CHAPTER 12 MULTIPLE COMPARISONS AMONG TREATMENT MEANS

OBJECTIVES

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

CONTENTS

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 which 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 be able 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 μ_6.

Many different techniques for making comparisons among means are available; here we will consider the most common and useful ones. A thorough discussion of this topic can be found in Miller (1981), and in Hochberg and Tamhane (1987), and Toothaker (1991). 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.

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.[1] 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.

We will distinguish two basic ways of specifying error rates, or the probability of Type I errors.[2] In doing so, we shall use the terminology that has become more or less standard since an extremely important unpublished paper by Tukey in 1953.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] Some authors choose among tests on the basis of power and are concerned with the probability of finding any or all significant differences among pairs of means (any-pairs power and all-pairs power). In this chapter, however, we will focus on the probability of Type I errors and the way in which different test procedures deal with these error rates.

[2] There is a third 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

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 $t_{.05}$, 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*)**.[3] 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.

---

attempt to control it directly. We can easily calculate it, however, as $PE = c\alpha$, where *c* is the number of comparisons and α is the per comparison error rate.

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 α' represent the error rate for any one comparison and *c* represent the number of comparisons, then

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

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 $PC \leq FW \leq c\alpha$ and in most reasonable cases *FW* is in the general vicinity of $c\alpha$. This fact becomes important when we consider the Bonferroni tests.

The null hypothesis and error rates

We have been speaking as if the null hypothesis in question were what is usually called the *complete null hypothesis* ($\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$). In fact, this is the null hypothesis tested by the overall analysis of variance. In many 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

[3] 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.

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.

A priori versus post hoc comparisons

It is often helpful to distinguish 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. To take a simple example, consider a situation in which you have five means. In this case, there are 10 possible comparisons involving pairs of means (e.g., $\bar{X}_1$ versus $\bar{X}_2$, $\bar{X}_1$ versus $\bar{X}_3$, and so on). Assume that the complete null hypothesis is true but that by chance two of the means are far enough apart to lead us erroneously to reject $H_0 : \mu_i = \mu_j$. In other words, the data contain one Type I error. If you have to plan your single comparison in advance, you have a probability of .10 of hitting on the 1 comparison out of 10 that will involve a Type I error. If you look at the data first, however, you are certain to make a Type I error, assuming that you are not so dim that you test anything other than the largest difference. In this case, you are implicitly making all 10 comparisons in your head, even though you perform the arithmetic for only the largest one. In fact, for some post hoc tests, we will adjust the error rate as if you literally made all 10 comparisons.

This simple example demonstrates that if comparisons are planned in advance (*and are a subset of all possible comparisons*), the probability of a Type I error is smaller than if the comparisons are arrived at on a post hoc basis. It should not surprise you, then, that we will treat a priori and post hoc comparisons separately. 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.

Significance of the overall *F*

Some controversy surrounds the question of whether one should insist that the overall *F* on groups 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.

The logic behind most of our post hoc tests, however, does not require overall significance before making specific comparisons. First of all, the hypotheses tested by the overall test and a multiple-comparison test are quite different, with quite different levels of power. For example, the overall *F* actually distributes differences among groups across the number of degrees of freedom for groups. This has the effect of diluting the overall *F* in the situation where several group means are equal to each other but different from

some other mean. Second, requiring overall significance will actually change the *FW*, making the multiple comparison tests conservative. The tests were designed, and their significance levels established, without regard to the overall *F*.

Wilcox (1987a) has considered this issue and suggested that "there seems to be little reason for applying the (overall) *F* test at all" (p. 36). Wilcox would jump straight to multiple-comparisons without even computing the *F*. Others have said much the same thing. That position may be a bit extreme, but it does emphasize the point. And perhaps it is not all that extreme. If you recognize that typical multiple-comparison procedures do not require a significant overall *F*, you will examine group differences regardless of the value of that *F*. Why, then, do we even need that *F* except to provide a sense of closure? The only reason I can think of is "tradition," and that is a powerful force.

12.2. 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 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) 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. If, however, after repeated injections of atropine, physiological saline (which should have no effect whatsoever) is suddenly injected (*in the same physical setting*), 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.)

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 as sensitive to pain as would be a naive animal (one who had never received morphine) because of the tolerance that has 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. Siegel also reasoned that if you gave the animal repeated morphine injections in one setting but then tested it 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 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.

Our version of Siegel's experiment is based on the prediction 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 levels of pain sensitivity on that fourth trial. Group M-S received morphine (in the test setting) on the first three trials but then received saline on the fourth trial. 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

S-M = Mc-M > M-M ? S-S > M-S

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

Table 12.1 Data and analysis on morphine tolerance

(a) Data

	M-S	**M-M**	**S-S**	**S-M**	**Mc-M**
	3	2	14	29	24
	5	12	6	20	26
	1	13	12	36	40
	8	6	4	21	32
	1	10	19	25	20
	1	7	3	18	33
	4	11	9	26	27
	9	19	21	17	30
Mean	4.00	10.00	11.00	24.00	29.00
St. Dev	3.16	5.13	6.72	6.37	6.16

(b) Summary Table

Source	***df***	**SS**	**MS**	***F***
Treatment	4	3497.60	874.40	27.33*
Error	35	1120.00	32.00	
Total	39	4617.60		

P < .05

The results of this experiment 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.

12.3. A PRIORI COMPARISONS

There are two reasons for starting our discussion with *t* tests. In the first place, standard *t* tests between pairs of means can, in a limited number of situations, 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, 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_{error} from the overall analysis of variance and evaluate the *t* on df_{error} degrees of freedom. When the variances are heterogeneous but the sample sizes are equal, we do not use MS_{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). (For an evaluation of this approach, albeit for a slightly different test statistic, see Games and Howell, 1976.)

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 $\alpha = .01$.

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.) 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 Mc-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_{error} as the pooled variance estimate and will evaluate the result on df_{error} degrees of freedom.

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

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{MS_{error}}{n} + \frac{MS_{error}}{n}}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{2MS_{error}}{n}}}$$

Substituting the data from our example, group M-S versus group S-S yields

$$\bar{X}_{M-S} = 4.00 \qquad \bar{X}_{S-S} = 11.00 \qquad MS_{error} = 32.00$$

$$t = \frac{\bar{X}_{M-S} - \bar{X}_{S-S}}{\sqrt{\frac{2MS_{error}}{n}}} = \frac{4.00 - 11.00}{\sqrt{\frac{2(32.00)}{8}}} = \frac{-7}{\sqrt{8}} = -2.47$$

And group Mc-M versus group M-M yields

$$\bar{X}_{Mc-M} = 29.00 \quad \bar{X}_{M-M} = 10.00 \quad \text{MS}_{\text{error}} = 32.00$$

$$t = \frac{\bar{X}_{Mc-M} - \bar{X}_{M-M}}{\sqrt{\frac{2\text{MS}_{\text{error}}}{n}}} = \frac{29.00 - 10.00}{\sqrt{\frac{2(32.00)}{8}}} = \frac{19}{\sqrt{8}} = 6.72$$

Both of these obtained values of *t* would be evaluated against $t_{.025}$ (35) = $\pm$ 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 α = .05 per comparison, the *FW* will approach .10 if both null hypotheses are true, which seems quite unlikely.

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 under a priori tests because that is where they are most commonly used.

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

$$L = a_1\bar{X}_1 + a_2\bar{X}_2 + \cdots + a_k\bar{X}_k = \sum a_j\bar{X}_j$$

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**. 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, or the mean of one condition with the combined mean of several conditions. As an example, consider three means ($\bar{X}_1$, $\bar{X}_2$, and $\bar{X}_3$). Letting $a_1 = 1$, $a_2 = -1$, and $a_3 = 0$, $\sum a_j = 0$,

$$L = (1)\bar{X}_1 + (-1)\bar{X}_2 + 0\bar{X}_3 = \bar{X}_1 - \bar{X}_2$$

In this case, *L* 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

$$L = (1/2)\bar{X}_1 + (1/2)\bar{X}_2 + (-1)\bar{X}_3 = \frac{\bar{X}_1 + \bar{X}_2}{2} - \bar{X}_3$$

in which case *L* 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 let

$$L = a_1\bar{X}_1 + a_2\bar{X}_2 + \cdots + a_k\bar{X}_k = \sum a_j\bar{X}_j$$

it can be shown that

$$\mathrm{SS}_{\text{contrast}} = \frac{nL^2}{\Sigma a_j^2} = \frac{n\left(\Sigma a_j\bar{X}_j\right)^2}{\Sigma a_j^2}$$

is a component of the overall $\mathrm{SS}_{\text{treat}}$ on 1 *df*, where *n* represents the number of scores per treatment.[4]

Suppose we have three treatments such that

$$n = 10 \qquad \bar{X}_1 = 1.5 \qquad \bar{X}_2 = 2.0 \qquad \bar{X}_3 = 3.0$$

For the overall analysis of variance,

$$SS_{treat} = n\Sigma(\bar{X}_j - \bar{X}_{..})^2 = 10\left[(1.5-2.167)^2 + (2-2.167)^2 + (3-2.167)^2\right]$$
$$= 10\left[0.4449 + 0.0278 + 0.6939\right] = 11.667$$

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

$$L = \sum a_j \bar{X}_j = (1)(1.5) + (1)(2.0) + (-2)(3.0) = -2.5$$
$$\text{SS}_{\text{contrast}} = \frac{nL^2}{\sum a_j^2} = \frac{10(-2.5)^2}{6} = \frac{62.5}{6} = 10.417$$

This sum of squares is a component of the overall SS_{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 $a_1 = 1,\ a_2 = -1$, and $a_3 = 0$. Then

$$L = \sum a_j \bar{X}_j = (1)(1.5) + (-1)(2.0) + (0)(3.0) = -0.5$$
$$\text{SS}_{\text{contrast}} = \frac{nL^2}{\sum a_j^2} = \frac{10(-0.5)^2}{2} = \frac{2.5}{2} = 1.25$$

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

$$\text{SS}_{\text{treat}} = \text{SS}_{\text{contrast}_1} + \text{SS}_{\text{contrast}_2}$$
$$11.667 = 10.417 + 1.25$$

---

[4] For unequal sample sizes, $\text{SS}_{\text{contrast}} = \dfrac{L^2}{\sum\left(a_j^2/n_j\right)}$

and thus the two contrasts account for all of the SS_{treat} and all of the *df* attributable to treatments. We say that we have *completely* **partitioned** SS_{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 one set the number of treatment groups in the other set, and vice versa. One arbitrary set of coefficients is then given a minus sign. For example, take the means

$$\bar{X}_1 \quad \bar{X}_2 \quad \bar{X}_3 \quad \bar{X}_4 \quad \bar{X}_5$$

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 $\bar{X}_4$ and $\bar{X}_5$ the $a_j = 3$. The second set contains two means, and therefore for $\bar{X}_1$, $\bar{X}_2$, and $\bar{X}_3$ the $a_j = 2$. We will let the 3s be negative. Then we have

Means:	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{X}_5$	
a_j**:**	2	2	2	-3	-3	$\sum a_j = 0$

Then $\sum a_j \bar{X}_j$ reduces to $2\left(\bar{X}_1 + \bar{X}_2 + \bar{X}_3\right) - 3\left(\bar{X}_4 + \bar{X}_5\right)$.

