8

Orthogonal, Planned and Unplanned Comparisons

8.1 Introduction

In this chapter we discuss in greater detail the nature of a comparison.[1] In the sections that follow, we will assume equal sample sizes to make the description simpler. That is, we have k means with an equal number of observations for each treatment. Later, we will generalize the discussion to cases where there are a different number of observations across conditions.

Table 8.1 repeats the analysis of variance summary table from the preceding chapter, where we had $k = 4$ treatments with $n = 10$ observations for each treatment. Although in that example the omnibus $F = MS_T/MS_W = 9.15$ was statistically significant, significance is not a necessary condition for testing orthogonal, planned comparisons. Comparisons can be tested directly without conducting the omnibus test. Indeed, a fledgling view among methodologists is that omnibus tests should be avoided because they do not provide information about specific patterns between treatment means. Com-

[1] Comparisons are also referred to as contrasts.

Table 8.1: Summary of the analysis of variance of a between-subjects design with $k = 4$ treatments and $n = 10$ participants randomly assigned to each treatment.

Source of variation	Sum of squares	*df*	Mean square	*F*
Treatments	83.50	3	27.83	9.15
Within treatments	109.44	36	3.04	
Total	192.94	39		

parisons provide one way to test specific research questions and hypotheses among a set of treatment means, so comparisons extend the omnibus test on treatment means presented in Chapter 6.

8.2 Comparisons on Treatment Means

A comparison involves quantifying a particular research question by taking a linear combination of treatment means. For instance, a researcher might be interested in comparing three treatments where clients received therapies to a fourth condition where clients were given a placebo therapy. This question can be worded as "Does the average of the three treatments that received therapy differ from the single group that did not receive therapy?" Table 8.2 shows three of the many comparisons that might be made on a set of $k = 4$ treatment means. The values of each comparison are called **coefficients** of the treatment means, and we will use a with appropriate subscripts, as shown on the right-hand side of the table, to represent coefficients. The first subscript refers to a particular treatment mean and the second subscript corresponds to a particular comparison, or research question. In this way we can tailor comparisons to specific research questions targeted to specific patterns of means.

The first comparison in Table 8.2 involves the difference between Treatment 1 and Treatment 2. This comparison corresponds to the research question "Is the mean for Treatment 1 statistically different from the mean for

Table 8.2: Three comparisons on $k = 4$ treatment means.

	Coefficients				Notation				Value
Comp.	$\bar{X}_{1.}$	$\bar{X}_{2.}$	$\bar{X}_{3.}$	$\bar{X}_{4.}$	$\bar{X}_{1.}$	$\bar{X}_{2.}$	$\bar{X}_{3.}$	$\bar{X}_{4.}$	$\sum_{j=1}^{k} a_{ji}^2$
d_1	1	-1	0	0	a_{11}	a_{21}	a_{31}	a_{41}	2
d_2	0	0	-1	1	a_{12}	a_{22}	a_{32}	a_{42}	2
d_3	1/2	1/2	$-1/2$	$-1/2$	a_{13}	a_{23}	a_{33}	a_{43}	1

Treatment 2?" Treatments 1 and 2 receive coefficients of 1 and -1, respectively, but Treatments 3 and 4 are assigned coefficients of 0 because those two means are irrelevant to this particular research question. Multiplying the treatment means by the coefficients in the first row, we obtain the comparison

$$\begin{aligned} d_1 &= (1)\bar{X}_{1.} + (-1)\bar{X}_{2.} + (0)\bar{X}_{3.} + (0)\bar{X}_{4.} \\ &= \bar{X}_{1.} - \bar{X}_{2.} \end{aligned}$$

The coefficients equal to zero eliminate the treatments that are not involved in that particular comparison. Note how the comparison yields a simple difference between the two means in question.

The second comparison in Table 8.2 involves the difference between the means of Treatment 3 and Treatment 4 because the coefficients are 0, 0, -1, 1. Multiplying the treatment means by the coefficients in the second row, we obtain the comparison

$$\begin{aligned} d_2 &= (0)\bar{X}_{1.} + (0)\bar{X}_{2.} + (-1)\bar{X}_{3.} + (1)\bar{X}_{4.} \\ &= \bar{X}_{4.} - \bar{X}_{3.} \end{aligned}$$

The third comparison in Table 8.2 involves a more complicated research question: Is the average of Treatments 1 and 2 statistically different from the average of Treatments 3 and 4? If we multiply the treatment means by the coefficients in the last row, we have the comparison

$$d_3 = \tfrac{1}{2}\left(\bar{X}_{1.} + \bar{X}_{2.}\right) - \tfrac{1}{2}\left(\bar{X}_{3.} + \bar{X}_{4.}\right)$$

or the difference between the average of the means for Treatments 1 and 2 and the average of the means for Treatments 3 and 4. The first two comparisons are pairwise comparisons, but the third is not. Thus, comparisons need not be limited to pairwise tests, and so the material in the present chapter generalizes the pairwise tests in Chapter 7. Many types of research questions can be converted into comparisons and tested in a very simple way.

Comparisons of the kind shown in Table 8.2 are linear functions of the treatment means. Any linear function of the treatment means such as

$$d_i = a_{1i}\bar{X}_{1.} + a_{2i}\bar{X}_{2.} + \cdots + a_{ki}\bar{X}_{k.}$$

is called a **comparison**, if at least two of the coefficients are not equal to zero and if the sum of the coefficients is equal to zero, that is, if

$$\sum_{j=1}^{k} a_{ji} = 0 \tag{8.1}$$

For Equation 8.1 to be true under the conditions stated, then it is obvious that for the sum of the coefficients to be 0 at least one of the coefficients must be negative and at least one must be positive.

Under the standard null hypothesis of the omnibus analysis of variance, all of the k treatment means have the same expected value μ. Then because the coefficients have the property that $\sum_{j=1}^{k} a_{ji} = 0$ for any comparison d_i, under the null hypothesis the expected value of d_i (the weighted sum of means where the values of the comparisons are the weights) will also be equal to zero. For example, if the comparison is (3, −1, −1, −1) so that on four treatment means we have

$$d_i = 3\bar{X}_{1i} - \left(\bar{X}_{2i} + \bar{X}_{3i} + \bar{X}_{4i}\right)$$

then under the null hypothesis the population value of d_i is 0.

