Chapter

11 Comparing the Means of Many Independent Samples

Objectives

In this chapter we study analysis of variance (ANOVA). We will

- discuss when and why an analysis of variance may be conducted.
- develop the intuition behind the ANOVA model.
- demonstrate how ANOVA calculations are carried out.
- describe and examine the conditions under which ANOVA is valid.
- see how blocking is used and how to conduct randomized blocks ANOVA.
- describe interactions and main effects in factorial ANOVA models.
- construct contrasts and other linear combinations of means.
- introduce and compare several methods for dealing with multiple comparisons.

11.1 Introduction

In Chapter 7 we considered the comparison of two independent samples with respect to a quantitative variable Y. The classical techniques for comparing the two sample means $\overline{Y}_1$ and $\overline{Y}_2$ are the test and the confidence interval based on Student's t distribution. In the present chapter we consider the comparison of the means of I independent samples, where I may be greater than 2. The following example illustrates an experiment with $I = 5$.

Example 11.1.1

Sweet Corn When growing sweet corn, can organic methods be used successfully to control harmful insects and limit their effect on the corn? In a study of this question researchers compared the weights of ears of corn under five conditions in an experiment in which sweet corn was grown using organic methods. In one plot of corn a beneficial soil nematode was introduced. In a second plot a parasitic wasp was used. A third plot was treated with both the nematode and the wasp. In a fourth plot a bacterium was used. Finally, a fifth plot of corn acted as a control; no special treatment was applied here. Thus, the treatments were

Treatment 1: Nematodes
Treatment 2: Wasps
Treatment 3: Nematodes and wasps
Treatment 4: Bacteria
Treatment 5: Control

Ears of corn were randomly sampled from each plot and weighed. The results are given in Table 11.1.1 and plotted in Figure 11.1.1.[1] Note that in addition to the differences between the treatment means, there is also considerable variation within each treatment group. ■

We will discuss the classical method of analyzing data from I independent samples. The method is called an **analysis of variance**, or **ANOVA**. In applying analysis of variance, the data are regarded as random samples from I populations. We will denote the means of these populations as $\mu_1, \mu_2, \ldots, \mu_I$ and the standard deviations as $\sigma_1, \sigma_2, \ldots, \sigma_I$.

Table 11.1.1 Weights (ounces) of ears of sweet corn

	Treatment				
	1	2	3	4	5
	16.5	11.0	8.5	16.0	13.0
	15.0	15.0	13.0	14.5	10.5
	11.5	9.0	12.0	15.0	11.0
	12.0	9.0	10.0	9.0	10.0
	12.5	11.5	12.5	10.5	14.0
	9.0	11.0	8.5	14.0	12.0
	16.0	9.0	9.5	12.5	11.0
	6.5	10.0	7.0	9.0	9.5
	8.0	9.0	10.5	9.0	18.5
	14.5	8.0	10.5	9.0	17.0
	7.0	8.0	13.0	6.5	10.0
	10.5	5.0	9.0	8.5	11.0
Mean	11.5	9.6	10.3	11.1	12.3
SD	3.5	2.4	2.0	3.1	2.9
n	12	12	12	12	12

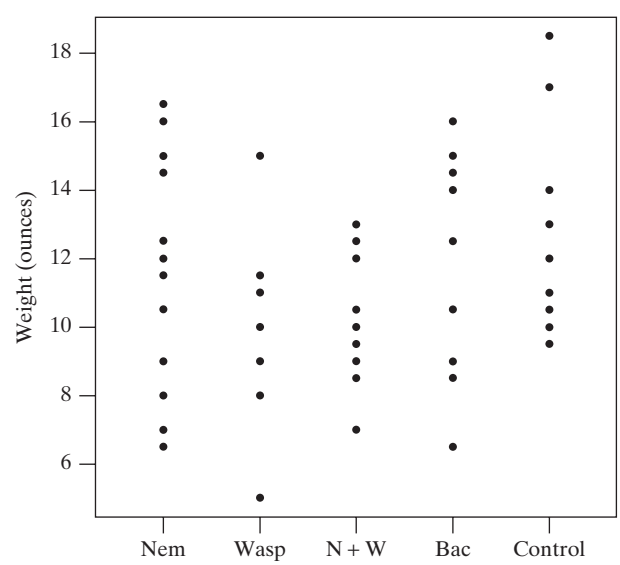


Figure 11.1.1 Weights of ears of corn receiving five different treatments

Why Not Repeated *t* Tests?

It is natural to wonder why the comparison of the means of I samples requires any new methods. For instance, why not just use a two-sample t test on each pair of samples? There are three reasons why this is not a good idea.

1. *The problem of multiple comparisons* The most serious difficulty with a naive "repeated t tests" procedure concerns Type I error: The probability of false rejection of a null hypothesis may be much higher than it appears to be. For instance, suppose $I = 4$ and consider the null hypothesis that all four population means are equal ($H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$) versus the alternative hypothesis that the four means are not all equal.* Among four means there are six possible pairs to compare. The pairings are displayed in Figure 11.1.2. The six resulting hypotheses are

$$H_0: \mu_1 = \mu_2 \quad H_0: \mu_1 = \mu_3 \quad H_0: \mu_1 = \mu_4$$
$$H_0: \mu_2 = \mu_3 \quad H_0: \mu_2 = \mu_4 \quad H_0: \mu_3 = \mu_4$$

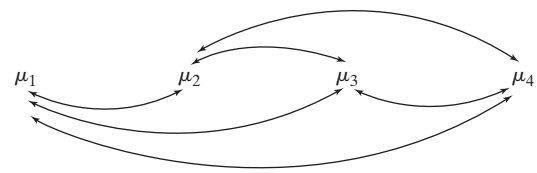


Figure 11.1.2 Comparing four population means requires six comparisons

Let's consider the risk of a Type I error for testing our primary null hypothesis that all four means are equal by conducting six separate t tests. If *any* of the six t tests finds a significant difference between a pair of means, we would reject our primary null hypothesis that all four means are equal. A Type I error would occur if *any* of the six t tests found a significant difference between a pair of means when in fact all four means are equal. Thus, using $\alpha = 0.05$ for each of the individual t tests carries an overall risk of a Type I error that is greater than 5%.

Our intuition might suggest that the risk of an overall Type I error in the preceding example should be $6 \times 0.05 = 0.3 = 30\%$ (in each of six tests we had a 5% chance of wrongly finding evidence for a difference), but this is not the case. The computation of this overall Type I error rate is more complex. Table 11.1.2 displays the overall risk of Type I error,† that is,

Overall Type I error risk = Probability that at least one of the t tests will reject its null hypothesis, when in fact $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_I$.

Table 11.1.2 Overall risk of Type I error in using repeated t tests at $\alpha = 0.05$

I	Overall risk
2	0.05
3	0.12
4	0.20
6	0.37
8	0.51
10	0.63

*In Section 11.2 we will elaborate more on the form of this alternative hypothesis.
†Table 11.1.2 was computed assuming that the sample sizes are large and equal and that the population distributions are normal with equal standard deviations.

If $I = 2$, then the overall risk is 0.05, as it should be, but with larger I the risk increases rapidly; for $I = 6$ it is 0.37. It is clear from Table 11.1.2 that the researcher who uses repeated t tests is highly vulnerable to Type I error unless I is quite small.

The difficulties illustrated by Table 11.1.2 are due to **multiple comparisons**—that is, many comparisons on the same set of data. These difficulties can be reduced when the comparison of several groups is approached through ANOVA.

2. *Estimation of the standard deviation.* The ANOVA technique combines information on variability from all the samples simultaneously. This global sharing of information can yield improved precision in the analysis.
3. *Structure in the groups.* In many studies the logical structure of the treatments or groups to be compared may inspire questions that cannot be answered by simple pairwise comparisons. For example, we may wish to study the effects of two experimental factors simultaneously. ANOVA can be used to analyze data in such settings (see Sections 11.6, 11.7, and 11.8).

A Graphical Perspective on ANOVA

When data are analyzed by analysis of variance, the usual first step is to test the following global null hypothesis:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_I$$

which asserts that all the population means are equal. A statistical test of H_0 will be described in Section 11.4. However, we will first consider analysis of variance from a graphical perspective.

Consider the dotplots shown in Figure 11.1.3(a). These dotplots were generated in a setting in which H_0 is true. The sample means, which are shown as lines on the graph, differ from one another only as a result of chance error. For the data shown in Figure 11.1.3(b), H_0 is false. The sample means are quite different—there is substantial variability between the group means, which provides evidence that the corresponding population means (μ_1, μ_2, μ_3, and μ_4) are not all equal. In this particular case, it appears that μ_1 and μ_2 differ from μ_3 and μ_4.

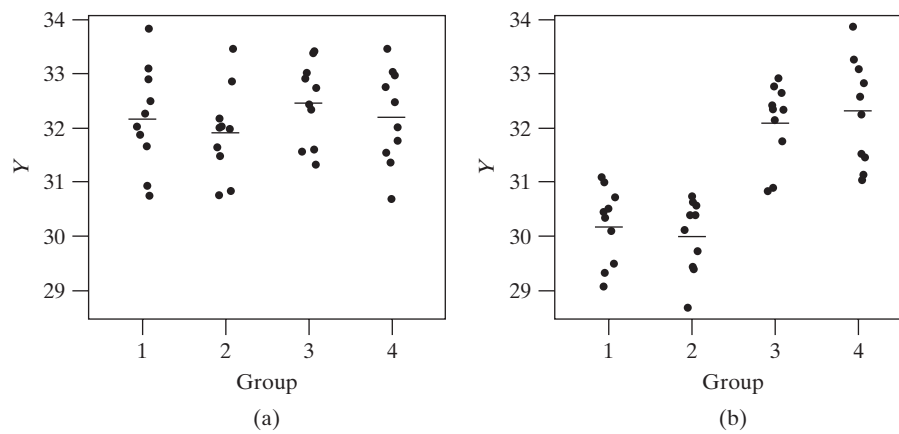


Figure 11.1.3 (a) H_0 true, (b) H_0 false, with small SDs for the groups

Figure 11.1.4 shows a situation that is less clear. In fact, H_0 is false here—the means in Figure 11.1.4 are identical to those in Figure 11.1.3(b). However, the individual group standard deviations are quite large, which makes it hard to tell that the population means differ.*

*Note the change in scale on the vertical axis in Figure 11.1.4.

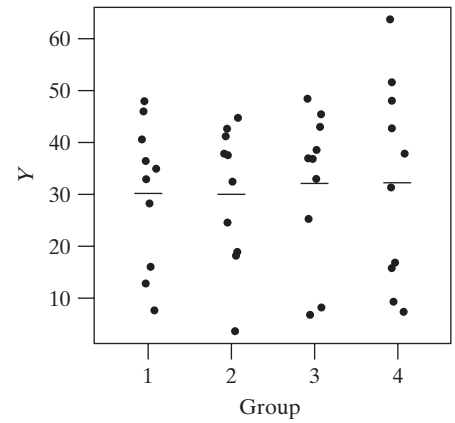


Figure 11.1.4 H_0 false, with large SDs for the groups

We need to know how much inherent variability there is in the data before we can judge whether a difference in sample means is fairly small and attributable to chance or whether it is too large to be due to chance alone. As Figures 11.1.3 and 11.1.4 illustrate, in order to find compelling evidence for a difference in population means, not only must there be (1) variation among the group means, but it must be large *relative* to (2) the inherent variability in the groups. It is through comparing the relative magnitudes of these two kinds of *variability*—this "analysis of variance"—that we are able to make an inference about *means*.

A Look Ahead

If the global null hypothesis that $\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_I$ is rejected, then the data provide sufficient evidence to conclude that at least *some* of the μ's are unequal; the researcher would usually proceed to detailed comparisons to determine the *pattern* of differences among the μ's. If there is a lack of evidence against the global null hypothesis, then the researcher might choose to construct one or more confidence intervals to characterize the lack of significant differences among the μ's.

All the statistical procedures of this chapter—the test of the global null hypothesis and various methods of making detailed comparisons among the means—depend on the same basic calculations. These calculations are presented in Section 11.2.

11.2 The Basic One-Way Analysis of Variance

The ANOVA model presented in Section 11.1 that compares the means of three or more groups is called a **one-way ANOVA**. The term "one-way" refers to the fact that there is one variable that defines the groups or treatments (e.g., in the sweet corn example the treatments were based on the type of harmful insect/bacteria). Later in this chapter we will examine other ANOVA models such as the randomized complete block ANOVA (Section 11.6) and the two-way ANOVA model (Section 11.7), which consider the impact of having more than one variable defining the groups or how treatments are assigned to experimental units.

In this section we present the basic one-way ANOVA calculations that are used to describe the data and to facilitate further analysis. In the previous section we noted that if the between-group mean variability is large relative to within-group

variability, we will take this as evidence against the null hypothesis that the population means are all equal. Hence, the analysis of variance of I samples, or groups, begins with the calculation of quantities that describe the variability of the data *between* the groups and *within* the groups.* (For clarity, in this chapter we will often refer to the samples as "groups" of observations.)

Notation

To describe several groups of quantitative observations, we will use two subscripts: one to keep track of group membership and the other to keep track of observations within the groups. Thus, we will denote observation j in group i as

$$y_{ij} = \text{observation } j \text{ in group } i$$

Thus, the first observation in the first group is y_{11}, the second observation in the first group is y_{12}, the third observation in the second group is y_{23}, and so on.

We will also use the following notation:

$$\begin{aligned} I &= \text{number of groups} \\ n_i &= \text{number of observations in group } i \\ \bar{y}_i &= \text{mean for group } i \\ s_i &= \text{standard deviation for group } i \end{aligned}$$

The total number of observations is

$$n_\bullet = \sum_{i=1}^{I} n_i$$

Finally, the **grand mean**—the mean of all the observations—is

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}}{n_\bullet}$$

Equivalently we can express $\bar{\bar{y}}$ as a weighted average of the group means

$$\bar{\bar{y}} = \frac{\sum_{i=1}^{I} n_i \bar{y}_i}{\sum_{i}^{I} n_i} = \frac{\sum_{i=1}^{I} n_i \bar{y}_i}{n_\bullet}$$

The following example illustrates this notation.

Example 11.2.1

Weight Gain of Lambs Table 11.2.1 shows the weight gains (in two weeks) of young lambs on three different diets. (These data are fictitious, but are realistic in all respects except for the fact that the group means are whole numbers.)[2]

The total number of observations is

$$n_\bullet = 3 + 5 + 4 = 12$$

*Grammatically speaking, the word *among* should be used rather than between when referring to three or more groups; however, we will use "between" because it more clearly suggests that the groups are being compared against each other.