(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.) [5]

One final word about coefficients. If you are doing the computations by hand, you can save yourself a lot of arithmetic if you divide through by a common factor. For example, suppose that the steps we took had left us with

$a_j =$ 2 2 -2 -2

You can divide through by 2 and have

$a_j =$ 1 1 -1 -1

which simplifies the arithmetic considerably. In a similar vein, some authors use fractional coefficients, which make it clearer that we are really averaging sets of means. However, I think the arithmetic is cumbersome, and that it is easier to work with whole numbers. Take your choice.

5 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. (You may even send me a note if you need the information, and I will send it to you.)

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, and can thus be divided by $\mathrm{MS}_{\text{error}}$ to produce an *F*. Because *all* contrasts have 1 degree of freedom

$$F = \frac{\mathrm{MS}_{\text{contrast}}}{\mathrm{MS}_{\text{error}}} = \frac{nL^2 \Big/ \sum a_j^2}{\mathrm{MS}_{\text{error}}} = \frac{nL^2}{\sum a_j^2 \, \mathrm{MS}_{\text{error}}}$$

This *F* will have one and df_{error} degrees of freedom.

For our example, suppose we had planned (a priori) to compare the two groups receiving saline on trial 4 with three groups receiving morphine on trial 4. 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, this contrast is not independent of the first three contrasts.)

Groups:	**M-S**	**M-M**	**S-S**	**S-M**	**Mc-M**
Means:	**4.00**	**10.00**	**11.00**	**24.00**	**29.00**

						$\sum c_j^2$	$L = \sum a_j \bar{X}_j$
	Coefficient						
a_j	-3	2	-3	2	2	30	81
b_j	0	-1	0	0	1	2	19
c_j	-1	0	1	0	0	2	7
d_j	0	1	-1	0	0	2	-1

$$\mathrm{SS}_{\mathrm{contrast}_1} = \frac{n\left(\sum a_j \bar{X}_j\right)^2}{\sum a_j^2} = \frac{8(81)^2}{30} = \frac{52488}{30} = 1749.60$$

$$F = \frac{\mathrm{MS}_{\mathrm{contrast}}}{\mathrm{MS}_{\mathrm{error}}} = \frac{1749.6}{32.00} = 54.675$$

$$\mathrm{SS}_{\mathrm{contrast}_2} = \frac{n\left(\sum b_j \bar{X}_j\right)^2}{\sum b_j^2} = \frac{8(19)^2}{2} = \frac{2888}{2} = 1444.00$$

$$F = \frac{\mathrm{MS}_{\mathrm{contrast}}}{\mathrm{MS}_{\mathrm{error}}} = \frac{1444.00}{32.00} = 45.125$$

$$\mathrm{SS}_{\mathrm{contrast}_3} = \frac{n\left(\sum c_j \bar{X}_j\right)^2}{\sum c_j^2} = \frac{8(7)^2}{2} = \frac{392}{2} = 196.00$$

$$F = \frac{\mathrm{MS}_{\mathrm{contrast}}}{\mathrm{MS}_{\mathrm{error}}} = \frac{196.00}{32.00} = 6.125$$

$$\mathrm{SS}_{\mathrm{contrast}_4} = \frac{n\left(\sum d_j \bar{X}_j\right)^2}{\sum d_j^2} = \frac{8(-1)^2}{2} = \frac{8}{2} = 4.00$$

$$F = \frac{\mathrm{MS}_{\mathrm{contrast}}}{\mathrm{MS}_{\mathrm{error}}} = \frac{4.00}{32.00} = 0.125$$

Each of these *F* values can be evaluated against $F_{.05}(1,35) = 4.12$. 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. Note that contrasts 2 and 3 test the same hypotheses that we tested using individual *t* tests—and, as you should recall, when there is 1 *df* between groups, $F = t^2$. 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.

With four contrasts, we have an *FW* approaching .20. 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 α; 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 *FW* is to run fewer contrasts. For example, the comparison of M-M with S-S is probably not very important. Whether 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 $\bar{X}_4$ is likely to be greater than $\bar{X}_5$. These two 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.

Orthogonal coefficients

Given that sample sizes are equal, for contrasts to be orthogonal the coefficients must meet the following 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 $SS_{contrast}$ 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 number of treatments on the right, and vice versa, with one of the sets being negative. Thus, the coefficients are (3, 3,-2,-2,-2) 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, 2,-1,-1). Treatments 1 and 2 have coefficients of 0 because they are not part of this comparison. Treatment 3 has a coefficient of 2 because it is compared with two other treatments. Treatments 4 and 5 received coefficients of -1 because they are compared with one other treatment. The negative signs can be arbitrarily assigned to either side of the comparison.

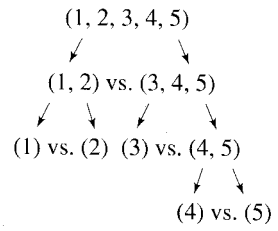
The previous procedure could be carried on until we have exhausted all possible sets of comparisons. This will occur when we have made as many comparisons as there are *df* for treatments. As a result of this procedure, we might arrive at the comparisons and coefficients shown in Figure 12.1. To show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

$$\sum a_j b_j = (3)(1) + (3)(-1) + (-2)(0) + (-2)(0) + (-2)(0) = 0$$

and

$$\sum a_j c_j = (3)(0) + (3)(0) + (-2)(2) + (-2)(-1) + (-2)(-1) = 0$$

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.



	Coefficients				
	3	3	−2	−2	−2
	1	−1	0	0	0
	0	0	2	−1	−1
	0	0	0	1	−1

Figure 12.1 Tree diagram illustrating orthogonal partition of SS_{treat}

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.

The actual computation of *F* 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. Authors went out of their way to impress on you the importance of orthogonality, and the need to feel somewhat guilty if you ran comparisons that were not orthogonal. That attitude has changed over the years. While 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. 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 α 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. (We do not usually have critical values of *t* for α between .05 and .01.) 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*.05 = .15. In more formal terms, if *c* represents the number of comparisons and α' represents the probability of a Type I error for each comparison, then *FW* is less than or equal to $c\alpha'$. From this it follows that if we set $\alpha' = \alpha/c$ for each comparison, where α = the desired maximum *FW*, then $FW \le c\alpha' = c(\alpha/c) = \alpha$. Dunn (1961) used this

inequality to design a 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 α. If you want to run each of three comparisons at α' $= \alpha/c = .05/3 = .0167$, you would need tables of critical values of *t* at $\alpha = .0167$. Dunn's major contribution was to provide such tables. (Although such tables are less crucial now that virtually all computer programs report exact probability values for each *F*, they still have a role to play, and her table can be found in the appendix of this book.)

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

$$t' = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{\text{MS}_{\text{error}}}{n} + \frac{\text{MS}_{\text{error}}}{n}}} = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2\text{MS}_{\text{error}}}{n}}}$$

and evaluate t' against the critical value of t' taken from Dunn's tables in Appendix t'. 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 evaluation. With unequal sample sizes but homogeneous variances, replace the *n*s 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.

$$L = \sum a_j \bar{X}_j \text{ and } t' = \frac{L}{\sqrt{\frac{\sum a_j^2 \, \text{MS}_{\text{error}}}{n}}}$$

This represents the most general form for the Bonferroni *t*, and it can be shown that if *L* is *any* linear combination (not necessarily even a linear contrast, requiring $\sum a_j = 0$), the *FW* with *c* comparisons is at most α (Dunn, 1961).[6] 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*.

A variation on the Bonferroni procedure was proposed by Šidák (1967). His test is based on the multiplicative inequality $p(FW) \le 1-(1-\alpha)^c$ 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 this test. [For four comparisons, the Šidák approach would test each comparison at $\alpha' = 1-(1-\alpha)^{1/4} = 1-.95^{.25} = 0.0127$ level, whereas the Bonferroni approach would test at $\alpha/c = .05/4 = .0125$. You can see that there is not a lot of difference in power.]

[6] Note the similarity between the right side of the equation and our earlier formula for *F* with linear contrasts. The resemblance is not accidental; one is just the square of the other.

When we considered linear contrasts, we ran four comparisons, which had an *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*(.05) = .20, we can use the Bonferroni *t* test. In each case, we will solve for t' and refer that to Dunn's tables. Taking the pairwise tests first, the calculations follow.

Mc-M versus M-M:

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

S-S versus M-S:

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

M-M versus S-S:

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

The calculations for the more complex contrast, letting the a_j = 2, 2, 2, -3, -3 as before, follow.

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

$$t' = \frac{\sum a_j \bar{X}_j}{\sqrt{\frac{\sum a_j^2 \text{MS}_{\text{error}}}{n}}} = \frac{(2)(24) + \cdots + (-3)(4)}{\sqrt{\frac{(30)(32.00)}{8}}} = \frac{81}{\sqrt{120}} = 7.39$$

From Appendix t', with $c = 4$ and $df_{\text{error}} = 35$, we find by interpolation $t'_{.05}(35) = 2.64$. In this case, the first and last contrasts are significant, but the other two are not.[7] 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*, as we usually should be, 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 covered 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 familywise error rate. These procedures have the advantage of setting a limit on the *FW*

error rate at α 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, multistage procedures would not be used as a substitute when making all pairwise comparisons among a set of means.