This is a more general null hypothesis than the usual "population treatment means are equal" because the individual treatment means need not have identical population means in order for a weighted sum to be 0 under the null hypothesis. That is, the requirement under the null hypothesis for the comparison (3, −1, −1, −1) is

$$3\mu_1 - (\mu_2 + \mu_3 + \mu_4) \quad = \quad 0$$

There are many combinations of those four means that could result in a weighted sum of 0.

The data analyst can convert just about any research question about means into a comparison over those means. Thus, comparisons offer a direct way to test research questions. We now turn to describing how to perform statistical tests on comparisons and then return to the problem of how to convert a research idea into a comparison. We will also discuss constraints that are imposed on the number and types of comparisons one can make.

8.3 Standard Error of a Comparison

The estimated standard error of any comparison d_i, that is, the standard error of the corresponding weighted sum obtained by multiplying the means by the coefficients for the comparison, will be given by

$$s_{d_i} = \sqrt{MS_W \left(\frac{a_{1i}^2}{n_1} + \frac{a_{2i}^2}{n_2} + \cdots + \frac{a_{ki}^2}{n_k} \right)} \tag{8.2}$$

where MS_W is the mean square within treatments from the analysis of variance. Recall that MS_W is estimated as a pooled variance, so the homogeneity of variance assumption is applicable here as well. If the number of observations is the same for each mean, then Equation 8.2 may be written more succinctly as

$$s_{d_i} = \sqrt{\frac{MS_W}{n} \sum_{j=1}^{k} a_{ji}^2} \tag{8.3}$$

where n is the number of observations for a single mean. We note the special case that for any comparison between two means $\bar{X}_{l.}$ and $\bar{X}_{m.}$ the corresponding coefficients will be 1 and -1, and $\sum_{j=1}^{k} a_{ji}^2 = 2$. Then Equation 8.3 reduces to $\sqrt{\frac{2MS_W}{n}}$, which is similar to the usual standard error of the difference between two means when $n_1 = n_2 = n$. The difference however is in the computation of the MS_W term. When there are more than two treatment groups, all groups enter into the computation of the MS_W term, even though some treatment groups are weighted 0. This is justified under the equality of variance assumption. When the equality of variance assumption holds, then the pooled error term leads to a more powerful statistical test.

8.4 The *t* Test of Significance for a Comparison

Under the null hypothesis for the comparison, we have the population value of d equal to 0. Then, the statistical significance of the difference represented by any comparison d_i can be evaluated by finding the t value

$$t = \frac{d_i}{s_{d_i}} \tag{8.4}$$

The degrees of freedom for this t value is equal to the number of degrees of freedom associated with the mean square within treatments from the analysis of variance (that is, the denominator of the omnibus F test). This computed t value is compared to the critical value found in Table B.1 in Appendix B.

In Table 8.3, we give the means for the treatments of the analysis of variance reported in Table 8.1 and the coefficients for the three comparisons

Table 8.3: Application of the comparisons of Table 8.2. The means are those obtained in the experiment summarized in the source table shown in Table 8.1.

	$\bar{X}_{1.}$	$\bar{X}_{2.}$	$\bar{X}_{3.}$	$\bar{X}_{4.}$	
Comparison	17.2	19.4	15.8	19.0	Value of d_i
d_1	1	−1	0	0	−2.2
d_2	0	0	−1	1	3.2
d_3	1/2	1/2	−1/2	−1/2	0.9

of Table 8.2. Multiplying the means by the corresponding coefficients for each comparison, we obtain $d_1 = -2.2$, $d_2 = 3.2$, and $d_3 = 0.9$.

Summing the squares of the coefficients for each comparison, we have

$$\sum_{j=1}^{k} a_{j1}^2 = 2 \qquad \sum_{j=1}^{k} a_{j2}^2 = 2 \qquad \sum_{j=1}^{k} a_{j3}^2 = 1$$

From the analysis of variance in Table 8.1, the pooled error term is $MS_W = 3.04$. The standard errors given by Equation 8.3 for each of the three comparisons are as follows:

$$s_{d_1} = \sqrt{\frac{3.04}{10}(2)} = 0.78$$
$$s_{d_2} = \sqrt{\frac{3.04}{10}(2)} = 0.78$$
$$s_{d_3} = \sqrt{\frac{3.04}{10}(1)} = 0.55$$

Dividing each d_i by its standard error, we obtain the corresponding t tests

$$t_1 = \frac{-2.2}{0.78} = -2.82$$
$$\mathrm{t}_2 = \frac{3.2}{0.78} = 4.10$$
$$\mathrm{t}_3 = \frac{0.9}{0.55} = 1.64$$

Each of these t's has 36 degrees of freedom, the number of degrees of freedom associated with the MS_W from the analysis of variance. If we use a two-sided $\alpha = 0.05$ with a t critical of 2.028, the first two comparisons d_1 and d_2 are statistically significant, whereas the third d_3 is not.

Confidence limits for d_i may be established in the usual way as $d_i \pm ts_{d_i}$, where the t used in the confidence interval formula is the t from the Table for a two-tailed test at the desired interval. For a 95% confidence interval the two-tailed t-value corresponds to the tabled value for $\alpha = 0.05$ with degrees of freedom corresponding to the MS_W term. For each of the three comparisons in the example we have

$$\begin{aligned} -2.2 &\pm (2.028)(0.78) \\ 3.2 &\pm (2.028)(0.78) \\ 0.9 &\pm (2.028)(0.55) \end{aligned}$$

The lower and upper 95% confidence limits are, respectively, $(-3.78, -0.62)$, $(1.62, 4.78)$, and $(-0.22, 2.02)$. Confidence intervals that do not include zero, the typical value of the null hypothesis, correspond to rejecting the null hypothesis in the context of a statistical test.

