

Analysis of Variance

12

Chapter Outline

- 12.1 Assumptions for Use of the ANOVA
- 12.2 One-Way ANOVA
- 12.3 Multiple Comparisons
- 12.4 Two-Way ANOVA for the Randomized Block Design with m Replicates
- 12.5 Two-Way ANOVA with Interaction
- 12.6 Linear Model Representation of the ANOVA
- 12.7 ANOVA with Unequal Numbers of Observations in Subgroups

In Chapter 8, we used the t test for testing the equality of two population means based on data from two independent samples. In this chapter, we introduce a procedure for testing the equality of two or more means. The two experimental designs discussed in Chapter 6 — the completely randomized and the randomized block designs — will be considered.

The comparison of two or more means is based on partitioning the variation in the dependent variable into its components — hence, the method is called the analysis of variance (ANOVA). It was introduced by Sir Ronald A. Fisher and has been used in many fields of research. We begin this chapter with a presentation of the assumptions made when the ANOVA is used. This section is followed by an introduction to the one-way ANOVA. In conjunction with this analysis, we present three methods used in multiple comparison analysis. These topics are followed by the analysis of the randomized block design, an example of a two-way ANOVA, and a two-way ANOVA with interaction. We next provide a linear model representation of the ANOVA, followed by the use of the linear model with unequal group sizes.

12.1 Assumptions for Use of the ANOVA

The ANOVA is used to determine whether or not there is a statistically significant difference among the population means of two or more groups. The theoretical basis of the ANOVA requires that the data being analyzed are independent and normally distributed. We must also assume that the population variances in each of the groups have the same value, σ^2 . The ANOVA procedure works reasonably well if there are small departures from the normality assumption. However, if the variances are very different, there is concern about the significance levels reported in the analysis (Scheffé 1959). This concern is consistent with the material presented in Chapters 7 and 8, where we saw different methods for comparing two means, depending on whether or not we assumed that the population variances were equal. One method for protecting against the effects of different values for the variances is to have approximately equal numbers of observations in each of the groups being analyzed. Another approach involves trans-

formations of the dependent variable (Kleinbaum, Kupper, and Muller 1988; Lin and Vonesh 1989). One further assumption is that the groups being compared are the only groups of interest. This assumption means that the factors — the independent variables — are fixed factors. Fixed and random factors and their implications are discussed later in this chapter, and further information is available elsewhere (Steel and Torrie 1980).

There are no firm rules for the number of observations required by the ANOVA. It is possible to perform power calculations or to use the size of confidence intervals to estimate the required sample size. In general, we recommend that there be a minimum of 5 to 10 observations for each of the combinations of levels of the independent variables used in the analysis. For example, with two independent variables, if one has 3 levels and the other independent variable has 4 levels, there are 12 combinations of levels.

12.2 One-Way ANOVA

In a one-way ANOVA, there is only one independent variable. The data to be analyzed are obtained from either (1) a random sample of subjects who belong to different groups — for example, different racial groups — or (2) an experiment in which the subjects are randomly assigned to one of several groups. The latter situation arises when we use the completely randomized design discussed in Chapter 6. In the completely randomized design, subjects are randomly allocated to groups and the groups represent the levels of the independent variable. Observations of the continuous variable of interest, the dependent variable, are taken on the subjects and the subject's group membership is also recorded. In the following example, we consider data from a completely randomized design, and we wish to determine whether or not there is a difference in mean age among three groups.

Example 12.1

Data shown in Table 12.1 are based on an article by Kimball et al. (1986) and can be analyzed using a one-way ANOVA. In the article, the authors wished to evaluate ventricular performance after surgical correction of congenital coarctation of the aorta. The ventricular performance was compared to that found in two control groups. Because of the possible roles that age and gender play on ventricular performance, the authors wanted the age and sex distributions of the subjects who had undergone the surgery to be similar to those of the members of the two control groups. We wish to examine whether or not the authors were successful in obtaining groups that were similar on the age variable. The ages shown in Table 12.1 are hypothetical, based on the summary values reported by Kimball et al. In this example, the dependent variable is age, and the independent variable is the group to which the subjects belong.

The entries in Table 12.1 can be represented symbolically as y_{ij} , where the first subscript indicates the subject's group membership and the second subscript identifies the subject in the i th group. For example, y_{11} is 32 years old, y_{12} is 28 years old, y_{25} is 34 years old, y_{26} is 33 years old, and so on. The first subscript ranges from 1 to 3. When the first subscript has the value of 1, the range of the second subscript is

Table 12.1 Hypothetical ages for control and surgery subjects.

Group	Ages																	
Surgery	32	28	22	25	20	20	28	28	20	29	22	37	18	29	22	32	21	34
	19	23	23	26	41	20	33											
Control I	32	26	31	39	34	33	29	41	35	33	33	43	25	39	36	37	28	34
	27	45	22	29	51	28	35											
Control II	31	35	26	28	22	29	27	21	22	27	24	44	21	25	27	18	27	36

from 1 to 25, and this is also the case when the first subscript is 2. When the first subscript has the value of 3, the second subscript ranges from 1 to 18. In general, there are r groups and n_i observations in the i th group. We also use the \cdot notation introduced in Chapter 10. For example, $y_{i\cdot}$ is a shorthand notation for $\sum_j y_{ij}$ and $y_{\cdot\cdot}$ is shorthand for $\sum_i \sum_j y_{ij}$. Thus, $y_{1\cdot}$ represents the sum of all the ages for the subjects in the surgery group, and $y_{\cdot\cdot}$ is the sum of all the 68 ages in the sample. It follows that $\bar{y}_{i\cdot}$ is the sample mean of the i -th group, and $\bar{y}_{\cdot\cdot}$ is the overall sample mean.

In the following section, we show how the data in Example 12.1 can be analyzed.

12.2.1 Sums of Squares and Mean Squares

As was just mentioned, this method, the analysis of variance, is based on a partitioning of the variation in the dependent variable. In the one-way ANOVA, there are two possible sources of variation in the dependent variable. One source is variation among (or between) the groups — that is, the groups may have different means that vary about the overall mean. The other possible source is variation within the groups. Not all the subjects in the same group will have exactly the same values, and the within-group variation reflects this.

The null hypothesis being tested here is that the population group means are equal to one another. If this hypothesis is true, all the observations come, in effect, from the same population. Thus, any variation that remains among the group means really reflects the random variation among the observations — that is, the within-groups variation. Thus, the adjusted among and within variations should be similar if the null hypothesis is true. If the null hypothesis is false, the adjusted among-groups variation should be larger than the adjusted within-group variation because it includes variation between the populations as well as the within-group variation. Thus, we can use the adjusted among- and within-group variations as the basis of a test of the equality of the group means.

We can represent the above idea in symbols as

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2.$$

This equation shows the partitioning of the total variation in Y , the dependent variable, about its mean into an among (or between)-group component and a within-group component. These sum of squares are called the total sum of squares corrected for the mean (SST), the among (or between)-group sum of squares (SSB) and the within-group sum of squares (SSW).

If we adjust these two components for the number of independent observations used in their calculations — that is, divide each component sum of squares by its degrees of freedom — we have the mean square among (or between) and the mean square within. The mean square between is

$$MSB = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{r-1} = \frac{\sum_{i=1}^r n_i (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{r-1}$$

where the second expression reflects the fact that the terms in the parentheses do not vary with j . The mean square within is

$$MSW = \frac{\sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{y}_{i\cdot} - \bar{y}_{i\cdot})^2}{n-r}$$

where n is the total number of observations — that is, the sum of the n_i . The degrees of freedom for the mean square between, $r-1$, comes from the calculation of the variation in r means. The degrees of freedom for the mean square within, $n-r$, is the result of summing the $n_i - 1$ degrees of freedom associated with the i th group over the r groups.