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 α.

Holm and Larzelere and Mulaik tests

Both Holm (1979) and Larzelere and Mulaik (1977) have proposed a multistage test that adjusts the denominator (c) in $\alpha' = \alpha/c$ depending on the number of null hypotheses remaining to be tested. Holm's test is generally referred to when speaking about the analysis of variance, whereas the Larzelere and Mulaik test is best known as a test of significance for a large set of correlation coefficients. The logic of the two tests is the same, though the method of calculation is different.

[7] The actual probabilities would be .000, .073, 1.00, and .000.

In the Holm procedure we calculate values of t' just as we did with the Bonferroni *t* test. For the equal *n* case, we compute

$$t' = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2\,\mathrm{MS}_{\mathrm{error}}}{n}}}$$

For the unequal *n* case, or when we are concerned about heterogeneity of variance, we compute

$$t' = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}}$$

We calculate t' for all contrasts of interest and then arrange the t' values in increasing order without regard to sign. This ordering can be represented as $|t'_1| \le |t'_2| \le |t'_3| \le \cdots \le |t'_c|$, where *c* is the total number of contrasts to be tested.

The first significance test is carried out by evaluating t_c against the critical value in Dunn's table corresponding to *c* contrasts. In other words, t_c is evaluated at $\alpha' = \alpha/c$. If this largest t' is significant, then we test the next largest $t'\left(i.e.\ t'_{c-1}\right)$ against the critical value in Dunn's table corresponding to *c*-1 contrasts. Thus, t'_{c-1} is evaluated at $\alpha' = \alpha/(c-1)$. The same procedure continues for $t'_{c-2}, t'_{c-3}, t'_{c-4}, \ldots$ until the test returns a nonsignificant result. At that point we stop testing. Holm has shown that such a procedure continues to keep $FW \le \alpha$, while offering a more powerful test.

The rationale behind the test is that when we reject the null for t_c, we are declaring that null hypothesis to be false. If it is false, that only leaves c-1 possibly true null hypotheses, and so we only need to protect against c-1 contrasts. A similar logic applies as we carry out additional tests. This logic makes particular sense when you know, even before the experiment is conducted, that several of the null hypotheses are almost certain to be false. If they are false, there is no point in protecting yourself from erroneously rejecting them.

To illustrate the use of Holm's test, consider our example on morphine tolerance. With the standard Bonferroni t test, we evaluated four contrasts with the following results, arranged by increasing magnitude of t':

Contrast	**Order (*i*)**	t'	$\boldsymbol{t'_{crit}}$
M-M vs. S-S	1	t' =-0.35	2.03
S-S vs. M-S	2	t' = 2.47	2.35*
Mc-M vs. M-M	3	t' = 6.72	2.52*
S-M, Mc-M, M-M vs. S-S, M-S	4	t' = 7.39	2.64*

*p<.05

If we were using the Bonferroni test, each of these t's would be evaluated against $t'_{.05}$ = 2.64, which is actually Student's t at $\alpha = 0.0125$. For Holm's test we vary the critical value in stages, depending on the number of contrasts that have not been tested. This number is indexed by "Order (i)" in the table above. These critical values are presented in the right-hand column above. They were taken, with interpolation, from Dunn's tables for $c = i$ and 35 degrees of freedom. For example, the critical value of 2.35 corresponds to the entry in Dunn's tables for c = 2 and df = 35. For the smallest t', the critical value came from the standard Student t distribution (Appendix t).

From this table you can see that the test on the complex contrast S-M, Mc-M, M-M vs. S-S, M-S required a t' of 2.64 or above to reject H_0. Because t' was 7.39, the difference was significant. The next largest t' was 6.72 for Mc-M vs. M-M, and that was also significant, exceeding the critical value of 2.52. The contrast S-S vs. M-S is tested as if there were only two contrasts in the set, and thus t' must exceed 2.35 for significance. Again this test is significant. If it had not been, we would have stopped at this point. But because it is, we continue and test M-M vs S-S, which is not significant. Because of the increased power of Holm's test over the Bonferroni *t* test, we have rejected one null hypothesis (S-S vs. M-S) that was not rejected by the Bonferroni.

Larzelere and Mulaik Test

Larzelere and Mulaik (1977) proposed a test equivalent to Holm's test, but their primary interest was in using that test to control *FW* when examining a large set of correlation coefficients. As you might suspect, something that controls error rates in one situation will tend to control them in another. I will consider the Larzelere and Mulaik test with respect to correlation coefficients rather than the analysis of variance, because such an example will prove useful to those who conduct research that yields large numbers of such coefficients. However, as you will see when you look at the calculations, the test would be applied in the same way whenever you have a number of test statistics with their associated probability values. If you had never heard of Larzelere and Mulaik, you could still accomplish the same thing with Holm's test. However, the different

calculational approach is instructive. It is worth noting that when these tests are applied in an analysis of variance setting we usually have a small number of comparisons. However, when they are used in a regression/correlation setting, we commonly test all pairwise correlations.

Compas, Howell, Phares, Williams, and Giunta (1989) investigated the relationship between daily stressors, parental levels of psychological symptoms, and adolescent behavior problems [as measured by Achenbach's Youth Self-Report Form (YSR) and by the Child Behavior Checklist (CBCL)]. The study represented an effort to understand risk factors for emotional/behavioral problems in adolescents. Among the analyses of the study was the set of intercorrelations between these variables at Time 1. These correlations are presented in Table 12.2.

Table 12.2 Correlations among behavioral and stress measures

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Mother							
(1) Stress	1.00	.69	.48	.37	-.02	.30	.03
(2) Symptoms		1.00	.38	.42	.12	.39	.19
Father							
(3) Stress			1.00	.62	.07	.22	.07
(4) Symptoms				1.00	.00	.24	.20
Adolescent							
(5) Stress					1.00	.11	.44
(6) CBCL						1.00	.23
(7) YSR							1.00

Most standard correlation programs print out a *t* statistic for each of these correlations. However, we know that with 21 hypothesis tests, the probability of a Type I error based

on that standard *t* test, if all null hypotheses were true, would be high. It would still be high if only a reduced set of them were true. For this reason we will apply the modified Bonferroni test proposed by Larzelere and Mulaik. There are two ways to apply this test to this set of correlations. For the first method we could calculate a *t* value for each coefficient, based on

$$t = \frac{r\sqrt{(N-2)}}{\sqrt{(1-r^2)}}$$

(or take the *t* from a standard computer printout) and then proceed exactly as we did for the Holm procedure. Alternatively, we could operate directly on the two-tailed *p* values associated with the *t* test on each correlation. These *p* values can be taken from standard computer printouts, or they can be calculated using commonly available programs. For purposes of an example, I will use the *p*-value approach.

Table 12.3 shows the correlations to be tested from Table 12.2 as well as the associated *p* values. The *p* values have been arranged in increasing numerical order. (Note that the sign of the correlation is irrelevant—only the absolute value matters.)

Table 12.3 Significance tests for correlations in Table 12.2

Pair	*i*	Correlation	*p* value	$\alpha/(k-i+1)$
1 vs. 2	1	.69	.0000	.00238*
3 vs. 4	2	.62	.0000	.00250*
1 vs. 3	3	.48	.0000	.00263*
5 vs. 7	4	.44	.0000	.00278*
2 vs. 4	5	.42	.0000	.00294*
2 vs. 6	6	.39	.0001	.00313*
2 vs. 3	7	.38	.0001	.00333*
1 vs. 4	8	.37	.0002	.00357*
1 vs. 6	9	.30	.0028	.00385*
4 vs. 6	10	.24	.0179	.00417
6 vs. 7	11	.23	.0236	.00455
3 vs. 6	12	.22	.0302	.00500
4 vs. 7	13	.20	.0495	.00556
2 vs. 7	14	.19	.0618	.00625
2 vs. 5	15	.12	.2409	.00714
5 vs. 6	16	.11	.2829	.00833
3 vs. 5	17	.07	.4989	.01000
3 vs. 7	18	.07	.4989	.01250
1 vs. 7	19	.03	.7724	.01667
1 vs. 5	20	-.02	.8497	.02500
4 vs. 5	21	.00	1.0000	.05000

$P < .05$

The right-hand column gives the value of α' required for significance. For example, if we consider 21 contrasts to be of interest, $\alpha' = \alpha/(k-i+1) = .05/21 = .00238$. By the time we have rejected the first four correlations and wish to test the fifth largest, we are going to behave as if we want a Bonferroni *t* adjusted for just the *k*-*i* + 1 = 21-5 + 1 = 21-4 = 17 remaining correlations. This correlation will be tested at $\alpha' = \alpha/(k-i+1) = .05/17 = .00294$.

Each correlation coefficient is tested for significance by comparing the *p* value associated with that coefficient with the entry in the final column. For example, for the largest

correlation coefficient out of a set of 21 coefficients to be significant, it must have a probability (under $H_0 : \rho = 0$) less than .00238. Because the probability for r = .69 is given as .0000 (there are no nonzero digits until the sixth decimal place), we can reject H_0 and declare that correlation to be significant.

Having rejected H_0 for the largest coefficient, we then move down to the second row, comparing the obtained p value against p = .00250. Again we reject H_0 and move on to the third row. We continue this procedure until we find a row at which the obtained p value in column 4 exceeds the critical p value in column 5. At that point we declare that correlation to be nonsignificant and stop testing. All correlations below that point are likewise classed as nonsignificant. For our data, those correlations equal to or greater than .30 are declared significant, and those below .30 are nonsignificant. The significant correlations are indicated with an asterisk in the table.

Had we used a standard Bonferroni test, we would have set α' = .05/21 = .00238, and a correlation less than .37 would not have been significant. In this particular case the multistage test made only a small difference. But often the difference is substantial in terms of the number of coefficients that are declared significant.

One more comment

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 α by the number of tests he ran, and declare a test to be significant only if its resulting probability was less than α /c. You don't even need to know anything about logistic regression to do that.

12.4. 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. However, many experiments involve many hypotheses[8] and/or hypotheses that are arrived at only after the data have been examined. In this situation, a number of a posteriori or post hoc techniques are available.

