerable amount of power by using the variance estimates from only the groups in question. If the sample variances appear to be homogeneous I would use the overall \( MS_{\text{error}} \) and its degrees of freedom, in computing my \( t \) values, but there is room for disagreement on this.)

The indiscriminate use of multiple \( t \) tests is typically brought up as an example of a terrible approach to multiple comparisons. In some ways, this is an unfair criticism. It is a terrible thing to jump into a set of data and lay waste all around you with \( t \) tests on each and every pair of means that looks as if it might be interesting. The familywise error rate will be outrageously high. However, if you have only one or two comparisons to make and if those comparisons were truly planned in advance (you cannot cheat and say, “Oh well, I would have planned to make them if I had thought about it”), the \( t \)-test approach has much to recommend it. With only two comparisons, for example, the maximum \( FW \) would be approximately 0.10 if each comparison were run at \( \alpha = .05 \), and would be approximately 0.02 if each comparison were run at \( \alpha = .01 \). For a discussion of the important role that individual contrasts can play in an analysis, see Howell (2008c).

In the study on morphine tolerance described previously, we would probably not use multiple \( t \) tests simply because too many important comparisons should be considered. (In fact, we would probably use one of the post hoc procedures for making all pairwise comparisons unless we can restrict ourselves to relatively few comparisons.) For the sake of an example, however, consider two fundamental comparisons that were clearly predicted by the theory and that can be tested easily with a \( t \) test. The theory predicted that a rat that had received three previous morphine trials and was then tested in the same environment using a saline injection would show greater pain sensitivity than would an animal that had always been tested using saline. This involves a comparison of group M-S with group S-S. Furthermore, the theory predicted that group M-M would show less sensitivity to pain than
would group M-M, because the former would be tested in an environment different from the one in which it had previously received morphine. Because the sample variances are similar and the sample sizes are equal, we will use $MS_{\text{error}}$ as the pooled variance estimate and will evaluate the result on $df_{\text{error}}$ degrees of freedom.

Our general formula for $t$, replacing individual variances with $MS_{\text{error}}$, will then be

$$t = \frac{X_i - X_j}{\sqrt{MS_{\text{error}}/n} + \sqrt{MS_{\text{error}}/n}} = \frac{X_i - X_j}{\sqrt{2MS_{\text{error}}/n}}$$

Substituting the data from our example, the contrast of group M-S with group S-S yields

$$X_{M-S} = 4.00 \quad X_{S-S} = 11.00 \quad MS_{\text{error}} = 32.00$$

$$t = \frac{X_{M-S} - X_{S-S}}{\sqrt{2MS_{\text{error}}/n}} = \frac{4.00 - 11.00}{\sqrt{2(32.00)/8}} = \frac{-7}{\sqrt{8}} = -2.47$$

And group Mc-M versus group M-M yields

$$X_{Mc-M} = 29.00 \quad X_{M-M} = 10.00 \quad MS_{\text{error}} = 32.00$$

$$t = \frac{X_{Mc-M} - X_{M-M}}{\sqrt{2MS_{\text{error}}/n}} = \frac{29.00 - 10.00}{\sqrt{2(32.00)/8}} = \frac{19}{\sqrt{8}} = 6.72$$

Both of these obtained values of $t$ would be evaluated against $t_{0.025}(35) = 2.03$, and both would lead to rejection of the corresponding null hypothesis. We can conclude that with two groups of animals tested with saline, the group that had previously received morphine in the same situation will show a heightened sensitivity to pain. We can also conclude that changing the setting in which morphine is given significantly reduces, if it does not eliminate, the conditioned morphine-tolerance effect. Because we have tested two null hypotheses, each with $\alpha = .05$ per comparison, the $FW$ will approach $0.10$ if both null hypotheses are true, which seems quite unlikely. In fact, given the position of Jones and Tukey (2000) that it is highly unlikely that either null hypothesis would be true, or that we can only incorrectly find a significant difference in the wrong direction, the probability of an error in this situation is at most .05. That is important to keep in mind when we speak of the advantages and disadvantages of individual contrasts on pairs of means.

The basic $t$ test that we have just used is the basis for almost everything to follow. I may tweak the formula here or there, and I will certainly use a number of different tables and decision rules, but it remains your basic $t$ test—even when I change the formula and call it $q$.

### Linear Contrasts

The use of individual $t$ tests is a special case of a much more general technique involving what are known as linear contrasts. In particular, $t$ tests allow us to compare one group with another group, whereas linear contrasts allow us to compare one group or set of groups with another group or set of groups. Although we can use the calculational procedures of linear contrasts with post hoc tests as well as with a priori tests, they are discussed here

---

1. The words “contrast” and “comparison” are used pretty much interchangeably in this context.
under a priori tests because that is where they are most commonly used. Keep in mind that
tests on contrasts are just an extension of standard $t$ tests.

To define linear contrasts, we must first define a **linear combination**. A linear combi-
nation of means takes the form

$$L = a_1X_1 + a_2X_2 + \cdots + a_kX_k = \sum a_jX_j$$

This equation simply states that a linear combination is a weighted sum of treatment means.
If, for example, the $a_j$ were all equal to 1, $L$ would just be the sum of the means. If, on the
other hand, the $a_j$ were all equal to 1/$k$, then $L$ would be the mean of the means.

When we impose the restriction that $\sum a_j = 0$, a linear combination becomes what is
called a **linear contrast**. By convention we designate the fact that it is a linear contrast by
replacing “$L$” with the Greek psi ($\psi$). With the proper selection of the $a_j$, a linear contrast
is very useful. It can be used, for example, to compare one mean with another mean, giving
the same result as a $t$ test, or the mean of one condition with the combined mean of several
conditions. As an example, consider three means ($X_1$, $X_2$, and $X_3$). Letting $a_1 = 1$, $a_2 = -1$, and $a_3 = 0$, $\sum a_j = 0$,

$$\psi = (1)X_1 + (\text{-}1)X_2 + (0)X_3 = X_1 - X_2$$

In this case, $\psi$ is simply the difference between the means of group 1 and group 2, with the
third group left out. If, on the other hand, we let $a_1 = 1/2$, $a_2 = 1/2$, and $a_3 = -1$, then

$$\psi = (1/2)X_1 + (1/2)X_2 + (\text{-}1)X_3 = \frac{X_1 + X_2}{2} - X_3$$

in which case $\psi$ represents the difference between the mean of the third treatment and the
average of the means of the first two treatments.

**Sum of Squares for Contrasts**

One of the advantages of linear contrasts is that they can be converted to sums of squares
very easily and can represent the sum of squared differences between the means of sets of
treatments. If we write

$$\psi = a_1X_1 + a_2X_2 + \cdots + a_kX_k = \sum a_jX_j$$

it can be shown that

$$SS_{\text{contrast}} = \frac{n\psi^2}{\sum a_j^2} = \frac{n(\sum a_jX_j)^2}{\sum a_j^2}$$

is a component of the overall $SS_{\text{treat}}$ on 1 $df$, where $n$ represents the number of scores per
treatment.\(^4\)

Suppose we have three treatments such that

$$n = 10 \quad X_1 = 1.5 \quad X_2 = 2.0 \quad X_3 = 3.0$$

\(^4\) For unequal sample sizes, $SS_{\text{contrast}} = \frac{\psi^2}{\sum (a_j^2/n_j)}$
For the overall analysis of variance,

\[ SS_{\text{treat}} = n \sum (X_j - \bar{X})^2 = 10 \left[ (1.5 - 2.167)^2 + (2 - 2.167)^2 + (3 - 2.167)^2 \right] \]

\[ = 10[0.4449 + 0.0278 + 0.6939] = 11.667 \]

Suppose we wanted to compare the average of treatments 1 and 2 with treatment 3. Let \( a_1 = 1/2, a_2 = 1/2, a_3 = -1. \) Then

\[ \psi = \sum a_j \bar{X}_j = \left( \frac{1}{2} \right)(1.5) + \left( \frac{1}{2} \right)(2.0) + (-1)(3.0) = -1.25 \]

\[ SS_{\text{contrast}} = \frac{n \psi^2}{\sum a_j^2} = \frac{10(-1.25)^2}{1.5} = \frac{15.625}{1.5} = 10.417 \]

This sum of squares is a component of the overall \( SS_{\text{treat}} \) on 1 df. We have 1 df because we are really comparing two quantities (the mean of the first two treatments with the mean of the third treatment).

Now suppose we obtain an additional linear contrast comparing treatment 1 with treatment 2. Let \( a_1 = 1, a_2 = -1, a_3 = 0. \) Then

\[ \psi = \sum a_j \bar{X}_j = (1)(1.5) + (-1)(2.0) + (0)(3.0) = -0.5 \]

\[ SS_{\text{contrast}} = \frac{n \psi^2}{\sum a_j^2} = \frac{10(-0.5)^2}{2} = \frac{2.5}{2} = 1.25 \]

This \( SS_{\text{treat}} \) is also a component of \( SS_{\text{treat}} \) on 1 df. In addition, because of the particular contrasts that we chose to run,

\[ SS_{\text{treat}} = SS_{\text{contrast}_1} + SS_{\text{contrast}_2} \]

\[ 11.667 = 10.417 + 1.25 \]

We say that we have completely partitioned \( SS_{\text{treat}}. \)

**The Choice of Coefficients**

In the previous example, it should be reasonably clear why we chose the coefficients we did. They weight the treatment means in what seems to be a logical way to perform the contrast in question. Suppose, however, that we have five groups of equal size and wish to compare the first three with the last two. We need a set of coefficients \( a_j \) that will accomplish this task and for which \( \sum a_j = 0. \) The simplest rule is to form the two sets of treatments and to assign as weights to each set the reciprocal of the number of treatment groups in that set. One arbitrary set of coefficients is then given a minus sign. For example, take the means

\[ \bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4, \bar{X}_5 \]

We want to compare \( \bar{X}_1, \bar{X}_2, \) and \( \bar{X}_3 \) combined with \( \bar{X}_4 \) and \( \bar{X}_5 \) combined. The first set contains three means, so for \( \bar{X}_1, \bar{X}_2, \) and \( \bar{X}_3 \) the \( a_j = 1/3. \) The second set contains two means, so for \( \bar{X}_4 \) and \( \bar{X}_5 \) the \( a_j = 1/2. \) We will let the 1/2s be negative. Then we have

**Means:** \[ \bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_4, \bar{X}_5 \]

**\( a_j: \)\)** \[ 1/3, 1/3, 1/3, -1/2, -1/2 \]

\[ \sum a_j = 0 \]
Then $\sum a_i \overline{X}_i$ reduces to $\frac{1}{3}(\overline{X}_1 + \overline{X}_2 + \overline{X}_3) - \frac{1}{2}(\overline{X}_4 + \overline{X}_5)$, which you can see is the mean of the first three conditions minus the mean of the last two conditions, which is what we want.

(If you go back to Siegel’s experiment on morphine, lump the first three groups together and the last two groups together, and look at the means of the combined treatments, you will get an idea of why this system makes sense.)

There are other ways of setting up the coefficients using whole numbers, and for many purposes you will arrive at the same result. I used to like alternative approaches because I find fractions messy, but using fractional values as I did here, where the sum of the absolute values of all coefficients is equal to 2, has some important implications when it comes to estimating effect sizes. The set of coefficients whose sum of absolute values equals 2 is often referred to as a standard set.

The Test of Significance

We have seen that linear contrasts can be easily converted to sums of squares on 1 degree of freedom. These sums of squares can be treated exactly like any other sums of squares. They happen also to be mean squares because they always have 1 degree of freedom (you are always comparing two quantities), and can thus be divided by $MS_{\text{error}}$ to produce an $F$. Because all contrasts have 1 degree of freedom

$$F = \frac{MS_{\text{contrast}}}{MS_{\text{error}}} = \frac{\frac{1}{n} \sum a_i^2}{MS_{\text{error}}} = \frac{\frac{1}{n} \sum a_i^2 MS_{\text{error}}}{MS_{\text{error}}}$$

This $F$ will have one and $df_{\text{error}}$ degrees of freedom. And if you feel more comfortable with $t$, you can take the square root of $F$ and have a $t$ on $df_{\text{error}}$ degrees of freedom.

For our example, suppose we had planned (a priori) to compare the mean of the two groups for whom the morphine should be maximally effective, either because they had never had morphine (Condition S-M) or because they had received morphine in a different context (Mc-M), with the mean of the other three groups (M-M, S-S, and M-S). We also planned to compare group Mc-M with group M-M, and group M-S with group S-S, for the same reasons given in the discussion of individual $t$ tests. Finally, we planned to compare group M-M with group S-S to see whether morphine tolerance developed to such an extent that animals that always received morphine were no different after only four trials from animals that always received saline. (As we will see shortly, these four contrasts are not independent, but they answer substantive questions.) The analysis is shown in Table 12.2.