8.5 Orthogonal Comparisons

We now define the useful concept of orthogonality. If we make two comparisons d_i and d_j on the same set of k treatment means where all treatments have the same sample size, then d_i and d_j are said to be **orthogonal** if the sum of the products of the corresponding coefficients for the two comparisons is equal to zero. That is, two comparisons are orthogonal when the weights satisfy this equation

$$\sum_{t=1}^{k} a_{ti}a_{tj} = a_{1i}a_{1j} + a_{2i}a_{2j} + \cdots + a_{ki}a_{kj} = 0$$

The comparisons shown in Table 8.2 are **mutually orthogonal** because the sum of the products of the coefficients for all possible pairs of comparisons are equal to zero. For example, for comparisons d_1 and d_2, we have

$$(1)(0) + (-1)(0) + (0)(-1) + (0)(1) = 0$$

The sum of the products of the coefficients for comparisons d_1 and d_3, and, also for comparisons d_2 and d_3, totals zero. As we will see below, orthogonality permits an interesting connection between a set of comparisons and the sum of squares for treatment in the analysis of variance.

When sample sizes are unequal, then orthogonality between two comparisons should be defined as follows:

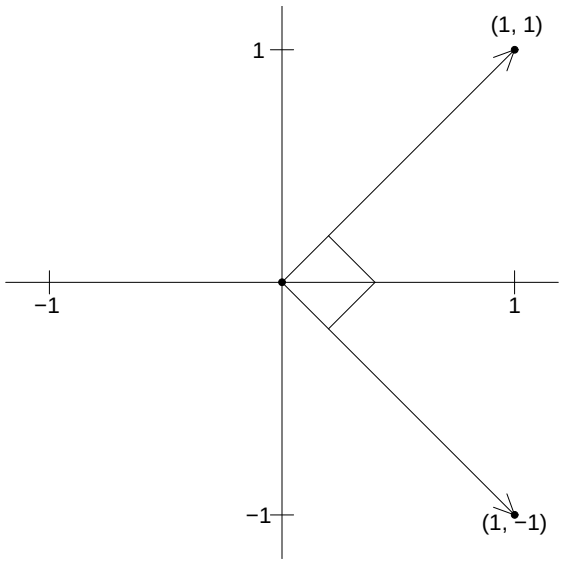
$$\sum \frac{a_{ti}a_{tj}}{n_t} = \frac{a_{1i}a_{1j}}{n_1} + \frac{a_{2i}a_{2j}}{n_2} + \cdots + \frac{a_{ki}a_{kj}}{n_k} = 0$$

Orthogonality can be given a geometric interpretation. Consider the two comparisons (1, 1) and (1, −1). If we plot these two points as vectors (an arrow from the origin to the point), we can see that the two comparisons are at 90 degrees to each other, as shown in Figure 8.1. These two comparisons are orthogonal because $(1)(1) + (1)(-1) = 0$. Indeed, orthogonality refers to comparisons that are at right angles when represented as vectors. When there are more than three treatments, the geometric picture is difficult to draw or see because we have trouble seeing in more than three spatial dimensions; but the concept of right angles extends to any number of treatments.

8.6 Choosing a Set of Orthogonal Comparisons

Orthogonality is a useful constraint on possible comparisons, but there are many sets of orthogonal comparisons that are possible. For example, with $k = 4$ means, the three sets of orthogonal comparisons given in Table 8.4 differ from one another and also from the set of orthogonal comparisons given in Table 8.2. Each of the three sets of comparisons given in Table 8.4

Figure 8.1: Geometric interpretation of orthogonality—the comparisons (1, 1) and (1, −1).



is orthogonal within the set, which can easily be verified by calculating the sum of the products of the corresponding coefficients for each possible pair of comparisons within each set. In each case, the sum of these products is equal to zero.

Because more than one set of orthogonal comparisons is possible for a given group of $k \geq 3$ means, the particular set of comparisons used in a study depends on the researcher's interests and should be planned at the same time the study is planned. Consider the particular three orthogonal comparisons shown in Table 8.2. Suppose the dependent variable of interest was a measure of maze performance and the four treatments were:

Group 1: a treatment tested after 12 hours of water deprivation

Table 8.4: Three different sets of orthogonal comparisons on $k = 4$ treatment means.

	$\bar{X}_{1.}$	$\bar{X}_{2.}$	$\bar{X}_{3.}$	$\bar{X}_{4.}$
Set 1	Coefficients			
d_1	-3	1	1	1
d_2	0	-2	1	1
d_3	0	0	1	-1
Set 2	Coefficients			
d_1	1	1	-1	-1
d_2	-1	1	-1	1
d_3	-1	1	1	-1
Set 3	Coefficients			
d_1	-3	-1	1	3
d_2	1	-1	-1	1
d_3	-1	3	-3	1

Group 2: a treatment tested after 24 hours of water deprivation

Group 3: a treatment tested after 12 hours of food deprivation

Group 4: a treatment tested after 24 hours of food deprivation

In this design the first comparison in Table 8.2 tests for the difference between the 12- and 24-hour water-deprived treatments; the second comparison tests for the difference between the 12- and 24-hour food-deprived treatments; and the third comparison tests for the difference between the average performance of the water-deprived and the food-deprived treatments. This third comparison tests whether the means of the water deprivation groups differ from the means of the food deprivation groups, ignoring the specific time interval of the deprivation (that is, 12 or 24 hours). We will return to examples such as these when we discuss factorial analysis of variance. These types of comparisons are given special names such as main effect comparison, or interaction comparison, or special main effect comparison, as we will see in later chapters.

We do not need to make all of the possible $k-1$ orthogonal comparisons in a given set. In some cases, the experimenter may only be interested

in a few of the possible comparisons. Again, it is not necessary that the omnibus $F = MS_T/MS_W$ be statistically significant (or even tested, for that matter) prior to testing planned orthogonal comparisons. Researchers may test individual comparisons without testing the omnibus F test.

8.7 Protection Levels with Orthogonal Comparisons

If $k-1$ orthogonal comparisons are made on a set of k treatment means, the numerators of the t ratios will be independent due to orthogonality of the comparison. The t ratios themselves will not be independent because the tests of significance are all made by using a common denominator MS_W. However, if α is small, say 0.05 or 0.01 for a single test, and if k is not large, then the protection level, $(1-\alpha)^{k-1}$, based on the assumption that the k tests are independent, will be approximately equal to the lower-bound protection level, $1-(k-1)\alpha$, based on the Bonferroni inequality. For example, with $\alpha = 0.01$ for a single test and with $k = 16$, we have $(1-0.01)^{16-1} = 0.86$ as the protection level for a set of 15 independent tests. With the Bonferroni inequality, we have $1-(16-1)0.01 = 0.85$ as the lower-bound estimate of the protection level. For other values of α or k, the approximation may not work as well. Thus, for all practical purposes, the set of $k-1$ orthogonal comparisons can be regarded as independent in evaluating the protection level and $P(E)$.