Fisher's least significant difference procedure

One of the oldest methods for making post hoc comparisons is known as **Fisher's least significant difference (LSD)** test (also known as Fisher's protected *t*). 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, the overall *F* no longer affords protection for *FW*. For this reason, many people 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 a good 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 only 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.

[8] 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.

The Studentized range statistic (q)

Because many of the post hoc tests 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

$$q_r = \frac{\bar{X}_l - \bar{X}_s}{\sqrt{\frac{\text{MS}_{\text{error}}}{n}}}$$

where $\bar{X}_l$ and $\bar{X}_s$ 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

$$q_r = \frac{\bar{X}_l - \bar{X}_s}{\sqrt{\frac{\text{MS}_{\text{error}}}{n}}}$$

$$t = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{\frac{2\left(\text{MS}_{\text{error}}\right)}{n}}}$$

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.[9] 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 rank the means from smallest to largest. We then take into account the number of steps between the means to be compared. For adjacent means, no change is made and $q_{.05} = t_{.05}\sqrt{2}$. 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

$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{X}_5$
4	10	11	24	29

with $n = 8$, $df_{error} = 35$, and $MS_{error} = 32.00$. 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).

[9] 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.

$$q_5 = \frac{\bar{X}_1 - \bar{X}_s}{\sqrt{\frac{\text{MS}_{\text{error}}}{n}}} = \frac{29 - 4}{\sqrt{\frac{32.00}{8}}} = \frac{25}{\sqrt{4}} = 12.5$$

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 $df_{\text{error}} = 35$, $q_{.05}(5,35) = 4.07$. 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_{obt} and referring q_{obt} 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 approach is frequently taken by post hoc procedures, so I cover it here, but I really don't find that it saves any time. Since

$$q_r = \frac{\bar{X}_1 - \bar{X}_s}{\sqrt{\frac{\text{MS}_{\text{error}}}{n}}}$$

then

$$\bar{X}_1 - \bar{X}_s = q_{.05}(r, df_{\text{error}})\sqrt{\frac{\text{MS}_{\text{error}}}{n}}$$

where $\bar{X}_1 - \bar{X}_s$ is the minimum difference between two means that will be found to be significant.

We know that with five means the critical value of $q_{.05}(5,35) = 4.07$. Then, for our data,

$$\bar{X}_1 - \bar{X}_s = 4.07\sqrt{\frac{32}{8}} = 8.14$$

Thus, a difference in means equal to or greater than 8.14 would be judged significant, whereas a smaller difference would not. Because the difference between the largest and smallest means in the example is 25, we would reject H_0.

Although *q* could be used in place of an overall *F* (i.e., instead of running the traditional analysis of variance, we would test the difference between the two extreme means), there is rarely an occasion to do so. In most cases, *F* is more powerful than *q*. However, where you expect several control group means to be equal to each other but different from an experimental treatment mean (i.e., $\mu_1 = \mu_2 = \mu_3 = \mu_4 \neq \mu_5$), *q* might well be the more powerful statistic.

Although *q* is seldom a good substitute for the overall *F*, it is a very important statistic when it comes to making multiple comparisons among individual treatment means. It forms the basis for the next several tests.

The Newman–Keuls test

The Newman–Keuls is a controversial test, for reasons that will become clear shortly. However, it is important to discuss it here if only because it is an excellent example of a whole class of multiple-comparison procedures. The basic goal of the **Newman–Keuls test** (sometimes called the Student-Newman-Keuls test) is to sort all the treatment means

into subsets of treatments. These subsets will be homogeneous in the sense that they do not differ among themselves, but they do differ from other subsets.

Because most people make comparisons using computer software, we might be able to get away with omitting calculations of the Newman–Keuls test. But if you understand the calculations you will more easily make sense of computer output and you will better understand how the various tests fit together. So take out your calculator and a piece of paper.

We will again use the data on morphine tolerance and will start by arranging the treatment means in ascending order from smallest to largest. We will designate these means as $\bar{X}_1 \ldots \bar{X}_5$, *where the subscript now refers to the position of that mean in the ordered series*. For the data in our example,

Treatment				
M-S	M-M	S-S	S-M	Mc-M
$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{X}_5$
4	10	11	24	29

We will define the *range* of means as the number of steps in an ordered series between those means. Adjacent means will be defined as being two steps apart, means that have one other mean intervening between them will have a range of three, and so on. In general, the range between $\bar{X}_i$ and $\bar{X}_j$ is $i - j + 1$ (for $i > j$).

If we wished to test the difference $\bar{X}_5 - \bar{X}_1$, the range would be 5-1 + 1 = 5 = *r*. For this example, we have *r* = 5 and 35 *df* for error, and thus (from Appendix *q*) would require $q_{\text{obt}} \geq 4.07$ if the difference is to be significant. Given this information, we could now solve for the smallest difference that would be judged significant. As we saw previously, this critical difference would be 8.14. Thus, when the means are five steps apart, a difference of at least 8.14 is required for significance. If we were concerned with means that are four steps apart (e.g., $\bar{X}_4 - \bar{X}_1$ or $\bar{X}_5 - \bar{X}_2$), then *r* = 4, *df* = 35, and $q_{.05}(4,35) = 3.815$ (by linear interpolation). Thus, the minimum difference that would be significant is

$$\bar{X}_1 - \bar{X}_s = q_{.05}(4,35)\sqrt{\frac{\text{MS}_{\text{error}}}{n}} = 3.815\sqrt{\frac{32.00}{8}} = 3.815(2) = 7.63$$

Thus, four-step differences greater than 7.63 will be classified as significant.[10]

This procedure will be repeated for all ranges possible with our data. If we define W_r as the smallest width or difference between means *r* steps apart that will be statistically significant, then

$$W_r = q_{.05}(r, df)\sqrt{\frac{\text{MS}_{\text{error}}}{n}}$$

For our example,

[10] The Newman–Keuls is sometimes referred to as a layered test because it adjusts the critical difference as a function of the number of means contained within a subset of means. This is in contrast to the Tukey HSD and Scheffé tests (to be described shortly), which essentially use a constant critical difference for all contrasts.

$$W_2 = q_{.05}(2,35)\sqrt{\frac{\text{MS}_{\text{error}}}{n}} = 2.875(2) = 5.75$$

$$W_3 = q_{.05}(3,35)\sqrt{\frac{\text{MS}_{\text{error}}}{n}} = 3.465(2) = 6.93$$

$$W_4 = q_{.05}(4,35)\sqrt{\frac{\text{MS}_{\text{error}}}{n}} = 3.815(2) = 7.63$$

$$W_5 = q_{.05}(5,35)\sqrt{\frac{\text{MS}_{\text{error}}}{n}} = 4.07(2) = 8.14$$

The Newman–Keuls test employs the values of W_r and a set of rules that are designed to control *FW* and to prevent inconsistent conclusions. If we did not adopt a set of rules governing the tests that will be made and the order of the testing, we would lose complete control of *FW* and, in addition, we might find ourselves making contradictory statements. For example, if three means are ordered $\bar{X}_1$, $\bar{X}_2$, and $\bar{X}_3$, it would be embarrassing if we found that $\bar{X}_1$ was different from $\bar{X}_2$, but not from $\bar{X}_3$ (since $\bar{X}_3$ is larger than $\bar{X}_2$).

To make the procedure systematic, we will form a matrix with treatment means on the rows and columns, and differences between means as the cell entries. Such a matrix is presented in Table 12.4a for the data in Table 12.1. The dashed lines in this table connect differences between means *r* steps apart with values of *r* and W_r. Thus, any values along a dashed line that are greater than the corresponding value of W_r are *potentially* significant.

Table 12.4 Newman-Keuls test applied to the data in Table 12.1

(a) Difference Between Means

			M-S	M-M	S-S	S-M	Mc-M		
			$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{X}_5$		
			4	10	11	24	29	r	W_r
M-S	$\bar{X}_1$	4	—	6	7	20	25	5	8.14
M-M	$\bar{X}_2$	10		—	1	14	19	4	7.63
S-S	$\bar{X}_3$	11			—	13	18	3	6.93
S-M	$\bar{X}_4$	24				—	5	2	5.75
Mc-M	$\bar{X}_5$	29					—		

(b) Pattern of Significant Difference

			M-S	M-M	S-S	S-M	Mc-M
			$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$	$\bar{X}_5$
			4	10	11	24	29
M-S	$\bar{X}_1$	4		*	*	*	*
M-M	$\bar{X}_2$	10				*	*
S-S	$\bar{X}_3$	11				*	*
S-M	$\bar{X}_4$	24					
Mc-M	$\bar{X}_5$	29					

We now start in the upper right corner and test along the first row until we reach a difference that is not significant. The upper right entry is 25, which represents a five-step difference. Because 25 > 8.14, this difference is significant, and an asterisk is placed in the corresponding cell of Table 12.4b. Moving to the left, we find an entry of 20, which represents a difference of four steps. We thus compare 20 against $W_4 = 7.63$, and again reject H_0. Once again, we place an asterisk in the corresponding cell of Table 12.4b. Moving farther to the left, we find an entry of 7, which represents a three-step difference and is significant since $7 > W_3 = 6.93$. Again we place an asterisk in Table 12.4b. When we come to the far left, we find that a two-step difference of 6 is also significant (6 > 5.75), and thus enter an asterisk in the matrix.

When we either reach a point at which a difference is not significant or else exhaust a row, we move to the next row. We now start working from right to left across the second row, but stop under one of three conditions: (1) we exhaust the row; (2) we reach a nonsignificant difference; or (3) we reach a column at which a nonsignificant difference was found in an earlier row.

In row 2, 19 > 7.63, 14 > 6.93, but 1 < 5.75. Thus, we place two asterisks in Table 12.4b and continue to the next row. In row 3, both differences are significant, because 18 > 6.93 and 13 > 5.75. Going to row 4, our one difference is not significant (5 < 5.75).