Each of these $F$ values can be evaluated against $F_{0.05}(1,35) = 4.12$. As expected, the first three contrasts are significant. The fourth contrast, comparing M-M with S-S, is not significant, indicating that complete morphine tolerance seems to develop in as few as four trials. (Be careful here, as I am acting as if I can prove the null hypothesis, when we know that is not possible.) Note that contrasts 2 and 3 test the same hypotheses that we tested using individual $t$ tests. If you take the square root of the $F$s for these two contrasts, they will equal 6.72 and 2.47, which are precisely the values we obtained for $t$ earlier. This simply illustrates the fact that $t$ tests are a special case of linear contrasts.

---

If we have different numbers of subjects in the several groups, we may need to obtain our coefficients somewhat differently. If the sample sizes differ in non-essential ways, such as when a few subjects are missing at random, the approach above will be the appropriate one. It will not weight one group mean more than another just because the group happens to have a few more subjects. However, if the sample sizes are systematically different, not just different at random, and if we want to give more weight to the means from the larger groups, then we need to do something different. Because there really are very few cases where I can imagine wanting the different sample sizes to play an important role, I have dropped that approach from this edition of the book. However, you can find it in earlier editions and on the Web pages referred to earlier.
With four contrasts, we have an $F_W$ approaching .20 if all null hypotheses are true, which seems highly unlikely.\(^6\) This error rate is uncomfortably high, although some experimenters would accept it, especially for a priori contrasts. One way of reducing the error rate would be to run each comparison at a more stringent level of $\alpha$; for example, $\alpha = .01$. Another alternative would be to use a different a priori procedure, the Bonferroni procedure, which amounts to almost the same thing as the first alternative but is conducted in a more precise manner. We will consider this procedure after we briefly discuss a special type of linear contrast, called orthogonal contrasts. Yet a third way to control $F_W$ is to run fewer contrasts. For example, the comparison of M-M with S-S is probably not very important. Whether

---

\(^6\) I should elaborate on that statement. We know that morphine tolerance is a well-established fact. So, for example, does it seem likely that rats receiving morphine for the first time behave like rats receiving it for the fourth time? I don’t think so. So right off the bat there weren’t four true null hypotheses that could be falsely rejected, so a probability as high as .20 is unreasonable. Without knowing anything more about the study I would be surprised if there are more than two true null hypotheses, in which case the actual $F_W$ error rate should not be above .10.
complete tolerance develops on the fourth trial or on the sixth or seventh trial is of no great theoretical interest. By eliminating that contrast, we could reduce the maximum $FW$ to .15. You should never choose to run contrasts the way you eat peanuts or climb mountains—just because they are there. In general, if a contrast is not important, do not run it.

**Orthogonal Contrasts**

Linear contrasts as they have been defined allow us to test a series of hypotheses about treatment differences. Sometimes contrasts are independent of one another, and sometimes they are not. For example, knowing that $\bar{X}_1$ is greater than the average of $\bar{X}_2$ and $\bar{X}_3$ tells you nothing about whether $\bar{X}_3$ is greater than $\bar{X}_2$ nor whether $\bar{X}_4$ is likely to be greater than $\bar{X}_5$. These contrasts are independent. However, knowing that $\bar{X}_1$ is greater than the average of $\bar{X}_2$ and $\bar{X}_3$ suggests that there is a better than 50:50 chance that $\bar{X}_1$ is greater than $\bar{X}_2$. These two contrasts are not independent. When members of a set of contrasts are independent of one another, they are called **orthogonal contrasts**, and the sums of squares of a complete set of orthogonal contrasts sum to $SS_{treat}$. (If the contrasts are not orthogonal, they contain overlapping amounts of information and do not have this additivity property.) From a calculational point of view, what sets orthogonal contrasts apart from other types of contrasts we might choose is the relationship between the coefficients for one contrast and the coefficients for other contrasts in the set. Other than that, the computations are exactly the same.

**Orthogonal Coefficients**

Given that sample sizes are equal, for contrasts to be orthogonal the coefficients must meet the following three criteria:

1. $\sum a_j = 0$
2. $\sum a_j b_j = 0$

where $a_j$ and $b_j$ are the sets of coefficients for different contrasts. Furthermore, for the $SS_{contr}$ to sum to $SS_{treat}$, we need to add a third criterion:

3. Number of comparisons = number of df for treatments

The first restriction has been discussed already; it results in the contrast’s being a sum of squares. The second restriction ensures that the contrasts are independent of (or orthogonal to) one another, and thus that we are summing nonoverlapping components. The third restriction says nothing more than that if you want the parts to sum to the whole, you need to have all the parts.

At first glance, it would appear that finding sets of coefficients satisfying the requirement $\sum a_j b_j = 0$ would require that we either undertake a frustrating process of trial and error or else solve a set of simultaneous equations. In fact, a simple rule exists for finding orthogonal sets of coefficients; although the rule will not find all possible sets, it will lead to most of them. The rule for forming the coefficients visualizes the process of breaking down $SS_{treat}$ in terms of a tree diagram. The overall $F$ for five treatments deals with all five treatment means simultaneously. That is the trunk of the tree. If we then compare the combination of treatments 1 and 2 with the combination of treatments 3, 4, and 5, we have formed two branches of our tree, one representing treatments 1 and 2 and the other representing treatments 3, 4, and 5. As discussed earlier, the value of $a_j$ for the treatment means on the left will be equal to the reciprocal of the number of treatments in that set, and vice versa, with one of the sets being negative. In this case the coefficients are $(\frac{1}{2}, \frac{1}{2}, -\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3})$ for the five treatments, respectively.

Now that we have formed two limbs or branches of our tree, we can never compare treatments on one limb with treatments on another limb, although we can compare treatments on the same limb. Thus, comparing treatment 3 with the combination of treatments 4 and 5 is an...
example of a legitimate comparison. The coefficients in this case would be \((0, 0, 1, -\frac{1}{2}, -\frac{1}{2})\). Treatments 1 and 2 have coefficients of 0 because they are not part of this comparison. Treatment 3 has a coefficient of 1 because it contains one treatment. Treatments 4 and 5 received coefficients of \(-\frac{1}{2}\) because there are two treatments in that set. The negative signs can be arbitrarily assigned to either side of the comparison.

The previous procedure could be carried on until we have exhausted all possible sets of comparisons. This will occur when we have made as many comparisons as there are \(df\) for treatments. As a result of this procedure, we might arrive at the comparisons and coefficients shown in Figure 12.1. To show that these coefficients are orthogonal, we need to show only that all pairwise products of the coefficients sum to zero. For example,

\[
\sum a_j b_j = \left(\frac{1}{2}\right)(1) + \left(\frac{1}{2}\right)(-1) + \left(-\frac{1}{2}\right)(0) + \left(-\frac{1}{2}\right)(0) + \left(-\frac{1}{2}\right)(0) = 0
\]

and

\[
\sum a_j c_j = \left(\frac{1}{2}\right)(0) + \left(\frac{1}{2}\right)(0) + \left(-\frac{1}{2}\right)(2) + \left(-\frac{1}{2}\right)(-1) + \left(-\frac{1}{2}\right)(-1) = 0
\]

Thus, we see that the first and second and the first and third contrasts are both independent. Similar calculations will show that all the other contrasts are also independent of one another.

These coefficients will lead to only one of many possible sets of orthogonal contrasts. If we had begun by comparing treatment 1 with the combination of treatments 2, 3, 4, and 5, the resulting set of contrasts would have been entirely different. It is important for the experimenter to decide which contrasts she considers important and to plan accordingly. Keep in mind that just because you can arrange coefficients to yield a legitimate contrast doesn’t mean that you actually have to carry out that contrast.

The actual computation of \(F\) or \(t\) with orthogonal contrasts is the same as when we are using nonorthogonal contrasts. Because of this, there is little to be gained by working through an example here. It would be good practice, however, for you to create a complete set of orthogonal contrasts and to carry out the arithmetic. You can check your answers by showing that the sum of the sums of squares equals \(SS_{treat}\).

When I first started teaching and writing about statistics, orthogonal contrasts were a big deal, just as was the distinction between a priori and post hoc tests. Authors went out of their way to impress on you the importance of orthogonality, and the need to feel guilty if you ran comparisons that were not orthogonal. That attitude has changed over the years. Although it is nice to have a set of orthogonal comparisons, in part because they sum to \(SS_{treat}\) people are far more willing to run nonorthogonal contrasts. I would certainly not suggest that you pass up an important contrast just because it is not orthogonal to others that you ran. In fact, the contrasts that I ran earlier are not orthogonal to each other, and that does not worry me much. They address important questions (well, possibly not S-S versus M-M, as I said). Nor should you use a contrast in which you have no interest, just because

\[
(1, 2, 3, 4, 5) \quad \text{Coefficients}
\]

\[
(1, 2) \quad (3, 4, 5) \quad \frac{1}{2} \quad \frac{1}{2} \quad -\frac{1}{3} \quad -\frac{1}{3} \quad -\frac{1}{3}
\]

\[
(1 \quad 2) \quad (3 \quad 4, 5) \quad 1 \quad -1 \quad 0 \quad 0 \quad 0
\]

\[
(4 \quad 5) \quad (1 \quad 2) \quad 0 \quad 0 \quad 1 \quad -\frac{1}{2} \quad -\frac{1}{2}
\]

\[
(4 \quad 5) \quad (3 \quad 4, 5) \quad 0 \quad 0 \quad 0 \quad 1 \quad -1
\]

\[
(1, 2, 3, 4, 5)
\]

\[
(1, 2) \quad (3, 4, 5)
\]

\[
(1 \quad 2) \quad (3 \quad 4, 5)
\]

\[
(4 \quad 5) \quad (1 \quad 2)
\]

\[
(4 \quad 5) \quad (3 \quad 4, 5)
\]

\[
(1, 2, 3, 4, 5)
\]

\[
(1, 2) \quad (3, 4, 5)
\]

\[
(1 \quad 2) \quad (3 \quad 4, 5)
\]

\[
(4 \quad 5) \quad (3 \quad 4, 5)
\]

Figure 12.1 Tree diagram illustrating orthogonal partition of \(SS_{treat}\)
it is part of an orthogonal set. But keep in mind that being nonorthogonal means that these contrasts are not independent of each other.

**Bonferroni t (Dunn’s test)**

I suggested earlier that one way to control the familywise error rate when using linear contrasts is to use a more conservative level of \( \alpha \) for each comparison. The proposal that you might want to use \( \alpha = .01 \) instead of \( \alpha = .05 \) was based on the fact that our statistical tables are set up that way. (Tables do not usually have many critical values of \( t \) for \( \alpha \) between .05 and .01, although statistical software to compute and print them is widely available.) A formal way of controlling \( FW \) more precisely by manipulating the per comparison error rate can be found in a test proposed by Dunn (1961), which is particularly appropriate when you want to make only a few of all possible comparisons. Although this test had been known for a long time, Dunn was the first person to formalize it and to present the necessary tables, and it is sometimes referred to as Dunn’s test. It now more commonly goes under the name Bonferroni \( t \). The Bonferroni \( t \) test is based on what is known as the Bonferroni inequality, which states that the probability of occurrence of one or more events can never exceed the sum of their individual probabilities. This means that when we make three comparisons, each with a probability of \( \alpha' = .05 \) of a Type I error, the probability of at least one Type I error can never exceed \( 3 \times .05 = .15 \). In more formal terms, if \( c \) represents the number of comparisons and \( \alpha' \) represents the probability of a Type I error for each comparison, then \( FW \) is less than or equal to \( ca' \). From this it follows that if we set \( \alpha' = \alpha/c \) for each comparison, where \( \alpha = \) the desired maximum \( FW \), then \( FW \leq ca' = c(\alpha/c) = \alpha \). Dunn (1961) used this inequality to design her test in which each comparison is run at \( \alpha' = \alpha/c \), leaving the \( FW \leq \alpha \) for the set of comparisons. This can be accomplished by using the standard \( t \) test procedure but referring the result to modified \( t \) tables.

The problem that you immediately encounter when you attempt to run each comparison at \( \alpha' = \alpha/c \) is that standard tables of Student’s \( t \) do not provide critical values for the necessary levels of \( \alpha \). If you want to run each of three comparisons at \( \alpha' = \alpha/c = .05/3 = .0167 \), you would need tables of critical values of \( t \) at \( \alpha = .0167 \), or software\(^7\) that will easily compute it. Dunn’s major contribution was to provide such tables. However, we no longer need those tables because all software solutions produce the actual \( p \) value, and if we do the calculations by hand, we can use readily available probability calculators. I will omit them from this edition.