Table 11.2.1 Weight gains of lambs (lb)*

	Diet 1	Diet 2	Diet 3
	8	9	15
	16	16	10
	9	21	17
		11	6
		18	
n_i	3	5	4
Sum $= \sum_{j=1}^{n_i} y_{ij}$	33	75	48
Mean $= \bar{y}_i$	11.000	15.000	12.000
SD $= s_i$	4.359	4.950	4.967

*Extra digits are reported for accuracy of subsequent calculations.

and the total of all the observations is

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij} = 33 + 75 + 48 = 156 \text{ or, equivalently } 3 \times 11 + 5 \times 15 + 4 \times 12 = 156$$

The grand mean is

$$\bar{\bar{y}} = \frac{156}{12} = 13 \text{ lb}$$

If the sample sizes (n_i's) are all equal, then the grand mean $\bar{\bar{y}}$ is just the ordinary average (i.e., mean) of the group means (the $\bar{y}_i$'s); but if the sample sizes are unequal, this is not the case. For instance, in Example 11.2.1 note that

$$\frac{11 + 15 + 12}{3} \neq 13$$

Measuring Variation within Groups

A combined measure of variation within the I groups is the pooled standard deviation s_{pooled}, often simply denoted as just s, which is computed as follows.*

Pooled Standard Deviation

$$s_{\text{pooled}} = s = \sqrt{\frac{\sum_{i=1}^{I}(n_i - 1)s_i^2}{\sum_{i=1}^{I}(n_i - 1)}} = \sqrt{\frac{\sum_{i=1}^{I}(n_i - 1)s_i^2}{n_\cdot - I}}$$

*There is no ambiguity in this notation since s_i (i.e., s with a subscript) denotes an individual group sample standard deviation.

We call $s^2_{\text{pooled}} = s^2$ the pooled variance*

$$s^2_{\text{pooled}} = s^2 = \frac{\sum_{i=1}^{I}(n_i - 1)s_i^2}{\sum_{i=1}^{I}(n_i - 1)}$$

Examining the formula we can see that the pooled variance is a weighted average of the group sample variances, and thus the pooled standard deviation can be very loosely interpreted as a weighted average of the group standard deviations.

The following example illustrates the computation of the pooled standard deviation, s.

Example 11.2.2

Weight Gain of Lambs Table 11.2.1 shows the group sample sizes and standard deviations for the lamb weight-gain data. The pooled variance and standard deviation are calculated as

$$s^2 = \frac{(3-1)4.359^2 + (5-1)4.950^2 + (4-1)4.967^2}{12-3} = \frac{210.025}{9} = 23.336$$

$$s = \sqrt{23.336} = 4.831$$

Observe that the pooled standard deviation, 4.831 lb, is a sensible representative value for the three group standard deviations, 4.359, 4.950, and 4.967 lb. If we assume that the population standard deviation of weight gains is the same for all three diets, then we would estimate this common value to be 4.83 lb. This estimate depends only on the variability within the groups and not on their mean values. Figure 11.2.1(a) displays the data from Table 11.2.1 while Figure 11.2.1(b) displays a modified version of the data for which 7 has been added to each Diet 2 observation and 5 has been subtracted from each Diet 3 observation. We see that while the group means are different for these two data sets, the pooled standard deviation—the inherent variability in each group—is the same.

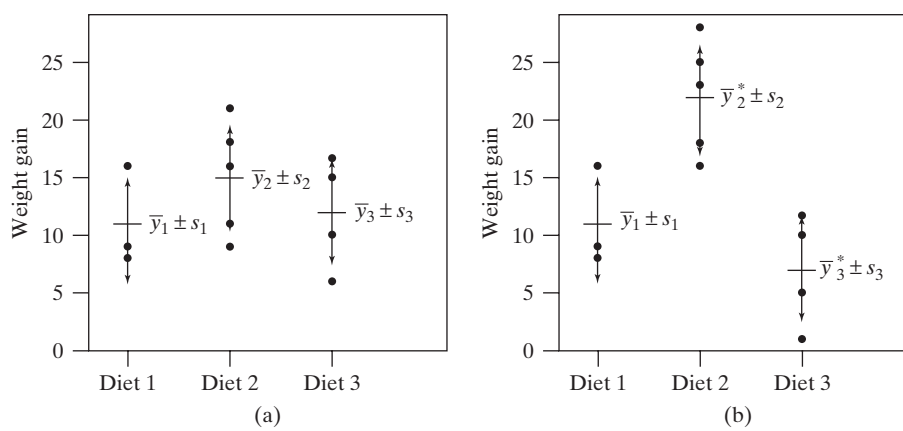


Figure 11.2.1 Examining within-group standard deviations. Plot (a) displays the weight gain data from Table 11.2.1 with $s = 4.831$. Plot (b) displays modified data with the same individual group standard deviations, and thus the same pooled standard deviation $s = 4.831$

ANOVA Notation

While our preceding formulas use familiar notation and terms, we will find it convenient to decompose the pooled variance into parts and subsequently define new terms to be used in the context of analysis of variance.

*Recall from Chapter 2 that the variance is simply the standard deviation squared.

The numerator of the pooled variance is known as the **sum of squares within groups**, **SS(within)**, while the denominator is known as the **degrees of freedom within groups**, **df(within)**. The formulas for these are displayed in the following box.*

Sum of Squares and df within Groups

$$\text{SS(within)} = \sum_{i=1}^{I}(n_i - 1)s_i^2$$

$$\text{df(within)} = n_\bullet - I$$

Their ratio is defined as the **mean square within groups**, or **MS(within)**. Note that MS(within) is just another name for the pooled variance.

Mean Square within Groups

$$\text{MS(within)} = \frac{\text{SS(within)}}{\text{df(within)}}$$

Hence, the quantity MS(within) measures the variability within the groups.†

The following example illustrates the calculation of SS(within), df(within), and MS(within).

Example 11.2.3

Weight Gain of Lambs In Example 11.2.2 when computing the pooled variance, we found

$$s^2 = \frac{(3-1)4.359^2 + (5-1)4.950^2 + (4-1)4.967^2}{12-3} = \frac{210.025}{9} = 23.336$$

Thus, SS(within) $= 210.025$, df(within) $= 9$, and MS(within) $= 23.336$. ■

Variation between Groups

For two groups, the difference between the groups is simply described by $(\bar{y}_1 - \bar{y}_2)$. How can we describe between-group variability for more than two groups? One naive idea is to simply compute the sample variance of the group means. The **mean square between groups**, or **MS(between)** is motivated by this idea. In fact, were it not for the n_i in the numerator of the following expression (to adjust for the sample sizes of the groups), the MS(between) would indeed be the sample variance of the group means.

Mean Square between Groups

$$\text{MS(between)} = \frac{\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{\bar{y}})^2}{I - 1}$$

*A popular but less intuitive formula for SS(within) is given by $\text{SS(within)} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$.

†If there were only one group, with n observations, then df(within) would be $n - 1$ and the SS(within) would be $(n-1)s^2$. MS(within) would then simply be $\frac{(n-1)s^2}{(n-1)} = s^2$, the sample variance.

As with the measures used for the within-group variation, MS(within), it is convenient to define the numerator of MS(between) as the **sum of squares between groups** or **SS(between)** and the denominator as the **degrees of freedom between groups** or **df(between)** so that

$$\text{MS(between)} = \frac{\text{SS(between)}}{\text{df(between)}}$$

where SS(between) and df(between) are explicitly defined as follows.

Sum of Squares and df between Groups

$$\text{SS(between)} = \sum_{i=1}^{I} n_i(\bar{y}_i - \bar{\bar{y}})^2$$

$$\text{df(between)} = I - 1$$

The following example illustrates these definitions.

Example 11.2.4

Weight Gain of Lambs For the data of Example 11.2.1, the quantities that enter SS(between) are shown in Table 11.2.2.

Table 11.2.2 Calculation of SS(between) for lamb weight gains

	Diet 1	Diet 2	Diet 3
Mean: $\bar{y}_i$	11	15	12
n_i	3	5	4
Grand mean $\bar{\bar{y}} = 13$			

From Table 11.2.2 we calculate

$$\text{SS(between)} = 3(11 - 13)^2 + 5(15 - 13)^2 + 4(12 - 13)^2 = 36$$

Since $I = 3$, we have

$$\text{df(between)} = 3 - 1 = 2$$

so that

$$\text{MS(between)} = \frac{36}{2} = 18$$

■

The SS(between) and MS(between) measure the variability between the samples means of the groups. This variability is shown graphically in Figure 11.2.2.

A Fundamental Relationship of ANOVA

The name *analysis of variance* derives from a fundamental relationship involving SS(between) and SS(within). Consider an individual observation y_{ij}. It is obviously true that

$$y_{ij} - \bar{\bar{y}} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{y}})$$

This equation expresses the deviation of an observation from the grand mean as the sum of two parts: a within-group deviation $(y_{ij} - \bar{y}_i)$ and a between-group deviation

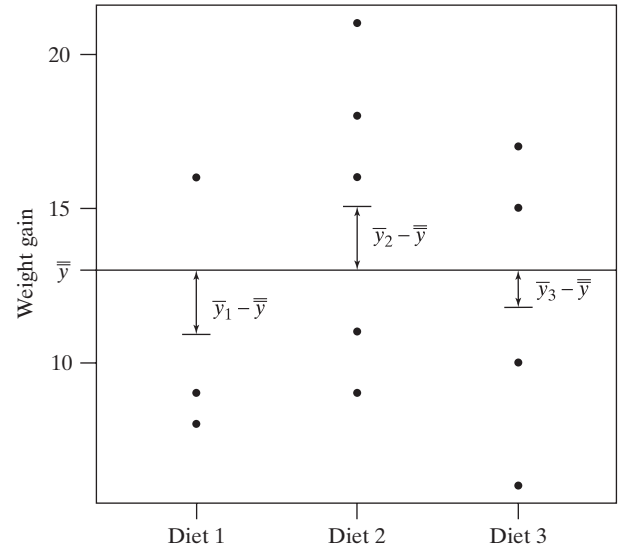


Figure 11.2.2 Measuring the differences between group means

$(\overline{y}_i - \overline{\overline{y}})$. It is also true (but not at all obvious) that the analogous relationship holds for the corresponding sums of squares; that is,

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \overline{\overline{y}})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{y}_i - \overline{\overline{y}})^2 \qquad (11.2.1)$$

which, by rewriting each of the sums on the right-hand side can be expressed as

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \overline{\overline{y}})^2 = \sum_{i=1}^{I}(n_i - 1)s_i^2 + \sum_{i=1}^{I}n_i(\overline{y}_i - \overline{\overline{y}})^2$$
$$= \text{SS(within)} + \text{SS(between)}$$

The quantity on the left-hand side of formula (11.2.1) is called the **total sum of squares**, or **SS(total)**:

Definition of Total Sum of Squares

$$\text{SS(total)} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \overline{\overline{y}})^2$$

Note that SS(total) measures variability among all n observations in the I groups. The relationship [formula (11.2.1)] can be written as

Relationship between Sums of Squares

$$\text{SS(total)} = \text{SS(between)} + \text{SS(within)}$$

The preceding fundamental relationship shows how the total variation in the data set can be analyzed, or broken down, into two interpretable components: between-sample variation and within-sample variation. This partition is an analysis of variance.

The **total degrees of freedom**, or **df(total)**, is defined as follows:

Total df

$$\text{df(total)} = n_\bullet - 1$$

With this definition, the degrees of freedom add, just as the sums of squares do; that is,

$$\text{df(total)} = \text{df(within)} + \text{df(between)}$$
$$n_\bullet - 1 = (n_\bullet - I) + (I - 1)$$

Notice that, if we were to consider all $n_\bullet$ observations as a single sample, then the SS for that sample (that is, the numerator of the variance) would be SS(total) and the associated df (that is, the denominator of the variance) would be df(total). Consequently, $\sqrt{\frac{\text{SS(total)}}{\text{df(total)}}}$ is the standard deviation of the entire data set when group membership is ignored.

The following example illustrates the fundamental relationships between the sums of squares and degrees of freedom.

Example 11.2.5

Weight Gain of Lambs For the data of Table 11.2.1, we found $\bar{\bar{y}} = 13$; we calculate SS(total) as

$$\begin{aligned}\text{SS(total)} &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \bar{\bar{y}})^2 \\ &= \left[(8-13)^2 + (16-13)^2 + (9-13)^2\right] \\ &\quad + \left[(9-13)^2 + (16-13)^2 + (21-13)^2 + (11-13)^2 + (18-13)^2\right] \\ &\quad + \left[(15-13)^2 + (10-13)^2 + (17-13)^2 + (6-13)^2\right] \\ &= 246\end{aligned}$$

For these data, we found that SS(between) = 36 and SS(within) = 210. We verify that

$$246 = 36 + 210$$

Also, we found that df(within) = 9 and df(between) = 2. We verify that

$$\text{df(total)} = 12 - 1 = 11 = 9 + 2$$

■

The ANOVA Table

When working with the ANOVA quantities, it is customary to arrange them in a table. The following example shows a typical format for the ANOVA table.

Example 11.2.6

Weight Gain of Lambs Table 11.2.3 shows the ANOVA for the lamb weight-gain data. Notice that the ANOVA table clearly shows the additivity of the sums of squares and the degrees of freedom. ■

Comment on terminology. While the terms "between-groups" and "within-groups" are not technical terms, they are useful in describing and understanding the ANOVA model. Computer software and other texts commonly refer to these sources of variability as **treatment** (between groups) and **error** (within groups).

Table 11.2.3 ANOVA table for lamb weight gains

Source	df	SS	MS
Between diets	2	36	18.00
Within diets	9	210	23.33
Total	11	246	

Summary of Formulas

For convenient reference, we display in the box the definitional formulas for the basic ANOVA quantities.

ANOVA Quantities with Formulas

Source	df	SS (Sum of Squares)	MS (Mean Square)
Between groups	$I - 1$	$\sum_{i=1}^{I} n_i(\bar{y}_i - \bar{\bar{y}})^2$	SS/df
Within groups	$n_\cdot - I$	$\sum_{i=1}^{I} (n_i - 1)s_i^2$	SS/df
Total	$n_\cdot - 1$	$\sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{\bar{y}})^2$	

Exercises 11.2.1–11.2.7

11.2.1 The accompanying table shows fictitious data for three samples.

	SAMPLE		
	1	2	3
	48	40	39
	39	48	30
	42	44	32
	43		35
Mean	43.00	44.00	34.00
SD	3.74	4.00	3.92

(a) Compute SS(between) and SS(within).
(b) Compute SS(total), and verify the relationship between SS(between), SS(within), and SS(total).
(c) Compute MS(between), MS(within), and s_{pooled}.