The resulting pattern of significant differences is given in Table 12.4b. Here we can see that group M-S is different from all other groups, groups M-M and S-S are different from groups S-M and Mc-M but not from each other, and groups S-M and Mc-M do not differ. These results can be represented graphically by writing down the treatments and underlining homogeneous subsets. Thus

M-S	M-M	S-S	S-M	Mc-M
4	10	11	24	29

Treatments not underlined by a common line differ significantly from each other. This pattern of results would appear to confirm Siegel's theory, because the group that received three morphine injections and then was switched to saline (M-S) showed hypersensitivity, whereas the group that received morphine in an environment different

from the test environment (Mc-M) was no different from the group that had not previously received morphine (S-M). The other two groups were intermediate, as the theory would predict. Notice that these results differ from those we obtained using the Bonferroni *t* test, even though both tests attempt to limit *FW*. The differences are due partly to the different approaches of the two tests and partly to the fact that the Newman–Keuls is actually a less conservative test, as will be discussed shortly. (It does not always hold *FW* at α.)

Unequal sample sizes and heterogeneity of variance

The Newman–Keuls procedure (and the Tukey procedure that follows) were 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 want to carry out a Newman–Keuls or a related test on the means. Although little work has been done on this topic with respect to the Newman–Keuls itself, work has been done on the problem with respect to the Tukey HSD test (see particularly Games and Howell, 1976; Keselman and Rogan, 1977; Games, Keselman, and Rogan, 1981). Because the problems would be the same in the two tests, it is reasonable to generalize from the latter findings. One solution, known as the Tukey–Kramer approach, is to replace $\sqrt{\text{MS}_{\text{error}}/n}$ with

$$\sqrt{\frac{\frac{MS_{error}}{n_i} + \frac{MS_{error}}{n_j}}{2}}$$

This procedure has been proposed in conjunction with the Tukey HSD, to be discussed in Section 12.5. 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

$$df' = \frac{\left(\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}\right)^2}{\frac{\left(\frac{s_i^2}{n_i}\right)^2}{n_i - 1} + \frac{\left(\frac{s_j^2}{n_j}\right)^2}{n_j - 1}}$$

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*. 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 vulnerable to violations of the assumption of homogeneity of variance, especially when the sample sizes are unequal. If there is no serious problem with heterogeneity of variance and if the n_i are nearly equal, then you are probably safe calculating the harmonic mean of the sample sizes [$\bar{n}_h = k/\sum(1/n_i)$] and using that in place of *n* in the standard approach. 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_{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.

Familywise error rate

Although the Newman–Keuls procedure was designed to control the familywise error rate, it does not control it completely. In fact, under certain conditions, the error rate can be quite high. If the *complete* null hypothesis is true (i.e., if $\mu_1 = \mu_2 = \cdots = \mu_k$), then the Newman–Keuls sets *FW* = .05, assuming that this is our chosen significance level. With, for example, five means and a true complete null hypothesis, our first test will compare μ_1 with μ_5, and because we used the Studentized range statistic, the probability of a Type I error will be .05. If we do not find a difference, we stop testing and thus do not have another chance to make a Type I error. If we do find a difference, we have already made a Type I error (because the complete null hypothesis is true), and *FW* is not affected by how many more errors we make. (*FW* is the probability of *at least* one Type I error.) Suppose, however, that some other state of affairs is true. For example, suppose $\mu_1 \neq \mu_2 = \mu_3 \neq \mu_4 = \mu_5$. Then our first test is likely to be significant, because $\mu_1 \neq \mu_5$. Now, however, there are two true null hypotheses to be tested and we have a probability of .05 of making a Type I error on each. Therefore, *FW* will be about $1-(1-.05)^2$, which is approximately .10. In general, the *maximum FW* for the Newman–

Keuls is approximately α times the maximum number of null hypotheses that could be true, which is equal to the number of pairs involving different means.

Therefore,

$$FW_{\max} \cong \begin{cases} \dfrac{\alpha k}{2} & \text{if } k \text{ is even} \\ \dfrac{\alpha(k-1)}{2} & \text{if } k \text{ is odd} \end{cases}$$

This means that with three means $FW = .05$ because there is at most one true null hypothesis to be falsely declared "significant," whereas with four or five means $FW \cong .10$ (there *at most* two true null hypotheses).

12.5. TUKEY'S TEST

Much of the work on multiple comparisons has been based on the original work of Tukey, and an important test bears his name.[11] The **Tukey test**, also called the **Tukey's HSD (honestly significant difference) test**, is similar to the Newman–Keuls, except that q_{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 α against all possible null hypotheses, not just the complete null hypothesis, although with a loss of power. The Tukey HSD is the favorite pairwise test for many people because of the control it exercises over α.

[11] A second test (the WSD), which is a modification on the HSD test, was proposed by Tukey as less conservative. However, I have never seen it used and have omitted it from discussion.

If we apply the Tukey HSD to the data on morphine tolerance, we apply a critical value of $W_r = 8.14$ to all differences. Thus, we declare all mean differences ($\bar{X}_i - \bar{X}_j$) to be significant if they exceed 8.14 and to be not significant if they are less than 8.14. For our data, the difference between $\bar{X}_{M-M}$ and $\bar{X}_{M-S}$ = 10-4 = 6, and the difference between $\bar{X}_{S-S}$ and $\bar{X}_{M-S}$ is 11-4 = 7. The Newman–Keuls declared these differences to be significant, but the Tukey HSD would declare them not significant because 6 and 7 are less than 8.14. On this basis, we would represent the set of homogeneous means as

M-S	M-M	S-S	S-M	Mc-M
4	10	11	24	29

This leads to quite a different interpretation of the results, because the group that was switched from morphine to saline (M-S) can no longer be declared significantly more sensitive to pain than is the group that had never received morphine (S-S) or the group that always received it (M-M).

12.6. THE RYAN PROCEDURE (REGWQ)

As we have seen, the Tukey procedure controls the familywise error rate at α regardless of the number of true null hypotheses (not just for the overall null hypothesis), whereas the Newman–Keuls allows the familywise error rate to rise as the number of true null

hypotheses increases. The Tukey test, then, provides a firm control over Type I errors, but at some loss in power. The Newman–Keuls tries to maximize power, but with some loss in control over the familywise error rate. A compromise, which holds the familywise error rate at α but which also allows the critical difference between means to shrink as *r* (the number of means in a set) decreases, was proposed by Ryan (1960) and subsequently modified by others.

What Newman and Keuls really did in creating their test was to hold the error rate at α for each set of *r* ordered means. The effect of this is to allow the critical values to grow as *r* increases, but they actually grow too slowly to keep the familywise error rate at α when multiple null hypotheses are true. Ryan (1960) also proposed modifying the value of α for each step size, but in such a way that the overall familywise error rate would remain unchanged at α. For *k* means and a step size of *r*, Ryan proposed using critical values of q_r at the

$$\alpha_r = \frac{\alpha}{k/r} = \frac{r\alpha}{k}$$

level of significance, rather than always using q_r at the α level of significance. This suggestion was then modified by Einot and Gabriel (1975) to set

$$\alpha_r = 1-(1-\alpha)^{1/(k/r)} = 1-(1-\alpha)^{r/k}$$

and then again by Welsch (1977) to keep the Einot and Gabriel suggestion but to allow α_r to remain at α for $r = k$, and $r = k$-1. These changes hold the overall familywise error rate at α while giving greater power than does Tukey to some comparisons. (Notice

the similarity in the first two of these suggestions to the way α is adjusted by the Bonferroni and the Dunn-Šidák procedures.)

What these proposals really do is to allow you to continue to use the tables of the Studentized Range Distribution, but instead of always looking for q_r at $\alpha = .05$, for example, you look for q_r at $\alpha = \alpha_r$, which is likely to be some unusual fractional value. The problem is that you don't have tables that give q_r at any values other than $\alpha = .05$ or $\alpha = .01$. That is no problem if you are using computer software that can calculate those values, but it is a problem if you are doing your analyses with your hand calculator.

One way that you can run the **Ryan procedure** (or the Ryan/Einot/Gabriel/Welsch procedure) is to use SPSS or SAS and request multiple comparisons using the **REGWQ** method. (The initials refer to the authors and to the fact that it uses the Studentized Range Distribution (*q*).) For those who have access to SPSS or other software that will implement this procedure, I recommend it over either the Newman–Keuls or the Tukey, because it appears to be the most powerful test generally available that still keeps the familywise error rate at α. Those who don't have access to the necessary software will have to fall back on one of the more traditional tests. The SAS output for the REGWQ procedure (along with the Student–Newman–Keuls, the Tukey, and the Scheffé tests) are presented later in the chapter so that you can examine the results. In this situation the conclusions to be drawn from the REGWQ and Tukey tests are the same, although you can see the difference in their critical ranges.

12.7. THE SCHEFFÉ TEST

The post hoc tests we have considered all primarily involve pairwise comparisons of means, although they can be extended to more complex contrasts. One of the best-known tests, which is both broader and more conservative, was developed by Scheffé. The **Scheffé test**, which uses the *F* distribution rather than the Studentized range statistic, sets the familywise error rate at α against all possible linear contrasts, not just pairwise contrasts. If we let

$$L = \sum a_j \bar{X}_j \text{ and } \mathrm{SS}_{\text{contrast}} = \frac{nL^2}{\sum a_j^2}$$

then

$$F = \frac{nL^2}{\sum a_j^2 \mathrm{MS}_{\text{error}}}$$

Scheffé has shown that if F_{obt} is evaluated against $(k-1)F_{\alpha}(k-1, df_{\text{error}})$—rather than against $F_{\alpha}(1, df_{\text{error}})$—the *FW* is at most α. (Note that all that we have done is to calculate *F* on a standard linear contrast, but we have evaluated that *F* against a modified critical value.) Although this test has the advantage of holding constant *FW* for all possible linear contrasts—not just pairwise ones—it pays a price; it has the least power of all the tests we have discussed. Partly to overcome this objection, Scheffé proposed that people may prefer to run his test at $\alpha = .10$. He further showed that the test is much less sensitive than the Tukey HSD for pairwise differences but is more sensitive than the Tukey HSD for complex comparisons (Scheffé, 1953, 1959). In general, the Scheffé test should never be used to make a set of solely pairwise comparisons, nor should it normally be used for

comparisons that were thought of a priori. The test was specifically designed as a post hoc test (as were the Newman–Keuls and Tukey tests), and its use on a limited set of comparisons that were planned before the data were collected would generally be foolish. Although most discussions of multiple-comparison procedures include the Scheffé, and many people recommend it, it is not often seen as the test of choice in research reports because of its conservative nature.