For the Bonferroni test on pairwise comparisons (i.e., comparing one mean with one other mean), define

\[
t' = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{MS_{\text{error}}}{n} + \frac{MS_{\text{error}}}{n}}} = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{2MS_{\text{error}}/n}} \quad \text{or} \quad F' = \frac{\psi^2}{\sum a_j^2MS_{\text{error}}/n}
\]

and evaluate \( t' \) against the critical value of \( t' \) at \( \alpha/c \), which is \( (t_{\alpha/c, \#}) \). Notice that we still use the standard formula for \( t \). The only difference between \( t' \) and a standard \( t \) is the tables used in their \( \alpha/c \) evaluation. With unequal sample sizes but homogeneous variances, replace the \( ns \) in the leftmost equation with \( n_i \) and \( n_j \). With heterogeneity of variance, see the solution by Games and Howell later in this chapter.

To write a general expression that allows us to test any comparison of means, pairwise or not, we can express \( t' \) in terms of linear contrasts.

\[
\psi = \sum a_j \bar{X}_j \quad \text{and} \quad t' = \frac{\psi}{\sqrt{\sum a_j^2MS_{\text{error}}/n}}
\]

This represents the most general form for the Bonferroni t, and it can be shown that if \( \psi \) is any linear combination (not necessarily even a linear contrast, requiring \( \sum a \neq 0 \)), the \( FW \) with \( c \) comparisons is at most \( \alpha \) (Dunn, 1961). To put it most simply, the Bonferroni t runs a regular t test but evaluates the result against a modified critical value of t that has been chosen so as to limit \( FW \).

I would offer one word of caution when it comes to the Bonferroni test and variations on it. These tests are appropriate when you have a limited number of planned contrasts, whether they are pairwise or complex. However SPSS and SAS offer the Bonferroni test only with pairwise post hoc tests, for which it is usually inappropriate. Under that system you automatically set \( c \) equal to the number of possible pairwise comparisons whether you care about all of them or not. If you want to apply such a correction to a planned set of contrasts, you need to specify those contrasts and then evaluate significance on your own in relation to \( a/c \). And to specify those contrast coefficients you will need to use Compare Means/One-way ANOVA and not the univariate procedure. In SAS you will need to use a contrast statement with Proc GLM.

A variation on the Bonferroni procedure was proposed by Šidák (1967). His test is based on the multiplicative inequality \( p(FW) \leq 1 - (1 - \alpha)^c \) and evaluates \( t' \) at \( \alpha' = 1 - (1 - \alpha)^{1/c} \). (This is often called the Dunn-Šidák test.) A comparison of the power of the two tests shows only very minor differences in favor of the Šidák approach, and we will stick with the Bonferroni test because of its much wider use. Many computer software programs, however, provide Šidák’s test.

When we considered linear contrasts earlier in this section, we ran four comparisons, which had a maximum \( FW \) of nearly .20. (Our test of each of those contrasts involved an F statistic but, because each contrast involves 1 df, we can go from t to F and vice versa by means of the relationship \( t = \sqrt{F} \).) If we wish to run those same comparisons but to keep \( FW \) at a maximum of \( .05 \) instead of \( 4 \times .05 = .20 \), we can use the Bonferroni t test. In each case, we will solve for \( t' \) and evaluate \( t' \) against the standard t at \( \alpha' = \alpha/4 \). Taking the pairwise tests first, the calculations follow. For four contrasts the critical value of \( t' \), adjusted for the number of contrasts, will be that value of t that cuts off \( 5/4 \times .0125 = .00625 \) in each tail. If you are using standard software, which prints out p values, you simple change your rejection level to .0125. Or, using the software mentioned in footnote 6, you will find that the cutoff value of \( df = 35 \) is \( t_{.00625,35} = 2.63 \).

The calculations for these tests are shown below. We want to identify contrasts with \( t \approx \pm 2.63 \).

**Mc-M versus M-M:**

\[
t' = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2MS_{\text{error}}}{n}}} = \frac{29.00 - 10.00}{\sqrt{\frac{(2)(32.00)}{8}}} = \frac{19}{\sqrt{8}} = 6.72
\]

**S-S versus M-S:**

\[
t' = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2MS_{\text{error}}}{n}}} = \frac{11.00 - 4.00}{\sqrt{\frac{(2)(32.00)}{8}}} = \frac{7}{\sqrt{8}} = 2.47
\]

**M-M versus S-S:**

\[
t' = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2MS_{\text{error}}}{n}}} = \frac{10.00 - 11.00}{\sqrt{\frac{(2)(32.00)}{8}}} = \frac{-1}{\sqrt{8}} = -0.35
\]

---

*To use Daniel Soper’s excellent probability calculators, choose the link to Student’s t distribution and click on Student’s t calculator.*
The calculations for the more complex contrast, letting the \( a_j = 1/3, 1/3, 1/3, -1/2, -1/2 \) as before, follow.

**S-M and Mc-M versus M-M, S-S and M-S:**

\[
 t' = \frac{\sum_{j} a_j x_j}{\sqrt{\frac{\sum_{j} a_j^2 MS_{error}}{n}}} = \frac{(\frac{1}{3})(24) + \cdots + (-\frac{1}{2})(4)}{\sqrt{(0.833)(32.00)}} = \frac{18.167}{\sqrt{3.3333}} = 9.95
\]

In this case, the first and last contrasts exceed 2.63 and are significant, but the other two are not.\(^9\) Whereas we earlier rejected the hypothesis that groups S-S and M-S were sampled from populations with the same mean, using the more conservative Bonferroni \( t \) test we are no longer able to reject that hypothesis. Here we cannot conclude that prior morphine injections lead to hypersensitivity to pain. The difference in conclusions between the two procedures is a direct result of our use of the more conservative familywise error rate. If we wish to concentrate on per comparison error rates, ignoring \( FW \), then we evaluate each \( t \) (or \( F \)) against the critical value at \( \alpha = .05 \). On the other hand, if we are primarily concerned with controlling \( FW \), then we evaluate each \( t \), or \( F \), at a more stringent level. The difference is not in the arithmetic of the test; it is in the critical value we choose to use. The choice is up to the experimenter.

**Multistage Bonferroni Procedures**

The Bonferroni multiple-comparison procedure has a number of variations. Although these are mentioned here in the context of the analysis of variance, they can be applied equally well whenever we have multiple hypothesis tests for which we wish to control the family-wise error rate. These procedures have the advantage of setting a limit on the \( FW \) error rate at \( \alpha \) against any set of possible null hypotheses, as does the Tukey HSD (to be discussed shortly), while at the same time being less conservative than Tukey’s test *when our interest is in a specific subset of contrasts*. In general, however, Bonferroni procedures would not be used as a substitute when making all pairwise comparisons among a set of means, though the multistage procedures, which change the critical value as null hypotheses are rejected, can be used for that purpose.

As you saw, the Bonferroni test is based on the principle of dividing up \( FW \) for a family of contrasts among each of the individual contrasts. Thus, if we want \( FW \) to be .05 and we want to test four contrasts, we test each one at \( \alpha = .05/4 = .0125 \). The multistage tests follow a similar principle, the major difference being in the way they choose to partition \( \alpha \). They basically test the largest difference (or correlation coefficient). If that is significant they move to the next largest contrast but reduce the “number of tests” by 1. This continues until the first nonsignificant difference. The logic behind this approach hinges on the fact that the familywise error rate is the probability of making at least one Type I error. Suppose that we have four contrasts to test. If you make an error on the first contrast, you already have your one error, and making more won’t hurt. (Well, any error hurts, but making more than one won’t affect the familywise error rate.) If you did not make a Type I error on that test, there are at most three possible true null hypotheses to reject, so we can set \( \alpha' = \alpha / 3 \).

Rather than elaborate here on these procedures, I have moved that discussion to this book’s Web site to save space. These are powerful procedures when you organize your tests in steps, and I recommend reading those pages.

---

\(^9\) The actual probabilities would be .000, .018, .726, and .000.
Trimmed Means

I want to include one more approach that is very general and can be shown to be more powerful than standard procedures when the data come from long-tailed distributions. This is the use of trimmed means. The nice thing about this approach is that it can be adapted to carry out any of the procedures in this chapter, simply by substituting the appropriate trimmed means and squared standard errors.

I will assume that you have reasonably large sample sizes because we will trim those samples from each end. Wilcox recommends 20% trimming, which results in a sizable drop in the effective sample size, but with a corresponding gain in power. For convenience, assume that we have 40 observations in each of several groups and that we will go along with Wilcox’s suggestion of 20% trimming. That means that we will omit the lowest (.20)(40) = 8 observations and the highest 8 observations, leaving us with a sample of 24 observations for each condition. The trimmed means will be the means of those 24 observations in each group. To calculate the variance, we will use Winsorized samples, in which the lowest 8 scores are replaced with the 9th lowest score and the highest 8 scores are replaced with the 9th highest score. This leaves us with samples of \( n_i = 40 \) scores, but only \( h_j = 24 \) of those are independent observations from the \( i^{th} \) sample. If we let \( s^2_{W_i} \) represent the variance of the Winsorized sample of 40 scores, then the squared standard error of the mean for that sample would be

\[
s^{2}_{w_i} = \frac{(n_j - 1)s^2_{W_i}}{h_j(h_j - 1)}
\]

and the robust pairwise \( t \) test on the difference between two means can be written as

\[
t_w = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{s^2_{W_i} + s^2_{W_j}}}
\]

Notice that we are not doing anything very surprising here. We are replacing means with trimmed means and variances with variances that are based on Winsorized samples, but using \( h_j \) (the size of the trimmed sample) to adjust \( n_i \) to account for the trimming. Other than that, we have a standard \( t \) test, and it can be used as a replacement for the \( t \) in any of the procedures we have discussed, or will discuss, in this chapter. There is one complication, however, and that refers to the estimated degrees of freedom. The degrees of freedom are estimated as

\[
df_W = \frac{(s^2_{W_i} + s^2_{W_j})^2}{(h_i(h_i - 1) + s^2_{W_i}(h_j - 1))}
\]

That is a messy formula, but not very difficult to work out. As Keselman et al. (2005) noted, “When researchers feel they are dealing with nonnormal data, they can replace the usual least squares estimators of central tendency and variability with robust estimators and apply these estimators in any of the previously recommended” multiple comparison procedures.

NOTE

I want to emphasize one more time that the Bonferroni test and its variants are completely general. They are not the property of the analysis of variance or of any other statistical procedure. If you have several tests that were carried by any statistical procedure (and perhaps by different procedures), you can use the Bonferroni approach to control \( FW \). For example, I recently received an e-mail message in which someone asked how they might go about applying the Bonferroni to logistic regression. He would do it the same way he would do it for the analysis of variance. Take the set of statistical tests that came from his logistic regression, divide \( \alpha \) by the number of tests he ran, and declare a test to be significant only if its resulting probability was less than \( \alpha /c \). You don’t even need to know anything about logistic regression to do that.
12.4 Confidence Intervals and Effect Sizes for Contrasts

Having run a statistical significance test on the data from an experiment, and looked at individual comparisons, often called “individual contrasts,” we will generally want to look at some measure of the amount of difference between group means. In Chapter 11 we saw that when we have the omnibus $F$, which compares all means together, the most commonly used measure is a member of the $r$-family measures, such as $\eta^2$ or $\omega^2$. However, when we are looking at comparisons of individual means, or sets of means, it generally makes more sense to calculate confidence limits on our differences and/or to use a $d$-family measure of the effect size.

There are several ways that we could approach $d$-family measures. One very simple way is to go back to Chapter 7, which discussed $t$ tests, and apply the measures that were discussed there. We will come out at the same place, however, if we approach the problem through linear contrasts. Remember that when you are looking at two groups, it makes no difference whether you run a $t$ test between those groups, or compute a linear contrast and then an $F$, and take the square root of that $F$. The advantage of going with linear contrasts is that they are more general, allowing us to compare means of sets of groups rather than just two individual groups.

We will take an example from our morphine study by Siegel. One contrast that really interests me is the contrast between Group M-M and Group Mc-M. If their means are statistically significantly different, then that tells us that there is something important about changing the physical context in which the morphine is given. The group statistics are given below.

<table>
<thead>
<tr>
<th>Condition</th>
<th>M-M</th>
<th>Mc-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>10.00</td>
<td>29.00</td>
</tr>
<tr>
<td>St. Dev</td>
<td>5.13</td>
<td>6.06</td>
</tr>
<tr>
<td>Variance</td>
<td>26.32</td>
<td>37.95</td>
</tr>
<tr>
<td>$MS_{\text{error}}$</td>
<td>32.00</td>
<td></td>
</tr>
</tbody>
</table>

The coefficients for the linear contrast of these two groups would be “−1” for M-M, “+1” for Mc-M, and “0” for the other three conditions.