11.2.2 Proceed as in Exercise 11.2.1 for the following data:

	SAMPLE		
	1	2	3
	23	18	20
	29	12	16
	25	15	17
	23		23
			19
Mean	25.00	15.00	19.00
SD	2.83	3.00	3.16

11.2.3 For the following data, SS(within) = 116 and SS(total) = 338.769.

SAMPLE		
1	2	3
31	30	39
34	26	45
39	35	39
32	29	37
	30	

(a) Find SS(between).

(b) Compute MS(between), MS(within), and s_{pooled}.

11.2.4 The following ANOVA table is only partially completed.

SOURCE	DF	SS	MS
Between groups	3		45
Within groups	12	337	
Total		472	

(a) Complete the table.

(b) How many groups were there in the study?

(c) How many total observations were there in the study?

11.2.5 The following ANOVA table is only partially completed.

SOURCE	DF	SS	MS
Between groups	4		
Within groups		964	
Total	53	1123	

(a) Complete the table.

(b) How many groups were there in the study?

(c) How many total observations were there in the study?

11.2.6 The following ANOVA table is only partially completed.

SOURCE	DF	SS	MS
Between groups		258	
Within groups	26		
Total	29	898	

(a) Complete the table.

(b) How many groups were there in the study?

(c) How many total observations were there in the study?

11.2.7 Invent examples of data with

(a) SS(between) $= 0$ and SS(within) > 0

(b) SS(between) > 0 and SS(within) $= 0$

(c) For each example, use three samples, each of size 5.

11.3 The Analysis of Variance Model

In Section 11.2 we introduced the notation y_{ij} for the jth observation in group i. We think of y_{ij} as a random observation from group i, where the population mean of group i is μ_i. We use analysis of variance to investigate the null hypothesis that $\mu_1 = \mu_2 = \cdots = \mu_I$. It can be helpful to think of ANOVA in terms of the following model:

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

In this model, μ represents the grand population mean—the population mean when all the groups are combined. If the null hypothesis is true, then μ is the common population mean. If the null hypothesis is false, then at least some of the μ_i's differ from the grand population mean of μ.

The term τ_i represents the effect of group i—that is, the difference between the population mean for group i, μ_i, and the grand population mean, μ. (τ is the Greek letter "tau.") Thus,

$$\tau_i = \mu_i - \mu$$

The null hypothesis

$$H_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_I$$

is equivalent to

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_I = 0$$

If H_0 is false, then at least some of the groups differ from the others. If τ_i is positive, then observations from group i tend to be greater than the overall average; if τ_i is negative, then data from group i tend to be less than the overall average.

The term e_{ij} in the model represents random error associated with observation j in group i. Thus, the model

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

can be stated in words as

$$\text{observation} = \text{overall average} + \text{group effect} + \text{random error}$$

We estimate the overall average, μ, with the grand mean of the data:

$$\hat{\mu} = \bar{\bar{y}}$$

Likewise, we estimate the population average for group i with the sample average for group i:

$$\hat{\mu}_i = \bar{y}_i$$

Since the group effect is

$$\tau_i = \mu_i - \mu$$

we estimate τ_i as

$$\hat{\tau}_i = \bar{y}_i - \bar{\bar{y}}$$

Finally, we estimate the random error, ε_{ij}, for observation y_{ij} as

$$\hat{\varepsilon}_{ij} = y_{ij} - \bar{y}_i$$

Putting these estimates together, we have

$$y_{ij} = \bar{\bar{y}} + (\bar{y}_i - \bar{\bar{y}}) + (y_{ij} - \bar{y}_i)$$

or

$$y_{ij} = \hat{\mu} + \hat{\tau}_i + \hat{\varepsilon}_{ij}$$

Note. Some authors use the terminology SS(error) for what we have called SS(within). This is due to the fact that the within-groups component $y_{ij} - \bar{y}_i$ estimates the random error term in the ANOVA model.

Example 11.3.1

Weight Gain of Lambs For the data of Example 11.2.1, the estimate of the grand population mean is $\hat{\mu} = 13$. The estimated group effects are

$$\hat{\tau}_1 = \bar{y}_1 - \bar{\bar{y}} = 11 - 13 = -2$$

$$\hat{\tau}_2 = 15 - 13 = 2$$

and

$$\hat{\tau}_3 = 12 - 13 = -1$$

Thus, we estimate that Diet 2 increases weight gain by 2 lb on average (when compared to the average of the three diets), Diet 1 decreases weight gain by an average of 2 lb, and Diet 3 decreases weight gain by 1 lb, on average. ■

When we conduct an analysis of variance, we are comparing the sizes of the sample group effects, the $\hat{\tau}_i$'s, to the sizes of the random errors in the data, the $\hat{\varepsilon}_{ij}$'s. We can see that

$$\text{SS(between)} = \sum_{i=1}^{I} n_i \hat{\tau}_i^2$$

and

$$\text{SS(within)} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2$$

11.4 The Global *F* Test

The global null hypothesis is

$$H_0\text{: } \mu_1 = \mu_2 = \cdots = \mu_I$$

We consider testing H_0 against the nondirectional (or omnidirectional) alternative hypothesis

$$H_A\text{: The } \mu_i\text{'s are not all equal}$$

Note that H_0 is compound (unless $I = 2$), and so rejection of H_0 does not specify *which* μ_i's are different. If we reject H_0, then we conduct a further analysis to make detailed comparisons among the μ_i's. Testing the global null hypothesis may be likened to looking at a microscope slide through a low-power lens to see if there is anything on it; if we find something, we switch to a greater magnification to examine its fine structure.

The *F* Distributions

The ***F* distributions**, named after the statistician and geneticist R. A. Fisher, are probability distributions that are used in many kinds of statistical analysis. The form of an *F* distribution depends on two parameters: the **numerator degrees of freedom** and the **denominator degrees of freedom**. Figure 11.4.1 shows an *F* distribution with numerator df = 4 and denominator df = 20. Critical values for the *F* distribution are given in Table 10 at the end of this book. Note that Table 10 occupies 10 pages, each page having a different value of the numerator df. As a specific example, for numerator df = 4 and denominator df = 20, we find in Table 10 that $F(4, 20)_{0.05} = 2.87$; this value is shown in Figure 11.4.1.

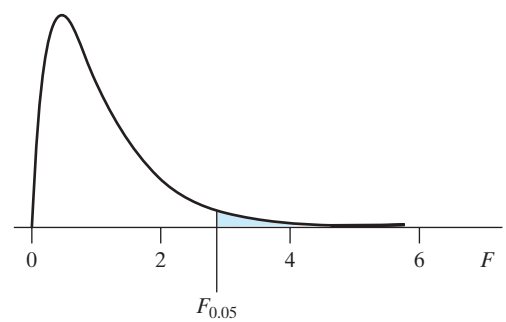


Figure 11.4.1 The *F* distribution with numerator df = 4 and denominator df = 20

The *F* Test

The ***F* test** is a classical test of the global null hypothesis. The test statistic, the ***F* statistic**, is calculated as follows:

$$F_s = \frac{\text{MS(between)}}{\text{MS(within)}}$$

From the definitions of the mean squares (Section 11.2), it is clear that F_s will be large if the discrepancies among the group means ($\overline{Y}_i$'s) are large relative to the variability within the groups. Thus, large values of F_s tend to provide evidence against H_0—evidence for a difference among the group means.

To carry out the F test of the global null hypothesis, critical values are obtained from an F distribution (Table 10) with

$$\text{Numerator df} = \text{df(between)}$$

and

$$\text{Denominator df} = \text{df(within)}$$

It can be shown that (when suitable conditions for validity are met) the null distribution of F_s is an F distribution with df as given above.

The following example illustrates the global F test.

Example 11.4.1

Weight Gain of Lambs For the lamb feeding experiment of Example 11.2.1, the global null hypothesis and alternative can be stated verbally as

H_0: Mean weight gain is the same on all three diets.

H_A: Mean weight gain is not the same on all three diets.

or symbolically as

H_0: $\mu_1 = \mu_2 = \mu_3$

H_A: The μ_i's are not all equal

We saw in Figure 11.2.2 that the three sample means do not differ much when compared to the variability within the groups, which is not very strong evidence against H_0. Let us confirm this visual impression by carrying out the F test at $\alpha = 0.05$. From the ANOVA table (Table 11.2.3) we find

$$F_s = \frac{18.00}{23.33} = 0.77$$

The degrees of freedom can also be read from the ANOVA table as

$$\text{Numerator df} = 2$$
$$\text{Denominator df} = 9$$

From Table 10 we find $F(2,9)_{0.20} = 1.93$, so that $P > 0.20$. Thus, there is a lack of significant evidence against H_0; there is insufficient evidence to conclude that there is any difference among the diets with respect to population mean weight gain. The observed differences in the mean gains in the samples can readily be attributed to chance variation. Because this study was an experiment (as opposed to

an observational study), we can even make a slightly stronger summary of the results: There is insufficient evidence to conclude that among these three diets, diet *affects* weight gain. ■

Relationship between *F* Test and *t* Test

Suppose only two groups are to be compared ($I = 2$). Then one could test $H_0: \mu_1 = \mu_2$ against $H_A: \mu_1 \neq \mu_2$ using either the F test or the t test. The t test from Chapter 7 can be modified slightly by replacing each sample standard deviation by s_{pooled}, as defined in Section 11.2, before calculating the standard error of $(\overline{Y}_1 - \overline{Y}_2)$. It can be shown that the F test and this "pooled" t test are actually equivalent procedures. The relationship between the test statistics is $t_s^2 = F_s$; that is, the value of the F statistic for any set of data is necessarily equal to the square of the value of the (pooled) t statistic. The corresponding relationship between the critical values is $t_{0.025}^2 = F_{0.05}$, $t_{0.005}^2 = F_{0.01}$, and so on. For example, suppose $n_1 = 10$ and $n_2 = 7$. Then the appropriate t distribution has df $= n_1 + n_2 - 2 = 15$, and $t_{15,0.025} = 2.131$, whereas the F distribution has numerator df $= I - 1 = 1$ and denominator df $= n_\cdot - I = 15$, so that $F(1,15)_{0.05} = 4.54$; note that $(2.131)^2 = 4.54$. Because of the equivalence of the tests, the application of the F test to compare the means of two samples will always give exactly the same P-value as the pooled t test applied to the same data.

Exercises 11.4.1–11.4.7

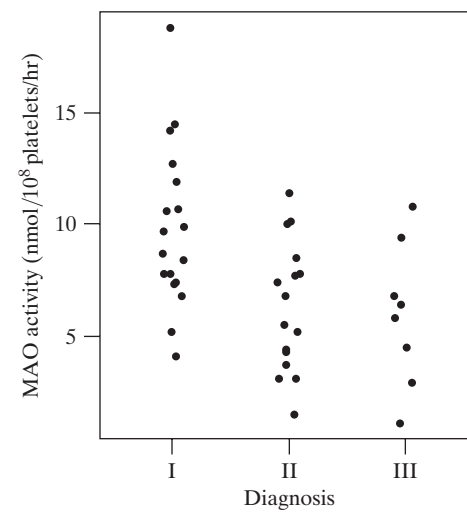
11.4.1 Monoamine oxidase (MAO) is an enzyme that is thought to play a role in the regulation of behavior. To see whether different categories of schizophrenic patients have different levels of MAO activity, researchers collected blood specimens from 42 patients and measured the MAO activity in the platelets. The results are summarized in the accompanying table. (Values are expressed as nmol benzylaldehyde product/10^8 platelets/hour.)[3] Calculations based on the raw data yielded SS(between) = 136.12 and SS(within) = 418.25.

	MAO ACTIVITY		
DIAGNOSIS	MEAN	SD	NO. OF PATIENTS
Chronic undifferentiated schizophrenic	9.81	3.62	18
Undifferentiated with paranoid features	6.28	2.88	16
Paranoid schizophrenic	5.97	3.19	8

(a) Dotplots of these data follow. Based on this graphical display, does it appear that the null hypothesis is true? Why or why not?

(b) Construct the ANOVA table and test the global null hypothesis at $\alpha = 0.05$.

(c) Calculate the pooled standard deviation, s_{pooled}.



11.4.2 It is thought that stress may increase susceptibility to illness through suppression of the immune system. In an experiment to investigate this theory, 48 rats were randomly allocated to four treatment groups: no stress, mild stress, moderate stress, and high stress. The stress conditions involved various amounts of restraint and electric shock. The concentration of lymphocytes (cells/ml $\times 10^{-6}$) in the peripheral blood was measured for each rat with the results given in the accompanying table.[4] Calculations based on the raw data yielded SS(between) = 89.036 and SS(within) = 340.24.

	NO STRESS	MILD STRESS	MODERATE STRESS	HIGH STRESS
$\bar{y}$	6.64	4.84	3.98	2.92
s	2.77	2.42	3.91	1.45
n	12	12	12	12

(a) Construct the ANOVA table and test the global null hypothesis at $\alpha = 0.05$.

(b) Calculate the pooled standard deviation, s_{pooled}.

11.4.3 Human beta-endorphin (HBE) is a hormone secreted by the pituitary gland under conditions of stress. An exercise physiologist measured the resting (unstressed) blood concentration of HBE in three groups of men: 15 who had just entered a physical fitness program, 11 who had been jogging regularly for some time, and 10 sedentary people. The HBE levels (pg/ml) are shown in the following table.[5] Calculations based on the raw data yielded SS(between) = 240.69 and SS(within) = 6,887.6.

	FITNESS PROGRAM ENTRANTS	JOGGERS	SEDENTARY
Mean	38.7	35.7	42.5
SD	16.1	13.4	12.8
n	15	11	10

(a) State the appropriate null hypothesis in words, in the context of this setting.

(b) State the null hypothesis in symbols.

(c) Construct the ANOVA table and test the null hypothesis. Let $\alpha = 0.05$.

(d) Calculate the pooled standard deviation, s_{pooled}.

11.4.4 An experiment was conducted in which the antiviral medication zanamivir was given to patients who had the flu. The length of time until the alleviation of major flu symptoms was measured for three groups: 85 patients who were given inhaled zanamivir, 88 patients who were given inhaled and intranasal zanamivir, and 89 patients who were given a placebo. Summary statistics are given in the following table.[6] The ANOVA SS(between) is 53.67 and the SS(within) is 2034.52.

	INHALED ZANAMIVIR	INHALED AND INTRANASAL ZANAMIVIR	PLACEBO
Mean	5.4	5.3	6.3
SD	2.7	2.8	2.9
n	85	88	89

(a) State the appropriate null hypothesis in words, in the context of this setting.

(b) State the null hypothesis in symbols.

(c) Construct the ANOVA table and test the null hypothesis. Let $\alpha = 0.05$.

(d) Calculate the pooled standard deviation, s_{pooled}.