12.8. DUNNETT'S TEST FOR COMPARING ALL TREATMENTS WITH A CONTROL

In some experiments the important comparisons are between one control treatment and each of several experimental treatments. In this case, the most appropriate test is **Dunnett's test**. This is more powerful (in this situation) than are any of the other tests we have discussed that seek to hold the familywise error rate at or below α.

We will let t_d represent the critical value of a modified *t* statistic. This statistic is found in tables supplied by Dunnett (1955, 1964) and reproduced in Appendix t_d. We can either run a standard *t* test between the appropriate means (using MS_{error} as the variance estimate and evaluating the *t* against the tables of t_d) or solve for a critical difference between means. For a difference between means $\bar{X}_c$ and $\bar{X}_j$ (where $\bar{X}_c$ represents the mean of the control group) to be significant, the difference must exceed

$$\text{Critical value } \left(\bar{X}_c - \bar{X}_j\right) = t_d \sqrt{\frac{2\text{MS}_{\text{error}}}{n}}$$

Applying this test to our data, letting group S-S from Table 12.1 be the control group,

$$\text{Critical value } \left(\bar{X}_c - \bar{X}_j\right) = t_d \sqrt{\frac{2(32.00)}{8}}$$

We enter Appendix t_d with $k = 5$ means and $df_{\text{error}} = 35$. The resulting value of t_d is 2.56.

$$\text{Critical value } \left(\bar{X}_c - \bar{X}_j\right) = 2.56\sqrt{\frac{2(32.00)}{8}} = 2.56(2.828) = 7.24$$

Thus, whenever the difference between the control group mean (group S-S) and one of the other group means exceeds ±7.24, that difference will be significant. The *k*-1 statements we will make concerning this difference will have an *FW* of $\alpha = .05$.

S-S versus M-S = 11 - 4 = 7
S-S versus M-M = 11 - 10 = 1
S-S versus S-M = 11 - 24 = -13
S-S versus Mc-M = 11 - 29 = -18

Because we have a two-tailed test (t_d was taken from two-tailed tables), the sign of the difference is irrelevant. The last two differences exceed ±7.24 and are therefore declared to be significant.

In the case in which the groups have unequal sample sizes or heterogeneous variances, a test on the difference in treatment means is given by the same general procedure we used with the Newman–Keuls.

12.9. COMPARISON OF DUNNETT'S TEST AND THE BONFERRONI *T*

Because the Bonferroni *t* test allows the experimenter to make any a priori test, it is reasonable to ask what would happen if we decided a priori to apply that test to the differences between the control mean and the experimental treatment means. If we did this for our data, we would find that the required critical difference would be 7.47 for the Bonferroni instead of the 7.24 required for Dunnett's test. Thus, we would have a less powerful test, because a larger difference is needed for rejection of H_0. Both the Bonferroni *t* and Dunnett's test are based on inequalities of the form $FW \leq \alpha$, but Dunnett's test uses a sharper inequality (Miller, 1981). To put this rather crudely, in Dunnett's case there is more of the *equal to* and less of the *less than* involved in the relationship between *FW* and α. For this reason, it is a more powerful test whenever you want simply to compare one treatment (it does not really have to be called a "control" treatment) with each of the others.

12.10. COMPARISON OF THE ALTERNATIVE PROCEDURES

Because the multiple-comparison techniques we have been discussing were designed for different purposes, there is no truly fair basis on which they can be compared. There is something to be gained, however, from summarizing their particular features and comparing the critical differences they require for the same set of data. Table 12.5 lists the tests, the error rate most commonly associated with them, the kinds of comparisons

they are primarily designed to test, and the type of test (range test, *F* test, or *t*—modified or not in each case).

Table 12.5 Comparison of alternative multiple-comparison procedures

Test	Error Rate	Comparison	Type	A Priori/ Post Hoc
1. Individual *t* tests	PC	Pairwise	*t*	A priori
2. Linear contrasts	PC	Any contrasts	*F*	A priori
3. Bonferroni *t*	FW	Any contrasts	*t*‡	
4. Holm: Larzelere & Mulaik	FW	Any contrasts	*t*‡	
5. Fisher's LSD	FW†	Pairwise	*t*	
6. Newman-Keuls	FW†	Pairwise	Range	
7. Ryan (REGWQ)	FW	Pairwise	Range	
8. Tukey HSD	FW	Pairwise§	Range‡	
9. Sheffé's test	FW	Any contrasts	*F*‡	
10.Dunnett's test	FW	With control	*F*‡	

†Against complete null hypothesis
‡Modified
§Tukey HSD con be used for all contrasts, but is poor for this purpose

If we compare the tests in terms of the critical values they require, we are being somewhat unfair to the a priori tests. To say that the Bonferroni *t* test, for example, requires a large critical value when making all possible pairwise comparisons is not really doing the test justice, because it was designed to make relatively few individual comparisons and not to be limited to pairwise contrasts. With this word of caution, Table 12.6 compares the critical differences (W_r) for each test. Linear contrasts have been omitted because they are not appropriate to the structure of the table, and the critical values for pairwise comparisons would be the same as for the individual *t* tests. Dunnett's test has also been omitted because it does not fit with the structure of the table.

Table 12.6 Comparison of critical differences for alternative procedures

	W_2	W_3	W_4	W_5
Individual *t* tests	5.74	5.74	5.74	5.74
Bonferroni *t* tests*	7.47	7.47	7.47	7.47
Holm**	5.74	6.64	7.13	7.47
Newman-Keuls	5.74	6.93	7.63	8.13
Ryan (REGWQ)	6.88	7.54	7.63	8.13
Tukey HSD	8.13	8.13	8.13	8.13
Scheffé test	9.19	9.19	9.19	9.19

* Assuming only four pairwise comparisons are desired.
** Assuming significance at each preceding level

12.11. 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* proposed by Holm 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 better off using the Ryan REGWQ if you have it available or the Tukey. In the past I

recommended the Newman–Keuls, because it does a fairly good job when you have five or fewer groups, but I have found myself in a distinct minority and have decided to bail out. With three groups the Newman–Keuls and the REGWQ test will be the same anyway, given Welsch's modification to that test, which earned him a place in its initials. 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.

People often fail to realize that in selecting a test, it is perfectly acceptable to compare each of several tests on your own data in terms of the size of the critical values, and to select a test on that basis. For example, if you are going to run only a few pairwise comparisons, the critical values for the Holm-modified Bonferroni test may be smaller than the critical values for the REGWQ. In that case, go with modified Bonferroni. On the other hand, you may discover that even though you do not wish to make all possible pairwise comparisons, the REGWQ (or the Tukey) gives smaller critical values than the modified Bonferroni, in which case you would waste power to go with the Bonferroni. The important point is that these decisions have to be based on a consideration of the critical values, and not the final results. You can't just try out every test you have and choose the one that gives you the answers you like.

12.12. 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 do SPSS and Minitab, and is thus somewhat more useful for our purpose.

Exhibit 12.1 begins with the program commands and the overall analysis of variance. ("Condition" is spelled as "Condtion" because SAS allows only 8 characters in a name.) This analysis agrees with the summary table shown in Table 12.1, and with an $R^2 = \eta^2 = .757$. 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.

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. These are analogous to the underlining of condition names that you saw earlier in this chapter. 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. These are the same conclusions we drew when we performed the Newman–Keuls by hand.

EXHIBIT 12.1

```
Options LineSize = 78;

Data Siegel;
     Infile 'Alexander:SAS610:Data Files:Siegel.dat';
     Infile Condtion Latency;
run;

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

```
                              The SAS System                              1
                                                  20:12 Tuesday, October 17, 1995
                        General Linear Models Procedure

Dependent Variable: LATENCY
                                  Sum of          Mean
Source                     DF     Squares        Square    F Value     Pr > F

Model                       4  3497.600000   874.400000      27.33     0.0001
Error                      35  1120.000000    32.000000
Corrected Total            39  4617.600000

                     R-Square         C.V.        Root MSE          LATENCY Mean
                     0.757450     36.26189        5.656854              15.60000

Dependent Variable: LATENCY

Source                     DF     Type III SS     Mean Square   F Value     Pr > F
CONDTION                    4  3497.600000   874.400000      27.33     0.0001
```

η^2 → R-Square 0.757450

F Value 27.33 ← *F* for Condition

```
          Student-Newman-Keuls test for variable: 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  df= 35  MSE= 32

Number of Means          2          3          4          5
Critical Range   5.7420598  6.921941  7.6279952  8.1319061

Means with the same letter are not significantly different.