The mean square within is particularly useful as it also provides an adjusted estimate of the variation within groups — that is, of σ^2 , the variance of the dependent variable. It is based on the assumption that the variance of the dependent variable is the same within each group. If there is no difference between the group means, then the mean square between also estimates σ^2 .

Example 12.2

For the data in Table 12.1, we have the following values of means and sums of squares. First, $\bar{y}_{1\cdot}$, the sample mean of the first group, is 26.08, $\bar{y}_{2\cdot}$ is 33.80, and $\bar{y}_{3\cdot}$ is 27.22. The overall sample mean, $\bar{y}_{\cdot\cdot}$, is 29.22 years. The sum of squares between is

$$SSB = 25(26.08 - 29.22)^2 + 25(33.80 - 29.22)^2 + 18(27.22 - 29.22)^2 = 842.9.$$

The sum of squares within involves too many terms to show, but its sum of squares is 2660.8 and the total sum of squares (corrected) is 3503.7.

12.2.2 The F Statistic

The comparison of these two mean squares provides information about whether or not the null hypothesis is true. One way of comparing the mean squares is to take their difference. If the null hypothesis were true, then the difference would be zero. However, the probability distribution of the difference is not widely available. Another way of comparing the mean squares is to take the ratio of the mean square between to the mean square within. If the null hypothesis were true, the ratio would equal one. If the null hypothesis were false, the ratio would be larger than one. Fortunately, the probability distribution of the ratio has been worked out, and it is an F distribution with $r-1$ and $n-r$ degrees of freedom. Tables of the F distribution, named in honor of Sir Ronald Fisher, are shown in Appendix B11 for the 0.01, 0.05, and 0.10 significance levels for values of the numerator (f_1) and denominator (f_2) degrees of freedom parameters.

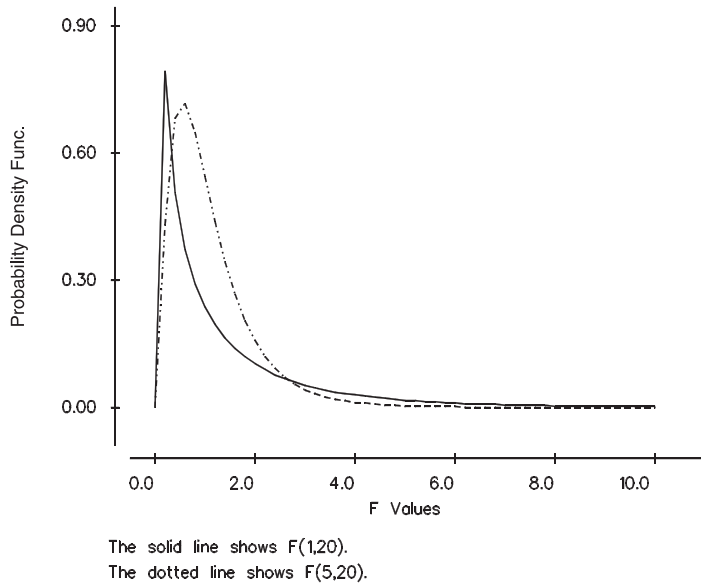


Figure 12.1 Plot of the probability density functions of the F distribution for $F_{1,20}$ (solid line) and $F_{5,20}$ (broken line).

The F distribution has many different shapes, depending on the values of the degrees of freedom parameters. Figure 12.1 shows the shape of the F distribution for degrees of freedom pairs (1 and 20) and (5 and 20). We can see that the shapes are different, but most of the probability (area) is associated with values of F close to one.

There is also a relation between the t and F distributions that can be seen from the t and F tables. The relation is $t_{k,1-\alpha/2}^2$ is equal to $F_{1,k,1-\alpha}$. For example, when k is 10, $t_{10,0.95}$ is 1.8125, and its square is 3.2852. Examination of the F tables in Appendix B11 shows that $F_{1,10,0.90}$ is 3.29. This equivalence when there are two groups leads us to think that there may be a relation between the ANOVA and t test approaches in the two-group situation.

12.2.3 The ANOVA Table

The preceding sums of squares and mean squares are usually presented in tabular format, as shown in Table 12.2.

Table 12.2 Typical ANOVA table for a one-way analysis.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Between Groups	$r - 1$	SSB	$SSB/(r - 1) = MSB$	MSB/MSW
Within Groups	$n - r$	SSW	$SSW/(n - r) = MSW$	
Total (Corrected)	$n - 1$	SST		

The degrees of freedom and sums of squares associated with the between and within groups sum to the corresponding total values. If these values do not sum to the total, a mistake has been made in the calculations.

The F statistic is then used to test the null hypothesis that the group means are equal against the alternative hypothesis that the group means are not all equal. When the null

hypothesis is true, the F statistic follows an F distribution with $r - 1$ and $n - r$ degrees of freedom. If the calculated F statistic is greater than $F_{r-1, n-r, 1-\alpha}$, found in Appendix Table B11, we reject the null hypothesis in favor of the alternative hypothesis at the α significance level. If the calculated F statistic is less than this critical value, we do not have sufficient evidence to reject the null hypothesis.

Example 12.3

Based on the sums of squares presented in Example 12.2, we can complete the ANOVA table for the ages shown in Table 12.1. Let us test the hypothesis of the equality of the mean ages at the 0.01 significance level. Table 12.3 is the ANOVA table for the age data.

There are 68 observations in the three groups. Hence, there are 2 degrees of freedom for the factor (between groups), variable, 65 degrees of freedom for error (within groups), and 67 degrees of freedom for the total sum of squares. The table shows the sums of squares and mean squares as well as the F ratio. The exact critical value of this test is not shown in Table B11, but the closest value shown for F is 4.98 for $F_{2,60,0.99}$. From the table, we see that the exact F value for $F_{2,65,0.99}$ is slightly less than 4.98.

The calculated F statistic (10.29) is greater than the approximate critical value of 4.98. Therefore, we reject the equality of the mean ages in favor of the alternative hypothesis. It appears that the three groups differ on age. This finding means that it may be necessary to take age into account in the analysis of ventricular performance. The square root of the mean square for error is 6.395 ($=\sqrt{40.9}$), and it is an estimate of σ .

Computer packages can be used to perform the analysis of variance (see **Program Note 12.1** on the website). The computer output shows the ANOVA table with the p -value associated with the F ratio along with group means and standard deviations.

Figure 12.2 shows box plots for the data in Table 12.1. The group means are represented by dots in the box. It appears that the difference is due mainly to the first control group having a mean age that is much greater than the other two groups. When there is a statistically significant difference among the group means, we can perform additional tests to see if we can determine the source of the differences in the means. The next section describes three approaches to this additional testing.

Table 12.3 ANOVA table for the ages shown in Table 12.1.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Between Groups	2	842.9	421.4	10.29
Within Groups	65	2660.8	40.9	
Total (Corrected)	67	3503.7		

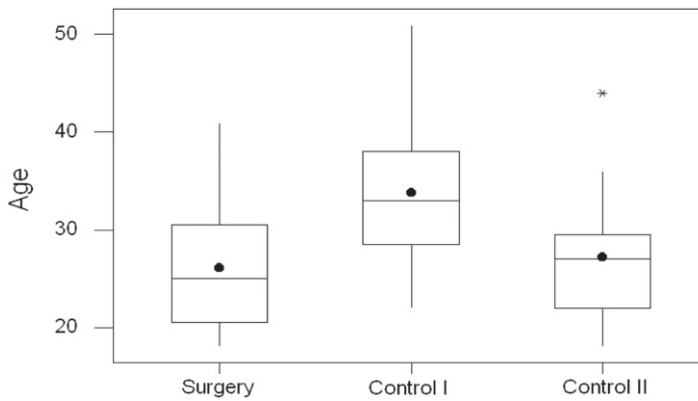


Figure 12.2 Box plots for the data in Table 12.1.