8.8 Treatments as Values of an Ordered Variable

In some studies the treatments may consist of different values of an ordered variable. For example, we might test different treatments after 0, 6, 12, and 18 hours of food or water deprivation. In other cases, the treatments may

consist of increasing intensities of shock, of increasing amounts of reward, of decreasing numbers of reinforcements, or of decreasing dosages of a drug.

If the treatments consist of different values of an ordered variable and the differences between the values are equal, then we may be interested in determining whether the treatment means are functionally related to the values of the treatment variable. We may, for example, be interested in testing whether the treatment means are linearly related to the values of the treatment variable or whether the treatment means deviate significantly from a linear relation (that is, deviate from a straight-line relation). If the deviations from linearity are statistically significant, then we may wish to determine whether there is a significant curvature in the trend of the means.

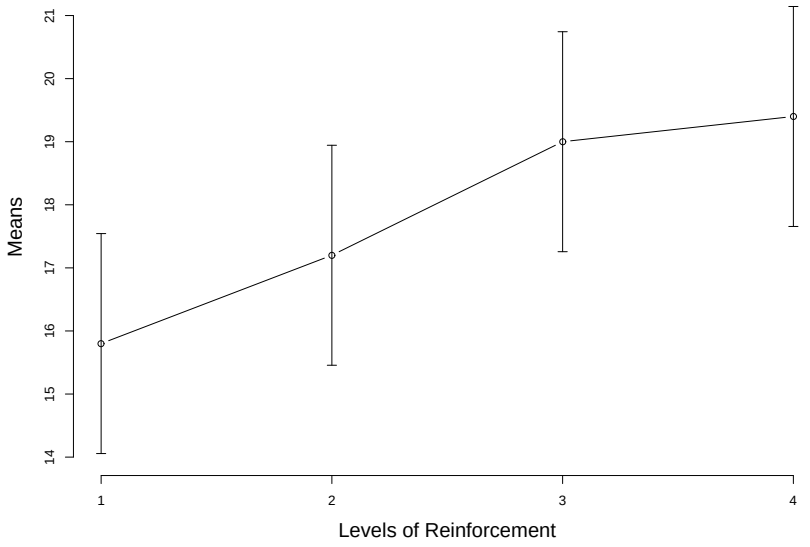
Assume, for example, that the treatments in an experiment consist of four equally increasing levels of reward, which we designate by 1, 2, 3, and 4. With $n = 10$ participants assigned to each treatment, assume that the analysis of variance for the experiment is as given in Table 8.1. The ordered treatment means, 15.8, 17.2, 19.0, and 19.4, represent the average performance on a game of skill at each of the four successive reinforcement levels. Figure 8.2 plots the treatment means against the levels of reinforcement. By visual inspection it appears that the trend of the means is approximately linear. We will next develop a statistical test to assess linearity and deviations from linearity.

8.9 Coefficients for Orthogonal Polynomials

To determine whether the linear component of the trend of the means is statistically significant and also whether the treatment means deviate significantly from linearity, we make use of a table of coefficients for orthogonal polynomials, Table B.5.[2] This table gives the coefficients to use for the linear, quadratic, and cubic components of the treatment sum of squares. The

[2]The coefficients for orthogonal polynomials given in Table B.5 are for the case of equal intervals in the values of the quantitative variable and for equal n's. If the intervals or n's

Figure 8.2: Treatment means for each of four levels of reinforcement. Error bars depict plus/minus 1 standard error where the standard error is based on the pooled MS_W, so the bars are identical across the four conditions.



coefficients in each row of Table B.5 sum to zero, and for any fixed value of k the sum of the products of the coefficients for the linear and quadratic comparisons is also zero. This result is true also for the linear and cubic coefficients and for the quadratic and cubic coefficients. The linear, quadratic, and cubic comparisons, therefore, meet the requirements for mutual orthogonality discussed earlier. The coefficients for the linear, quadratic, and cubic components for $k = 4$ treatments are shown in Table 8.5.

As another example, if $k = 5$, the successive sets of coefficients would correspond to the linear, quadratic, cubic, and quartic components of the

are unequal, the coefficients given in Table B.5 should not be used. For procedures to be used with unequal intervals and/or unequal n's, see Grandage (1958) or Gaito (1965).

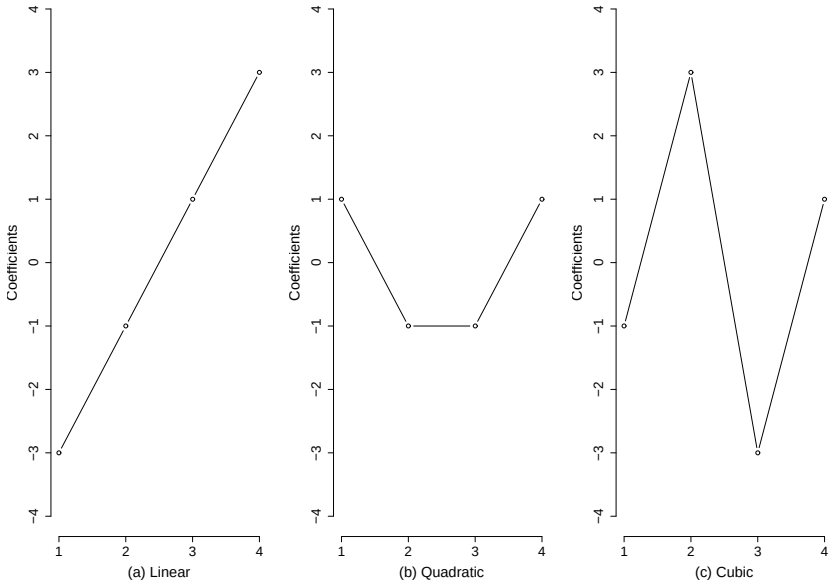
Table 8.5: Coefficients for the linear, quadratic, and cubic components for $k = 4$ treatments.