Confidence Interval

Let us first compute a confidence interval on the difference between conditions. The general formula for a confidence interval on a contrast of two means is

$$CI_{95} = (\bar{X}_i - \bar{X}_j) \pm t_{0.025}s_{\bar{X}_i - \bar{X}_j}$$

or, if we let “$\psi_j$” represent the value of the contrast, where $\psi_j = \Sigma a_i \bar{X}_i$, then

$$CI_{95} = (\psi_j) \pm t_{0.025}s_{\text{error}}$$

The standard error of the contrast ($s_{\text{error}}$), is

$$\sqrt{\frac{2MS_{\text{error}}}{n}}$$

For our confidence interval on the difference between the two conditions of interest I have

$$CI_{95} = (-1(10) + 1(29)) \pm 2.03\sqrt{8.00}$$

$$= 19 \pm 2.03(2.828) = 19 \pm 5.74$$

$$13.26 \leq \mu_{M-M} - \mu_{M-M} \leq 24.74$$
The probability is .95 that the interval will include the true difference between the population means.

When it comes time to form our effect size measure, we have a choice of what we will use as the error term—the standard deviation in the equation. I could choose to use the square root of \( \text{MSE} \) from the overall analysis, because that represents the square root of the average variance within each of the five groups. Kline (2004) recommends this approach. I have two other perfectly reasonable alternatives, however. First I could take the square root of the average sample variance of the two groups in question (perhaps weighted if the sample sizes were unequal). In this case it would be \((26.32 + 37.95)/2 = 32.155\) and \(\sqrt{32.155} = 5.669\). This would make sense if I were worried about heterogeneity of variance among the full set of five groups. Alternatively, I could consider one of the groups to be a control group and use its standard deviation as my error term. Here I might argue that \(M-M\) is like a control group because the conditions don’t change on trial 4. In this case I would let \(s_{\text{err}} = 5.13\). I think that my preference in general would be to base my estimate on the average of the variances of the groups in question. If the variances are homogeneous across all five groups, then the average of the groups in question won’t deviate much from the average of the variances of all five groups, so I haven’t lost much. Others might take a different view.

**Effect Size**

We have just seen that the confidence interval on the difference between \(M_c-M\) and \(M-M\) is \(13.26 \leq (\mu_{M_c-M} - \mu_{M-m}) \leq 24.74\). Both limits are on the same side of 0, reflecting the fact that the difference was statistically significant. However, the dependent variable here is the length of time before the animal starts to lick its paws, and I don’t suppose that any of us has a strong intuitive understanding of what a long or short interval is for this case. A difference of at least thirteen seconds seems pretty long, but I would like some better understanding of what is happening. One way to compute that would be to calculate an effect size on the difference between these means.

Our effect size measure will be essentially the same as it was in the case for \(t\) tests for independent samples. However, I will write it slightly differently because doing so will generalize to more complex comparisons. We have just seen that \(\psi\) represents a contrast between two means or sets of means, so it is really just a difference in means. We will take this difference and standardize it, which simply says that we want to represent the difference in group means in standard deviation units. (That is what we did in Chapter 7 as well.)

In Chapter 7 we defined

\[
\hat{d} = \frac{\bar{X}_i - \bar{X}_j}{s_p}
\]

where \(s_p\) is the square root of our pooled variance estimate and is a measure of the average standard deviation within the groups. We are going to calculate essentially the same thing here, but I will write its expression as

\[
\hat{d} = \frac{\psi}{s_e} = \frac{\sum (a_i\bar{X}_i)}{s_e}
\]

The numerator is a simple linear contrast, while the denominator is some estimate of the within groups standard deviation.

The preceding formula raises two points. In the first place, the coefficients must form what we have called a “standard set.” This simply means that the absolute values of the coefficients must sum to 2. For example, if we want to compare the mean of two groups with the mean of a third, we could use coefficients of \((\frac{1}{2}, \frac{1}{2}, -1)\) to form our contrast. We would...
get to the same place as far as our test of significance is concerned by using \((1, 1, -2)\) or \((3, 3, -6)\); the resulting \(F\) would be the same. But only the first would give us a numerical answer for the contrast that is the difference between the mean of the first two groups and the mean of the third. This is easily seen when you write

\[
\psi = \left(\frac{1}{2}\right) \overline{X}_1 + \left(\frac{1}{2}\right) \overline{X}_2 + \left(-\frac{1}{2}\right) \overline{X}_3
\]

You can see clearly that we are taking the difference between the means of sets of groups.

The second issue raised by our equation for \(\hat{d}\) is the choice of the denominator. As I mentioned a few paragraphs back, there are at least three possible estimates. We could use the square root of \(MSE_{error}\), or the square root of the average of the variances in the groups being contrasted, or we could conceive of one of the groups as a control group, and use its standard

\[
\text{Table 12.3  Means of conditions in our morphine example}
\]

<table>
<thead>
<tr>
<th>Groups:</th>
<th>M-S</th>
<th>M-M</th>
<th>S-S</th>
<th>S-M</th>
<th>Mc-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Means:</td>
<td>4.00</td>
<td>10.00</td>
<td>11.00</td>
<td>24.00</td>
<td>29.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>(\sum a_j^2)</th>
<th>(\psi = \sum a_j \overline{X}_j)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a_j)</td>
<td>-1/2</td>
<td>1/3</td>
</tr>
<tr>
<td>(b_j)</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>(c_j)</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>(d_j)</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>

\[
\text{M-M, S-M, Mc-M versus M-S, S-S}
\]

\[
\hat{d}_1 = \frac{\sum a_j \overline{X}_j}{s_{error}} = \frac{\left( -\frac{1}{2}\right) \overline{X}_{M-S} + \left( -\frac{1}{2}\right) \overline{X}_{M-M} + \left( -\frac{1}{2}\right) \overline{X}_{S-S} + \left( \frac{1}{2}\right) \overline{X}_{S-M} + \left( \frac{1}{2}\right) \overline{X}_{Mc-M}}{\sqrt{MS_{error}}}
\]

\[
= \frac{\left(-\frac{1}{2}\right)4.00 + \left(-\frac{1}{2}\right)10.00 + \left(-\frac{1}{2}\right)11.00 + \left(\frac{1}{2}\right)24.00 + \left(\frac{1}{2}\right)29}{\sqrt{32}}
\]

\[
= \frac{4.00 + 10.00 + 11.00 + 24.00 + 29.00}{3} \times \frac{2}{\sqrt{32}} = -8.333 + 26.5 = 18.167 = 3.21
\]

\[
\text{M-M versus Mc-M}
\]

\[
\hat{d}_2 = \frac{\sum b_j \overline{X}_j}{s_{error}} = \frac{\left(-1\right) \overline{X}_{M-M} + \left(1\right) \overline{X}_{Mc-M}}{\sqrt{MS_{error}}}
\]

\[
= \frac{\left(-1\right)10.00 + \left(1\right)29.00}{\sqrt{32}}
\]

\[
= \frac{-10.00 + 29.00}{32} = \frac{19}{5.657} = 3.36
\]
deviation as our estimate. The most common approach is to use the square root of $MS_{error}$ and that is what I will do here because the variances in our example are quite similar.

Earlier we looked at four contrasts that seemed to be of interest for theoretical reasons. The traditional Bonferroni procedure showed that two of the contrasts were statistically significant, while the other two were not. Computation of the effect sizes for significant contrasts are shown in Table 12.3. In these calculations I have used the square root of $MS_{error}$ as my denominator for consistency.

Because our tests showed that the last two contrasts were not nearly statistically significant, our best approach would probably be to treat these effect sizes as 0.00. There are no differences between groups. An interesting question arises as to what we would do if the test statistic had been nearly large enough to be significant. In that case I would present my effect size measure but caution that the corresponding hypothesis test was not significant.

You can see that the other effect sizes are substantial, all showing a difference of at least three standard deviations. I will speak about these effects in the following section.

### 12.5 Reporting Results

We have run several different tests on these data, and the following is a report based on the Bonferroni procedure.

This experiment examined the phenomenon of morphine tolerance in rats placed on a warm surface. The underlying hypothesis was that with repeated injections of morphine animals develop a hypersensitivity to pain, which reduces the effect of the drug. When animals are then tested without the drug, or with the drug in a different context, this hypersensitivity will be expressed in a shorter paw-lick latency.

The omnibus $F$ from the overall analysis was statistically significant ($F(4,35) = 27.33$, $p < .05$). Subsequent contrasts on important comparisons using the Bonferroni test revealed that morphine’s effects were as predicted. The groups receiving morphine on the test trial after having received either saline or morphine in the same or different context on trials 1–3 showed longer reaction times than the average of groups who received saline on the test trial ($t(35) = 9.95$, $t_{.006,35} = 2.64$). The standardized effect size was 3.21, indicating a difference of nearly $3\frac{1}{2}$ standard deviations between the means of the two sets of groups.

The effect of context is seen in a statistically longer mean paw-lick latency in the Mc-M ($\bar{X} = 29$) condition than in the M-M condition ($\bar{X} = 10$) ($t(35) = 6.72$, $t_{2.52} = 1.67$ $= \pm 2.52$). The standardized effect size here was 3.36.

### 12.6 Post Hoc Comparisons

There is much to recommend the use of linear contrasts and the Bonferroni $t$ test when a relatively small number of comparisons can be specified a priori. In fact, my strong preference would be to ask a few very pointed questions, which would best be approached by setting up linear contrasts. The use of broader post hoc comparisons may lose power by asking questions that you don’t care about, but that approach does have advantages in experiments involving many hypotheses\textsuperscript{10} and/or hypotheses that are clearly arrived at only after the data have been examined. In this situation, a number of a posteriori or post hoc techniques are available.

\textsuperscript{10} If there are many hypotheses to be tested, regardless of whether they were planned in advance, the procedures discussed here are usually more powerful than is the Bonferroni $t$ test.
Chapter 12  Multiple Comparisons Among Treatment Means

Fisher’s Least Significant Difference (LSD) Procedure

Fisher’s Least Significant Difference (LSD) test (also known as Fisher’s protected t). People sometimes (rightfully) complain about the use of this procedure when there are many means, but when we only have a few means to compare (particularly when we only have three), this is a very legitimate and useful procedure. The procedure consists of simply running pairwise comparisons among the means using a standard Student’s t test. The only difference between the post hoc LSD procedure and the a priori multiple t test procedure discussed earlier is that the LSD requires a significant F for the overall analysis of variance. When the complete null hypothesis is true (all population means are equal), the requirement of a significant overall F ensures that the familywise error rate will equal α. Unfortunately, if the complete null hypothesis is not true but some other more limited null hypotheses involving subsets of means are true, which is most likely to be the case, the overall F may no longer affords protection for FW. For this reason, some recommend that you not use this test, although Carmer and Swanson (1973) have shown it to be the most powerful of the common post hoc multiple-comparison procedures. If your experiment involves three means, the LSD procedure is an excellent one because FW will stay at α, and you will gain the added power of using standard t tests. (The FW error rate will be α with three means because if the complete null hypothesis is true, you have a probability equal to α of making a Type I error with your overall F, and any subsequent Type I errors you might commit with a t test will not affect FW. If the complete null is not true but a more limited one is, with three means there can be at most one null difference among the means and, therefore, only one chance of making a Type I error, again with a probability equal to α.) You should generally be reluctant to use the LSD for more than three means unless you have good reason to believe that there is at most one true null hypothesis hidden in the means.

The Studentized Range Statistic (q)

Studentized range statistic (q)

Because many of the post hoc are based on the Studentized Range Statistic or special variants of it, we will consider this statistic before proceeding. The Studentized range statistic (q) is defined as

$$ q_r = \frac{X_l - X_s}{\sqrt{MS_{\text{error}}} / n} $$

where $X_l$ and $X_s$ represent the largest and smallest of a set of treatment means and $r$ is the number of treatments in the set. You probably have noticed that the formula for $q$ is very similar to the formula for $t$. In fact

$$ q_r = \frac{X_l - X_s}{\sqrt{MS_{\text{error}}} / n} $$

$$ t = \frac{X_i - X_j}{\sqrt{2(\text{MS}_{\text{error}}) / n}} $$

and the only difference is that the formula for $t$ has a “$\sqrt{2}$” in the denominator. Thus, $q$ is a linear function of $t$ and we can always go from $t$ to $q$ by the relation $q = t\sqrt{2}$. The real difference between $q$ and $t$ tests comes from the fact that the tables of $q$ (Appendix q) are
set up to allow us to adjust the critical value of $q$ for the number of means involved, as will become apparent shortly. When there are only two treatments, whether we solve for $t$ or $q$ is irrelevant as long as we use the corresponding table.

When we have only two means or when we wish to compare two means chosen at random from the set of available means, $t$ is an appropriate test.