12.3 Multiple Comparisons

If the overall F statistic from the ANOVA is statistically significant, *multiple comparisons* procedures can be used in an attempt to discover the source of the significant differences among the group means. Most of these procedures are designed to examine the pairwise differences among group means, although there are more general procedures. The comparison of the group means is accomplished through the presentation of confidence intervals for pairwise differences of group means. The use of the multiple comparison procedures is generally not recommended when we fail to reject the null hypothesis. However, exceptions may occur when certain comparisons have been planned in the course of the experiment.

There are many different multiple comparison procedures, and we shall present three: the Tukey-Kramer method, Fisher's least significant difference (LSD) method, and Dunnett's method. The Tukey-Kramer method is the recommended procedure when one wishes to estimate simultaneously all pairwise differences among the means in a one-way ANOVA assuming that the variances are equal (Stoline 1981). We present the LSD method because it is frequently used in the literature. Dunnett's procedure is used when we wish to compare several groups with a specific group selected before the data were obtained (or the control group designated in the design). For example, if there were several new treatments and a standard treatment, we would use Dunnett's procedure to compare each of the new treatments with the standard. The multiple comparison procedures presented here use the mean square within as the estimator of σ^2 . Before presenting these methods, we shall discuss error rates associated with the methods.

12.3.1 Error Rates: Individual and Family

In the pairwise comparison of the group means, many confidence intervals are formed. For example, when there are three groups, we form confidence intervals for the differences of groups 1 and 2, groups 1 and 3, and groups 2 and 3. When there are r groups, there are rC_2 confidence intervals for the pairwise comparisons. Thus, we see that there are two probabilities of errors in multiple comparison procedures. One probability of

error is associated with each individual confidence interval — the individual error rate. The other is associated with the ${}_rC_2$ intervals — the family of confidence intervals — the family error rate. This is the rate that is usually of primary interest — the rate that we want to be less than or equal to α .

It is clear that if we use the $t_{1-\alpha/2}$ value in the creation of the confidence intervals, the family error rate will be larger than α . If we wish to control the family error rate to be less than or equal to α , then we must use some value other (greater) than $t_{1-\alpha/2}$ in the calculation of the confidence intervals.

12.3.2 The Tukey-Kramer Method

The Tukey-Kramer method focuses on the family error rate. It replaces $t_{\frac{n-r}{2}, 1-\alpha/2}$ in the confidence interval for the difference of two group means by $q_{r, n-r, 1-\alpha}/\sqrt{2}$, where q is the upper α value from the studentized range distribution (r is equivalent to p in Table B12). Table B12 shows the upper α value from the studentized range distribution (at $\alpha = 0.01$ and 0.05). Note that the q value takes the number of possible comparisons into account, since its value depends on r , the number of groups.

The confidence interval for the difference of μ_i and μ_j is

$$(\bar{y}_i - \bar{y}_j) \pm \frac{q_{r, n-r, 1-\alpha}}{\sqrt{2}} \sqrt{MSW} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

Example 12.4

Let us calculate the confidence intervals for the three pairwise comparisons for the hypothetical age data shown in Table 12.1. We shall set the family error rate to be 0.05. The value of $q_{3,65,0.95}$ is not found in Table B12. Since there is little variation in the value of q as $n - r$ changes from 40 to 60 to 120 in the table, we shall use 3.40 ($= q_{3,60,0.95}$) as an approximation to the desired value. The confidence interval for the difference of groups 1 and 2 is

$$(26.08 - 33.80) \pm \left(\frac{3.40}{\sqrt{2}} \sqrt{40.9} \sqrt{\frac{1}{25} + \frac{1}{25}} \right)$$

which yields

$$-7.72 \pm 4.35$$

and the interval ranges from -12.07 to -3.37 . The corresponding interval for $\mu_1 - \mu_3$ is -5.89 to 3.61 , and the interval for $\mu_2 - \mu_3$ is from 1.83 to 11.33 . Both of the intervals involving μ_2 fail to contain zero, suggesting that the first control group differs significantly from both the study group and the second control group.

12.3.3 Fisher's Least Significant Difference Method

Fisher's LSD method focuses on the individual error rate. When the n_i are all equal, there is a value — the least significant difference — such that if any of the differences in sample means are greater than that value, the difference is statistically significant. If

a difference is greater than that value, the corresponding confidence interval for the difference does not contain zero. If the number of sample observations differ across the groups, there is not a single least significant difference.

The LSD confidence interval looks like the ordinary confidence interval for the difference of two means with one exception. The mean square within is used as the estimator for the population variance instead of an estimator based on only data from the two groups being compared. The LSD confidence interval for $\mu_i - \mu_j$ is

$$(\bar{y}_i - \bar{y}_j) \pm t_{n-r, 1-\alpha/2} \sqrt{MSW} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

Example 12.5

Let us calculate the 0.05 individual error rate LSD confidence interval for $\mu_1 - \mu_2$. We have

$$(26.08 - 33.80) \pm \left(2.00 \sqrt{40.9} \sqrt{\frac{1}{25} + \frac{1}{25}} \right)$$

which yields

$$-7.72 \pm 3.62$$

and the interval ranges from -11.34 to -4.10 . This interval is narrower than the corresponding Tukey-Kramer interval as it must be, since it is based on the individual error rate, not the family error rate used by the Tukey-Kramer procedure. The corresponding LSD interval for $\mu_1 - \mu_3$ ranges from -5.10 to 2.82 , and the interval for $\mu_2 - \mu_3$ ranges from 2.62 to 10.54 .

12.3.4 Dunnett's Method

Dunnett's method is used in situations when we wish to compare the means of several groups with the mean of another group that was selected in advance. For example, we may wish to compare the means of different dosage levels of a new medication with the mean of a placebo group. In our example, there are two control groups and one treatment group. We wish to see if there is a difference between the two control groups and the treatment group (group 1). Thus, the comparisons of interest are $\mu_2 - \mu_1$ and $\mu_3 - \mu_1$.

The confidence interval for $\mu_i - \mu_j$ using Dunnett's procedure is given by

$$(\bar{y}_i - \bar{y}_j) \pm d_{r-1, n-r, 1-\alpha/2} \sqrt{MSW} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where the upper 0.005 and 0.025 levels of d are given in Table B13.

Example 12.6

Let us now calculate the confidence intervals using a family error rate of 0.05 with Dunnett's method for the data used in previous examples. For the comparison of the first control group with the treatment group, we have

$$(33.80 - 26.08) \pm \left(2.27 \sqrt{40.9} \sqrt{\frac{1}{25} + \frac{1}{25}} \right)$$

where 2.27 is the value of $d_{2,60,0.975}$, and this is used as an approximation to $d_{2,65,0.975}$. This calculation yields

$$7.72 \pm 4.11$$

and the interval ranges from 3.61 to 11.83. The corresponding interval for $\mu_3 - \mu_1$ ranges from -3.35 to 5.63 . The confidence intervals using Dunnett's procedure are narrower than those provided by the Tukey-Kramer method. This is reasonable, since we are doing fewer comparisons with Dunnett's procedure. Based on these intervals, there is a statistically significant difference between the first control group and the treatment group but no significant difference between the second control group and the treatment group.

These calculations can be performed by the computer packages in conjunction with analysis of variance (see **Program Note 12.1** on the website).

12.4 Two-Way ANOVA for the Randomized Block Design with m Replicates

As discussed in Chapter 6, in many situations the same experiment is conducted in several sites or under different conditions. In these situations, the random allocation of subjects takes place separately at each site or for each condition. These experiments are using what is called a *randomized block design*. The random allocation of the subjects to the treatments is performed separately for each block (site or condition) because it is thought that there may be an effect of the blocks on the outcome variable. If the subjects were randomly assigned ignoring the blocks, as in a completely randomized design, there is a chance that the block effects might be confounded with the treatment effects. Hence the random assignment is done separately.