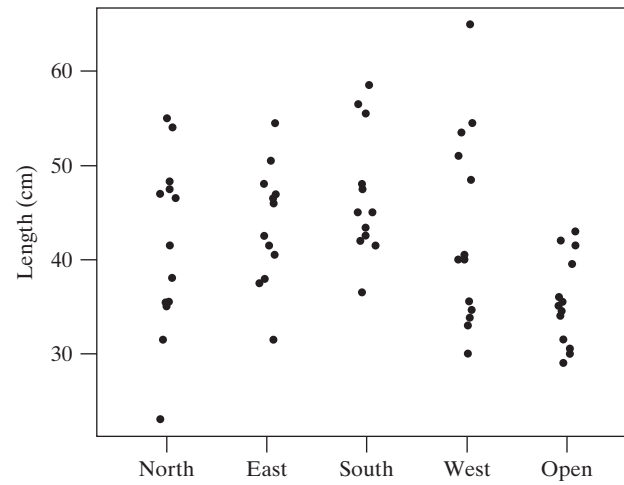
11.4.5 A researcher collected daffodils from four sides of a building and from an open area nearby. She wondered whether the average stem length of a daffodil depends on the side of the building on which it is growing. Summary statistics are given in the following table.[7] The ANOVA SS(between) is 871.408 and the SS(within) is 3588.54.

	NORTH	EAST	SOUTH	WEST	OPEN
Mean	41.4	43.8	46.5	43.2	35.5
SD	9.3	6.1	6.6	10.4	4.7
n	13	13	13	13	13

(a) Dotplots of these data follow. Based on the dotplots, does it appear that the null hypothesis is true? Why or why not?

(b) State the null hypothesis in symbols.

(c) Construct the ANOVA table and test the null hypothesis. Let $\alpha = 0.10$.



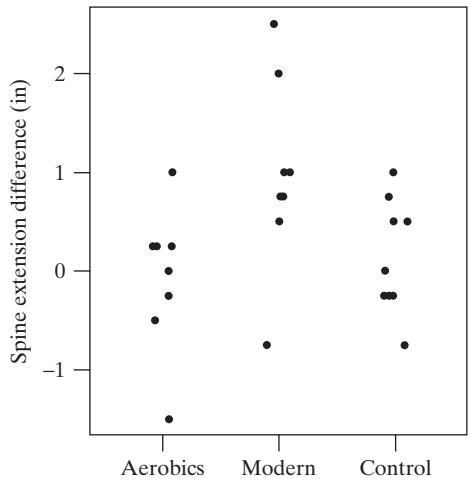
11.4.6 A researcher studied the flexibility of 10 women in an aerobic exercise class, 10 women in a modern dance class, and a control group of 9 women. One measurement she made on each woman was spinal extension, which is a measure of how far the woman could bend her back. Measurements were made before and after a 16-week training period. The change in spinal extension was recorded for each woman. Summary statistics are given in the following table.[8] The ANOVA SS(between) is 7.04 and the SS(within) is 15.08.

	AEROBICS	MODERN DANCE	CONTROL
Mean	−0.18	0.98	0.13
SD	0.80	0.86	0.57
n	10	10	9

(a) Dotplots of these data were are shown below. Based on the dotplots, does it appear that the null hypothesis is true? Why or why not?

(b) State the null hypothesis in symbols.

(c) Construct the ANOVA table and test the null hypothesis. Let $\alpha = 0.01$.



11.4.7 The following computer output is for an analysis of variance in which yields (bu/acre) of different varieties of oats were compared.[9]

SOURCE	DF	SUMS OF SQUARES	MEAN SQUARE	*F* RATIO	PROB
Group	2	76.8950	38.4475	0.40245	0.6801
Error	9	859.808	95.5342		
Total	11	936.703			

(a) How many varieties (groups) were in the experiment?

(b) State the conclusion of the ANOVA.

(c) What is the pooled standard deviation, s_{pooled}?

11.5 Applicability of Methods

Like all other methods of statistical inference, the calculations and interpretations of ANOVA are based on certain conditions.

Standard Conditions

The ANOVA techniques described in this chapter, including the global F test, are valid if the following conditions hold.

1. *Design conditions*
   (a) It must be reasonable to regard the groups of observations as random samples from their respective populations.
   (b) The I samples must be independent of each other.
2. *Population conditions* The I population distributions must be (approximately) normal with equal standard deviations:

$$\sigma_1 = \sigma_2 = \cdots = \sigma_I$$

These conditions are extensions of the conditions given in Chapter 7 for the independent-samples t test with the added condition that the standard deviations be equal. The condition of normal populations with equal standard deviations is less crucial if the sample sizes (n_i) are large and approximately equal.

Verification of Conditions

The design conditions may be verified as for the independent-samples t test. To check condition 1(a), one looks for biases or hierarchical structure in the collection of the data. A completely randomized design assures independence of the samples

[condition 1(b)]. If units have been allocated to treatment groups in a nonrandom manner (e.g., by a randomized blocks design to be discussed in Section 11.6), or if observations on the same experimental unit appear in different samples (e.g., for $I = 2$, paired data as seen in Chapter 9), then the samples are not independent.

As with the independent-samples t test, the population conditions can be roughly checked from the data. To check normality, a separate histogram or normal probability plot can be made for each sample. Another option is to make a single histogram or normal probability plot of the deviations $(y_{ij} - \overline{y}_i)$ from all the samples combined. In the context of analysis of variance we call these deviations from the group means **residuals**. Thus, a residual measures how far a data value falls from its respective group mean.

Equality of the population SDs is checked by comparing the sample SDs; one useful trick is to plot the SDs against the means ($\overline{y}_i$'s) to check for a trend. Another approach is to make a plot of the residuals $(y_{ij} - \overline{y}_i)$ against the means ($\overline{y}_i$'s). As a rule of thumb, we would like the largest sample SD divided by the smallest sample SD to be less than 2 or so. If this ratio is much larger than 2, then we cannot be confident in the P-value from the ANOVA, particularly if the sample sizes are small and unequal. In particular, if the sample sizes are unequal and the sample SD from a small sample is quite a bit larger than the other SDs, then the P-value can be quite inaccurate.

Example 11.5.1

Weight Gain of Lambs Consider the lamb feeding experiment of Example 11.2.2. Figure 11.2.1 (in Section 11.2) shows that the variability within groups is nearly equal across the three diets: The three sample SDs are 4.36, 4.95, and 4.97. Figure 11.5.1 is a normal probability plot of the 12 residuals $(y_{ij} - \overline{y}_i)$ (3 from Diet 1, 5 from Diet 2, and 4 from Diet 3). This plot is close to linear, which provides no evidence to cast doubt on the normality condition. ■

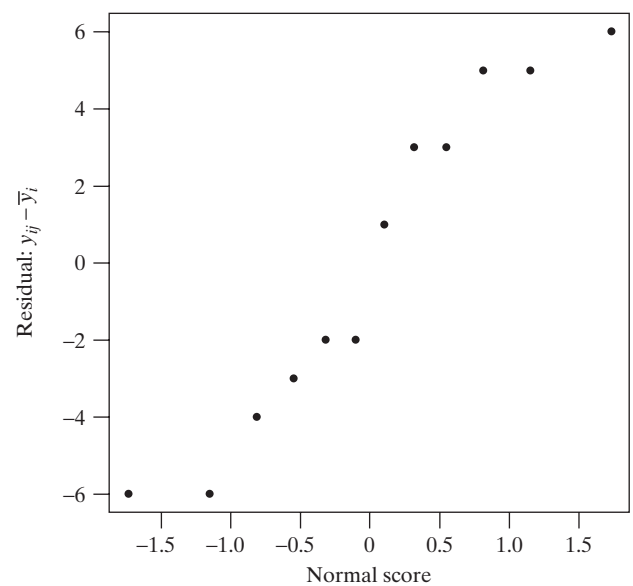


Figure 11.5.1 Normal probability plot of residuals $(y_{ij} - \overline{y}_i)$ in weight-gain data

Example 11.5.2

Sweet Corn Consider the sweet corn data of Example 11.1.1. Figure 11.5.2(a) shows the data with each group receiving its own plotting symbol. Using those same plotting symbols for each group, Figure 11.5.2(b) displays the residuals $(y_{ij} - \overline{y}_i)$ plotted

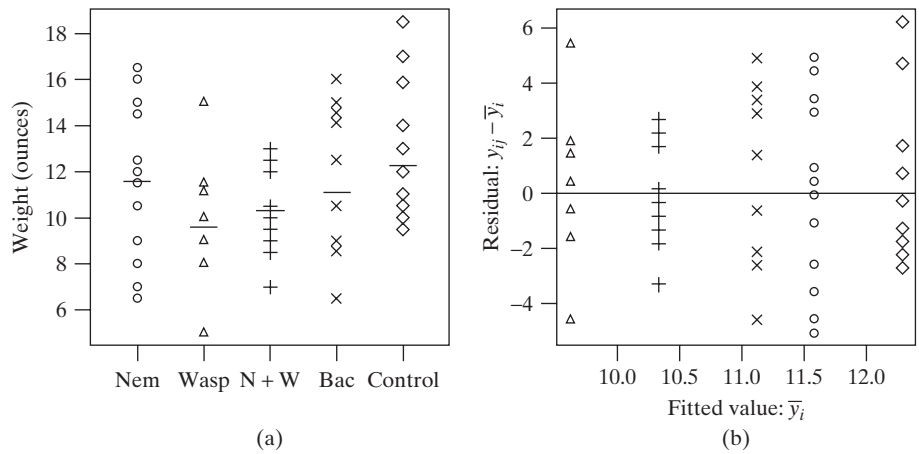


Figure 11.5.2 Plot of residuals versus sample mean for the sweet corn data

against each group's mean ($\overline{y}_i$) (also known as a **fitted value** in the context of analysis of variance). This second graph shows that the variability (as measured visually by the vertical spread) does not appreciably change as the mean changes (which is good—if the variability increased as the mean increased, then condition 2 would be violated). ■

While one could look at a basic plot of the data, as in Figure 11.5.2(a), to visually inspect that the SDs are similar across all groups, plotting the data as in Figure 11.5.2(b) provides some visual advantages. First, by examining the residuals (Figure 11.5.2(b)) and not the raw data (Figure 11.5.2(a)), one can scan the graph from left to right allowing the eyes to more clearly compare the variability among the groups without being distracted by the changing means. Furthermore, a common violation of the equal SD requirement is that the group SDs grow with the means. To illustrate this violation, consider the fictitious data graphed in Figure 11.5.3(a) consisting of five treatment groups and seven observations per group. Clearly the variability is not the same in all five groups. The plot of the residuals versus means in Figure 11.5.3(b) exposes this problem more clearly and shows that the SD (represented by vertical spread) increases with the mean. We often describe this as *funnel* or *horn* shape in the residuals.

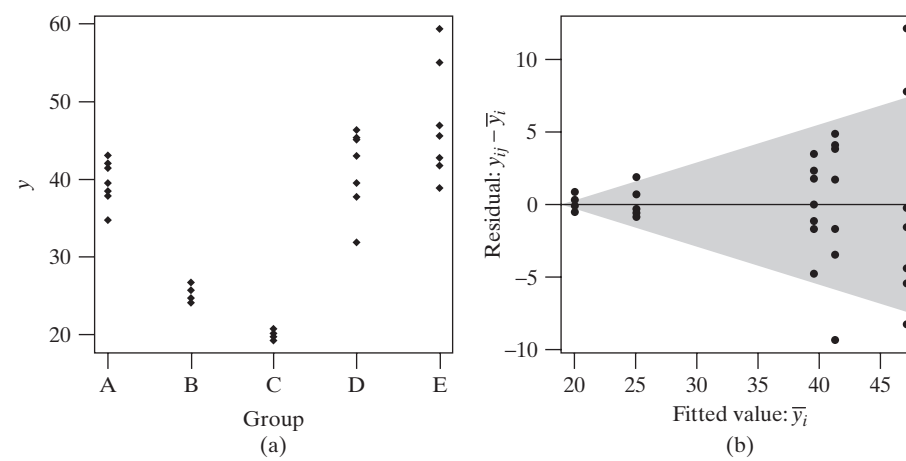


Figure 11.5.3 Plot of residuals versus sample means for a fictitious data set for which the standard deviation increases with the mean

Example 11.5.3

Sweet Corn Again considering the sweet corn data of Example 11.2.1, we examine the normality of the groups through examination of the residuals. Figure 11.5.4 contains a histogram and a normal probability plot of the 60 residuals $(y_{ij} - \overline{y}_i)$. The bell-shaped nature of plot (a) and linearity of plot (b) cast little doubt upon the normality condition. ■

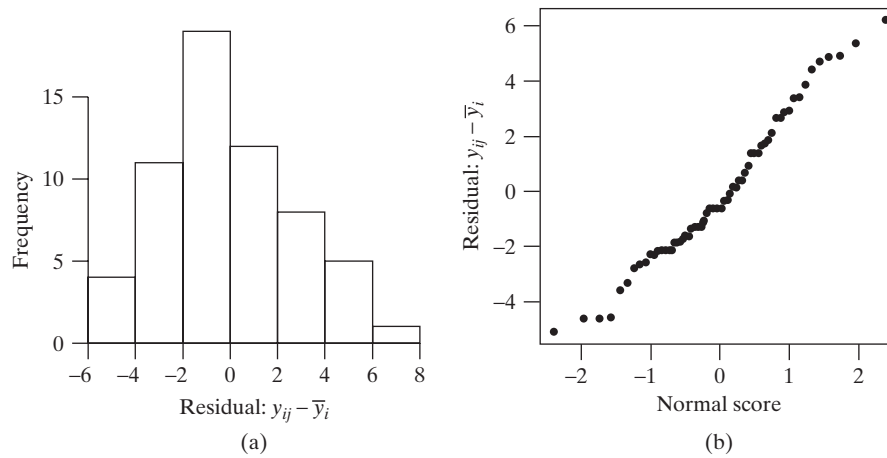


Figure 11.5.4 Histogram and normal probability plot of deviations $(y_{ij} - \overline{y}_i)$ in sweet corn data

Further Analysis

In addition to their relevance to the F test, the standard conditions underlie many classical methods for further analysis of the data.

If the I populations have the same SD, then a pooled estimate of that SD from the data is

$$s_{\text{pooled}} = \sqrt{\text{MS(within)}}$$

from the ANOVA. This pooled standard deviation s_{pooled} is a better estimate than any individual sample SD because s_{pooled} is based on more observations.