	Treatment means			
Comparison	15.8	17.2	19.0	19.4
Linear	−3	−1	1	3
Quadratic	1	−1	−1	1
Cubic	−1	3	−3	1

treatment sum of squares. Successive application of these coefficients would enable one to determine how well the trend of the treatment means is represented by a polynomial of the first, second, third, and fourth degree, respectively. Table B.5 in Appendix B gives only the coefficients for the linear, quadratic, and cubic components because seldom will the comparisons involving polynomials of degree greater than 3 be of interest. Coefficients for the higher-degree polynomials can be found in Fisher and Yates (1948) tables.

A graphical display of these polynomial comparisons may help illustrate the patterns they test. The coefficients for the linear component or comparison change signs only once, from minus to plus. These coefficients for $k = 4$ treatments are plotted in Figure 8.3(a), and the trend represented by the coefficients is a straight line. For the quadratic comparison, the coefficients change signs twice, from plus to minus to plus, and, as plotted in Figure 8.3(b), correspond to one reversal in the trend such as a U-shaped pattern. For the cubic coefficients, there are three sign changes in the coefficients, from minus to plus to minus to plus, and, as shown in Figure 8.3(c), these coefficients correspond to two reversals in the trend such as in an S-shaped, or Z-shaped, pattern. Thus, the number of sign changes in the coefficients indicates the degree of the polynomial.

Figure 8.3: Plots of linear (a), quadratic (b), and cubic (c) coefficients for orthogonal polynomials against equally spaced values of a quantitative variable. These plots show the characteristic pattern of each term in the polynomial: the linear checks for trends that do not have bends, the quadratic checks for trends with one bend, and the cubic checks for trends with two bends.



8.10 Tests of Significance for Trend Comparisons

The test of statistical significance uses the same equation for comparisons presented earlier in this chapter. Multiplying the treatment means by the coefficients for the linear comparison, as given in Table 8.5, we have for this comparison

$$L = (-3)\,(15.8) \;+\; (-1)\,(17.2) \;+\; (1)\,(19.0) \;+\; (3)\,(19.4) \;= 12.6$$

Then, with $\sum_{j=1}^{k} a_{jL}^2 = 20$, $n = 10$ observations for each treatment and $MS_W = 3.04$, we use Equation 8.4 to find the t value

$$t_L = \frac{12.6}{\sqrt{20\frac{3.04}{10}}} = 5.11$$

This observed t is compared to the critical t value from Table B.1 in Appendix B, based on 36 degrees of freedom (that is, the degrees of freedom associated with MS_W). The test confirms the visual inspection of Figure 8.2 that the four treatment means have a linear trend. The test rejects the null hypothesis that the slope of the line is zero (that is, rejects a horizontal line) because the observed t exceeds the $t_{critical} = 2.028$.

Similarly, the t test for the quadratic term has a numerator of

$$Q = (1)(15.8) + (-1)(17.2) + (-1)(19.0) + (1)(19.4) = -1$$

and a denominator of

$$\sqrt{4\frac{3.04}{10}} = 1.1027$$

because there are 10 participants per treatment, $MS_W = 3.04$, and $\sum_{j=1}^{k} a_{ji}^2 = 4$. The resulting t ratio is

$$\frac{-1}{1.1027} = -0.91$$

which in absolute value terms does not exceed $t_{critical}$. This failure to reject the null hypothesis for the quadratic comparison suggests there is little evidence in these data for a quadratic trend, at least up to the statistical power afforded by the present sample size.

Finally, the comparison corresponding to the cubic trend on the treatment means is

$$C = (-1)(15.8) + (3)(17.2) + (-3)(19.0) + (1)(19.4) = -1.8$$

with a resulting test statistic of

$$\frac{-1.8}{\sqrt{20\frac{3.04}{10}}} = -.73$$

The cubic trend is not statistically significant because the absolute value of the observed t of $-.73$ is not more extreme than the $t_{critical} = 2.028$. Thus, in this experiment the linear trend comparison is statistically significant with $\alpha = 0.05$, but the quadratic and cubic comparisons are not.

8.11 The Relation between a Set of Orthogonal Comparisons and the Treatment Sum of Squares

A set of orthogonal comparisons decomposes the sum of squares for treatments into smaller parts, each part representing the portion of sum of squares treatment attributable to that comparison. We illustrate this idea with the three orthogonal trend comparisons performed in the preceding section, but this idea will hold for any complete set of $k-1$ orthogonal comparisons.

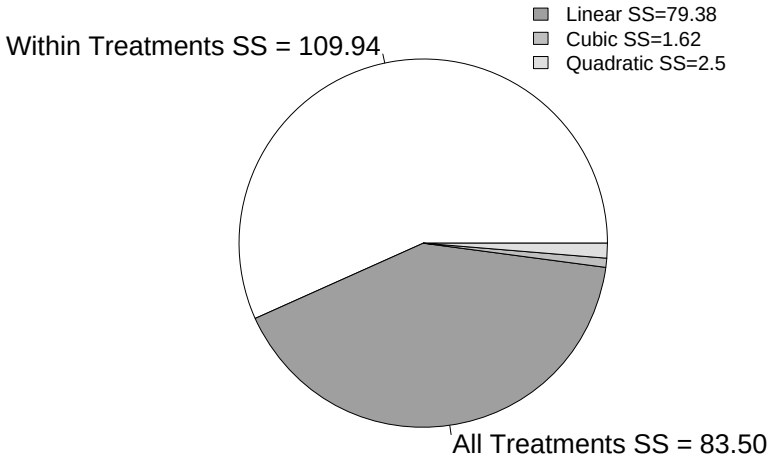
Recall that a complete set of orthogonal comparisons must involve $k-1$ comparisons, where k is the number of treatments and each pair of comparisons are orthogonal. The sum of squares for a single comparison i is defined as

$$SS_i = \frac{d_i^2}{\sum \frac{a^2}{n_i}}$$

Thus, for the linear, quadratic, and cubic comparisons in the preceding section we have

$$SS_L = \frac{12.6^2}{20}10 = 79.38$$
$$SS_Q = \frac{(-1)^2}{4}10 = 2.5$$
$$SS_C = \frac{(-1.8)^2}{20}10 = 1.62$$

Figure 8.4: Pie chart depicting the decomposition of sums of squares into treatments and within treatments, as well as the further decomposition of the sum of square treatments into orthogonal comparisons. The shaded regions together correspond to the entire treatment sum of squares (83.50), which is decomposed into separate portions by the particular orthogonal set of comparisons (in this example, the decomposition is based on the polynomial comparisons and their respective sum of squares SS_{linear}, $SS_{quadratic}$, and SS_{cubic}).