Example 12.7

The data in Table 12.4 are from a randomized block design with five replicates per cell. The data are the changes in weight for moderately overweight female employees who participated in weight reduction programs. The women worked at one of two company sites: either the headquarters or a manufacturing plant. At each site, after a semiannual health examination, the women were randomly given memberships to a diet clinic or to a health club or to both. There was a control for company site because it was thought that there may be a difference in the effects of the weight

Table 12.4 Difference of pre- and postintervention weights (pounds) after 6 months of participation by intervention program at two sites.

Program	Office Site					Factory Site				
	Diet Clinic	6	2	10	-1	8	3	15	4	8
	3	4	-2	6	-2	-4	6	8	-2	3
Both Programs	8	12	7	10	5	15	8	10	16	3

reduction programs for those who were less physically active — the headquarters group — compared to the women in the plant. After the next health examination, weight reduction was measured.

In this table, data are classified by program, the row variable, and site, the column variable. The type of intervention program is the treatment variable with three levels, and the site is considered to be the blocking variable with two levels. These two independent variables form six cells, and the cells all have the same number of observations. When there are the same numbers of observations in each cell, the design is said to be balanced. The analysis of unbalanced data is more complicated and will be discussed in the last section of this chapter.

The entries in Table 12.4 can be represented symbolically as y_{ijk} , where i is an indicator of the program (the row variable), j represents the site (the column variable), and k indicates the subject number within the i th program and j th site. The first subscript ranges from 1 to 3, the second subscript has the value 1 or 2, and the third subscript ranges from 1 to 5.

We continue to use the \cdot notation. For example, $y_{\cdot 1}$ represents $\sum_i \sum_k y_{i1k}$, the sum of weight losses for the female employees at the office site. Using this notation, the sample mean of the i th level of the program variable is $\bar{y}_{i\cdot}$, the sample mean of the j th level of the site variable is $\bar{y}_{\cdot j}$, and the overall sample mean is \bar{y}_{\dots} . These are the values of these sample means:

Program Means	Site Means		Overall Mean
Diet	6.10	Office	5.07
Exercise	2.00	Factory	6.60
Both	9.40		

To analyze this data set, we will use a two-way ANOVA. The method of analysis is called two-way because there are now two independent variables: the blocking variable with c levels and the treatment variable with r levels. The total sum of squares of the dependent variable about its mean is now partitioned into a sum of squares between treatment groups, a sum of squares between blocks and the within-cells (error or residual) sum of squares. This partitioning, based on m observations per cell, is

$$\sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^m (y_{ijk} - \bar{y}_{\dots})^2 = cm \sum_{i=1}^r (\bar{y}_{i\cdot} - \bar{y}_{\dots})^2 + rm \sum_{j=1}^c (\bar{y}_{\cdot j} - \bar{y}_{\dots})^2 + SSW.$$

The total variation of Y about its mean (SST) is partitioned into the sum of squares for the row or treatment variable (SSR), the sum of squares for the column or block

Table 12.5 ANOVA table for a randomized block design.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	<i>F</i>
Treatments	$r - 1$	SSR	$SSR/(r - 1) = MSR$	MSR/MSW
Blocks	$c - 1$	SSC	$SSC/(c - 1) = MSC$	MSC/MSW
Residual	$n - r - c + 1$	SSW	$SSW/(n - r - c + 1) = MSW$	
Total (Corrected)	$n - 1$	SST		

variable (SSC), and the within or residual sum of squares (SSW). SSW is found by subtracting the sum of SSR and SSC from SST. The value of SSR is

$$SSR = 2(5)[(6.10 - 5.83)^2 + (2.00 - 5.83)^2 + (9.40 - 5.83)^2] = 274.9.$$

The value of SSC is similarly found and is

$$SSC = 3(5)[(5.07 - 5.83)^2 + (6.60 - 5.83)^2] = 17.56.$$

Too many terms are involved to show the calculation of SST, but its value is 768.2 and SSW, found by subtraction, is 475.7.

We use the same approach to the analysis in the two-way ANOVA as was used in the one-way ANOVA. To test the hypothesis of no difference in the treatments, we use the F statistic calculated as the ratio of the mean square for treatment to the residual mean square. If the null hypothesis of no difference in the treatment means, adjusted for the blocking variable, is true, this F statistic follows the F distribution. The mean square for treatments has $r - 1$ degrees of freedom, and the residual mean square has $n - r - c + 1$ [$= n - (r - 1) - (c - 1) - 1$] degrees of freedom. Thus, the F statistic for the treatment variable will follow an F distribution with $r - 1$ and $n - r - c + 1$ degrees of freedom if there is no difference in the treatment group means. In the same way, we could also test the null hypothesis of no difference in the block means. The F statistic associated with this hypothesis follows the F distribution with $c - 1$ and $n - r - c + 1$ degrees of freedom if this null hypothesis is true. Usually, we are not as interested in the hypothesis about the block means as we are in the treatment group means.

The ANOVA table for a randomized block design with m replicates per cell is shown in Table 12.5. Let us perform the test of no treatment effect — that is, of no difference in the population means associated with the three interventions at the 0.05 significance level. The analysis for the change in weight data is shown in Table 12.6. As the calculated F -value of 7.51 is greater than the critical value of 3.37 ($= F_{2,26,0.95}$), we reject the null hypothesis and conclude that the intervention programs are significantly different. Alternatively, we can make the decision based on the p -value associated with 7.51. Since the p -value of 0.003 is smaller than 0.05, we draw the same conclusions. We are not interested in the site difference.

Since there is a difference in the treatment group means at the 0.05 significance level, we are interested in finding the source of the significant differences among the group means. Applying the Tukey-Kramer method of multiple comparisons, we find that the 95 confidence intervals for $(\mu_2 - \mu_1)$ is $(-8.85, 0.65)$, $(\mu_3 - \mu_1)$ is $(-1.45, 8.05)$, and $(\mu_3 - \mu_2)$ is $(2.65, 12.15)$. It appears that using both types of intervention is more effective than the intervention using exercise only.

What would have happened had we ignored the site variable in the preceding analysis? If we assume that we would have had the same assignment of the subjects to the

Table 12.6 ANOVA table for weight change data from Table 12.4: Three intervention programs at two sites.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	<i>F</i>	<i>p</i> -value
Between Programs	2	274.9	137.4	7.51	0.003
Between Sites	1	17.6	17.6		
Residual	26	475.7	18.3		
TOTAL	29	768.2			

different treatments, we can examine the effect of the use of the blocking variable. The residual sum of squares in the two-way ANOVA is less than or equal to the residual sum of squares in the corresponding one-way ANOVA, reflecting the removal of the between blocks sum of squares. If the sum of squares between the blocks is large and its degrees of freedom are small, then the residual mean square is much smaller in the two-way ANOVA. This means that if the blocking variable is important, there is a greater chance of detecting a difference in the treatment group means using the two-way ANOVA than using the corresponding one-way ANOVA. The computer packages can be used to perform the preceding analysis including the multiple comparisons (see **Program Note 12.2** on the website).

In the next section, we show a more general two-way analysis of variance that includes the interaction of the two independent variables.

12.5 Two-Way ANOVA with Interaction

In some instances, a researcher is interested in studying the effects of two factors. In these instances, the experimental subjects are randomly allocated to all combinations of levels of both factors. For example, if both the row and column factors have two levels each, then the subjects are randomly allocated to four groups. This type of experimental design is especially useful when we want to study the effects of each factor as well as the interaction effect of the factors with one another. *Interaction* exists when the differences in responses to the levels of one factor depend on the level of another factor. For example, in a study of byssinosis (brown-lung disease) in textile workers in North Carolina (Higgins and Koch 1977), two variables of interest were whether or not the worker smoked and whether or not the worker was exposed to dust in the workplace. Both of these variables were important — that is, both smoking and being exposed were associated with a higher occurrence of byssinosis. In addition, if a worker smoked and also was exposed to the dust, the occurrence of byssinosis was much higher than would have been expected by simply adding the effects of the smoking and exposure variables. In this case, there is a synergistic effect — that is, an interaction of these two independent variables.