Suppose, however, that we looked at a set of means and deliberately selected the largest and smallest means for testing. It is apparent that we have drastically altered the probability of a Type I error. Given that $H_0$ is true, the largest and smallest means certainly have a greater chance of being called “significantly different” than do means that are adjacent in an ordered series of means. This is the point at which the Studentized range statistic becomes useful. It was designed for just this purpose.

To use $q$, we first order the means from smallest to largest. We then take into account the number of steps between the means to be compared. For adjacent means $r = 2$ and no change is made. Thus $q_{0.05} = t_{0.05} \sqrt{2}$. For means that are not adjacent, however, the critical value of $q$ increases, growing in magnitude as the number of intervening steps between means increases.

As an example of the use of $q$, consider the data on morphine tolerance. The means are

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>10</td>
<td>11</td>
<td>24</td>
<td>29</td>
</tr>
</tbody>
</table>

with $n = 8$, $df_{\text{error}} = 35$, and $MS_{\text{error}} = 32.00$. The largest mean is 29 and the smallest is 4, and there are a total ($r$) of 5 means in the set (in the terminology of most tables, we say that these means are $r = 5$ steps apart).

$$q_s = \frac{X_1 - X_5}{\sqrt{\frac{MS_{\text{error}}}{n}}} = \frac{29 - 4}{\sqrt{32.00 / 8}} = \frac{25}{\sqrt{4}} = 12.5$$

Notice that $r$ is not involved in the calculation. It is involved, however, when we go to the tables. From Appendix $q$, for $r = 5$ and $df_{\text{error}} = 35$, $q_{0.05}(5,35) = 4.07$. Because 12.5 > 4.07, we will reject $H_0$ and conclude that there is a significant difference between the largest and smallest means.

An alternative to solving for $q_{\text{obs}}$ and referring $q_{\text{obs}}$ to the sampling distribution of $q$ would be to solve for the smallest difference that would be significant and then to compare our actual difference with the minimum significant difference. This seems like an unnecessary test, but that approach is frequently taken by computer based post hoc procedures, such as those used by SPSS. That explains the way some of your computer printout is displayed. Either way leads to the same results.

### 12.7 Tukey’s Test

Much of the work on multiple comparisons has been based on the original work of John Tukey, and an important test bears his name. The Tukey test, also called the Tukey’s HSD (Honestly Significant Difference) test or the WSD (Wholly Significant Difference) test, uses the Studentized $q$ statistic for its comparisons, except that $q_{\text{HSD}}$ is always taken as the maximum value of $q_r$. In other words, if there are five means, all differences are tested as if they were five steps apart. The effect is to fix the familywise error rate at $\alpha$ against all possible null hypotheses, not

---

11 With only two means we obtain all of the information we need from the $F$ in the analysis of variance table and would have no need to run any contrast.
just the complete null hypothesis, although with some loss of power. The Tukey HSD is the favorite pairwise test for many people because of the control it exercises over $\alpha$.

If we apply the Tukey HSD to the data on morphine tolerance, we first arrange the means in the order of increasing magnitude, as follows.

<table>
<thead>
<tr>
<th>M-S</th>
<th>M-M</th>
<th>S-S</th>
<th>S-M</th>
<th>Mc-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>10</td>
<td>11</td>
<td>24</td>
<td>29</td>
</tr>
</tbody>
</table>

From Appendix $q$ we find that with 35 df for $MS_{\text{error}}$ and $r$ set at 5, the critical value of $q$ equals 4.07. We will use that critical value for all contrasts.

We could run the necessary $q$ test between each pair of means, but that is not the way that most software goes about it. We know $MS_{\text{error}}, df, q$, and the critical value of $q$ (4.07) will be the same for all contrasts. In that case it is easiest to decide what minimal difference between means will be significant.

$$ q = \frac{\bar{X}_I - \bar{X}_J}{\sqrt{\frac{MS_{\text{error}}}{n}}} $$

therefore

$$ q\sqrt{\frac{MS_{\text{error}}}{n}} = \bar{X}_I - \bar{X}_J $$

$$ = 4.07\sqrt{\frac{32.00}{8}} = 8.14 $$

This result means that if the difference between two means is greater than 8.14, a $q$ test would be significant, otherwise it would not. From our means given above, we can see that the following differences are significant because they are all greater than 8.14.

- M-S versus S-M = 4.00 – 24.00 = –20
- M-S versus Mc-M = 4.00 – 29.00 = –25
- M-M versus S-M = 10.00 – 24.00 = –14
- M-M versus Mc-M = 10.00 – 29.00 = –19
- S-S versus S-M = 11.00 – 24.00 = –13
- S-S versus Mc-M = 11.00 – 29.00 = –18

Thus we can write

(M-S = M-M = S-S) $\neq$ (S-M = Mc-M)

The equal signs indicate simply that we could not reject the null hypothesis of equality, not that we have proven the means to be equal.

### Unequal Sample Sizes and Heterogeneity of Variance

The Tukey procedure was designed primarily for the case of equal sample sizes ($n_1 = n_2 = \cdots = n_k = n$). Frequently, however, experiments do not work out as planned, and we find ourselves with unequal numbers of observations and still want to carry out a comparison of means. A good bit of work has been done on this problem with respect to the Tukey HSD test (see particularly Games and Howell, 1976; Keselman and Rogan, 1977; Games, Keselman, and Rogan, 1981).

One solution, known as the Tukey–Kramer approach, is to replace $\sqrt{MS_{\text{error}}/n}$ with

$$ \sqrt{\frac{MS_{\text{error}}}{n_1} + \frac{MS_{\text{error}}}{n_2}} \over 2 $$
and otherwise conduct the test the same way you would if the sample sizes were equal. This is the default solution with SPSS.

An alternative, and generally preferable, test was proposed by Games and Howell (1976). The Games and Howell procedure uses what was referred to as the Behrens–Fisher approach to \( t \) tests in Chapter 7. The authors suggest that a critical difference between means (i.e., \( W_r \)) be calculated separately for every pair of means using

\[
W_r = \bar{X}_i - \bar{X}_j = q_{.05}(r, df^*) \sqrt{\frac{s_i^2/n_i + s_j^2/n_j}{2}}
\]

where \( q_{.05}(r, df^*) \) is taken from the tables of the Studentized range statistic on

\[
df^* = \frac{\left( \frac{s_i^2}{n_i} + \frac{s_j^2}{n_j} \right)^2}{\left( \frac{s_i^2}{n_i} \right)^2/n_i - 1 + \left( \frac{s_j^2}{n_j} \right)^2/n_j - 1}
\]

degrees of freedom. This is basically the solution referred to earlier in the discussion of multiple \( t \) tests, although here we are using the Studentized range statistic instead of \( t \), and it is an optional solution in SPSS. This solution is laborious, but the effort involved is still small compared to that of designing the study and collecting the data. The need for special procedures arises from the fact that the analysis of variance and its attendant contrasts are especially vulnerable to violations of the assumption of homogeneity of variance when the sample sizes are unequal. Moreover, regardless of the sample sizes, if the sample variances are nearly equal you may replace \( s_i^2 \) and \( s_j^2 \) in the formula for \( W_r \) with \( MS_{\text{error}} \) from the overall analysis of variance. And regardless of the sample size, if the variances are heterogeneous you should probably use the Games and Howell procedure.

**Other Range-Based Tests**

There are many other tests that have been developed on the basis of the Studentized Range Statistic, and they can be found on the Web site for this book. I just want to mention what they are and how they differ.

**Newman–Keuls Test**

The Newman–Keuls test has long been controversial, but it is still in use and can be obtained from SPSS. The basic difference between the Newman–Keuls test and Tukey’s test is that the latter fixes \( r \) at the number of levels of the independent variable, whereas the former continually adjusts \( r \) to equal the number of ordered means of from which we are testing the largest and smallest. (In many ways the logic here is similar to the logic for the adjusted Bonferroni test discussed a few pages back.) For example, if we tested M-S versus S-S, they are the largest and smallest of three means, so \( r = 3 \). The advantage of doing this is that we have a more powerful test. The disadvantage is that we lose some control over \( \alpha \).

**The Ryan Procedure**

The Ryan procedure (now known as the REGWQ test) is a modification of the Tukey in the direction of the Newman–Keuls. This test adjusts \( r \) as we go along, but it does so in a way that keeps the maximum FEW at \( \alpha \). I prefer this test over the Tukey, but it is rarely used.
Chapter 12  Multiple Comparisons Among Treatment Means

The Scheffé Test

The Scheffé test is one of our oldest tests and was based on Tukey’s procedure. But this test allows us to hold \( \alpha \) at .05 against any and all comparisons, not just pairwise comparisons. Scheffé acknowledged that the test lacked power when applied to pairwise comparisons, and he recommended that it not be used to make pairwise comparisons. Unfortunately SPSS and others do not take his advice and, when they use this test, are robbing themselves of power. (Don’t blame Scheffé, he told them not to do that.) See Howell (2010) for more information.

Dunnett’s Test

Dunnett’s test is a test designed to compare one treatment (usually a control treatment) with each of the other treatment means. It is more powerful for this purpose, though it lacks the flexibility of other tests. It uses its own tables, which can also be found on the Web site.

Benjamini–Hochberg Test

I have given only a brief sketch of some competing tests for pairwise comparisons. But I feel the need to bring in one more approach, in part because statisticians like it and it will likely play a more important role in the behavioral sciences over time. Each of the post hoc tests that we have discussed has focused on controlling the familywise error rate (FW), and several of them have been sequential tests, which change the critical value as you move through a series on comparisons. Benjamini and Hochberg (1995, 2000) have developed tests that are becoming more popular, are sequential, and are not based on the FW. They advocate using what they call the False Discovery Rate (FDR) instead of the familywise error rate. When Tukey began advocating FW in the early 1950s he, perhaps unintentionally, oriented our thinking almost exclusively toward controlling the probability of even one Type I error. When you compute a familywise rate, you are dealing with the probability of one or more Type I errors. In effect you are saying that your whole set of conclusions are erroneous when you make even one Type I error. (Curiously we don’t consider our conclusions to be erroneous if we make Type II errors.) Hochberg and Benjamini have looked at the problem somewhat differently and asked “What percentage of the significant results (“discoveries”) that we have found are false discoveries?” Suppose that we carry out nine comparisons (either simple contrasts, complex contrasts, tests on a single mean, or any other test). We find that there are four significant effects but, unknown to us, one of those significant effects is really a Type I error. The FDR is then defined as

\[
\text{FDR} = \frac{\text{Number of False Rejections}}{\text{Number of Total Rejections}} = \frac{1}{4} = .25
\]

I will take an example of a simple “thought experiment” from Maxwell and Delaney (2004), who have an excellent discussion of the FDR. Imagine that we have a situation in which we test 10 null hypotheses, three of which are known to be false and the others true. Suppose that we mentally run our experiment 100 times, testing all 10 hypotheses for each run. Further suppose that we have very considerable power to reject false null hypotheses, so that we nearly always reject the three false null hypotheses. Finally assume that we have chosen a critical value so as to set the familywise error rate at .20. (You probably think that .20 is too high, but bear with me.) Then out of our 100 hypothetical experiments, 80% of the time (1 – .20) we will make no Type I errors (we will nearly always reject the three truly significant null hypotheses and retain the other seven). Furthermore 20% of the time we will make one Type I error (assuming that we don’t make two type I errors in any experiment) because we set the familywise error rate at .20. Because we have a great deal of power, we will almost always reject the three false null hypotheses. Here our FW is .20, which perhaps made you wince. But what about the FDR? Given the description above, we will make no
errors in 80% of the experiments. In the other 20 experiments we will make one Type I error
and three correct rejections, for an FDR of \( \frac{1}{4} = .25 \) for those 20 experiments and an FDR of
0 for the other 80 experiments. Over the long haul of 100 experiments, the average FDR will
be .05, while the FWE will be .20. Thus the critical value that sets the familywise FWE at
.20 leaves the FDR at only .05. The problem is how we choose that critical value. Unfortu-
nately, that choice is quite complicated in the general case, but fortunately it is fairly simple
in the case of either independent contrasts or pairwise contrasts. See Keselman, Cribbie,
and Holland (1999). In this chapter I have been a strong advocate of pairwise contrasts, so
restricting ourselves to that case is not particularly onerous.

The procedure we will follow is called the **Benjamini and Hochberg’s Linear Step
Up (LSU) procedure**.\(^\text{12}\) I will not take the space to develop the logic of this test, but the
paper by Benjamini and Hochberg (1995) and the chapter by Maxwell and Delaney (2004)
are reasonably clear. Thissen, Steinberg, and Kuang (2002) present a simple method to
carry out the Benjamini Hochberg test using spreadsheets. I will frame this discussion in
terms of the steps needed to perform the test.