We now illustrate the decomposition of treatment sum of squares: the sum of the three sum of squares for each comparison ($79.38 + 2.5 + 1.62$) equals the treatment sum of squares from the analysis of variance, or 83.50, shown in Table 8.1. This is depicted graphically in Figure 8.4. Thus, a set of orthogonal comparisons decomposes the omnibus question into smaller chunks that test specific patterns in the treatment means.

8.12 Tests of Significance for Planned Comparisons

Planned comparisons provide information that is relevant to the interpretation of the outcome of a well-designed study. They are usually limited in number and planned prior to the examination of the data obtained in the experiment. In almost all cases, the comparisons will be based on theoretical or practical considerations of importance. In the drug experiment, for example, determining whether the combination of drugs A and B is or is not more effective than either drug A or drug B alone would be of practical, if not theoretical, importance. The comparison of the difference between the mean of drug A and the mean of drug B would also be of interest.

In testing planned orthogonal comparisons, one sets the protection level by the number of degrees of freedom associated with the MS_T (that is, $k - 1$). From an experimental point of view, it is difficult to perceive any great difference between testing $k - 1$ planned orthogonal comparisons on the one hand and $k - 1$ planned comparisons, not all of which are necessarily orthogonal, on the other hand. If the number of planned comparisons to be tested exceeds $k - 1$, then some experimenters may also consider it reasonable to perform these tests in the same manner in which they would perform planned orthogonal tests. Other experimenters may be more concerned about Type I errors and decide to use a more conservative test, such as the Bonferroni test, for planned but not necessarily orthogonal comparisons when the number of such comparisons is greater than $k - 1$. Recall that the Bonferroni procedure involves changing the criteria of the individual tests so that the overall Type I error remains at the desired level, say $\alpha = 0.05$. If one performs c comparisons, then each comparison can be tested using a criterion of $0.05/c$. Keep in mind that if the comparisons are not orthogonal, the Bonferroni procedure provides an upper-bound approximation.

Our suggestion is that if the experimenter is concerned about the Type I error rate, then he or she should replicate the study. Replication is a better way of dealing with concerns over Type I error rates than tinkering with the

α criterion level. But when replication is costly or not feasible (for example, when conducting a 30-year longitudinal study), then the Bonferroni correction provides one way to alleviate concerns about Type I errors that emerge when performing multiple tests. For a discussion of issues surrounding replication see Greenwald, Gonzalez, Harris, and Guthrie (1996).

8.13 Effect Size for Comparisons

In this section we present a measure of effect size for a specific comparison. The definitional formula is given by

$$r = \sqrt{\frac{SS_C}{SS_C + SS_W}} \tag{8.5}$$

where SS_C refers to the sum of squares for comparison (as given in Section 8.11) and SS_W refers to the sum of squares within treatments. For example, in Section 8.11 we presented the linear trend comparison, which had a sum of squares equal to 79.38. Recall that the sum of squares within treatments for this example was equal to 109.44. Thus, the effect size r for the linear comparison is equal to

$$0.648 = \sqrt{\frac{79.38}{79.38 + 109.44}}$$

The definition of effect size presented here compares the sum of squares of the specific comparison to the sum of squares within treatments. A more convenient and equivalent version of Equation 8.5 is

$$r = \sqrt{\frac{t^2}{t^2 + df}} \tag{8.6}$$

where the t corresponds to the observed t of the comparison and df corresponds to the degrees of freedom associated with the within-treatment MS_W

term. Both Equations 8.5 and 8.6 yield identical results. Equation 8.6 is more versatile because it can be applied readily to computer output and to published papers. Applying Equation 8.6 to the linear comparison example, we verify that it yields the same answer as Equation 8.5; that is, $t = 5.11$ and $df = 36$, thus

$$0.648 = \sqrt{\frac{5.11^2}{5.11^2 + 36}}$$

8.14 The Equality of Variance Assumption

In this chapter we reviewed the topic of comparisons under the assumption of equal variances. The homogeneity of variance assumption should be checked whenever comparisons are tested. Fortunately, there is a generalization of the Welch t test presented in an earlier chapter that permits testing of comparisons even when the pooling assumption may not be justified (Brown & Forsythe, 1974). The logic is the same as for the Welch t test—the degrees of freedom are adjusted to take into account the discrepancy in the treatment variances. This more general test of a comparison is now implemented in many statistical packages, usually under the label "separate variance" test to indicate that the variances are not pooled.

8.15 Unequal Sample Size

Unequal sample sizes are not a problem for the comparison test presented in this chapter (though an ANOVA purist would be careful to define orthogonality to take into account differences in sample size). The computation of the standard error uses Equation 8.2 to take into account the different sample sizes. Unequal sample sizes will become an issue when we discuss factorial designs, and we will return to this issue in a later chapter.

8.16 Unplanned Comparisons

We describe a procedure developed by Scheffé (1953) that can be used to test the significance of any and all comparisons on a set of k means, including comparisons that may be suggested after observing the treatment means themselves. In other words, we do not need to plan the comparisons in advance, nor do the comparisons need to be limited in number, nor do the comparisons need to be orthogonal. The Scheffé test provides such flexibility because of a very clever idea that Scheffé formalized. Scheffé's test can, of course, be used for testing planned or orthogonal comparisons as well as pairwise comparisons of the kind described in the preceding chapter. The test will be more conservative than the procedures described for testing planned or orthogonal comparisons; that is, larger observed differences from the null hypothesis will be required for statistical significance by the Scheffé criterion, so one may not want to use the Scheffé test in all settings. For example, if the investigator only wants to test all possible pairwise comparisons, then the Tukey test is sufficient.