A simple way to see the advantage of s_{pooled} is to consider the standard error of an individual sample mean, which can be calculated as

$$\text{SE}_{\overline{Y}} = \frac{s_{\text{pooled}}}{\sqrt{n}}$$

where n is the size of the individual sample. The df associated with this standard error is df(within), which is the sum of the degrees of freedom of all the samples. By contrast, if the individual SD were used in calculating $\text{SE}_{\overline{Y}}$, it would have only $(n - 1)$ df. When the SE is used for inference, larger df yield smaller critical values (see Table 4), which in turn lead to improved power and narrower confidence intervals.

In optional Sections 11.7 and 11.8 we will consider methods for detailed analysis of the group means $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_I$. Like the F test, these methods were designed for independent samples from normal populations with equal standard deviations. The methods use standard errors based on the pooled standard deviation estimate s_{pooled}.

Exercises 11.5.1–11.5.2

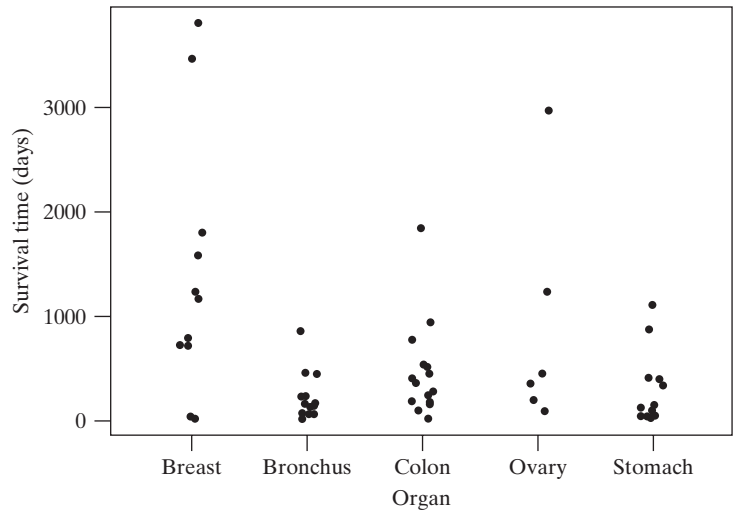
11.5.1 Refer to the lymphocyte data of Exercise 11.4.2. The global F test is based on certain conditions concerning the population distributions.

(a) State the conditions.

(b) Which features of the data suggest that the conditions may be doubtful in this case?

11.5.2 Patients with advanced cancers of the stomach, bronchus, colon, ovary, or breast were treated with ascorbate. The purpose of the study was to determine if the survival times differ with respect to the organ affected by the cancer. The variable of interest is survival time (in days).[10] Here are parallel dotplots of the raw data.

An ANOVA was done after a square root transformation was applied to the raw data. There were two (related) reasons that the data were transformed. What were those two reasons?



11.6 One-Way Randomized Blocks Design

The completely randomized design makes no distinctions among the experimental units. Often an experiment can be improved by a more refined approach, one that takes advantage of known patterns of variability in the experimental units.

In a **randomized blocks design**, we first group the experimental units into sets, or **blocks**, of relatively similar units and then we randomly allocate treatments within each block. Here is an example.

Example 11.6.1

Alfalfa and Acid Rain Researchers were interested in the effect that acid has on the growth rate of alfalfa plants. They created three treatment groups in an experiment: low acid, high acid, and control. The response variable in their experiment was the height of the alfalfa plants in a Styrofoam cup after five days of growth.* They had 5 cups for each of the 3 treatments, for a total of 15 observations. However, the cups were arranged near a window and they wanted to account for the effect of differing amounts of sunlight. Thus, they created 5 blocks—each block was a fixed

*More precisely, the response variable was the average height of plants within a cup, so that the observational unit was a cup, rather than individual plants.

Figure 11.6.1 Design of the alfalfa experiment

Window	Block 1	Block 2	Block 3	Block 4	Block 5
	high	control	control	control	high
	control	low	high	low	low
	low	high	low	high	control

Organization of blocks for alfalfa experiment

distance away from the window (block 1 being the closest through block 5, the farthest). Within each block the three treatments were randomly assigned, as shown in Figure 11.6.1.[11] ■

Example 11.6.1 is an illustration of a randomized blocks design. To carry out a randomized blocks design, the experimenter creates or identifies suitable blocks of experimental units and then randomly assigns treatments within each block in such a way that each treatment appears in each block.* In Example 11.6.1, the rows of cups at each of the five distances from the window serve as blocks. In general, we create blocks in order to reduce or eliminate variability caused by extraneous variables, so that the precision of the experiment is increased. We want the experimental units within a block to be homogenous; we want the extraneous variability to occur *between* the blocks. Here are more examples of randomized blocks designs in biological experiments.

Example 11.6.2

Blocking by Litter How does experience affect the anatomy of the brain? In a typical experiment to study this question, young rats are placed in one of three environments for 80 days:

T_1: *Standard environment.* The rat is housed with a single companion in a standard lab cage.

T_2: *Enriched environment.* The rat is housed with several companions in a large cage, furnished with various playthings.

T_3: *Impoverished environment.* The rat lives alone in a standard lab cage.

At the end of the 80-day experience, various anatomical measurements are made on the rats' brains.

Suppose a researcher plans to conduct the above experiment using 30 rats. To minimize variation in response, all 30 animals will be male, of the same age and strain. To reduce variation even further, the researcher can take advantage of the similarity of animals from the same litter. In this approach, the researcher would obtain three male rats from each of 10 litters. The three littermates from each litter would be assigned at random: one to T_1, one to T_2, and one to T_3.[12] ■

Another way to visualize the experimental design is in tabular form, as shown in Table 11.6.1. Each "Y" in the table represents an observation on one rat. Using the layout of Table 11.6.1, the experimenter can compare the responses of rats that received *different* treatments but are in the *same* litter. Such comparisons are not affected by any difference (genetic and other) that may exist between one litter and another.

*Strictly speaking, the design we discuss is termed a *randomized complete blocks design* because every treatment appears in every block. In an *incomplete blocks design*, each block contains some, but not necessarily all, of the treatments.

Table 11.6.1 Format for rat brain data

	Treatment		
	T_1	T_2	T_3
Litter 1	Y	Y	Y
Litter 2	Y	Y	Y
Litter 3	Y	Y	Y
.	.	.	.
.	.	.	.
.	.	.	.
Litter 10	Y	Y	Y

Example 11.6.3

Within-Subject Blocking (Pairing) A dermatologist is planning a study to compare two medicated lotions for their effectiveness in treating acne. Twenty patients are to participate in the study. Each patient will use lotion A on one side of his or her face and lotion B on the other; the dermatologist will observe the improvement on each side during a three-month period. For each patient, the side of the face to receive lotion A is randomly selected; the other side receives lotion B. The bottles of medication have coded labels so that neither the patient nor the physician knows which bottle contains A and which contains B—that is, in addition to blocking, the experiment also makes use of blinding.[13] This example, with blocks of size 2, is an example of pairing: The left side of the face is paired with the right side of the face. We have considered the analysis of paired data in Chapter 8. ■

Example 11.6.4

Blocking in an Agricultural Field Study When comparing several varieties of grain, an agronomist will generally plant many field plots of each variety and measure the yield of each plot. Differences in yields may reflect not only genuine differences among the varieties, but also differences among the plots in soil fertility, pH, water-holding capacity, and so on. Consequently, the spatial arrangement of the plots in the field is important. An efficient way to use the available field area is to divide the field into large regions—the blocks—and to subdivide each block into several plots. Within each block the various varieties of grain are then randomly allocated to the plots, with a separate randomization done for each block. For instance, suppose we want to test four varieties of barley. Then each block would contain four plots. The resulting randomized allocation might look like Figure 11.6.2, which is a schematic map of the field. The "treatments" T_1, T_2, T_3, and T_4 are the four varieties of barley. ■

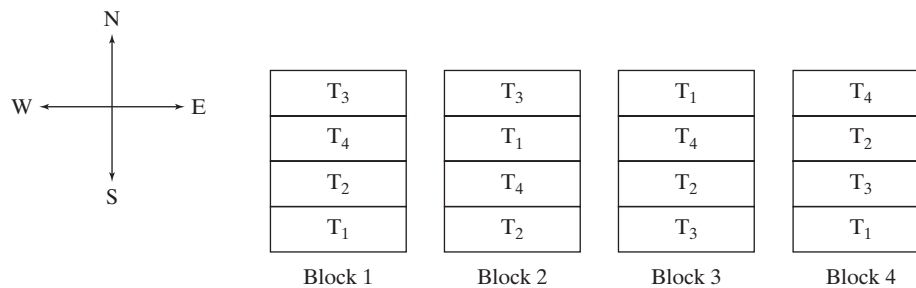


Figure 11.6.2 Layout of an agricultural randomized blocks design

Creating the Blocks

As the preceding examples show, blocking is a way of *organizing* the inherent variation that exists among experimental units. Ideally, the blocking should be arranged so as to increase the information available from the experiment. To achieve this goal, *the experimenter should try to create blocks that are as homogeneous within themselves as possible, so that the inherent variation between experimental units becomes, as far as possible, variation between blocks rather than within blocks.* This principle was illustrated in the preceding examples (e.g., in Example 11.6.2, where blocking by litter exploits the fact that littermates are more similar to each other than to nonlittermates). The following is another illustration.

Example 11.6.5

Agricultural Field Study For the barley experiment of Example 11.6.4, how would agronomists determine the best arrangement or layout of blocks in a field? They would design the blocks to take advantage of any prior knowledge they may have of fertility patterns in the field. For instance, if they know that an east–west fertility gradient exists in the field (perhaps the field slopes from east to west, with the result that the west end has a thicker layer of good soil or receives better irrigation), then they might choose blocks as in Figure 11.6.2; the layout maximizes soil differences between the blocks and minimizes differences between plots within each block. (But even if a field appears to be uniform, blocking is usually used in agronomic experiments, because plots closer together in the field are generally more similar than plots farther apart.) ■

To add solidity to this example, let us look at a set of data from a randomized blocks experiment on barley. Each entry in Table 11.6.2 shows the yield (bushels of barley per acre) of a plot 3.5 ft wide by 80 ft long.[14]

Table 11.6.2 Yield (lb) of barley

	Block 1	Block 2	Block 3	Block 4	Variety mean
Variety 1	93.5	66.6	50.5	42.4	63.3
Variety 2	102.9	53.2	47.4	43.8	61.8
Variety 3	67.0	54.7	50.0	40.1	53.0
Variety 4	86.3	61.3	50.7	46.4	61.2
Block Mean	87.4	59.0	49.7	43.2	

It appears from Table 11.6.2 that the yield potential of the blocks varies greatly; the data indicate a definite fertility gradient from block 1 to block 4. Because of the blocked design, comparison of the varieties is relatively unaffected by the fertility gradient. Of course, there also appears to be substantial variation within blocks. [You might find it an interesting exercise to peruse the data and ask yourself whether the observed differences between varieties are large enough to conclude that, for example, variety 1 is superior (in mean yield) to variety 3; use your intuition rather than a formal statistical analysis. The truth is revealed in Note 14.]

The Randomization Procedure

Once the blocks have been created, the blocked allocation of experimental units is straightforward: It is as if a mini-experiment is conducted within each block. Randomization is carried out for each block separately, as illustrated in the following example.

Example 11.6.6

Agricultural Field Study Consider the agricultural field experiment of Example 11.6.4. In block 1, let us label the plots 1, 2, 3, 4, from north to south (see Figure 11.6.2); we will allocate one plot to each variety. The allocation proceeds as for the completely randomized design, by choosing plots at random from the four, and assigning the first plot chosen to T_1, the second to T_2, and so on. For instance, using a computer to randomly permute the numbers 1 through 4 (or even shuffled cards numbered 1 through 4) we might obtain the sequence 4, 3, 1, 2 which would lead to the following treatment allocation.

Block 1

T_1: Plot 4

T_2: Plot 3

T_3: Plot 1

T_4: Plot 2

This is in fact the assignment shown in Figure 11.6.2 for block 1. We can then repeat this procedure for blocks 2, 3, and so on. ■

Analyzing Data from a Randomized Block Experiment

In the same way we cannot use a two-sample t test when data are paired, when an experiment has been blocked, we no longer can use our ANOVA methods of Section 11.4. Instead, we will use a **randomized blocks ANOVA** model. We will illustrate the concepts as we reconsider the alfalfa and acid rain experiment of Example 11.6.1 in which the researchers blocked the experiment based on rows of cups placed parallel to a window so that each block has roughly the same light exposure. The data are given in Table 11.6.3 and are graphed in Figure 11.6.3.

Table 11.6.3 Alfalfa plant height after five days (cm)

	High acid	Low acid	Control	Block mean
Block 1	1.30	1.78	2.67	1.917
Block 2	1.15	1.25	2.25	1.550
Block 3	0.50	1.27	1.46	1.077
Block 4	0.30	0.55	1.66	0.837
Block 5	1.30	0.80	0.80	0.967
Treatment mean $= \bar{y}_i$	0.910	1.130	1.768	
n	5	5	5	

Our usual ANOVA null hypothesis for comparing I populations or treatments is

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_I$$

Example 11.6.7

Alfalfa and Acid Rain The null hypothesis for the alfalfa growth experiment is that acid has no effect on five-day growth. (We can make a strong causal claim like this because this was an experiment.) More directly, the null hypothesis is that the mean five-day growth is the same for all three treatments (high acid, low acid, and control).

$$H_0: \mu_1 = \mu_2 = \mu_3$$ ■

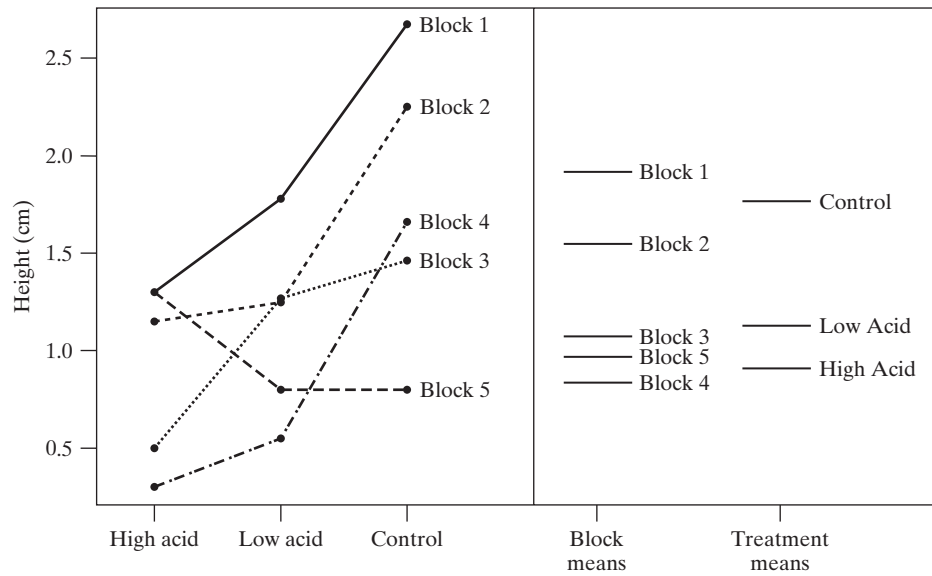


Figure 11.6.3 Dotplots of the alfalfa growth data with a summary of block and treatment means

This hypothesis can be tested with an analysis of variance F test, but first we want to remove the variability in the data that is due to differences between the blocks. To do this, we extend the ANOVA model presented in Section 11.3 to the following model:

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

In this model y_{ijk} is the kth observation when treatment i is applied in block j. (In Example 11.6.1 there is only one observation for each treatment in each block, but in general there might be more than one.) Here, as before, μ represents the grand population mean and the term τ_i represents the effect of group i (that is, treatment i). The new term in the model is β_j, which represents the effect of the jth block.