Assume that we have performed 10 pairwise contrasts on Siegel’s morphine data. The
results are shown in Table 12.4 ordered by \( p \) value. The column labeled “\( i \)” is the index of
the comparison and simply ranks the \( p \) values from highest to lowest. The critical part of
the table is labeled \( p_{\text{crit}} \), the critical value for our test. We define

\[
p_{\text{crit}} = \left( \frac{i}{k} \right) \alpha
\]

where “\( i \)” is the index, “\( k \)” is the number of tests (here \( k = 10 \)), and \( \alpha \) is the desired FDR
(here \( \alpha = .05 \)). To carry out the test we work our way down the table. If \( p > p_{\text{crit}} \) we retain
the null hypothesis and move on to the next row. As soon as \( p < p_{\text{crit}} \) we reject that null
hypothesis and all subsequent ones.

Using the Benjamini-Hochberg test we would declare that M-M vs. S-S, S-M vs.
Mc-M, and M-S vs. M-M are not different from each other pairwise. All other contrasts
are judged statistically significant. With Tukey’s test we only rejected six null hypotheses,
retaining M-S vs. S-S, which was rejected by Benjamin-Hochberg. That is a contrast that
we would like to find significant, because it says that if you get saline in the same condition
in which you had received morphine your sensitivity will increase.

<table>
<thead>
<tr>
<th>Group</th>
<th>( t )</th>
<th>( p )</th>
<th>( i )</th>
<th>( p_{\text{crit}} )</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-M vs. S-S</td>
<td>-0.354</td>
<td>.726</td>
<td>10</td>
<td>0.05</td>
<td>No</td>
</tr>
<tr>
<td>S-M vs. Mc-M</td>
<td>-1.768</td>
<td>.086</td>
<td>9</td>
<td>0.045</td>
<td>No</td>
</tr>
<tr>
<td>M-S vs. M-M</td>
<td>-2.121</td>
<td>.041</td>
<td>8</td>
<td>0.040</td>
<td>No</td>
</tr>
<tr>
<td>M-S vs. S-S</td>
<td>-2.475</td>
<td>.018</td>
<td>7</td>
<td>0.035</td>
<td>Yes</td>
</tr>
<tr>
<td>S-S vs. S-M</td>
<td>-4.596</td>
<td>.00007</td>
<td>6</td>
<td>0.030</td>
<td>Yes</td>
</tr>
<tr>
<td>M-M vs. S-M</td>
<td>-4.950</td>
<td>.00003</td>
<td>5</td>
<td>0.025</td>
<td>Yes</td>
</tr>
<tr>
<td>S-S vs. Mc-M</td>
<td>-6.364</td>
<td>.00000</td>
<td>4</td>
<td>0.020</td>
<td>Yes</td>
</tr>
<tr>
<td>M-M vs. Mc-M</td>
<td>-6.717</td>
<td>.00000</td>
<td>3</td>
<td>0.015</td>
<td>Yes</td>
</tr>
<tr>
<td>M-S vs. S-M</td>
<td>-7.071</td>
<td>.00000</td>
<td>2</td>
<td>0.010</td>
<td>Yes</td>
</tr>
<tr>
<td>M-S vs. Mc-M</td>
<td>-8.839</td>
<td>.00000</td>
<td>1</td>
<td>0.005</td>
<td>Yes</td>
</tr>
</tbody>
</table>

\(^{12}\) Benjamini and Hochberg (2000) recommended a variation on the test given here, sometimes called their
“adaptive” test. It is more powerful than the LSU test, but somewhat more cumbersome. Both of these tests are
different from the Hochberg GT2 test produced by SPSS. A program in R (BenjaminiHochbergLSU.R) is avail-
able at the book’s Web site.
12.8 Which Test?

Choosing the most appropriate multiple-comparison procedure for your specific situation is not easy. Many tests are available, and they differ in a number of ways. The choice is a bit easier if we consider the two extremes first.

If you have planned your test in advance and you want to run only one comparison, I would suggest that you run a standard $t$ test (correcting for heterogeneity of variance if necessary), or, if you have a complex comparison, a linear contrast. If you have several a priori contrasts to run, not necessarily pairwise, the multistage Bonferroni $t$ does a good job of controlling $FW$ while at the same time maximizing power.

If you have a large number of groups and wish to make many comparisons, whether or not you are interested in all of the possible pairwise comparisons, you would probably be well advised to use Tukey’s test (or the REGWQ if possible). I can’t think of a situation where I would personally recommend the Scheffé, but I presented it here because it is a common test and real hard-liners like it.

What about the Benjamini-Hochberg test? This is a difficult test to place in a table because it controls an entirely different error rate. It is not fair to say that one test is more powerful than another when they are working on different error rates. I have considerable fondness for the Benjamini-Hochberg test just because it is not based on the idea that one false rejection invalidates a family of conclusions. If you are willing to accept an occasional Type I error to gain power for other contrasts, there is much to recommend this test.

12.9 Computer Solutions

Most software packages will perform multiple comparison procedures, but not all packages have all procedures available. Exhibit 12.1 contains the results of an analysis of the morphine data using SAS. I chose SAS because it has a broad choice of procedures and is one of the major packages. It also has more information in its printout than does SPSS and is thus somewhat more useful for our purpose. I have included the Scheffé test for comparison even though I have already said that it is totally inappropriate for simple pairwise comparisons.

Exhibit 12.1 begins with the SAS program commands and the overall analysis of variance. This analysis agrees with the summary table shown in Table 12.1. The $R^2 = .757$ is simply $\eta^2$. You can see that our experimental manipulation accounts for a substantial portion of the variance. The remainder of the exhibit includes the results of the Newman-Keuls, Ryan, Tukey, and Scheffé tests, some of which I have mentioned only briefly.

The Newman-Keuls, as the least conservative test, reports the most differences between conditions. If you look first at the means and “SNK Grouping” at the end of that portion of the printout, you will see a column consisting of the letters A, B, and C. Conditions that share the same letter are judged to not differ from one another. Thus the means of Conditions Mc-M and S-M are not significantly different from one another, but, because they don’t have a letter in common with other conditions, they are different from the means of S-S, M-M, and M-S. Similarly, Conditions S-S and M-M share the letter B and their means are thus not significantly different from each other, but are different from the means of the other three conditions. Finally, the mean of Condition M-S is different from the means of all other conditions.
Data Siegel;
Input 'Siegel.dat';
Input Subject Condition latency;
Run;

Proc GLM Data = Siegel;
Class Condition;
Model Latency = Condition/SS3;
Means Condition/ SNK Tukey REGWQ Scheffe;
Run;

The SAS System
11:15 Wednesday
August 18, 2010

The GLM Procedure

Dependent Variable: LATENCY

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>3497.600000</td>
<td>874.400000</td>
<td>27.33</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>35</td>
<td>1120.000000</td>
<td>32.000000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>39</td>
<td>4617.600000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \eta^2 \rightarrow R\text{-Square} = 0.757450 \]

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONDITION</td>
<td>4</td>
<td>3497.600000</td>
<td>874.400000</td>
<td>27.33</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

**Student-Newman-Keuls** Test for LATENCY

NOTE: This test controls the Type I experimentwise error rate under the complete null hypothesis but not under partial null hypotheses.

Alpha = 0.05  Error Degrees of Freedom = 35  Error Mean Square = 32

<table>
<thead>
<tr>
<th>Number of Means</th>
<th>Critical Range</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5.7420599</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6.9219411</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7.6279954</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8.1319062</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>SNK Grouping</th>
<th>Mean</th>
<th>N</th>
<th>Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>29.000</td>
<td>8</td>
<td>Mc-M</td>
</tr>
<tr>
<td>B</td>
<td>10.000</td>
<td>8</td>
<td>S-S</td>
</tr>
<tr>
<td>C</td>
<td>4.000</td>
<td>8</td>
<td>M-S</td>
</tr>
</tbody>
</table>

Exhibit 12-1

(Continues)
Ryan-Einot-Gabriel-Welsch Multiple Range Test for LATENCY

NOTE: This test controls the Type I experimentwise error rate.
Alpha = 0.05  Error Degrees of Freedom = 35  Error Mean Square = 32

<table>
<thead>
<tr>
<th>Number of Means</th>
<th>Critical Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>6.8765475</td>
</tr>
<tr>
<td>3</td>
<td>7.5391919</td>
</tr>
<tr>
<td>4</td>
<td>7.6279954</td>
</tr>
<tr>
<td>5</td>
<td>8.1319062</td>
</tr>
</tbody>
</table>

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>REGWQ Grouping</th>
<th>Mean</th>
<th>N</th>
<th>CONDITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>29.000</td>
<td>8</td>
<td>Mc-M</td>
</tr>
<tr>
<td>A</td>
<td>24.000</td>
<td>8</td>
<td>S-M</td>
</tr>
<tr>
<td>B</td>
<td>11.000</td>
<td>8</td>
<td>S-S</td>
</tr>
<tr>
<td>B</td>
<td>10.000</td>
<td>8</td>
<td>M-M</td>
</tr>
<tr>
<td>B</td>
<td>4.000</td>
<td>8</td>
<td>M-S</td>
</tr>
</tbody>
</table>

Tukey’s Studentized Range (HSD) Test for LATENCY

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha = 0.05  Error Degrees of Freedom = 35  Error Mean Square = 32

Critical Value of Studentized Range = 4.06595
Minimum Significant Difference = 8.1319

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>Tukey Grouping</th>
<th>Mean</th>
<th>N</th>
<th>CONDITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>29.000</td>
<td>8</td>
<td>Mc-M</td>
</tr>
<tr>
<td>A</td>
<td>24.000</td>
<td>8</td>
<td>S-M</td>
</tr>
<tr>
<td>B</td>
<td>11.000</td>
<td>8</td>
<td>S-S</td>
</tr>
<tr>
<td>B</td>
<td>10.000</td>
<td>8</td>
<td>M-M</td>
</tr>
<tr>
<td>B</td>
<td>4.000</td>
<td>8</td>
<td>M-S</td>
</tr>
</tbody>
</table>

Scheffe’s Test for LATENCY

NOTE: This test controls the Type I experimentwise error rate.

Alpha = 0.05  Error Degrees of Freedom = 35  Error Mean Square = 32

Critical Value of F = 2.64147
Minimum Significant Difference = 9.1939

Means with the same letter are not significantly different.

<table>
<thead>
<tr>
<th>Scheffe Grouping</th>
<th>Mean</th>
<th>N</th>
<th>CONDITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>29.000</td>
<td>8</td>
<td>Mc-M</td>
</tr>
<tr>
<td>A</td>
<td>24.000</td>
<td>8</td>
<td>S-M</td>
</tr>
<tr>
<td>B</td>
<td>11.000</td>
<td>8</td>
<td>S-S</td>
</tr>
<tr>
<td>B</td>
<td>10.000</td>
<td>8</td>
<td>M-M</td>
</tr>
<tr>
<td>B</td>
<td>4.000</td>
<td>8</td>
<td>M-S</td>
</tr>
</tbody>
</table>

Exhibit 12-1 (Continued)
If you look a bit higher in the table you will see a statement about how this test deals with the familywise (here called “experimentwise”) error rate. As I said earlier, the Newman-Keuls holds the familywise error rate at \( \alpha \) against the complete null hypothesis, but allows it to rise in the case where a subset of null hypotheses is true. You next see a statement saying that the test is being run at \( \alpha = 0.05 \), that we have 35 \( df \) for the error term, and that \( MS_{\text{error}} = 32.00 \). Following this information you see the critical ranges. These are the minimum differences between means that would be significant for different values of \( r \). The critical ranges are equal to

\[
W_r = q_{0.05}(r, df_e)\sqrt{\frac{MS_{\text{error}}}{n}}
\]

For example, when \( r = 3 \) (a difference between the largest and smallest of three means)

\[
W_3 = q_{0.05}(3, df_e)\sqrt{\frac{MS_{\text{error}}}{n}} = 3.46\sqrt{\frac{32}{8}} = 3.46(2) = 6.92
\]

Because all three step differences (e.g., 29 – 11 = 18; 24 – 10 = 14; 11 – 4 = 7) are greater than 6.92, they will all be declared significant.

The next section of Exhibit 12.1 shows the results of the Ryan (REGWQ) test. Notice that the critical ranges for \( r = 2 \) and \( r = 3 \) are larger than they were for the Newman–Keuls (though smaller than they will be for the Tukey). As a result, for \( r = 3 \) we need to exceed a difference of 7.54, whereas the difference between 11 and 4 is only 7. Thus this test will not find Group 1 (M-S) to be different from Group 3 (S-S), whereas it was different for the more liberal Newman–Keuls. However, the maximum familywise error rate for this set of comparisons is \( \alpha = 0.05 \), whereas it would be nearly \( \alpha = 0.10 \) for the Newman–Keuls.