We have previously been concerned about interaction, although we did not use the term *interaction* when we considered the Cochran-Mantel-Haenszel procedure. We said that the procedure should not be used when the odds ratios were not consistent across the subtables. If the odds ratios are not consistent, this means that the relation between the dependent and independent variables depends on the levels of an extraneous or confounding variable — that is, there is interaction between the independent and extraneous variable. If the interaction exists, it does not make sense to talk about an overall

effect of the independent variable because its effect varies with the level of the extraneous or confounding variable.

Example 12.8

The data in Table 12.7 are from a two-factor experiment in a health education teacher-training program. Three new textbooks (factor A) were tested with two methods of instruction (factor B), and 36 trainees were randomly allocated to the six groups with six subjects per group. The trainees were tested before and after four weeks of instruction, and the increases in test scores were recorded as shown in the table. As in the randomized block design, data are classified by textbook, the row factor, and method of instruction, the column factor. In this experiment, the random allocation of subjects was done simultaneously to all combinations of the two sets of levels, whereas the randomization took place separately in each block in the randomized block design.

Table 12.7 Increase in test scores after four weeks of instruction using three textbooks and two teaching methods.

Textbook	Method of Instruction											
	Lecture						Discussion					
1	30	43	12	18	22	16	36	34	15	18	40	45
2	21	26	10	14	17	16	33	31	28	15	29	26
3	42	30	18	10	21	18	41	46	19	23	38	48

The entries in this table are also represented symbolically by y_{ijk} as in the randomized block design with replicates. Several means again will be used in the analysis. The means here include the cell means (\bar{y}_{ij}), two sets of marginal means — row ($\bar{y}_{i..}$) and column ($\bar{y}_{.j}$) — and the overall mean ($\bar{y}_{...}$). The values of these means are as follows:

Textbook	Methods of Instruction			Marginal Book Means
	Lecture	Discussion		
1	23.50	31.33		27.42
2	17.33	27.00		22.17
3	23.17	35.83		29.50
Marginal Method Means	21.33	31.39		26.36 (Overall Mean)

We analyze this data set using a two-way ANOVA with interaction. For the randomized block design, we used a two-way ANOVA, ignoring interaction. The researcher for this experiment could have used two separate completely randomized experiments (one-way ANOVAs) — one to compare the three textbooks and the other to compare the two types of instructional methods. However, based on these two separate experiments, the researcher would not know whether any textbook works better with one instructional method than the other. The effects of the textbooks may differ across the instructional methods. *Interaction* measures the difference in the textbook effects across the two

instructional methods. If the distribution of the mean increase in test scores for the three textbook types for those taught by lecture differs from the corresponding distribution for those taught by discussion, there is interaction. The average effects of textbooks across both types of instruction and the average instructional effects across all textbooks are measures of the *main effects* of the two independent variables.

If there is an interaction of the two independent variables, then usually the interaction terms are of more interest than the main effects of the two independent variables. This is because, if there is an interaction, the effect of one independent variable depends — it changes — as the level of the other independent variable changes. Hence, in our analysis, we must first examine the test of hypothesis that there is no interaction before considering the test of no main effects of the independent variables.

If there is interaction, we can examine the cell means in an attempt to discover the nature of the interaction. If there is no evidence of an interaction, then we consider the hypotheses about the main effects. In this case, some statisticians would remove the interaction term from the analysis — that is, incorporate its sum of squares and degrees of freedom into the error term before calculating the F statistics for the main effects. The decision to incorporate or not to incorporate the nonsignificant interaction term into the error term usually has little effect on the results.

In order to include interaction in the analysis, the total sum of squares (SST) of the dependent variable about its mean is now partitioned into a sum of squares for the row factor R (SSR), a sum of squares for the column factor C (SSC), a sum of squares for interaction between factor R and factor C (SSRC), and the error sum of squares (SSE). As before, we shall use the symbols r and c for the numbers of levels for factors R and C, respectively, and use m to represent the number of replicates in each of the cells formed by the crosstabulation of factors R and C. This partitioning of the total sum of squares is expressed symbolically as

$$\begin{aligned} \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^m (y_{ijk} - \bar{y}_{...})^2 &= cm \sum_{i=1}^r (\bar{y}_{i..} - \bar{y}_{...})^2 + rm \sum_{j=1}^c (\bar{y}_{.j.} - \bar{y}_{...})^2 \\ &+ m \sum_{i=1}^r \sum_{j=1}^c (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{i=1}^r \sum_{j=1}^c \sum_{k=1}^m (y_{ijk} - \bar{y}_{ij.})^2. \end{aligned}$$

The rest of the analytic approach is the same as before. The mean squares for the main effects and the interaction are calculated by dividing the sums of squares by appropriate degrees of freedom. The mean squares for factors R and C have $r - 1$ and $c - 1$ degrees of freedom, respectively. The mean square for interaction has $(r - 1)(c - 1)$ degrees of freedom, and the error mean square has $n - rc [= rc(m - 1)]$ degrees of freedom. The error mean square is then used as the denominator in the calculation of the F statistics for the two main effects and interaction. The ANOVA table for a two-factor experimental design with interaction is shown in Table 12.8.

The calculations of the sums of squares, similar to those shown previously in the randomized block analysis, are not shown here but are summarized in Table 12.9.

Let us perform the tests of hypotheses at the 0.05 significance level. The F statistic and its associated p -value for interaction indicate that there is no statistically significant interaction of the two independent variables. Since this is the case, we can now examine

Table 12.8 ANOVA table for a two-factor design with interaction.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	<i>F</i>
Factor R	$r - 1$	SSR	$SSR/(r - 1) = MSR$	MSR/MSE
Factor C	$c - 1$	SSC	$SSC/(c - 1) = MSC$	MSC/MSE
Interaction	$(r - 1)(c - 1)$	SSRC	$SSRC/(r - 1)(c - 1) = MSRC$	MSRC/MSE
Error	$n - rc$	SSE	$SSE/(n - rc) = MSE$	
Total (Corrected)	$n - 1$	SST		

Table 12.9 ANOVA table for test score increase data in Table 12.6 by combinations of three textbooks and two methods of instruction.

Source	DF	SS	MS	<i>F</i>	<i>p</i> -value
Textbooks	2	342.7	171.4	1.66	0.207
Methods of Instruction	1	910.0	910.0	8.81	0.006
Interaction	2	35.7	17.9	0.17	0.842
Error	30	3099.8	103.3		
Total	35	4388.3			

the *F* statistics associated with the hypotheses of no difference in the test score improvement between the two methods of instruction and among the three textbooks. There is a statistically significant effect for the methods of instruction — a *p*-value less than 0.05 — but no significant effect associated with the textbooks.

If we had removed the interaction term from the analysis after finding that it was not important, the error sum of squares would have been 3135.5 ($= 35.7 + 3099.8$), and there would have been 32 degrees of freedom associated with this error sum of squares. The error mean square would have been 97.98 instead of 103.3, and the *F* ratios for textbooks and methods of instruction would have been 1.75 and 9.29, respectively.

Let us explore further the preceding analytical results in relation to the cell means that were just calculated and are repeated here for our convenience. The lack of a significant main effect for textbooks is reflected in the marginal means for textbooks. The first and third textbooks appear to be a little more effective than the second book, but the ANOVA results indicated that these differences are not statistically significant. On the other hand, the discussion method was associated with a much greater increase — about 10 points — in test scores than the lecture method and this difference was statistically significant. The lack of an interaction effect is reflected in the cell means that are plotted in Figure 12.3.