Visualizing the Block Effects

To visualize how blocking affects our ANOVA, we can think of our model in a slightly different way:

$$(y_{ijk} - \tau_i) = \mu + \beta_j + \varepsilon_{ijk}$$

The left-hand side of the equation describes the data after treatment effects have been removed. With our data we estimate this left-hand side as

$$y_{ijk} - \hat{\tau}_i = y_{ijk} - \bar{y}_{i\bullet}$$

That is, within each treatment group, the treatment mean is subtracted from each data value.* We've seen this before—in the context of a one-way ANOVA (Section 11.2) we called these deviations or residuals. Figure 11.6.4 is a plot of the deviations from the treatment means for the alfalfa data broken down by block. We can see that there is still a lot of structure in the data: The mean deviations in blocks 1 and 2 are greater than zero while blocks 3, 4, and 5 are below zero (corresponding to above average growth near the window and below average growth farther from the

*Here we write $\bar{y}_{i\bullet}$ rather than $\bar{y}_i$ to distinguish the treatment means from the block means $\bar{y}_{\bullet j}$.

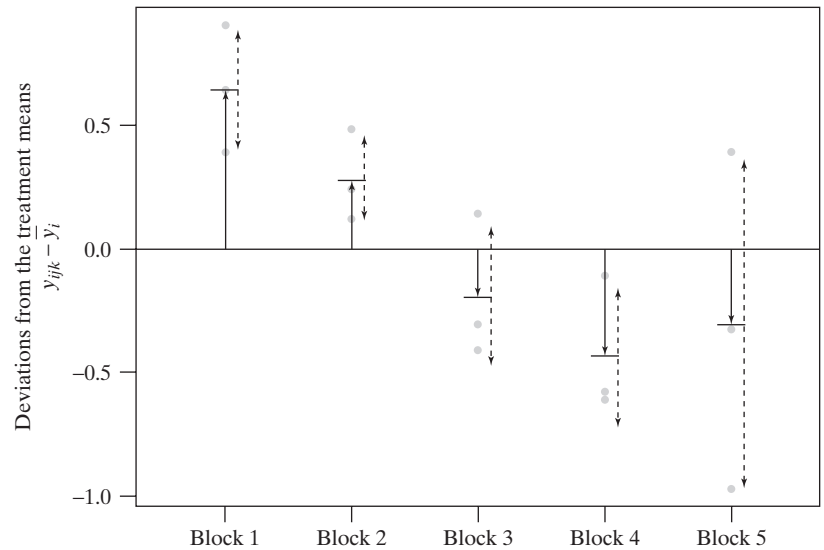


Figure 11.6.4 Deviations from the treatment means for the alfalfa growth data by blocks

window). The fact that these mean deviations are not all zero is a consequence of the variability due to the blocks. Soon we shall describe how to measure the variability of these mean deviations for the blocks through the **mean squares for blocks**, or **MS(blocks)**.

To visualize how accounting for this block-to-block variation improves our ability to detect treatment effects, consider the alfalfa and acid rain data graphed in Figure 11.6.5. Figure 11.6.5(a) displays the growth data for each treatment group and simply ignores the blocks entirely while Figure 11.6.5(b) displays the growth data after adjusting for the estimated block effects.* While the variability among the

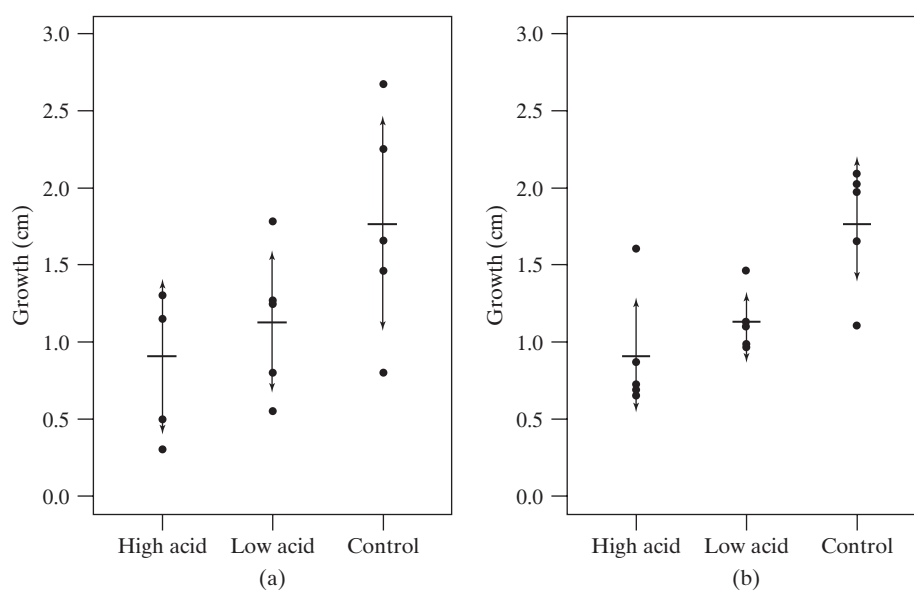


Figure 11.6.5 Visualizing the effect of blocking when comparing mean growth under the three acid treatments in the alfalfa experiment. Plot (a) displays the raw growth data while (b) displays the growth data after adjusting for the estimated block effects. Treatment means are indicated by horizontal lines and within-group standard deviations by arrows

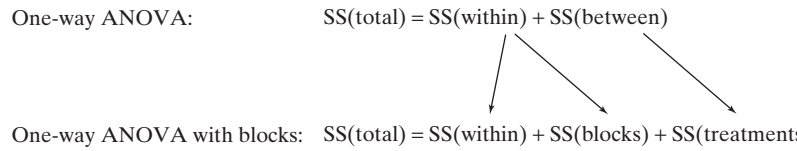
*To account for the blocking, the adjusted growth data on the y-axis for each treatment group is computed as $y_{ijk} - \bar{y}_{\cdot j}$.

treatment means is unchanged between the plots, we observe that the variability within the treatment groups is much smaller after accounting for the blocks and thus the differences among the treatments are more pronounced.

The One-Way Randomized Complete Block *F* Test

Recall that the ANOVA *F* test is a ratio that compares the variability among the treatment means to the within-group variability. As seen in Figure 11.6.5, accounting for the blocks has reduced the within-group variability and will thus increase the *F* statistic value. We now briefly discuss the computations involved in computing the ANOVA table for the randomized complete block *F* test.

In Section 11.2 for a one-way ANOVA, we discussed how the total sum of squares, SS(total), is broken down into SS(between), which measures variability attributed to differences among the treatment means, and SS(within), which measures unexplained random variation in the data. For a randomized blocks experiment, we write SS(treatments) rather than SS(between) to describe the variability between treatment means to be clear that we're measuring variability between treatments and not blocks. For a randomized blocks experiment we also split the one-way ANOVA SS(within) into two parts: SS(blocks), which measures variability among the block means, and SS(within), which measures the remaining unexplained variation in the data. Thus, we have



Usually we are not interested in testing a hypothesis about the blocks, but nonetheless we want to take into consideration the effect that blocking has on the response variable. Refining the one-way ANOVA by calculating SS(blocks) accomplishes this goal and furthermore, if blocks are chosen wisely, can lead to more powerful tests.

Computing the sums of squares is typically left to a computer and rarely performed by hand. Nonetheless, the formulas are worth noting as they mathematically reveal how the blocks are being accounted for.

The **mean squares between blocks** is calculated in a manner similar to our computation of MS(between) from the one-way ANOVA of Section 11.2. Roughly speaking, we compute a sort of weighted variance of the block means in which we weight the differences between a block mean and the overall mean by the block sample size. If we define the average of the observations in block j to be $\bar{y}_{\bullet j}$ and we let m_j denote the number of observations in block j, then the mean squares due to blocks is defined as follows:

Mean Squares between Blocks

$$\text{MS(blocks)} = \frac{\sum_{j=1}^{J} m_j(\bar{y}_{\bullet j} - \bar{\bar{y}})^2}{J - 1}$$

Analogous to our formulas in Section 11.2 we define SS(blocks) and df(blocks) as the numerator and denominator of MS(blocks) as follows:

Sum of Squares and df between Blocks

$$\text{SS(blocks)} = \sum_{j=1}^{J} m_j(\bar{y}_{\bullet j} - \bar{\bar{y}})^2$$

$$\text{df(blocks)} = J - 1$$

As noted previously, the blocking reduces MS(within). To compute MS(within) for the randomized complete block experiment we compute

$$\text{SS(within)} = \text{SS(total)} - \text{SS(treatment)} - \text{SS(blocks)}$$

where SS(treatment) and SS(total) are computed as in Section 11.2. As sums of squares are always nonnegative values, the preceding formula shows directly how the blocks reduce the within-group variability.

Similarly, to compute df(within) for the randomized complete block experiment, we have

$$\begin{aligned}\text{df(within)} &= \text{df(total)} - \text{df(treatment)} - \text{df(blocks)}\\ &= (n_\bullet - 1) - (I - 1) - (J - 1)\\ &= n_\bullet - I - J + 1\end{aligned}$$

Example 11.6.8

Alfalfa and Acid Rain For the alfalfa growth data in Table 11.6.2, the total of all the observations is $1.30 + 1.15 + \cdots + 0.80 = 19.04$ and the grand mean is

$$\bar{\bar{y}} = \frac{19.04}{15} = 1.269$$

We calculate

$$\text{SS(treatments)} = 5(0.910 - 1.269)^2 + 5(1.130 - 1.269)^2 + 5(1.768 - 1.269)^2 = 1.986$$

Since $I = 3$, we have

$$\text{df(treatments)} = 3 - 1 = 2$$

so that

$$\text{MS(treatments)} = \frac{1.986}{2} = 0.993$$

We calculate

$$\begin{aligned}\text{SS(blocks)} &= 3(1.917 - 1.269)^2 + 3(1.550 - 1.269)^2\\ &\quad + 3(1.077 - 1.269)^2 + 3(1.837 - 1.269)^2\\ &\quad + 3(1.967 - 1.269)^2\\ &= 2.441\end{aligned}$$

Since $J = 5$, we have

$$\text{df(blocks)} = 5 - 1 = 4$$

and

$$\text{MS(blocks)} = \frac{2.441}{4} = 0.610$$

The total sum of squares is found as $(1.30 - 1.269)^2 + \cdots + (0.80 - 1.269)^2 = 5.879$.

By subtraction, we compute SS(within):

$$\begin{aligned}\text{SS(within)} &= \text{SS(total)} - \text{SS(treatments)} - \text{SS(blocks)} \\ &= 5.879 - 1.986 - 2.441 = 1.452\end{aligned}$$

Similarly, we compute df(within) as

$$\text{df(within)} = \text{df(total)} + \text{df(treatments)} + \text{df(blocks)}$$

which in this case gives us $14 - 2 - 4 = 8$.

$$\text{Thus, MS(within)} = \frac{1.452}{8} = 0.182.$$ ■

The sums of squares, degrees of freedom, and resulting mean squares are collected in an expanded ANOVA table, which includes a line for the effect of the blocks.

To test the null hypothesis, we calculate

$$F_s = \frac{\text{MS(treatments)}}{\text{MS(within)}}$$

and reject H_0 if the P-value is too small.

Example 11.6.9

Alfalfa and Acid Rain For the alfalfa growth data of Example 11.6.1, the ANOVA summary is given in Table 11.6.4. The F statistic is $0.993/0.182 = 5.47$, with degrees of freedom 2 for the numerator and 8 for the denominator. From Table 10 we bracket the P-value as $0.02 < P\text{-value} < 0.05$. (Using a computer gives P-value $= 0.0318$.) The P-value is small, indicating that the differences between the three sample means are greater than would be expected by chance alone. There is significant evidence that acid affects the growth of alfalfa plants. (It is worth noting that if we ignore the blocks and conduct an erroneous one-way ANOVA, we would find P-value $= 0.0842$, which would not provide significant evidence for an acid effect at $\alpha = 0.05$). ■

Table 11.6.4 ANOVA table for alfalfa experiment

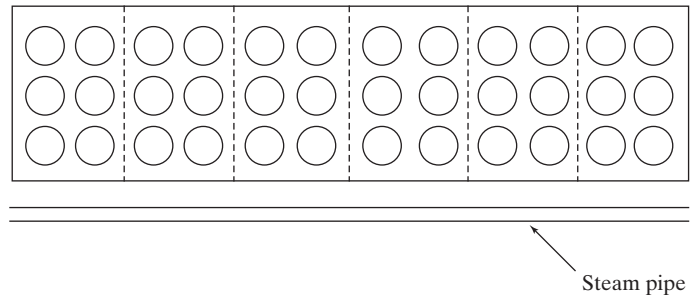
Source	df	SS	MS	F ratio
Between treatments	2	1.986	0.993	5.47
Between blocks	4	2.441	0.610	
Within groups	8	1.452	0.182	
Total	14	4.278		

Exercises 11.6.1–11.6.10

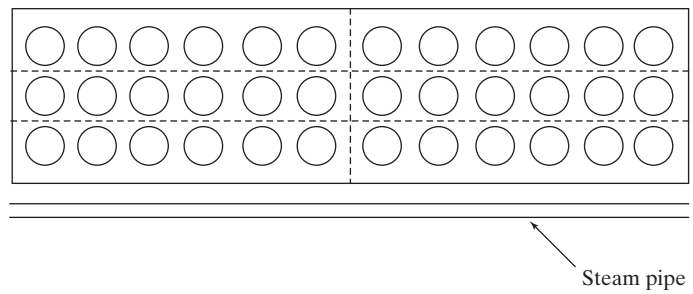
(*Note:* In several of these exercises you are asked to prepare a randomized allocation. For this purpose you can use either Table 1, random digits from your calculator, or a computer.)