      SNK Grouping              Mean      N  CONDTION

                A            29.000      8  M-S
                A            24.000      8  M-M
              B              11.000      8  S-S
              B              10.000      8  S-M
            C                 4.000      8  Mc-M
```

$w_5 = q_5\sqrt{\frac{MS_{error}}{n}}$

(the Condtion labels are wrong throughout—the order is Mc-M, S-M, S-S, M-M, M-S)

EXHIBIT 12.1 (Cont.)

Ryan-Einot-Gabriel-Welsch Multiple Range Test for variable: LATENCY

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

Alpha= 0.05 df= 35 MSE= 32

Number of Means 2 3 4 5
Critical Range 6.8765473 7.5391917 7.6279952 8.1319061 ← Same as SNK

Larger than for SNK

Means with the same letter are not significantly different.

REGWQ Grouping	Mean	N	CONDTION
A	29.000	8	M-S
A	24.000	8	M-M
B	11.000	8	S-S
B	10.000	8	S-M
B	4.000	8	Mc-M

Tukey's Studentized Range (HSD) for variable: LATENCY

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

Alpha= 0.05 df= 35 MSE= 32
Critical Value of Studentized Range= 4.066
Minimum Significant Difference= 8.1319 ← Critical range for all differences

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	CONDTION
A	29.000	8	M-S
A	24.000	8	M-M
B	11.000	8	S-S
B	10.000	8	S-M
B	4.000	8	Mc-M

Scheffe's test for variable: LATENCY

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than REGWF for all pairwise comparisons

Alpha= 0.05 df= 35 MSE= 32
Critical Value of F= 2.64147
Minimum Significant Difference= 9.1939 ← Critical range for all differences

Means with the same letter are not significantly different.

Scheffe Grouping	Mean	N	CONDTION
A	29.000	8	M-S
A	24.000	8	M-M
B	11.000	8	S-S
B	10.000	8	S-M
B	4.000	8	Mc-M

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 was said earlier, the Newman-Keuls holds the familywise error rate at α against the complete null hypothesis, but allows it to rise in the case where a subset of null hypotheses are true. You next see a statement saying that the test is being run at $\alpha = .05$, that we have 35 *df* for the error term,

and that $MS_{error} = 32.00$. 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

$$W_r = q_{.05}(r, df_e)\sqrt{\frac{MS_{error}}{n}}$$

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

$$W_3 = q_{.05}(3, df_e)\sqrt{\frac{MS_{error}}{n}} = 3.46\sqrt{\frac{32}{8}} = 3.46(2) = 6.92$$

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 familywise error rate for this set of comparisons is $\alpha = .05$, whereas it would be nearly $\alpha = .10$ 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.

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. But I'd first have to "cross my heart and hope to die," and even then I'm not sure if I'd trust myself.

12.13. 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 comparisons. 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 grains of aspirin, where 1 grain is equivalent to what used to be called "baby aspirin" and 5 grains is the standard tablet. In this study we would not be concerned so much with whether a 4-grain dose was better than a 2-grain 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 low doses being less effective and high doses adding little, if anything, to the effect.

Typical linear and **quadratic functions** are illustrated in Figure 12.2. (They were produced using JMP on a Macintosh.) 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 upon a rising linear trend.

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 the both the group means and the particular ordering of those means. To put this slightly differently using the aspirin example, a Newman–Keuls test between the largest and the smallest means will not be affected by which group happens to have each 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.)

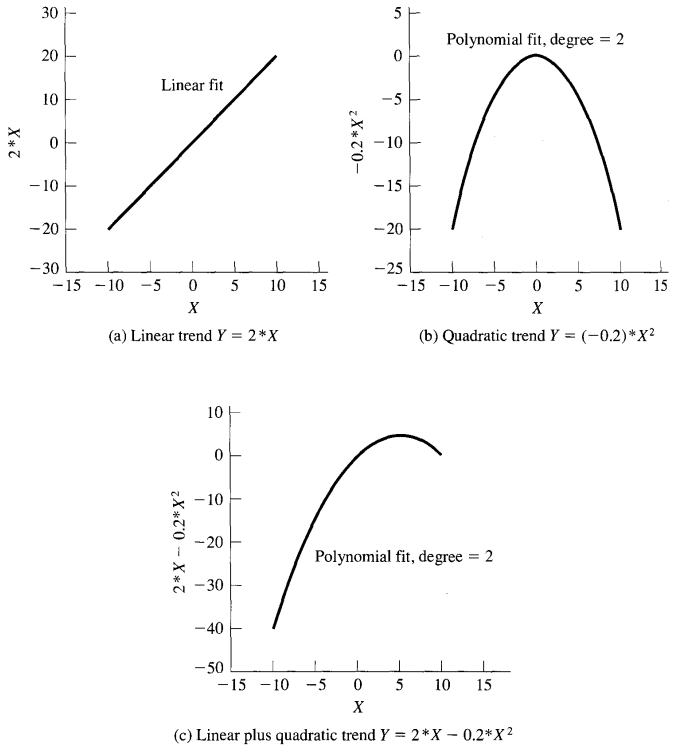


Figure 12.2 Typical linear and quadratic functions

Boring is attractive

A useful example of trend analysis comes from a study by Langlois and Roggman (1990), which examined the question of what makes a human face attractive. They approached the problem from both an evolutionary and a cognitive perspective. Modern evolutionary theory would suggest that average values of some trait would be preferred to extreme ones, and cognitive theory suggests that both adults and children respond to prototypes of objects more positively than to objects near the extremes on any dimension. A prototype, by definition, possesses average values of the object along important dimensions. (A prototype of a cat is one that is not too tall or too short, not too fat or too thin, and doesn't purr too loudly or too quietly.)

Langlois and Roggman took facial photographs of 336 males and 214 females. They then created five groups of composite photographs by computer-averaging the individual faces. Thus, for one group the computer averaged 32 randomly selected same-gender faces, producing a quite recognizable face with average width, height, eyes, nose length, and so on. For the other groups the composite faces were averaged over either 2, 4, 8, or 16 individual faces. An example of composite faces can be seen in Figure 12.3. The label Composite will be used to represent the five different groups. That is not an ideal name for the independent variable, but neither I nor the study's authors have a better suggestion. Within each group of composite photographs were three male and three female faces, but we will ignore gender for this example. (There were no significant gender differences, and the overall test on group differences is not materially affected by ignoring that variable.)

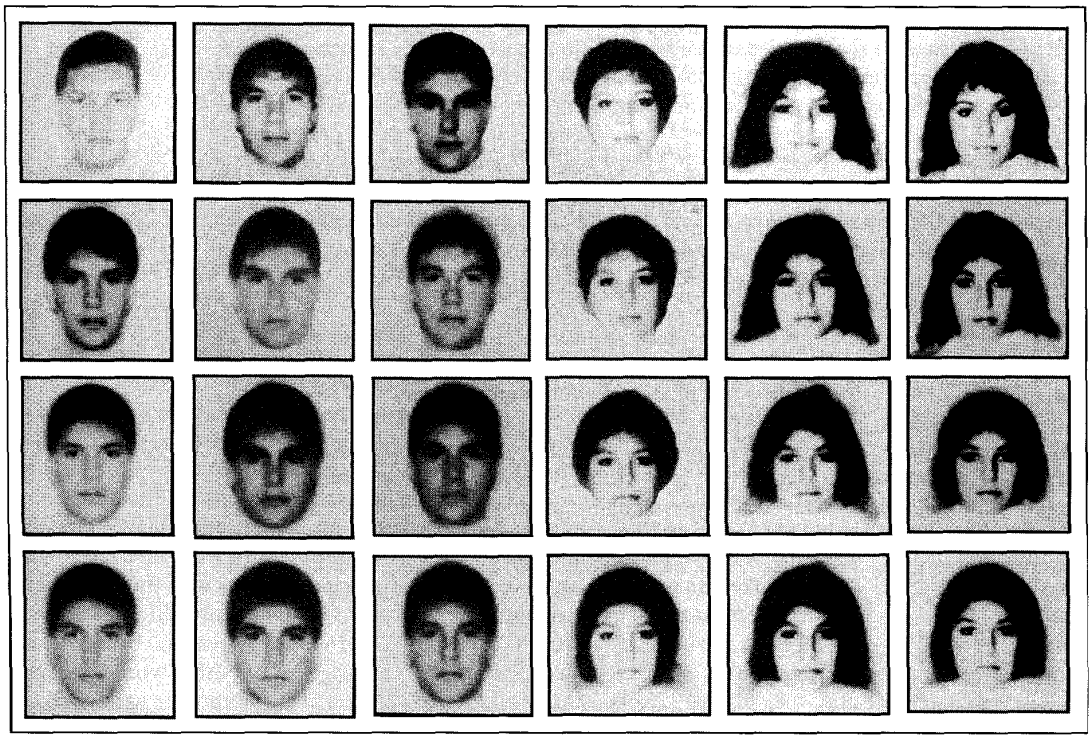


FIGURE 12.3 Composite faces. Faces from left to right represent the six different composite sets. Faces from top to bottom represent composite levels of 4 faces, 8 faces, 16 faces, and 32 faces.

Langlois and Roggman presented different groups of subjects with composite faces and asked them to rate the attractiveness of the faces on a 1–5 scale, where 5 represents "very attractive." The individual data points in their analysis were actually the means averaged across raters for the six different composites in each condition. The data are given in Table 12.7. These data are fictional, but they have been constructed to have the same mean and variance as those reported by Langlois and Roggman, so the overall F and the tests on trend will be the same as those they reported.

Table 12.7 Data on rated attractiveness

	Group 1	Group 2	Group 3	Group 4	Group 5
	2.201	1.893	2.906	3.233	3.200
	2.411	3.102	2.118	3.505	3.253
	2.407	2.355	3.226	3.192	3.357
	2.403	3.644	2.811	3.209	3.169
	2.826	2.767	2.857	2.860	3.291
	3.380	2.109	3.422	3.111	3.290
Mean	2.6047	2.6450	2.8900	3.1850	3.2600

A standard one-way analysis of variance on these data would produce the following summary table:

Source	***df***	**SS**	**MS**	***F***
Composite	4	2.1704	0.5426	3.13*
Error	25	4.3281	0.1731	
Total	29	6.4985		

*$p<.05$

From the summary table it is apparent that there are significant differences among the five groups, but it is not clear how these differences are manifested. One way to examine these differences would be to plot the group means as a function of the number of individual pictures that were averaged to create the composite. An important problem that arises if we try to do this concerns the units on the abscissa. We could label the groups as "2, 4, 8, 16, and 32," on the grounds that these values correspond to the number of elements over which the average was taken. However, it seems unlikely that rated attractiveness would increase directly with those values. We might expect that a picture averaged over 32 items would be more attractive than one averaged over 2 items, but I doubt that it would be 16 times more attractive. But notice that each value of the independent variable is a power of 2. In other words, the values of 2, 4, 8, 16, and 32

correspond to $2^1, 2^2, 2^3, 2^4,$ and 2^5. (Put another way, taking the $\log_2$ of 2, 4, 8, 16, and 32 would give us 1, 2, 3, 4, and 5.) For purposes of analyzing these data, I am going to represent the groups with the numbers 1 to 5 and refer to these as measuring the degree of the composite. (If you don't like my approach, and there is certainly room to disagree, be patient and we will soon see a solution using unequally spaced values of the independent variable. The example will be simpler statistically if the units on the abscissa are evenly spaced.) The group means using my composite measure on the abscissa are plotted in Figure 12.4, where you can see that the rated attractiveness does increase with increasing levels of Composite.

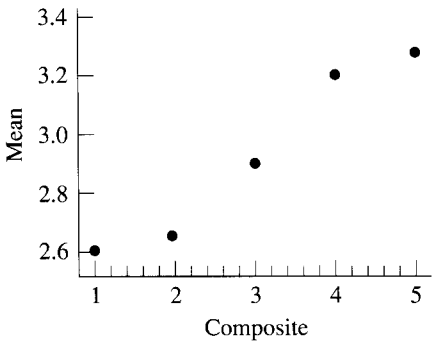


Figure 12.4 Scatterplot of mean versus composite group

Our first question asks whether a nonhorizontal straight line provides a good fit to the data. A glance at Figure 12.4 would suggest that this is the case. 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

$$L = a_1\bar{X}_1 + a_2\bar{X}_2 + a_3\bar{X}_3 + \cdots + a_k\bar{X}_k = \sum a_j\bar{X}_j$$

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 shown in the appendix, 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.

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

$$\text{SS}_{\text{contrast}} = \frac{nL^2}{\sum a_j^2}$$

In our case,

$$\begin{aligned} L_{\text{linear}} &= (-2)2.6047 + (-1)2.6450 + (0)2.8900 + (1)3.1850 + (2)3.2600 \\ &= 1.8506 \end{aligned}$$

$$\begin{aligned} \text{SS}_{\text{linear}} &= \frac{nL^2}{\sqrt{\sum a_j^2}} = \frac{6\left(1.8506^2\right)}{10} \\ &= 2.0548 \end{aligned}$$

Like all contrasts, this contrast has a single degree of freedom, and therefore $\text{SS}_{\text{linear}} = \text{MS}_{\text{linear}}$. 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_{error}:

$$\begin{aligned} F &= \frac{\text{MS}_{\text{linear}}}{\text{MS}_{\text{error}}} \\ &= \frac{2.0548}{0.1731} \\ &= 11.8706 \end{aligned}$$

This is an *F* on 1 and 24 degrees of freedom, and from Appendix *F* we find that $F_{.05}(1, 24) = 4.26$. Because the *F* for the linear component (11.87) exceeds 4.26, we will reject H_0 and conclude that there is a significant linear trend in our means. In other words, we will conclude that attractiveness varies linearly with increasing levels of Composite. Notice here that a significant *F* means that the trend component we are testing is significantly different from 0.

It is conceivable that we could have a significant linear trend in our data and still have residual variance that can be explained by a higher-order term. For example, it is possible that we might have both linear and quadratic, or linear and cubic, components. In fact, it would be reasonable to expect a quadratic component in addition to a linear one, because it seems unlikely that judged attractiveness will keep increasing indefinitely as we increase the number of individual photographs we average to get the composite. There will presumably be some diminishing returns, and the curve should level off.

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_{linear} accounted for virtually all of $SS_{Composite}$, there would be little or nothing left over for higher-order terms to explain. On the other hand, if SS_{linear} were a relatively small part of $SS_{Composite}$, then it would make sense to look for higher-order components. From our previous calculations we obtain

$$
\begin{aligned}
SS_{residual} &= SS_{Composite} - SS_{linear} \\
&= 2.1704 - 2.0548 \\
&= 0.1156 \\
df_{residual} &= df_{Composite} - df_{linear} \\
&= 4 - 1 \\
&= 3 \\
MS_{residual} &= \frac{SS_{residual}}{df_{residual}} \\
&= \frac{0.1156}{3} \\
&= 0.0385 \\
F_{residual} &= \frac{MS_{residual}}{MS_{error}} \\
&= \frac{0.0385}{0.1731} \\
&< 1
\end{aligned}
$$

Because F for the residual is less than 1, we know automatically that it is not significant. This tells us that there is no significant variability left to be explained over and above that accounted for by the linear component. We would, therefore, normally stop here. However, for purposes of an example I will go ahead and calculate the quadratic component. The calculations will be shown without discussion, because the discussion would essentially be the same as above with the word *quadratic* substituted for *linear*.

$$
\begin{aligned}
L_{\text{quadratic}} &= (2)2.6047 + (-1)2.6450 + (-2)2.8900 + (-1)3.1850 + (2)3.2600 \\
&= 0.1194 \\
\text{SS}_{\text{quadratic}} &= \frac{nL^2}{\sum b_j^2} \\
&= \frac{6\left(0.1194^2\right)}{14} \\
&= 0.0061 \\
F &= \frac{\text{MS}_{\text{quadratic}}}{\text{MS}_{\text{error}}} \\
&= \frac{0.0061}{0.1731} \\
&< 1
\end{aligned}
$$

As our test on the residual suggested, there is no significant quadratic component on our plot of the group means. Thus there is no indication, over the range of values used in this study, that the means are beginning to level off. Therefore, we would conclude from these data that attractiveness increases linearly with Composite, at least given the definition of Composite used here.

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 ($\text{SS}_{\text{linear}}$, ... , $\text{SS}_{\text{quartic}}$) will sum to $\text{SS}_{\text{Composite}}$, and be very 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 linear components to the data at hand. A quadratic was stretching things a bit. Moreover, if you were to fit a fourth-order polynomial and found 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. In fact, I actually transformed the independent variable into a scale called Composite to fulfill that requirement. 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. The simplest thing to do, using the attractiveness data as an example, would be to code the independent variable as 1, 2, 3, 4, and 5, and then recode it as 2, 4, 8, 16, 32.

If the software is making appropriate use of the levels of the independent variable, you should get different answers. Then the problem is left up to you to decide which answer you want, when both methods of coding make sense. For example, if you use SPSS ONEWAY procedure and ask for polynomial contrasts, *where the independent variable is coded 1, 2, 3, 4, 5*, you will obtain the same results as above. If you code the variable 2, 4, 8, 16, 32, you will obtain slightly different results. However, if you use SPSS General Linear Model/Univariate procedure, the way in which you code the independent variable will not make any difference—both will produce results as if the coding were 1, 2, 3, 4, 5. It always pays to check.

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

Error rate per comparison (*PC*) (12.1)

Familywise error rate (*FW*) (12.1)

A priori comparisons (12.1)

Post hoc comparisons (12.1)

Contrasts (12.3)

Linear combination (12.3)

Linear contrast (12.3)

Partition (12.3)

Orthogonal contrasts (12.3)

Dunn's test (12.3)

Bonferroni *t* (12.3)

Bonferroni inequality (12.3)

Dunn-Šidák test (12.3)

Fisher's least significance difference (LSD) (12.4)

Studentized range statistic (*q*) (12.4)

Newman–Keuls test (12.4)

Tukey test (12.5)

HSD (honestly significant difference) test (12.5)

Ryan procedure (REGWQ) (12.6)

Scheffé test (12.7)

Dunnett's test (12.8)

Trend (12.13)

Quadratic function (12.13)

Polynomial trend coefficients (12.13)

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.

Ad Lib Control	Two per Day Control	Food Deprived	Water Deprived	Food and Water Deprived
18	20	6	15	12
20	25	9	10	11
21	23	8	9	8
16	27	6	12	13
15	25	11	14	11
90	120	40	60	55

a) Analyze the data using linear contrasts (*Note, I am not asking for linear polynomials (trend) here, just standard contrasts*).

b) Show that the contrasts are orthogonal.

c) Show that the sums of squares for the contrasts sum to SS_{treat}.

12.2) Using the data from Exercise 11.1, compute the linear contrasts for 5 versus (20 and 35) days and 20 versus 35 days, using $\alpha = .05$ for each contrast.

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. (*Note that this and subsequent exercises refer to exercises in Chapter 11, not this chapter*.). 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 $t\sqrt{2}$ (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) Repeat Exercise 12.7, using Holm's multistage test.

12.9) Apply Holm's multistage test to Exercise 12.1.

12.10) Run a Newman–Keuls test on the example given in Table 11.2 (page …)and interpret the results.

12.11) Calculate the Tukey test on the data in the example in Table 11.2, and compare your results to those you obtained for Exercise 12.10.

12.12) Consider the following data for five groups:

Group	**1**	**2**	**3**	**4**	**5**
$\bar{X}_j$	10	18	19	21	29
n_j	8	5	8	7	9
s_j^2	7.4	8.9	8.6	7.2	9.3

Run a Newman–Keuls test on these data.

12.13) Run Tukey's HSD procedure on the data in Exercise 12.12.

12.14) Use the Scheffé test on the data in Exercise 12.12 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). (*Hint*: You will need to go back to the section in which unequal sample sizes are discussed in conjunction with Table 12.2.)

12.15) Apply the Tukey procedure to the log transformed THC data from Table 11.5 (page …). What is the maximum *FW* for this procedure?

12.16) Apply Dunnett's test to the log transformed data in Table 11.5.

12.17) 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?

12.18) The Holm test is referred to as a modified sequentially rejective procedure. Why?

12.19) Fit linear and quadratic trend components to the Conti and Musty (1984) log transformed data in Table 11.5. The control condition received 0 μg 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 μg group and a 2.35 and 2.36 to the 1 μg group without altering the results significantly.) 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.

12.20) Use any statistical package to compute Fisher's LSD procedure on all three pairs of means (even though the overall *F* was not significant) for GSIT from Mireault's data (Mireault.dat). (This is based on the analysis of variance in Exercise 11.27.) Compare these results with the individual *t* tests that you ran for Exercise 7.46. Interpret the results.

12.21) Use any statistical package to apply the Newman–Keuls, Tukey, REGWQ (if available), and Scheffé procedures to the data from Introini-Collison and McGaugh (1986), described in the exercises for Chapter 11. Do these analyses for both Epineq.dat and Epinuneq.dat. Do not combine across the levels of the interval variable.

12.22) In Exercise 12.21 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?

12.23) 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].)

12.24) Interpret the results in Exercise 12.23.

12.25) 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 AM. 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.

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

a) Plot these data using both the actual dosage, and the values 1, 2, 3, 4, 5, 6 as the values of *X*.

b) Run a trend analysis using SPSS Oneway, if available, with the actual dosage as the independent variable.

c) Repeat part b) using the 1, 2, 3, 4, 5, 6 coding as the independent variable.

d) Interpret your results. How might these results have something to say to students who stay up all night studying for an exam?

e) Why might you, or Stone et al., prefer one coding system over another?

Discussion Question

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

12.27) Write an explanation of why the Newman–Keuls error rate is greater than $\alpha = .05$ when the overall null hypothesis is not true. How does Ryan's modification correct the problem?

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 interpret 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.]