The Tukey test is presented slightly differently, but you can see that Tukey requires all differences between means to exceed a critical range of 8.1319 to be declared significant, regardless of where they lie in an ordered series. For this specific set of data our conclusions are the same as they were for the Ryan test, although that will certainly not always be the case.

Although the Scheffé test is run quite differently from the others, it is possible to compute a critical range for all pairwise comparisons. From Exhibit 12.1 we can see that this range is 9.1939, almost a full point larger than the critical range for Tukey. This reflects the extreme conservatism of the Scheffé procedure, especially with just pairwise contrasts, and illustrates my major objection to the use of this test for this purpose.

SAS will also produce a number of other multiple comparison tests, including the Bonferroni and the Dunn–Šidák. I do not show those here because it is generally foolish to use either of those tests when you want to make all possible pairwise comparisons among means. The Ryan or Tukey test is almost always more powerful and still controls the familywise error rate. I suppose that if I had a limited number of pairwise contrasts that I was interested in, I could use the Bonferroni procedure in SAS (BON) and promise not to look at the contrasts that were not of interest.

12.10 Trend Analysis

The analyses we have been discussing are concerned with identifying differences among group means, whether these comparisons represent complex contrasts among groups or simple pairwise contrasts. Suppose, however, that the groups defined by the independent variable are ordered along some continuum. An example might be a study of the beneficial effects of aspirin in preventing heart disease. We could ask subjects to take daily doses of 1, 2, 3, 4, or 5 of 81mg aspirin, often referred to as “baby aspirin.” In this study we would not be concerned so much with whether a 4-pill dose was better than a 2-pill dose, for example, as with whether the beneficial
effects of aspirin increase with increasing the dosage of the drug. In other words, we are concerned with the trend in effectiveness rather than multiple comparisons among specific means.

To continue with the aspirin example, consider two possible outcomes. In one outcome we might find that the effectiveness increases linearly with dosage. In this case the more aspirin you take, the greater the effect, at least within the range of dosages tested. A second, alternative, finding might be that effectiveness increases with dosage up to some point, but then the curve relating effectiveness to dosage levels off and perhaps even decreases. This would be either a “quadratic” relationship or a relationship with both linear and quadratic components. It would be important to discover such relationships because they would suggest that there is some optimal dose, with lower doses being less effective and higher doses adding little, if anything, to the effect.

Typical linear and quadratic functions are illustrated in Figure 12.2. It is difficult to characterize quadratic functions neatly because the shape of the function depends both on the sign of the coefficient of $X^2$ and on the sign of $X$ (the curve changes direction when $X$ passes from negative to positive, and for positive values of $X$ the curve rises if the coefficient is positive and falls if it is negative). Also included in Figure 12.2 is a function with both linear and quadratic components. Here you can see that the curvature imposed by a quadratic function is superimposed on a rising linear trend.

![Figure 12.2 Typical linear and quadratic functions](image)

**Figure 12.2** Typical linear and quadratic functions
Tests of trend differ in an important way from the comparison procedures we have been discussing. In all of the previous examples, the independent variable was generally qualitative. Thus, for example, we could have written down the groups in the morphine-tolerance example in any order we chose. Moreover, the $F$ or $t$ values for the contrasts depended only on the numerical value of the means, not on which particular groups went with which particular means. In the analysis we are now considering, $F$ or $t$ values will depend on both the group means and the particular ordering of those means. To put this slightly differently using the aspirin example, a test between the largest and the smallest means will not be affected by which group happens to have which mean. However, in trend analysis the results would be quite different if the 1-grain and 5-grain groups had the smallest and largest means than if the 4- and 2-grain groups had the smallest and largest means, respectively. (A similar point was made in Section 6.7 in discussing the nondirectionality of the chi-square test.)

**Alcohol and Aggression**

We will take the example of the study by Giancola and Corman (2007) on the relationship between alcohol and aggression in the presence of a distracting task. This is the study that began Chapter 11, and the data and summary tables are presented in Table 12.5. The levels of the independent variable (D0, D2, D4, D6, and D8) refer to the number of illuminated squares that participants had to report in the distracting task. Notice that they increase linearly from 0 to 8. Obvious questions are whether aggression decreases linearly with distraction, whether it decreases to some level and then holds steady, or whether it decreases and then starts to increase again. A glance at the means would suggest the latter, but we want to test that.

A standard one-way analysis of variance produced the following summary table.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatments</td>
<td>4</td>
<td>62.460</td>
<td>15.615</td>
<td>6.90*</td>
</tr>
<tr>
<td>Error</td>
<td>55</td>
<td>124.458</td>
<td>2.263</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>59</td>
<td>189.918</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* $p < .05$

**Table 12.5** Level of shock administered as a function of task difficulty

<table>
<thead>
<tr>
<th></th>
<th>D0</th>
<th>D2</th>
<th>D4</th>
<th>D6</th>
<th>D8</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.28</td>
<td>-1.18</td>
<td>-0.41</td>
<td>-0.85</td>
<td>2.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.35</td>
<td>0.15</td>
<td>-1.25</td>
<td>0.14</td>
<td>0.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.31</td>
<td>1.36</td>
<td>-1.33</td>
<td>-1.38</td>
<td>2.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.06</td>
<td>2.61</td>
<td>-0.47</td>
<td>1.28</td>
<td>-1.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.59</td>
<td>0.66</td>
<td>-0.60</td>
<td>1.85</td>
<td>5.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.25</td>
<td>1.32</td>
<td>-1.72</td>
<td>-0.59</td>
<td>2.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.98</td>
<td>0.73</td>
<td>-1.74</td>
<td>-1.30</td>
<td>6.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.53</td>
<td>-1.06</td>
<td>-0.77</td>
<td>0.38</td>
<td>2.94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2.68</td>
<td>0.24</td>
<td>-0.41</td>
<td>-0.35</td>
<td>0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.64</td>
<td>0.27</td>
<td>-1.20</td>
<td>2.29</td>
<td>3.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.26</td>
<td>0.72</td>
<td>-0.31</td>
<td>-0.25</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.06</td>
<td>2.28</td>
<td>-0.74</td>
<td>0.51</td>
<td>-0.66</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Mean | 1.802 | 0.675 | -0.912 | 0.544 | 1.889 | 0.800 |
| Mean Mean | 1.656 | 1.140 | 0.515 | 1.180 | 2.370 | 1.800 |
| St. Dev. | 2.741 | 1.299 | 0.265 | 1.394 | 5.616 | 3.168 |
From the summary table it is apparent that there are significant differences among the five groups. One way to examine these differences would be to plot the group means as a function of the number of illuminated squares. This is shown in Figure 12.3, and it is apparent that the effectiveness of distraction in reducing aggression increases up to a point but with higher levels of distraction aggression returns to base levels.

The overall analysis of variance really asked if a horizontal straight line through \( Y = 0.800 \) (the grand mean) would fit the data adequately. The \( F \) led to rejection of that null hypothesis because sample means were as high as 1.9 and as low as \(-0.9\). Our next question asks whether a nonhorizontal straight line provides a good fit to the data. A glance at Figure 12.3 would suggest that this probably is not the case, or at least the fit would be poor. We will then follow that question by asking whether systematic residual (nonerror) variance remains in the data after fitting a linear function, and, if so, whether this residual variance can be explained by a quadratic function.

To run a trend analysis, we will return to the material we discussed under the headings of linear and orthogonal contrasts. (Don’t be confused by the use of the word linear in the last sentence. We will use the same approach when it comes to fitting a quadratic function. Linear in this sense simply means that we will form a linear combination of coefficients and means, where nothing is raised to a power.)

In Section 12.3 we defined a linear contrast as

\[
\psi = a_1 \bar{x}_1 + a_2 \bar{x}_2 + a_3 \bar{x}_3 + \cdots + a_k \bar{x}_k = \sum a_i \bar{x}_i
\]

The only difference between what we are doing here and what we did earlier will be in the coefficients we use. In the case in which there are equal numbers of subjects in the groups and the values on the abscissa are equally spaced, the coefficients for linear, quadratic, and higher-order functions (polynomial trend coefficients) are easily tabled and are found in Appendix Polynomial. From Appendix Polynomial we find that for five groups the linear and quadratic coefficients are

**Linear:** \([-2, -1, 0, 1, 2]\)

**Quadratic:** \([2, -1, -2, -1, 2]\)

We will not be using the cubic and quartic coefficients, but their use will be evident from what follows. Notice that like any set of orthogonal linear coefficients, the requirements that \( \sum a_j = 0 \) and \( \sum a_i b_j = 0 \) are met. The coefficients do not form a “standard set,” because the sum of the absolute values of the coefficients does not equal 2. That is not a problem here.

As you should recall from Section 12.3, we calculate a sum of squares for the contrast as

\[
SS_{\text{contrast}} = \frac{mb^2}{\sum a_j^2}
\]

**Figure 12.3** Aggression as a function of distraction
In our case,

\[ \psi_{\text{linear}} = (-2)1.802 + (-1)0.675 + (0) -1.912 + (1)0.544 + (2)1.889 \]
\[ = 0.0425 \]

\[ SS_{\text{linear}} = \frac{n\psi^2}{\sum a^2} = \frac{12(0.0425^2)}{10} = .002 \]

Like all contrasts, this contrast has a single degree of freedom, and therefore \( SS_{\text{linear}} = MS_{\text{linear}} \). As you probably suspect from what you already know, we can convert this mean square for the contrast to an \( F \) by dividing by \( MS_{\text{error}} \):

\[ F = \frac{MS_{\text{linear}}}{MS_{\text{error}}} \]
\[ = \frac{.002}{2.263} = 0.0009 \]

This is an \( F \) on 1 and 55 degrees of freedom; from Appendix \( F \) we find that \( F(0.05, 1, 55) = 4.015 \). Because the \( F \) for the linear component (0.0009) is far less than 4.015, we will retain \( H_0 \) and conclude that there is no significant linear trend in our means. In other words, we will conclude that aggressiveness does not vary linearly with increasing levels of distraction. Notice here that a nonsignificant \( F \) means that the trend component we are testing is not significantly different from 0.13

Although our data do not show a linear trend, it would appear that they do show a quadratic one—the line goes down and then up. It is also common to find data with both a linear and a quadratic trend, which would look like the plot in Figure 12.2c.

The next step is to ask whether the residual variance remaining after we fit the linear component is significantly greater than the error variance that we already know is present. If \( SS_{\text{linear}} \) accounted for virtually all of \( SS_{\text{treatment}} \), there would be little or nothing left over for higher-order terms to explain. On the other hand, if \( SS_{\text{linear}} \) were a relatively small part of \( SS_{\text{treatment}} \), then it would make sense to look for higher-order components. From our previous calculations we obtain

\[ SS_{\text{residual}} = SS_{\text{Treatment}} - SS_{\text{linear}} \]
\[ = 62.460 - 0.002 \]
\[ = 62.458 \]

\[ df_{\text{residual}} = df_{\text{Treatment}} - df_{\text{linear}} \]
\[ = 4 - 1 \]
\[ = 3 \]

\[ MS_{\text{residual}} = \frac{SS_{\text{residual}}}{df_{\text{residual}}} \]
\[ = \frac{62.458}{3} \]
\[ = 20.819 \]

13 I recently received a message from someone with similar data. He was studying the experimental hypothesis that drug effects increased with dosage. He had obtained a nonsignificant overall \( F \), but when he computed a test on linear trend, the result was “highly significant.” He wanted to know what to do. Because the linear trend tested his hypothesis directly, whereas the overall \( F \) did not, my recommendation was to rely solely on the test for trend.
Because $F$ for the residual is greater than $F_{(0.05, 3, 55)}$, we will reject the null hypothesis and conclude that there is significant variability left to be explained over and above that accounted for by the linear component. The calculations for the quadratic component are essentially the same as above with the exception that quadratic coefficients take the place of linear ones.

$$\psi_{\text{quadratic}} = (2)(1.802) + (-1)(0.675) + (-2)(-0.912) + (-1)(0.544) + (2)(1.889) = 7.989$$

$$SS_{\text{quadratic}} = \frac{n\psi^2}{\sum b_i^2} = \frac{12(7.989^2)}{14} = 54.706$$

$$F = \frac{MS_{\text{quadratic}}}{MS_{\text{error}}} = \frac{54.706}{2.263} = 24.17$$

This test is clearly significant, showing that there is a quadratic trend to our data. That should not come as any surprise because it is obvious from the plot.