Textbook	Methods of Instruction		Marginal Book Means
	Lecture	Discussion	
1	23.50	31.33	27.42
2	17.33	27.00	22.17
3	23.17	35.83	29.50
Marginal Method Means	21.33	31.39	26.36 (Overall Mean)

Interaction measures the degree of similarity between the responses to factor A at different levels of factor B. The lines connecting the three cell means for the discussion method are roughly parallel with the lines connecting cell means for the lecture method,

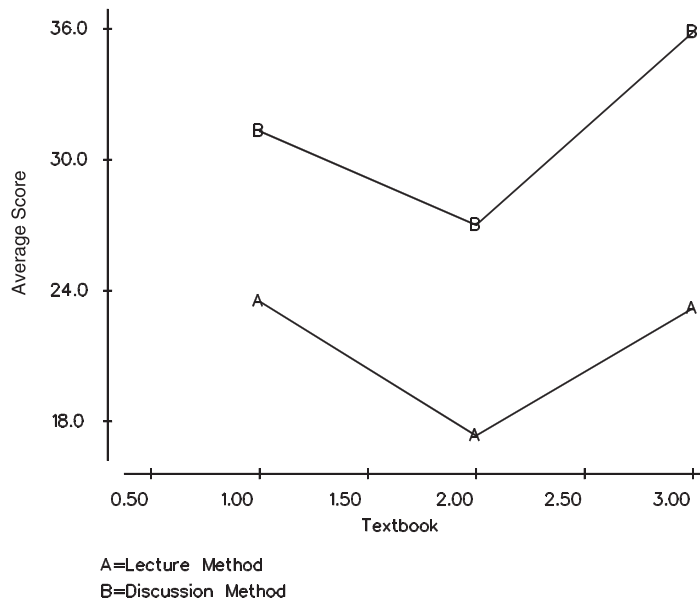


Figure 12.3 Plot of mean scores by methods of instruction on three textbooks (A = lecture method; B = discussion method).

reflecting the absence of interaction. If these two lines were not parallel or crossed each other, then the interaction effect would have been statistically significant. If a significant interaction is present, we need to examine the cell means carefully to draw appropriate conclusions. Computer packages can be used to conduct the preceding analysis (see **Program Note 12.2** on the website).

12.6 Linear Model Representation of the ANOVA

As shown in the last two sections, a two-way ANOVA can be used with or without interaction, which suggests that we need to specify the model to be used in the analysis. The choice of a model is dependent on how the data are collected and how we consider each effect to be specified. We consider this modeling aspect of ANOVA in this section.

In the ANOVA, we have partitioned the sum of squares of Y about its mean into within and between components in the completely randomized design or into treatment, blocks, and within components in the randomized block design. Underlying these partitions are linear models that show the relation between the dependent variable and the independent — treatment and/or blocking — variables. In the following sections, we show these models as well as the model with interaction. From these models, we can also see that it is possible to extend the ANOVA method of analysis to include combinations of the independent variables as well as including more than two independent variables.

12.6.1 The Completely Randomized Design

One representation of the linear model underlying the completely randomized design shows the dependent variable being equal to a constant plus a treatment effect plus individual variation — that is,

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

for i ranging from 1 to r and j going from 1 to n_i . The value of the j th observation of the dependent variable at the i th treatment level is y_{ij} . There are r levels of the treatment variable and n_i observations of Y at the i th treatment level. The constant is represented by μ , and the effect of the i th treatment level is represented by α_i . Since not everyone who has received the i th level of treatment will have the same value of the dependent variable, this individual variation, the departure from the sum of μ plus α_i , is represented by ε_{ij} .

Note that this model can be rewritten as follows:

$$y_{ij} = \mu + x_{ij} \alpha_i + \varepsilon_{ij}$$

where x_{ij} is an indicator variable which has the value of 1 if the ij th subject has received the i th level of the treatment and 0 otherwise. The X variable here simply indicates which level of treatment the person has received. We do not use this representation of the model here, but we shall refer to it in the next chapter.

In this linear model, there are $r + 1$ population parameters — the constant μ and the $r\alpha$'s; however, there are only r different treatment levels or groups. Since we can only estimate the same number of parameters as there are groups, to obtain estimators for r of the parameters, we must make some assumption about them. The appendix on the linear model in Forthofer and Lehnen (1981) provides a presentation of a number of assumptions that we could make. In this book, we shall measure the effect of the treatment levels from the effect of the r th treatment level. This means that α_r is assumed to be zero.

Now let us rewrite the linear model in terms of the population means. The equation for the i th level becomes

$$\mu_i = \mu + \alpha_i$$

and the representation of the model for all r levels is

$$\mu_1 = \mu + \alpha_1$$

$$\mu_2 = \mu + \alpha_2$$

...

$$\mu_{r-1} = \mu + \alpha_{r-1}$$

$$\mu_r = \mu.$$

From these equations, we can see that the constant term is the mean of the r th level, and the effect of the other levels — $\alpha_1, \alpha_2, \dots, \alpha_{r-1}$ — are measured from μ_r (or μ). For example, using the first of these equations to solve for α_1 , we have

$$\alpha_1 = \mu_1 - \mu = \mu_1 - \mu_r.$$

This equation makes it clear that we are measuring the effects of the first level relative to the effect of the r th level, and the same is true for levels 2 through $r - 1$.

The sample estimator of the i th effect, $\hat{\alpha}_i$, is obtained by substituting the sample means for the population means — that is,

$$\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{r..}$$

and the estimator of μ is simply $\bar{y}_{r..}$.

The t test for comparing the means of two populations, assuming equal variances, also fits into the ANOVA framework. In this case, r is 2, and the preceding model still applies.

12.6.2 The Randomized Block Design with m Replicates

A linear model underlying the randomized block design has the dependent variable being equal to a constant plus the effect of the i th level of the treatment variable plus the j th block effect plus the individual variation term. In symbols, this is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where i goes from 1 to r , j ranges from 1 to c , and k ranges from 1 to m .

Just as in the completely randomized situation, the effects of the levels of the treatment variable are measured relative to the r th level of the treatment variable. In the same way, the effects of the levels of the blocking variable are measured relative to the c th level of the blocking variable. The definition of the parameters in terms of the μ_{ij} is complicated and will not be shown for this model, but it will be shown for the model in the next section.

12.6.3 Two-Way ANOVA with Interaction

The model for this situation is similar to the preceding two-way ANOVA model, except that it includes the interaction term, denoted by β_{ij} , in the model. The model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$

where i goes from 1 to r , j ranges from 1 to c , and k ranges from 1 to m .

The main effect terms in the model, the α_i and the β_j , again are all measured relative to their last level. The representation of this model in terms of the cell means, the μ_{ij} , for the first row is

$$\mu_{11} = \mu + \alpha_1 + \beta_1 + \alpha\beta_{11}$$

$$\mu_{12} = \mu + \alpha_1 + \beta_2 + \alpha\beta_{12}$$

...

$$\mu_{1c} = \mu + \alpha_1.$$

Note that there is no β_c term or any $\alpha\beta_{1c}$ terms in the final equation. Since the c th level is the reference level for the column variable, β_c is taken to be zero. In addition, interaction terms having either an r or a c as a subscript are reference levels, and these interaction terms are also assumed to be zero. This pattern is repeated for the other rows except the last one.

$$\mu_{21} = \mu + \alpha_2 + \beta_1 + \alpha\beta_{21}$$

$$\mu_{22} = \mu + \alpha_2 + \beta_2 + \alpha\beta_{22}$$

$$\begin{aligned}
 & \dots \\
 \mu_{2c} &= \mu + \alpha_2 \\
 & \dots \\
 \mu_{r1} &= \mu + \beta_1 \\
 \mu_{r2} &= \mu + \beta_2 \\
 & \dots \\
 \mu_{rc} &= \mu.
 \end{aligned}$$

For the cells in the r th row, there is no α_r effect shown, since the r th level is the reference level for the row variable and α_r is taken to be zero. There are also no $\alpha\beta_{ij}$ terms in the last row, since the r th level is also a reference level for the interaction terms.