We emphasize, however, that if the omnibus $F = MS_T/MS_W$ is not statistically significant at a given α criterion, then no comparison will be judged significant by the Scheffé test at the same α criterion. It is useless, in other words, to apply the Scheffé test to comparisons when the omnibus test is not significant. The rationale for this assertion will be explained in more detail later in the chapter. On the other hand, if $F = MS_T/MS_W$ is significant with, say, $\alpha = 0.05$, then there will be at least one comparison on the treatment means that will also be statistically significant by the Scheffé criterion. There may, of course, be more than one statistically significant comparison.

8.16.1 Some Examples of the Scheffé Test

Table 8.6 shows some of the many possible comparisons that could be made on a set of $k = 4$ treatment means. The first 6 comparisons are pairwise

comparisons; the next 12 comparisons are between one mean and the average of two other means; the next 3 comparisons are between the average of two means and the average of two other means; and the last 4 comparisons are between one mean and the average of the other three means. Since there are at most $k - 1 = 3$ orthogonal comparisons possible, Table 8.6 has a high degree of redundancy across the 25 comparisons that are shown.

The comparisons shown in Table 8.6 do not exhaust all the possibilities. For example, the table does not show the linear, quadratic, and cubic comparisons that could be performed on $k = 4$ treatment means. Nor does the table show comparisons of the kind

$$2 \qquad 1 \qquad -3 \qquad 0$$

or

$$2 \qquad 7 \qquad -6 \qquad -3$$

of which there are an unlimited number.

We do not recommend that experimenters perform many comparisons simply because it is possible. Rather, we are pointing out that the Scheffé test gives the researcher the opportunity to test as many comparisons as he or she wants. The price paid, however, is a very conservative criterion for statistical significance. Further, there may be problems in interpretation when there are many comparison tests. Some of those comparisons will be redundant (that is, nonorthogonal); thus two comparisons may turn out to be significant because they overlap in how they weight the treatment means. For instance, these two comparisons on four treatment means overlap on the first two treatments: (1, −1, 0, 0) and (1, −1, 1, −1). If the four means are (6, 2, 3, 3), then both comparisons will be significant because they pick up the 6 versus 2 difference in the first two treatment means. The Scheffé test merely controls the Type I error rate; it does not identify which contrasts represent meaningful research questions—that task is left to the researcher.

Table 8.6: Some possible comparisons on $k = 4$ treatment means.

Comparison	$\bar{X}_{1.}$ 17.2	$\bar{X}_{2.}$ 19.4	$\bar{X}_{3.}$ 15.8	$\bar{X}_{4.}$ 19.0	$\sum a^2$	d
1 vs. 2	1	−1	0	0	2	−2.2
1 vs. 3	1	0	−1	0	2	1.4
1 vs. 4	1	0	0	−1	2	−1.8
2 vs. 3	0	1	−1	0	2	3.6
2 vs. 4	0	1	0	−1	2	0.4
3 vs. 4	0	0	1	−1	2	−3.2
1 vs. 2 + 3	2	−1	−1	0	6	−0.8
1 vs. 2 + 4	2	−1	0	−1	6	−4.0
1 vs. 3 + 4	2	0	−1	−1	6	−0.4
2 vs. 1 + 3	−1	2	−1	0	6	5.8
2 vs. 1 + 4	−1	2	0	−1	6	2.6
2 vs. 3 + 4	0	2	−1	−1	6	4.0
3 vs. 1 + 2	−1	−1	2	0	6	−5.0
3 vs. 1 + 4	−1	0	2	−1	6	−4.6
3 vs. 2 + 4	0	−1	2	−1	6	−6.8
4 vs. 1 + 2	−1	−1	0	2	6	1.4
4 vs. 1 + 3	−1	0	−1	2	6	5.0
4 vs. 2 + 3	0	−1	−1	2	6	2.8
1 + 2 vs. 3 + 4	1	1	−1	−1	4	1.8
1 + 3 vs. 2 + 4	1	−1	1	−1	4	−5.4
1 + 4 vs. 2 + 3	1	−1	−1	1	4	1.0
1 vs. 2 + 3 + 4	3	−1	−1	−1	12	−2.6
2 vs. 1 + 3 + 4	−1	3	−1	−1	12	6.2
3 vs. 1 + 2 + 4	−1	−1	3	−1	12	−8.2
4 vs. 1 + 2 + 3	−1	−1	−1	3	12	4.6
D_1	5	−4	0	0	−30	
D_2	0	0	8	−3	120	
D_3	11	11	−9	−9	−300	

Note: Columns 2–5 list the coefficients for each of four treatments, column 6 lists the sum of the squared coefficients, and column 7 lists the value of d, the sum of the products of the comparison coefficient with the respective treatment mean.

8.16.2 The Scheffé Test for Comparisons

If comparisons are made on the treatment means, then, as we saw before, the standard error of the comparison will be given by

$$s_{d_i} = \sqrt{MS_W \sum \frac{a_i^2}{n_i}} \tag{8.7}$$

The test of significance for the comparison is then made by finding

$$t = \frac{d_i}{s_{d_i}} \tag{8.8}$$

The numerator of Equation 8.8 (that is, d_i) is computed by multiplying the comparison coefficient with the respective treatment mean and summing the products.

The Scheffé test uses a special criterion to compare the observed t ratio of the comparison. The t defined by Equation 8.8 is evaluated for significance by comparing it with the Scheffé criterion

$$t' = \sqrt{(k - 1)\, F} \tag{8.9}$$

where k is the number of treatments and F is the critical value from Table B.2 in Appendix B for $(k-1)$ numerator degrees of freedom and the degrees of freedom for the denominator corresponding to MS_W.

Confidence limits for the comparison value d_i can also be constructed under the Scheffé framework by the formula

$$d_i \pm t' s_{d_i} \tag{8.10}$$

Defining confidence intervals in this manner for a set of comparison values d_i yields what is known as a **simultaneous confidence interval**. The Type I error rate for the set of confidence intervals is controlled by the Scheffé criterion.

Because there is an MS_W term in Equation 8.7, the equality of variance assumption is invoked as usual (that is, the homogeneity of variance assumption is what justifies the pooling of the treatment variances into MS_W). The Scheffé test has been generalized in a manner that relaxes the equality of variance assumption (Brown & Forsythe, 1974). However, this generalized Scheffé test has not been widely implemented in standard statistical packages.