A word of caution is in order at this point. You might be tempted to go ahead and apply the cubic and quartic coefficients that you find in Appendix Polynomial. You might also observe that having done this, the four sums of squares ($SS_{\text{linear}}, \ldots, SS_{\text{quartic}}$) will sum to $SS_{\text{treatment}}$ and be impressed that you have accounted for all of the sums of squares between groups. Before you get too impressed, think about how proud you would be if you showed that you could draw a straight line that exactly fit two points. The same idea applies here. Regardless of the data, you know before you begin that a polynomial of order $k - 1$ will exactly fit $k$ points. That is one reason why I was not eager to go much beyond fitting the quadratic components to the data at hand. Moreover, if you were to fit a fourth-order polynomial and find that the quartic component was significant, what would you have to say about the results? A linear or quadratic component would make some sense, but a quartic component could not be explained by any theory I know.

### Unequal Intervals

In the preceding section we assumed that the levels of the independent variable are equally spaced along some continuum. It is possible to run a trend analysis when we do not have equal intervals, and the arithmetic is the same. The only problem comes when we try to obtain the trend coefficients, because we cannot take our coefficients from Appendix Polynomial unless the intervals are equal.

Calculating quadratic coefficients is not too difficult and a good explanation can be found in Keppel (1973). For higher-order polynomials the calculations are more laborious,
but a description of the process can be found in Robson (1959). For most people, their analyses will be carried out with standard statistical software, and that software will often handle the problem of unequal spacing. Without diving deeply into the manuals, it is often difficult to determine how your software handles the spacing problem.

An example containing both a quadratic and a cubic component can be found in Exercise 12.25. Working through that exercise can teach you a lot about trend analysis.

### Key Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition/Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate per comparison (PC)</td>
<td>(12.1)</td>
</tr>
<tr>
<td>Familywise error rate (FW)</td>
<td>(12.1)</td>
</tr>
<tr>
<td>Omnibus null hypothesis</td>
<td>(12.1)</td>
</tr>
<tr>
<td>A priori comparisons</td>
<td>(12.1)</td>
</tr>
<tr>
<td>Post hoc comparisons</td>
<td>(12.1)</td>
</tr>
<tr>
<td>Contrasts</td>
<td>(12.3)</td>
</tr>
<tr>
<td>Linear combination</td>
<td>(12.3)</td>
</tr>
<tr>
<td>Linear contrast</td>
<td>(12.3)</td>
</tr>
<tr>
<td>Partition</td>
<td>(12.3)</td>
</tr>
<tr>
<td>Orthogonal contrasts</td>
<td>(12.3)</td>
</tr>
<tr>
<td>Dunn’s test</td>
<td>(12.3)</td>
</tr>
<tr>
<td>Bonferroni test</td>
<td>(12.3)</td>
</tr>
<tr>
<td>Bonferroni inequality</td>
<td>(12.3)</td>
</tr>
<tr>
<td>Dunn-Šidák test</td>
<td>(12.3)</td>
</tr>
<tr>
<td>Fisher’s least significance difference (LSD)</td>
<td>(12.6)</td>
</tr>
<tr>
<td>Studentized range statistic (q)</td>
<td>(12.6)</td>
</tr>
<tr>
<td>Tukey test</td>
<td>(12.6)</td>
</tr>
<tr>
<td>Tukey HSD (honestly significant difference) test</td>
<td>(12.6)</td>
</tr>
<tr>
<td>Newman–Keuls test</td>
<td>(12.6)</td>
</tr>
<tr>
<td>Ryan procedure (REGWQ)</td>
<td>(12.6)</td>
</tr>
<tr>
<td>Scheffé test</td>
<td>(12.6)</td>
</tr>
<tr>
<td>Dunnett’s test</td>
<td>(12.6)</td>
</tr>
<tr>
<td>False Discovery Rate (FDR)</td>
<td>(12.6)</td>
</tr>
<tr>
<td>Benjamini and Hochberg’s Linear Step Up (LSU) procedure</td>
<td>(12.6)</td>
</tr>
<tr>
<td>Trend</td>
<td>(12.9)</td>
</tr>
<tr>
<td>Quadratic function</td>
<td>(12.9)</td>
</tr>
<tr>
<td>Polynomial trend coefficients</td>
<td>(12.9)</td>
</tr>
</tbody>
</table>

### Exercises

12.1 Assume that the data that follow represent the effects of food and/or water deprivation on behavior in a learning task. Treatments 1 and 2 represent control conditions in which the animal received ad lib food and water (1) or else food and water twice per day (2). In treatment 3 animals were food deprived, in treatment 4 they were water deprived, and in treatment 5 they were deprived of both food and water. The dependent variable is the number of trials to reach a predetermined criterion. Assume that before running our experiment we decided that we wanted to compare the combined control groups (treatments 1 and 2) with the combined experimental groups, the control groups with each other, the singly deprived treatments with the doubly deprived treatment, and the singly deprived treatments with each other.

<table>
<thead>
<tr>
<th>Ad Lib Control</th>
<th>Two per Day Control</th>
<th>Food Deprived</th>
<th>Water Deprived</th>
<th>Food and Water Deprived</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>20</td>
<td>6</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>21</td>
<td>23</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>27</td>
<td>6</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>15</td>
<td>25</td>
<td>11</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>90</td>
<td>120</td>
<td>40</td>
<td>60</td>
<td>55</td>
</tr>
</tbody>
</table>

a. Analyze the data using linear contrasts.
b. Show that the contrasts are orthogonal.
c. Show that the sums of squares for the contrasts sum to $SS_{treat}$. 

© Cengage Learning 2013
12.2 Using the data from the first exercise in Chapter 11, compute the linear contrasts for Counting and Rhyming versus Adjective and Imagery, and then compare the Adjective versus Imagery conditions. Use \( \alpha = .05 \) for each contrast. (Note that this and subsequent exercises refer to exercises in Chapter 11, not this chapter.)

12.3 What would be the per comparison and familywise error rates in Exercise 12.2? (Hint: Are the contrasts orthogonal?)

12.4 Compute \( F \) for the linear contrast on the two groups in Exercise 11.2. Is this a waste of time? Why or why not?

12.5 Compute the Studentized range statistic for the two groups in Exercise 11.2 and show that it is equal to \( r \sqrt{2} \) (where \( t \) is taken from Exercise 11.2b).

12.6 Compute the \( F_s \) for the following linear contrasts in Exercise 11.3. Save the results for use in Chapter 13.
   a. 1 and 2 versus 3 and 4
   b. 1 and 3 versus 2 and 4
   c. 1 and 4 versus 2 and 3
   d. What questions do the contrasts in (a), (b), and (c) address?

12.7 Run the Bonferroni \( t \) test on the data for Exercise 11.1, using the contrasts supplied in Exercise 12.2. Set the maximum \( FW \) at .05.

12.8 Run a Tukey test on the example given in Table 11.2 (page 332) and interpret the results.

12.9 Why might you be more interested in running specific contrasts on the data referred to in Exercises 12.8?

12.10 Run the Games and Howell (1976) approach to Tukey’s HSD procedure for unequal sample sizes on the following data.

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{x}_j )</td>
<td>10</td>
<td>18</td>
<td>19</td>
<td>21</td>
<td>29</td>
</tr>
<tr>
<td>( n_j )</td>
<td>8</td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>( s^2_j )</td>
<td>7.4</td>
<td>8.9</td>
<td>8.6</td>
<td>7.2</td>
<td>9.3</td>
</tr>
</tbody>
</table>

12.11 Use the Scheffé test on the data in Exercise 12.10 to compare groups 1, 2, and 3 (combined) with groups 4 and 5 (combined). Then compare group 1 with groups 2, 3, and 4 (combined).

12.12 Apply the Tukey procedure to the log transformed THC data from Table 11.6 (page 348). What is the maximum \( FW \) for this procedure?

12.13 Using SPSS to apply Dunnett’s test to the log transformed data in Table 11.6.

12.14 How could a statistical package that did not have a Bonferroni command be used to run the Bonferroni \( t \) test on the data in Exercise 12.7?

12.15 The Bonferroni multistage test is sometimes referred to as a modified sequentially rejective procedure. Why?

12.16 Fit linear and quadratic trend components to the Conti and Musty (1984) log transformed data in Table 11.6. The control condition received 0 \( \mu g \) of THC. For purposes of this example, assume that there were 10 subjects in all groups. (You could add a 2.56 to the 0.5 \( \mu g \) group and a 2.35 and 2.36 to the 1 \( \mu g \) group without altering the results appreciably.) The linear coefficients (calculated with unequal spacing on the independent variable) are \([-0.72, -0.62, -0.22, 0.28, 1.28]\]. The quadratic coefficients are \([0.389, 0.199, -0.362, -0.612, 0.387]\). Verify your answers using SPSS ONEWAY if you have it available.

12.17 Write a brief report of the results computed for Exercise 12.17.

12.18 Use any statistical package to apply the Tukey and Scheffé procedures to the data from Introini-Collison and McGaugh (1986), described in the exercises for Chapter 11 (page 366).
Do these analyses for both Epineq.dat and Epinuneq.dat, which are on the book’s Web site.
Do not combine across the levels of the interval variable.

12.19 In Exercise 12.18 it would not have made much of a difference whether we combined the data across the three intervals or not. Under what conditions would you expect that it would make a big difference?

12.20 Using the data in Epineq.dat, compute both the linear and quadratic trend tests on the three drug dosages. Do this separately for each of the three intervals. (Hint: The linear coefficients are \([-0.597110, -0.183726, 0.780836]\), and the quadratic coefficients are \([0.556890, -0.795557, 0.238667]\).)

12.21 Interpret the results in Exercise 12.20.

12.22 Stone, Rudd, Ragozzino, and Gold (1992) investigated the role that glucose plays in memory. Mice were raised with a 12-hour light-on/light-off cycle, starting at 6:00 A.M. During training mice were placed in the lighted half of an experimental box and given foot shock when they moved into the dark half. The mice quickly learned to stay in the lighted half. The day/night cycle was then advanced by 4 hours for all mice, which is known to interfere with memory of the original training. Three days later mice were retested 30 minutes after being injected with 0, 1, 10, 100, 250, or 500 mg/kg of sucrose. The purpose was to see whether sucrose would reduce the disruptive effects of changing the diurnal cycle, and whether different doses would have different effects. Data that have been generated to loosely mimic the results of Stone et al., are given below, where the dependent variable is the latency to enter the dark chamber.

<table>
<thead>
<tr>
<th>Glucose Level in mg/kg</th>
<th>0</th>
<th>1</th>
<th>10</th>
<th>100</th>
<th>250</th>
<th>500</th>
</tr>
</thead>
<tbody>
<tr>
<td>295</td>
<td>129</td>
<td>393</td>
<td>653</td>
<td>379</td>
<td>521</td>
<td></td>
</tr>
<tr>
<td>287</td>
<td>248</td>
<td>484</td>
<td>732</td>
<td>530</td>
<td>241</td>
<td></td>
</tr>
<tr>
<td>91</td>
<td>350</td>
<td>308</td>
<td>570</td>
<td>364</td>
<td>162</td>
<td></td>
</tr>
<tr>
<td>260</td>
<td>278</td>
<td>112</td>
<td>434</td>
<td>385</td>
<td>197</td>
<td></td>
</tr>
<tr>
<td>193</td>
<td>150</td>
<td>132</td>
<td>690</td>
<td>355</td>
<td>156</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>195</td>
<td>414</td>
<td>679</td>
<td>558</td>
<td>384</td>
<td></td>
</tr>
</tbody>
</table>

12.23 Using the data from Exercise 12.1, compute confidence interval for the first comparison (contrast) described in that question. Interpret your answer. (If you use SPSS, use the Compare Means/One-Way ANOVA procedure, which allows you to specify coefficients.)

12.24 Using the data from Exercise 12.1, compute effect sizes on all of the contrasts that you ran with that question. How would you interpret these effect sizes? Why are these called standardized effect sizes, and what would an unstandardized effect size be?

12.25 Using the data from Exercise 11.27, perform the appropriate test(s) to draw meaningful conclusions from the study by Davey et al. (2003).

12.26 In Exercise 11.8 we considered a study by Foa et al. concerning therapy for victims of rape. The raw data can be found on the Web site at Ex12.26.dat. Apply the Benjamini & Hochberg LSU procedure to these data.
Discussion Questions

12.27 Students often have difficulty seeing why a priori and post hoc tests have different family-wise error rates. Make up an example (not necessarily from statistics) that would help to explain the difference to others.

12.28 Find an example in the research literature of a study that used at least five different conditions and create a data set that might have come from this experiment. Apply several of the techniques we have discussed, justifying their use and interpreting the results. (You would never apply several different techniques to a set of data except for an example such as this. 

*Hint:* You can generate data with a given mean and variance by taking any set of numbers [make them at least unimodal and symmetrical], standardizing them, multiplying the standard scores by the desired standard deviation, and then adding the desired mean to the result. Do this for each group separately and you will have your data.)