11.6.1 In an experiment to compare six different fertilizers for tomatoes, 36 individually potted seedlings are to be used, 6 to receive each fertilizer. The tomato plants will be grown in a greenhouse, and the total yield of tomatoes will be observed for each plant. The experimenter has decided to use a randomized blocks design: The pots are to be arranged in six blocks of 6 plants each on the greenhouse bench. Two possible arrangements of the blocks are shown in the accompanying figure.

Arrangement I:



Arrangement II:



One factor that affects tomato yield is temperature, which cannot be held exactly constant throughout the greenhouse. In fact, a temperature gradient across the bench is likely. Heat for the greenhouse is provided by a steam pipe that runs lengthwise under one edge of the bench, and so the side of the bench near the steam pipe is likely to be warmer.

(a) Which arrangement of blocks (I or II) is better? Why?

(b) Prepare a randomized allocation of treatments to the pots within each block. (Refer to Example 11.6.4 as a guide; assume that the assignments of seedlings to pots and of pots to positions within the block have already been made.)

11.6.2 An experiment on vitamin supplements is to be conducted on young piglets, using litters as blocks in a randomized blocks design. There will be five treatments: four types of supplement and a control. Thus, five piglets from each litter will be used. The experiment will include five litters. Prepare a randomized blocks allocation of piglets to treatments. (Refer to Example 11.6.4 as a guide.)

11.6.3 Refer to the vitamin experiment of Exercise 11.6.2. Suppose a colleague of the experimenter proposes an alternative design: All pigs in a given litter are to receive the same treatment, with the five litters being randomly allocated to the five treatments. He points out that his proposal would save labor and greatly simplify the record keeping. If you were the experimenter, how would you reply to this proposal?

11.6.4 In a pharmacological experiment on eating behavior in rats, 18 rats are to be randomly allocated to three treatment groups: T_1, T_2, and T_3. While under observation, the animals will be kept in individual cages in a rack. The rack has three tiers with six cages per tier. In spite of efforts to keep the lighting uniform, the lighting conditions vary somewhat from one tier to another (the bottom tier is darkest), and the experimenter is concerned about this because lighting is thought to influence eating behavior in rats. The following three plans are proposed for allocating the rats to positions in the rack (to be done after the allocation of rats to treatment groups):

Plan I. Randomly allocate the 18 rats to the 18 positions in the rack.

Plan II. Put all T_1 rats on the first tier, all T_2 rats on the second, and all T_3 rats on the third tier.

Plan III. On each tier, put two T_1 rats, two T_2 rats, and two T_3 rats.

Put these three plans in order, from best to worst. Explain your reasoning.

11.6.5 An experimenter is planning an agricultural field experiment to compare the yields of 25 varieties of corn.

She will use a randomized blocks design with six blocks; thus, there will be 150 plots, and the yield of each plot must be measured. The experimenter realizes that the time required to harvest and weigh all the plots is so long that rain might interrupt the operation. If rain should intervene, there could be a yield difference between the harvests before and after the rain. The experimenter is considering the following plans.

Plan I. Harvest all plots of variety 1 first, all of variety 2 next, and so on.

Plan II. Harvest all plots of block 1 first, all of block 2 next, and so on.

Which plan is better? Why?

11.6.6 For an experiment to compare two methods of artificial insemination in cattle, the following cows are available:

Heifers (14–15 months old): 8 animals

Young cows (2–3 years old): 8 animals

Mature cows (4–8 years old): 10 animals

The animals are to be randomly allocated to the two treatment groups, using the three age groups as blocks. Prepare a suitable allocation, randomly dividing each stratum into two equal groups.

11.6.7 True or false (and say why): The primary reason for using a randomized blocks design in an experiment is to reduce bias.

11.6.8 In an experiment to understand the impact of fish grazing on invertebrate populations in streams, researchers established nine observation channels in three streams (three channels per stream). Each of the three channels within a stream received one of three treatments: No fish were added, Galaxias fish were added, or Trout fish were added. (The channels were constructed with mesh to prevent fish from entering or leaving.) Twelve days after establishing the channels, the number of *Deleatidium* mayfly nymphs present in a specified region in the center of the channel were counted. The number of nymphs for each treatment in each creek follows.[15]

		CREEK		
		A	B	C
Treatment	No Fish	11	8	7
	Galaxias	9	4	4
	Trout	6	4	0

(a) Identify the blocking, treatment (i.e., the explanatory variable of interest), and response variables in this study.

(b) In the context of this problem, explain to someone who has never taken a statistics course how blocking may help better identify treatment differences should they exist.

11.6.9 (Continuation of 11.6.8)

(a) The accompanying table is an (improper) ANOVA table for the data in Exercise 11.6.8. This analysis does not account for the blocking that was performed in the experiment. Based on this analysis, is there evidence that fish affect the number of mayfly nymphs present in the channels? Use $\alpha = 0.05$.

	DF	SUM SQ	MEAN SQ	*F* VALUE
Between groups	2	42.889	21.444	2.924
Within groups	6	44.000	7.333	
Total	8	86.889		

(b) The proper ANOVA table for the data, which accounts for blocking, follows. Based on this proper analysis, is there evidence that fish affect the number of mayfly nymphs present in the channels? Use $\alpha = 0.05$.

	DF	SUM SQ	MEAN SQ	*F* VALUE
Between groups	2	42.889	21.444	16.783
Between blocks	2	38.889	19.444	15.217
Within groups	4	5.111	1.278	
Total	8	86.889		

(c) Compute and compare s_{pooled} using the ANOVA table from parts (a) and (b). Why is one estimate larger than the other? What is s_{pooled} measuring in part (a)? In part (b)?

11.6.10 Consider the experiment described in Exercise 11.6.8. In addition to measuring the number of mayfly nymphs at the end of 12 days, stones of the same size were removed from each channel and the algal ash free dry mass (mg/cm^2) was measured for each of nine stones. These data produced SS(blocks) = 0.889, SS(within) = 0.444, and SS(total) = 2.889.

(a) Construct an ANOVA table similar to Table 11.6.4 to summarize these data.

(b) Is there evidence that the presence or type of fish is associated with the mean algal ash free dry mass in the channels? Use $\alpha = 0.05$.

(c) Can a causal conclusion be drawn from the analysis performed in part (b) based on these data? If so, what causal conclusion can be made? If not, explain why no causal conclusion is appropriate.

11.7 Two-Way ANOVA

Factorial ANOVA

In a typical analysis of variance application there is a single explanatory variable or **factor** under study. For example, in the weight gain setting of Example 11.2.1, the factor is "type of diet," which takes on three **levels**: diet 1, diet 2, and diet 3. However, some analysis of variance settings involves the simultaneous study of two or more factors. The following is an example.

Example 11.7.1

Growth of Soybeans A plant physiologist investigated the effect of mechanical stress on the growth of soybean plants. Individually potted seedlings were randomly allocated to four treatment groups of 13 seedlings each. Seedlings in two groups were stressed by shaking for 20 minutes twice daily, while two control groups were not stressed. Thus, the first factor in the experiment was presence or absence of stress, with two levels: control or stress. Also, plants were grown in either low or moderate light. Thus, the second factor was amount of light, with two levels: low light or moderate light. This experiment is an example of a 2×2 *factorial experiment*; it includes four treatments:

Treatment 1: Control, low light
Treatment 2: Stress, low light
Treatment 3: Control, moderate light
Treatment 4: Stress, moderate light

After 16 days of growth, the plants were harvested, and the total leaf area (cm^2) of each plant was measured. The results are given in Table 11.7.1 and plotted in Figure 11.7.1.[16]

Table 11.7.1 Leaf area (cm^2) of soybean plants

	Treatment			
	Control, low light	Stress, low light	Control, moderate light	Stress, moderate light
	264	235	314	283
	200	188	320	312
	225	195	310	291
	268	205	340	259
	215	212	299	216
	241	214	268	201
	232	182	345	267
	256	215	271	326
	229	272	285	241
	288	163	309	291
	253	230	337	269
	288	255	282	282
	230	202	273	257
Mean	245.3	212.9	304.1	268.8
SD	27.0	29.7	26.9	35.2
n	13	13	13	13

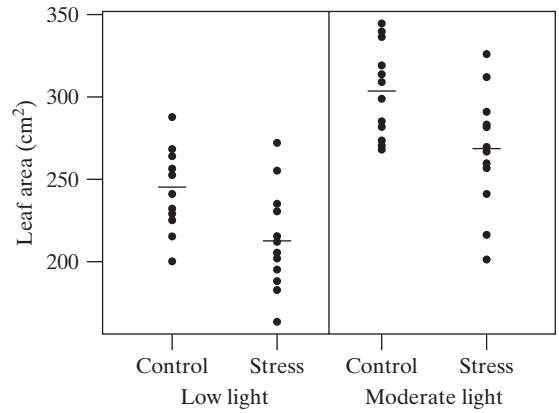


Figure 11.7.1 Leaf area of soybean plants receiving four different treatments. Group means indicated by (–)

There is evidence in Figure 11.7.1 that stress reduces leaf area. This is true under low light and under moderate light. Likewise, moderate light increases leaf area, whether or not the seedlings are stressed. ■

A model for this setting is

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

where y_{ijk} is the kth observation of level i of the first factor and level j of the second factor. The term τ_i represents the effect of level i of the first factor (stress condition in Example 11.7.1) and now the term β_j represents the effect of level j of the second factor (light condition in Example 11.7.1).

When studying two factors within a single experiment it helps to organize the sample means in a table that reflects the structure of the experiment and to present the means in a graph that features this structure.

Example 11.7.2

Growth of Soybeans Table 11.7.2 summarizes the data of Example 11.7.1. For example, when the first factor is at its first level (control) and the second factor is at its first level (low light), the sample mean is $\bar{y}_{11} = 245.3$. The format of this table permits us easily to consider the two factors—stress condition and light condition—separately and together. The last column shows the effect of light at each stress level. The numbers in this column confirm the visual impression of Figure 11.7.1: Moderate light increases average leaf area by roughly the same amount when the seedlings are stressed as it does when they are not stressed. Likewise, the last row (−32.4 versus −35.3) shows that the effect of stress is roughly the same at each level of light. ■

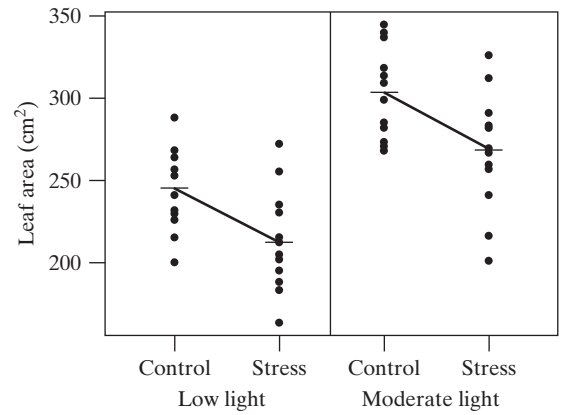
Table 11.7.2 Mean leaf areas for soybean experiment

		Light condition		
		Low light	Moderate light	Difference
Shaking condition	Control	245.3	304.1	58.8
	Stress	212.9	268.8	55.9
	Difference	−32.4	−35.3	

If the joint influence of two factors is equal to the sum of their separate influences, the two factors are said to be **additive** in their effects. For instance, consider the soybean experiment of Example 11.7.1. If stress reduces mean leaf area by the same

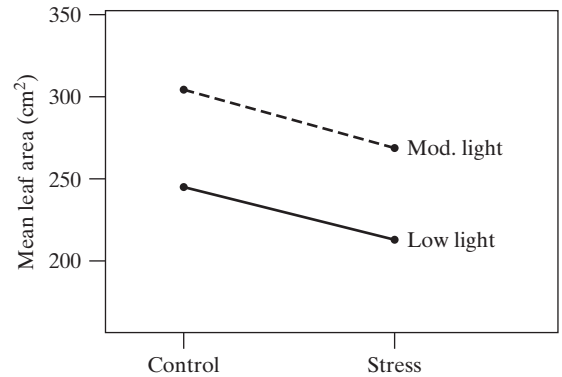
amount in either light condition, then the effect of stress (a negative effect in this case) is *added* to the effect of light. To visualize this additivity of effects, consider Figure 11.7.2, which shows the data with the four treatment means. The solid lines connecting treatment means are almost parallel because the data display a pattern of nearly perfect additivity.*

Figure 11.7.2 Data and treatment means for soybean experiment



When the effects of factors are additive we say that there is no **interaction** between the factors. A graph that displays only the treatment means is often called an interaction graph. Figure 11.7.3, which is a summary version of Figure 11.7.2, is an interaction graph highlighting the effect of stress on mean leaf area for the two light conditions. Analogous graphs can be made to draw the focus to comparing the effect of light on mean leaf area for the two stress conditions.

Figure 11.7.3 Interaction graph for soybean experiment



Sometimes the effect that one factor has on a response variable depends on the level of a second factor. When this happens we say that the two factors interact in their effect on the response. The following is an example.

*The difference between the mean leaf area for stress under low light (212.9) and the mean leaf area for control under low light of (245.3) is called the **simple effect** of shaking under low light. Thus, the simple effect of shaking under low light is $212.9 - 245.3 = -32.4$. Likewise, the simple effect of shaking under moderate light is $268.8 - 304.1 = -35.3$. A **main effect** is an average of simple effects. For example, the main effect of shaking is $(-32.4 + -35.3)/2 = -33.85$. The main effect of light is $(58.8 + 55.9)/2 = 57.35$.