Using these equations, we obtain the following definitions of the parameters (μ , α , β , and $\alpha\beta$) in terms of the cell means. For example, from the last equation, we see that the constant term in the model is simply the mean of the cell formed by the r th row and c th column — that is, $\mu = \mu_{rc}$.

Once we have expressed μ in terms of the cell means, we can find the estimate of α_i from the equation $\mu_{ic} = \mu + \alpha_i$. This gives the solution that $\alpha_i = \mu_{ic} - \mu_{rc}$, where we have replaced μ_{rc} for μ . This definition for α_i is reasonable, as it compares the mean of the cell in the i th row and c th column with the mean of the cell in the r th row and c th column. It is comparing a cell in the i th row with its reference cell in the r th row. The column effect, β_j , is similarly defined as $\beta_j = \mu_{rj} - \mu_{rc}$.

The definition of the interaction term is $\alpha\beta_{ij} = (\mu_{ij} - \mu_{ic}) - (\mu_{rj} - \mu_{rc})$. The rc th cell is the reference cell and the other parameters are defined in terms of it. The ij th interaction parameter focuses on the difference of the j th and c th columns, and compares that difference for the i th and r th rows. If there is no interaction, the difference of the j th and c th columns is the same over all the rows.

12.7 ANOVA with Unequal Numbers of Observations in Subgroups

In the preceding discussion, we allowed the number of observations in each treatment to vary for the one-way ANOVA model, but we assumed an equal number of observations in each cell for the two-way ANOVA models. However, it is not always possible to have an equal number of observations on all treatment combinations. Even balanced designs often become unbalanced because people may drop out of the study or some of the data are missing. This imbalance in the size of subclasses introduces complications in the analysis. The main difficulty is that there is no unique way of finding the sums-of-squares corresponding to each main effect and each interaction.

One method of calculating the sums of squares is to consider the factors and interaction(s) sequentially. The effect of the first factor entered into the model is calculated unconditionally, but the second factor is evaluated conditional upon whatever factor was entered first. Using the notation used for conditional probability, the effect of the second factor (s) conditional on the first factor (f) can be written as $s|f$. The par-

tioned sums of squares obtained by sequential fitting are labeled as Type I SS in the computer output. The sequential sums of squares will add up to the total sum of squares.

One problem with the sequential approach is that the factors are treated differently depending on the order of entry, and effects of factors are difficult to interpret. An alternative approach is to consider the effect of each factor adjusted for all the other factors in the model. This approach produces adjusted sums of squares (called Type III SS) that do not add up to the total SS unless the data are balanced. The adjusted sums of squares are generally used when testing the effect of each factor (Maxwell and Delaney 1990). We will use this approach in examining an example following.

Computer packages provide Type III and Type I sums of squares by default. There are two other types of sums of squares, but they are generally of lesser interest. In cases where the interaction terms are unimportant and a main effect needs to be examined at each level of the other factors, Type I sums of squares are recommended (Nelder 1977).

Example 12.9

Let us consider the data shown in Table 12.10. This table represents a cross-classification of creatinine measurements by sex and age (categorized into two groups: under 56 and 56 & over) for the 40 patients in DIG40 that was introduced in Chapter 3. The numbers of observations in the four cells are not equal, and, therefore, we will use the general linear models procedure for these unbalanced data.

Table 12.10 Creatinine measurements (mg/dL) by sex and age group, DIG40.

Cell	Sex	Age	Creatinine Measurements (mg/dL)									
1,1	Male	<56	1.600	1.300	1.159	1.307	1.886	1.034	0.900	1.398	1.307	
1,2	Male	≥56	2.682	1.091	1.250	1.705	2.000	1.227	1.100	2.239	1.300	1.614
			1.200	1.455	1.489	1.700	1.307	1.200	1.273	1.300	1.659	1.261
			0.900									
2,1	Female	<56	1.386	0.900	1.000	1.148	1.170					
2,2	Female	≥56	1.534	0.900	0.900	1.352	0.909					

Before analyzing the data by ANOVA, let us look at the cell means. These are shown here in a 2 by 2 table:

Sex	Ages		Difference
	<56	≥56	
Male	1.3212	1.4739	-0.1527
Female	1.1236	1.1113	0.0123
Difference	0.1976	0.3626	

The effect of the sex variable (difference between male and female) can be calculated at each level of the age variable. Likewise, the effect of the age variable can be seen in the difference between the two age groups at each level of the sex variable. The

Table 12.11 ANOVA by the general linear models procedure for Table 12.10 (with interaction).

Source	DF	Adjusted (Type III)				Sequential SS (Type I)
		SS	MS	F	P	
Sex	1	0.5519	0.5519	4.21	0.048	0.7124
Age	1	0.0407	0.0407	0.31	0.581	0.1042
Sex*Age	1	0.0427	0.0427	0.33	0.572	0.0427
Error	36	4.7247	0.1312			4.7247
Total	39	5.5839				5.5839

Table 12.12 ANOVA by general linear models procedure for Table 12.10 (without interaction).

Source	DF	Adjusted (Type III)				Sequential SS (Type I)
		SS	MS	F	P	
Sex	1	0.5951	0.5951	4.62	0.038	0.7124
Age	1	0.1042	0.1042	0.81	0.374	0.1042
Error	37	4.7674	0.1288			4.7674
Total	39	5.5839				5.5839

effect of the sex variable is considerably larger than the age effect. The effects of the sex variable at two levels of age are in the same direction, although the effect of the sex variable is larger for the older age group, suggesting that the interaction of the sex and age variables on creatinine is likely to be nonsignificant.

Table 12.11 shows the results of ANOVA by the general linear models procedure. The model includes two main effects and the interaction of sex and age. Adjusted sums of squares are shown, and the analysis of variance is carried out based on Type III sums of squares to assess the effect of each term. Sequential sums of squares are shown in the last column. Note that sequential sum of squares of the last term listed in the model (the interaction term in this case) is the same as the Type III sum of squares for the factor because it is adjusted to all other factors in the model. The effect of the sex variable is significant, while the effect of the age variable is not.

Since the interaction effect is unimportant, we dropped the interaction term from the model and repeated the analysis. The results are shown in Table 12.12. Note that the sequential sum of squares for age is the same as the adjusted sum of squares for the factor because it is the last term in the model. Note also that the sum of squares due to the interaction is now included in the error term. Again, the sex variable effect is significant, while the age effect is not.

It is usually necessary to use a computer package for performing a general linear models analysis (see **Program Note 12.3** on the website).

Choosing an appropriate model for ANOVA is not straightforward, especially for an unbalanced data. As we saw in Chapter 6, it is important to strive for a balanced design to alleviate the complications in the analysis.

Conclusion

In this chapter we presented several basic models of analysis of variance. The one-way ANOVA is used to analyze data from a completely randomized experimental design. The two-way ANOVA can be used for a randomized block design as well as for a two-factor design with interaction. To use these analytical methods properly, we must be aware of how the data were collected and make sure that the data meet the ANOVA assumptions. Finally, we discussed the problems and methods for analyzing unbalanced data. In the next chapter, we will expand the linear model to regression models.

EXERCISES

- 12.1** The data shown here, taken from Brogan and Kutner (1980), are the change in the maximal rate of urea synthesis (MRUS) level for cirrhotic patients who underwent either a standard operation (a nonselective shunt) or a new procedure (a selective shunt). The purpose of the operations was to improve liver function, measured by MRUS. A low value of MRUS is associated with poor liver function. Patients in the nonselective shunt group are divided into two groups based on the preoperative MRUS values (≤ 40 and > 40).