8.16.3 Properties of the Scheffé Test

The Scheffé test is a statistical test that permits the investigator to examine the data and to make an unlimited number of comparisons. Regardless of the number of comparisons tested, the protection level remains greater than or equal to $1 - \alpha$, and $P(E)$ remains less than or equal to α. That is, if comparisons are tested with the Scheffé procedure, the probability of making a Type I error will be less than or equal to α—regardless of how many comparisons one chooses to make.

As we pointed out earlier, the Scheffé test has the property that if the omnibus $F = MS_T/MS_W$ is not statistically significant at α, then no comparison that can be made on the k treatment means will be statistically significant. In the example, if the F test for the treatment mean square had not been equal to or greater than $F = 2.86$ (which is the critical value for 3 and 36 degrees of freedom with $\alpha = 0.05$), then there would not be any comparison on the $k = 4$ means that would result in $t \geq t' = 2.93$ (that is, no comparison would reach statistical significance). If the omnibus $F = MS_T/MS_W$ is statistically significant with $\alpha = 0.05$, then there will be at least one comparison that can be made on the treatment means that will also be significant, and, of course, there may be more than one comparison that will be statistically significant. It does not follow, however, that comparisons found to be statistically significant will necessarily be those that are of interest to the experimenter or even correspond to meaningful research questions.

The Scheffé test is based on Scheffé's observation that it is possible to find a single comparison that yields the same sum of squares as the overall treatment sum of squares. Recall that orthogonal comparisons decompose the treatment sum of squares. Scheffé showed that it is always possible to find a comparison that completely exhausts the sum of squares treatment SS_T. That is, there exists a comparison with sum of squares SS_C equal to the entire sum of squares for treatments SS_T. This comparison is called the "maximum comparison" because a comparison cannot be greater than this single comparison (otherwise it would be greater than the treatment sum of squares, which it cannot be). Scheffé derived the sampling distribution for this maximum comparison and thus proved the test that we now refer to as the Scheffé test. Because this test is based on sampling distribution of the maximum comparison, it has the property that it can be used for any number of comparisons, and it will automatically provide a correction for the Type I error problem.

8.17 Summary

The comparison of treatment means described in this chapter is an important tool in research. Frequently, a researcher conducts a study to test a particular hypothesis about a pattern of treatment means. The omnibus test we discussed in Chapter 6 is useless to the researcher who is not interested in the global question of whether the means differ but instead is interested in testing specific hypotheses about the treatment means. As long as the research question can be operationalized in terms of a weighted sum of treatment means, then a comparison to test that predicted pattern directly is most useful.

The Scheffé test is the ideal test for the experimenter who does not have planned comparisons and who wishes to explore thoroughly the outcome of an experiment, making any and all comparisons suggested by the data. The experimenter can do so knowing that, regardless of the number of compar-

isons made, $P(E)$ will be less than or equal to the chosen value of α, say $\alpha = 0.05$. Of course, this flexibility in statistical tests comes at the price of a very conservative criterion for statistical significance. This may lead to lower statistical power.

In the case where an experimenter will only test all possible pairwise comparisons, then we recommend the Tukey test (Chapter 7) over the Scheffé test. The Tukey test directly takes into account the sampling distribution relevant to pairwise comparisons and will not be as conservative as the Scheffé test, which takes into account the sampling distribution for a much larger number of possible comparisons.

8.18 Questions and Problems

1. A study includes a control group and 5 treatment groups, with 10 observations for each group. We have $MS_W = 36.00$ for the within-treatment mean square. The means for the six groups are given here:

Control	A	B	C	D	E
18.6	20.5	23.4	19.6	28.3	26.2

   Perform a comparison that tests whether the control group differs from the average of the five treatments. Explain why this comparison is not identical to a two-sample t test that compares the control group to all other participants (that is, calling participants in the other five treatments a single group).

2. We have a between-subjects experimental design in which the treatments consist of three equally spaced intervals of testing. One group is tested for retention of learned material after 12 hours, another group after 24 hours, and the third group after 36 hours. The means for the groups are 11.0, 9.0, and 5.0, respectively. We have $n = 10$ participants in each group, and MS_W is equal to 20.0 with 27 degrees of freedom.

(a) Do the treatment means differ significantly from one another? Explain how you interpret this question, and justify the particular statistical test you perform.

(b) Test the linear trend component of the means for significance. What is the effect size of the linear component?

3. In an experiment involving $k = 4$ treatments, $n = 10$ participants were assigned at random to each treatment. We have $MS_W = 16.0$. Find the standard error for each of the following comparisons:

(a)	1	1	−1	−1
(b)	1	−1	0	0
(c)	3	−1	−1	−1

4. What is meant by a comparison on a set of k means?

5. What is the condition for two comparisons to be called orthogonal?

6. We have $k = 5$ treatments with $n = 8$ participants assigned at random to each treatment. If the omnibus $F = MS_T/MS_W$ is not significant with $\alpha = 0.05$, explain why there will not be a comparison on the treatment means that will be significant according to the Scheffé test.

7. We have $n = 10$ participants assigned at random to each of $k = 8$ treatments. We wish to make $c = 5$ unplanned comparisons and to have $P(E) \leq 0.05$. What value of t (that is, $t_{critical}$) will be required for significance for these comparisons, using Scheffé's test?

8. Suppose that in the experiment described in Problem 7 the $c = 5$ comparisons are planned comparisons, decided upon prior to conducting the experiment. We are concerned, however, about Type I errors and want to have $P(E) \leq 0.05$ for the set of $c = 5$ comparisons. If the Bonferroni t statistic is used to control $P(E)$, what value of t will be required for significance in testing each of the five comparisons? Which test, Scheffé or Bonferroni, is more conservative in this example?

9. Use data presented as part of Question 2 in Chapter 6 (page 187), which presented raw data for six groups. Imagine that these six conditions are six levels of increasing dose of a drug, with Treatment 1 receiving the lowest dose and Treatment 6 receiving the highest dose.

   (a) Construct a set of polynomial comparisons for these six treatment groups.

   (b) Compute the t test for each of these comparisons.

   (c) Show that the sum of squares across the set of orthogonal comparisons equals the sum of squares from the one-way ANOVA.

Guilford Publications
72 Spring Street
New York, NY 10012
212-431-9800
800-365-7006
www.guilford.com