Change in Maximal Rate of Urea Synthesis (MRUS) Level (mg Urea N/hr/kg Body Weight) by Group								
Group	Change in MRUS Values							
Selective Shunt	-3	20	-6	-5	-3	-3	-6	12
Nonselective Shunt I	-18	-4	-18	-18	-6	-18		
Nonselective Shunt II	-24	-7	-15	4	-14	-8	-11	

Perform an analysis of variance of these data at the 0.05 significance level to determine if there is a difference in the three groups. If there is a significant difference, use an appropriate multiple comparison procedure to find the source of the difference.

- 12.2** In Chapter 8, we used the t test to compare the proportion of caloric intake from fat for fifth- and sixth-grade boys compared to seventh- and eighth-grade boys. The calculated t test statistic was -0.727 (Example 8.9). Perform a one-way ANOVA on these data in Table 7.7 and compare your results with the t test approach. How does the t statistic compare with the F statistic?
- 12.3** For the weight change data shown in Table 12.4, we were concerned about the level of physical activity of the women. Instead of using the site — headquarters or plant — as a way of controlling for physical activity, how else might we have controlled for the physical activity? Do you think that a control group — no intervention — should have been used? Explain your reasoning. Would you do anything to determine whether or not the women used the memberships? What, if any, other variables should be included in the analysis?
- 12.4** To investigate publication bias, 75 referees for one journal were randomly assigned to receive one of five versions of a manuscript (Dickersin 1990). All versions were identical in the Introduction and Methods sections but varied in either the Results or Discussion sections. The first and second groups received versions with either positive or negative results, respectively. The third and

fourth groups received versions with mixed results and either positive or negative discussion. The fifth group was asked evaluate the manuscript on the basis of the Methods section, and no data were provided. The referees used a scale of 0 to 6 (low to high) to rate different aspects of the manuscript. The average scores for three aspects are shown here:

Manuscript Version	No. of Referees	Mean Ratings		
		Methods	Scientific Contribution	Publication Merit
Positive Results	12	4.2	4.3	3.2
Negative Results	14	2.4	2.4	1.8
Mixed Results with Positive Discussion	13	2.5	1.6	0.5
Mixed Results with Negative Discussion	14	2.7	1.7	1.4
Methods Only	14	3.4	4.5	3.4

State an appropriate linear model for this experiment using scientific contribution as the dependent variable. What are the null and alternative hypotheses of interest for this model? Assuming that the standard deviations for the scientific contribution score for the five groups are 1.1, 0.9, 0.7, 0.8, and 1.1, respectively, perform an analysis of variance of these data at the 0.05 significance level to determine if there is a bias in refereeing scientific papers for this journal. If there is a significant difference, use an appropriate multiple comparison procedure to find the source of the bias. State your conclusions clearly.

- 12.5** In an investigation of the effect of smoking on work performance under different lighting conditions in a large company, a random sample of nine male workers was selected from each of the three smoking status groups: nonsmokers, moderate smokers, and heavy smokers. Each sample was randomly assigned to three working environments with different levels of lighting. The time to complete a standard assembling task was recorded in minutes. The sums of squares were as follows:

Source	df	SS	MS	F	p-value
Smoking Status		84.90			
Lighting Conditions		298.07			
Interaction		2.81			
Error		59.25			
Total		445.03			

Perform an analysis of variance for these data to examine the interaction of the variables at the 0.05 significance level. If there is no significant interaction, test whether or not the smoking and lighting conditions variables have significant effects on the workers' performance and state your conclusions.

- 12.6** The midterm and final test scores of 12 students in a class are recorded as follows:

Test	Students											
	1	2	3	4	5	6	7	8	9	10	11	12
Midterm	80	85	65	77	58	98	91	72	62	82	45	42
Final	78	90	72	80	71	92	93	70	73	85	60	61

These are paired data. First, perform the paired t test (discussed in Chapter 8) to compare the two sets of test scores. Now perform a two-way ANOVA for the randomized block design with two replicates (an additive model), considering students as blocks, and compare your results with the paired t test approach. How does the t statistic compare with the F statistic for the test variable? Can you draw the same conclusion?

- 12.7** A randomized study was conducted to compare the effects of two intervention procedures (tracking with outreach and provider prompting) to raise immunization in primary care clinics serving impoverished children (Rodewald et al. 1999). The study used a 2 by 2 factorial design, and each intervention had two levels (1 = no intervention; 2 = intervention). After 18 months of intervention the immunization status was assessed. Two major outcome measures were the percentage of immunization and the number of days of delay in immunization. The authors claim that the two-way ANOVA was used to test for effects of each of the interventions on the outcome measures. They stated that the interaction was insignificant. The number of children allocated to each group was slightly different and the number of children who completed the assessment of the outcomes also varied as shown here:

Prompting	Treatment Groups		Number Allocated	Number Completed	Percent Immunized	Mean Days of Delay ^a
		Outreach (Group)				
1		1 (no intervention)	769	719	74.0	140.0
1		2 (outreach only)	715	630	95.1	76.5
2		1 (prompting only)	801	744	75.9	133.2
2		2 (both interventions)	732	648	95.1	69.1

^aMean days of delay in immunization

Discuss whether the two-way ANOVA was appropriate for the two major outcome measures shown. If you think the ANOVA is inappropriate for any of the outcome measures, what statistical method would you recommend? If you think the ANOVA is appropriate for any of outcome measure, would you accept the claim of no interaction based on these data? A considerable number of subjects were lost during the course of the study, and the number of dropouts varies between the four groups. Discuss how the differential loss of the subjects might impact the study outcome.

REFERENCES

- Brogan, D. R., and M. H. Kutner. "Comparative Analyses of Pretest-Posttest Research Designs." *The American Statistician* 34:229–232, 1980.
- Dickersin, K. "The Existence of Publication Bias and Risk Factors for Its Occurrence." *Journal of American Medical Association* 263:1385–1389, 1990.

- Forthofer, R. N., and R. G. Lehnen. *Public Program Analysis: A New Categorical Data Approach*. Belmont, CA: Lifetime Learning Publications, 1981.
- Higgins, J. E., and G. G. Koch. "Variable Selection and Generalized Chi-Square Analysis of Categorical Data Applied to a Large Cross-Sectional Occupational Health Survey." *International Statistical Review* 45:51–62, 1977.
- Kimball, B. P., B. L. Shurvell, S. Houle, J. C. Fulop, H. Rakowski, and P. R. McLaughlin. "Persistent Ventricular Adaptations in Postoperative Coarctation of the Aorta." *Journal of the American College of Cardiology* 8:172–178, 1986.
- Kleinbaum, D. G., L. L. Kupper, and K. E. Muller. *Applied Regression Analysis and Other Multivariable Methods*, 2nd ed. Boston: PWS-Kent, 1988, Chapter 12.
- Lin, L. I., and E. F. Vonesh. "An Empirical Nonlinear Data-Fitting Approach for Transforming Data to Normality." *The American Statistician* 43:237–243, 1989.
- Maxwell, S. E., and H. D. Delaney. *Designing Experiments and Analyzing Data*. Belmont, CA: Wadsworth, 1990.
- Nelder, J. A. "A Reformulation of Linear Models." *Journal of the Royal Statistical Society* 140:48–63, 1977.
- Rodewald, L. E., P. G. Szilagyi, S. G. Humiston, R. Barth, R. Kraus, and R. F. Raubertas. "A Randomized Study of Tracking with Outreach and Provider Prompting to Improve Immunization Coverage and Primary Care." *Pediatrics* 103 (1):31–38, January 1999.
- Scheffé, H. *The Analysis of Variance*. New York: John Wiley & Sons, 1959, Chapter 10.
- Steel, R. G. D., and J. H. Torrie. *Principles and Procedures of Statistics: A Biometrical Approach*, New York: McGraw-Hill, 1980, Chapter 9.
- Stoline, M. R. "The Status of Multiple Comparisons: Simultaneous Estimation of All Pairwise Comparisons in One-Way ANOVA Designs." *The American Statistician* 35:134–141, 1981.