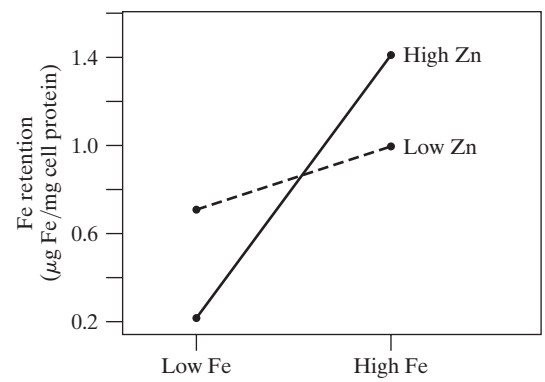
Example 11.7.3

Iron Supplements in Milk-Based Fruit Beverages Iron and zinc fortification of milk-based fruit drinks are common practice. To better understand the effects of drink fortification on the cellular retention of iron, researchers conducted an experiment by fortifying milk-based fruit drinks with low and high levels of iron (Fe) and zinc (Zn). The drinks were digested in a simulated gastrointestinal tract and cellular iron retention was measured (μg Fe/mg cell protein). Table 11.7.3 summarizes the data, which included eight observations for each combination of Fe and Zn supplementation levels.[17] Figure 11.7.4 is an interaction graph showing the four means. Note that when the Zn supplementation level is low, the effect of the Fe supplementation on cellular retention is much smaller than when the Zn supplementation level is high (i.e., the slopes of the two lines differ—the lines are not parallel). Thus, the effect of Fe supplementation on mean cellular retention depends on the amount of Zn supplementation used. We say that Fe and Zn interact in their effects on cellular retention. ■

Table 11.7.3 Mean iron retention (μg Fe/mg cell protein) for drink supplement experiment

		Zn Level		
		Lo	Hi	Difference
Fe	Lo	0.707	0.215	−0.492
Level	Hi	0.994	1.412	0.418
	Difference	0.287	1.197	

Figure 11.7.4 Interaction graph for drink supplementation experiment



When we suspect that two factors interact in an ANOVA setting, we can extend our model by adding an interaction term:

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

Here the term γ_{ij} is the effect of the interaction between level i of the first factor and level j of the second factor. As before, if there are $n_\bullet$ total observations, then df(total) $= n_\bullet - 1$. If there are I levels of the first factor, then it has $I - 1$ degrees of freedom. Likewise, if there are J levels of the second factor, then it has $J - 1$ degrees of freedom. There are $(I - 1) \times (J - 1)$ interaction degrees of freedom. With I levels of the first factor and J levels of the second factor there are IJ treatment combinations. Thus, df(within) $= n_\bullet - IJ$.*

*This is analogous to the definition of df(within) $= n_\bullet - I$ for one-way ANOVA from Section 11.2. In each setting df(within) = total number of observations − number of treatments.

A null hypothesis of interest is that all interaction terms are zero:

$$H_0: \gamma_{11} = \gamma_{12} = \cdots = \gamma_{IJ} = 0$$

To test this null hypothesis we calculate

$$F_s = \frac{\text{MS(interaction)}}{\text{MS(within)}}$$

and reject H_0 if the P-value is too small.

Example 11.7.4

Iron Supplements in Milk-Based Fruit Beverages Table 11.7.4 shows the analysis of variance results for the drink supplement experiment of Example 11.7.3. This table includes a line for the interaction term.* There were eight observations at each combination of Fe and Ze supplementation level; thus $n_{\cdot} = 32$ and df(total) $= 31$. In this example $I = J = 2$, so df(Fe levels) $=$ df(Zn levels) $=$ df(interaction) $= 1$. We can find df(within) by subtraction: df(within) $= 31 - 1 - 1 - 1 = 28$. (This agrees with the formula df(within) $= n_{\cdot} - IJ = 32 - 2 \times 2$.)

To test whether Fe and Zn supplementation levels interact we use the F ratio $1.6555/0.0019 = 871.3$, which has degrees of freedom 1 for the numerator and 28 for the denominator. From Table 10 we bracket the P-value as P-value < 0.0001. The P-value is extremely small, indicating that the interaction pattern seen in Figure 11.7.4 is more pronounced than would be expected by chance alone. Thus, we reject H_0. ■

Table 11.7.4 ANOVA table for drink supplement experiment

Source	df	SS	MS	F ratio
Between Fe levels	1	4.4023	4.4023	2317.0
Between Zn levels	1	0.0109	0.0109	5.736
Interaction	1	1.6555	1.6555	871.3
Within groups	28	0.0523	0.0019	
Total	31	6.1210		

The concept of interaction occurs throughout biology. The terms "synergism" and "antagonism" describe interactions between biological agents. The term "epistasis" describes interaction between genes at two loci.

When interactions are present, as in Example 11.7.3, the main effects of factors don't have their usual interpretations. Regarding Example 11.7.3, it is difficult to state the independent effect of Fe because the nature and magnitude of the effect depends on the particular level of Zn supplementation. Because of this, we usually test for the presence of interactions first. If interactions are present, as in the drink supplementation example, then we often stop the analysis at this stage. If no evidence for an interaction effect is found (that is, if we do not reject H_0), then we proceed to testing the main effects of the individual factors. The following example illustrates this process.

*The ANOVA formulas that are used to calculate the sum of squares due to interaction are rather messy and aren't presented here. In particular, it matters whether or not the design is "balanced." The drink supplementation experiment is balanced in that there are eight observations in each of the four combinations of factor levels shown in Table 11.7.3. However, unbalanced designs, which lead to complicated calculations and analyses, are possible. We rely here on computer software to calculate the necessary sums of squares.

Example 11.7.5 Growth of Soybeans Table 11.7.5 is an analysis of variance table for the soybean growth data of Example 11.7.1. The null hypothesis

$$H_0: \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = 0$$

is tested with the F ratio

$$F_s = \frac{\text{MS(interaction)}}{\text{MS(within)}} = \frac{26.3}{895.34} = 0.029$$

Looking in Table 10 with degrees of freedom 1 and 12, we see that the P-value is greater than 0.20; thus there is no significant evidence for an interaction and we do not reject H_0.

Since there is no evidence of interactions, we test the main effect of stress level. Here the F ratio is

$$F_s = \frac{\text{MS(between stress levels)}}{\text{MS(within)}} = \frac{14858.5}{895.34} = 16.6$$

This is highly significant (i.e., the P-value is very small) and we reject H_0.

Likewise, the test for the main effect of light levels has an F ratio of

$$F_s = \frac{\text{MS(between light levels)}}{\text{MS(within)}} = \frac{42751.6}{895.34} = 47.75$$

Again, this is highly significant and we reject H_0. ■

Table 11.7.5 ANOVA table for soybean growth experiment

Source	df	SS	MS	F ratio
Between stress levels	1	14858.5	14858.5	16.60
Between light levels	1	42751.6	42751.6	47.75
Interaction	1	26.3	26.3	0.029
Within groups	48	42976.3	895.34	
Total	51	100612.7		

Interaction graphs can be used when there are more than two levels for a factor, as in the next example.

Example 11.7.6 Toads Researchers studied the effect that exposure to ultraviolet-B radiation has on the survival of embryos of the western toad *Bufo boreas*. They conducted an experiment in which several *B. borea* embryos were placed at one of three water depths—10 cm, 50 cm, or 100 cm—and one of two radiation settings—exposed to UV-B radiation or shielded. The response variable was the percentage of embryos surviving to hatching. Table 11.7.6 summarizes the data, which

Table 11.7.6 Percent embryos surviving for toads experiment

		UV-B		
		Exposed	Shielded	Difference
Water depth	10 cm	0.425	0.759	0.334
	50 cm	0.729	0.748	0.019
	100 cm	0.785	0.766	−0.019

included four observations at each combination of depth and UV-B exposure. Figure 11.7.5 is an interaction graph showing the six means. The presence of interactions here is readily apparent. Table 11.7.7 summarizes the analysis of variance.[18] ■

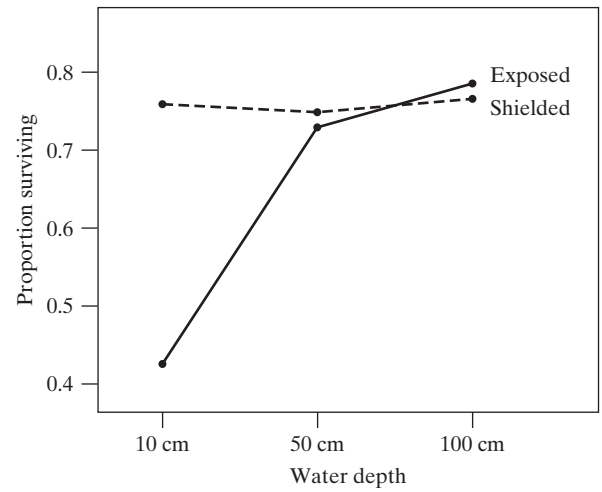


Figure 11.7.5 Interaction graph for toad experiment

Table 11.7.7 ANOVA table for toad experiment

Source	df	SS	MS	*F* ratio
Between water depths	2	0.150676	0.075338	13.92
Between UV-B levels	1	0.074371	0.074371	13.74
Interaction	2	0.150185	0.075093	13.88
Within groups	18	0.097401	0.005411	
Total	23	0.472633		

The topic of interactions is also discussed in Section 11.8.

Exercises 11.7.1–11.7.6

11.7.1 A plant physiologist investigated the effect of flooding on root metabolism in two tree species: flood-tolerant river birch and the intolerant European birch. Four seedlings of each species were flooded for one day and four were used as controls. The concentration of adenosine triphosphate (ATP) in the roots of each plant was measured. The data (nmol ATP per mg tissue) are shown in the table.[19]

	RIVER BIRCH		EUROPEAN BIRCH	
	FLOODED	CONTROL	FLOODED	CONTROL
	1.45	1.70	0.21	1.34
	1.19	2.04	0.58	0.99
	1.05	1.49	0.11	1.17
	1.07	1.91	0.27	1.30
Mean	1.19	1.785	0.2925	1.20

Prepare an interaction graph (like Figure 11.7.3).

11.7.2 Consider the data from Exercise 11.7.1. For these data, SS(species of birch) = 2.19781, SS(flooding) = 2.25751, SS(interaction) = 0.097656, and SS(within) = 0.47438.

(a) Construct the ANOVA table.

(b) Carry out an F test for interactions; use $\alpha = 0.05$.

(c) Test the null hypothesis that species has no effect on ATP concentration. Use $\alpha = 0.01$.

(d) Assuming that each of the four populations has the same standard deviation, use the data to calculate an estimate of that standard deviation.

11.7.3 A completely randomized double-blind clinical trial was conducted to compare two drugs, ticrynafen (T) and hydrochlorothiazide (H), for effectiveness in treatment of high blood pressure. Each drug was given at either a low or a high dosage level for six weeks. The accompanying table shows the results for the drop (baseline minus final value) in systolic blood pressure (mm Hg).[20]

	TICRYNAFEN (T)		HYDROCHLOROTHIAZIDE (H)	
	LOW DOSE	HIGH DOSE	LOW DOSE	HIGH DOSE
Mean	13.9	17.1	15.8	17.5
No. of patients	53	57	55	58

Prepare an interaction graph (like Figure 11.7.3).

11.7.4 Consider the data from Exercise 11.7.3. The difference in response between T and H appears to be larger for the low dose than for the high dose.

(a) Carry out an F test for interactions to assess whether this pattern can be ascribed to chance variation. Let $\alpha = 0.10$. For these data SS(interaction) = 31.33 and SS(within) = 30648.81.

(b) Based on your results in part (a), is it sensible to examine and interpret the main effects of drug and of dose?

11.7.5 Consider the data from Exercise 11.7.3. For these data, SS(drug) = 69.22, SS(dose) = 330.00, SS(interaction) = 31.33, and SS(within) = 30648.81.

(a) Construct the ANOVA table.

(b) Carry out a test of the null hypothesis that the effects of the two drugs (T and H) are equal. Let $\alpha = 0.05$.

11.7.6 In a study of lettuce growth, 36 seedlings were randomly allocated to receive either high or low light and to be grown in either a standard nutrient solution or one containing extra nitrogen. After 16 days of growth, the lettuce plants were harvested and the dry weight of the leaves was determined for each plant. The accompanying table shows the mean leaf dry weight (gm) of the 9 plants in each treatment group.[21]

	NUTRIENT SOLUTION	
	STANDARD	EXTRA NITROGEN
Low light	2.16	3.09
High light	3.26	4.48

For these data, SS(nutrient solution) = 10.4006, SS(light) = 13.95023, SS(interaction) = 0.18923, and SS(within) = 11.1392.

(a) Construct the ANOVA table.

(b) Carry out an F test for interactions; use $\alpha = 0.05$.

(c) Test the null hypothesis that nutrient solution has no effect on weight. Use $\alpha = 0.01$.

11.8 Linear Combinations of Means (Optional)

In many studies, interesting questions can be addressed by considering linear combinations of the group means. A **linear combination** L is a quantity of the form

$$L = m_1\bar{y}_1 + m_2\bar{y}_2 + \cdots + m_I\bar{y}_I$$

where the m's are multipliers of the $\bar{y}_i$'s.

Linear Combinations for Adjustment

One use of linear combinations is to "adjust" for an extraneous variable, as illustrated by the following example.

Example 11.8.1

Forced Vital Capacity One measure of lung function is forced vital capacity (FVC), which is the maximal amount of air a person can expire in one breath. In a public health survey, researchers measured FVC in a large sample of people. The results for male ex-smokers, stratified by age, are shown in Table 11.8.1.[22]

Table 11.8.1 FVC in male ex-smokers

	FVC (liters)		
Age (years)	n	Mean	SD
25–34	83	5.29	0.76
35–44	102	5.05	0.77
45–54	126	4.51	0.74
55–64	97	4.24	0.80
65–74	73	3.58	0.82
25–74	481	4.56	

Suppose it is desired to calculate a summary value for FVC in male ex-smokers. One possibility would be simply to calculate the grand mean of the 481 observed values, which is 4.56 liters. But the grand mean has a serious drawback: It cannot be meaningfully compared with other populations that may have different age distributions. For instance, suppose we were to compare ex-smokers with nonsmokers; the observed difference in FVC would be distorted because ex-smokers as a group are (not surprisingly) older than nonsmokers. A summary measure that does not have this disadvantage is the "age-adjusted" mean, which is an estimate of the mean FVC value in a reference population with a specified age distribution. To illustrate, we will use the reference distribution in Table 11.8.2, which is (approximately) the distribution for the entire U.S. population.[23]

Table 11.8.2 Age distribution in reference population

Age	Relative frequency
25–34	0.23
35–44	0.22
45–54	0.24
55–64	0.22
65–74	0.09

The "age-adjusted" mean FVC value is the following linear combination:

$$L = 0.23\bar{y}_1 + 0.22\bar{y}_2 + 0.24\bar{y}_3 + 0.22\bar{y}_4 + 0.09\bar{y}_5$$

Note that the multipliers (m's) are the relative frequencies in the reference population. From Table 11.8.1, the value of L is

$$\begin{aligned} L &= (0.23)(5.29) + (0.22)(5.05) + (0.24)(4.51) + (0.22)(4.24) + (0.09)(3.58) \\ &= 4.67 \text{ liters} \end{aligned}$$

This value is an estimate of the mean FVC in an idealized population of people who are biologically like male ex-smokers, but whose age distribution is that of the reference population. ■

Contrasts

A linear combination whose multipliers (m's) add to zero is called a **contrast**. The following example shows how contrasts can be used to describe the results of an experiment.

Example 11.8.2 Growth of Soybeans Table 11.8.3 shows the treatment means and sample sizes for the soybean growth experiment of Example 11.6.8. We can use contrasts to describe the effects of stress in the two temperature conditions.

Table 11.8.3 Soybean growth data

Treatment	Mean leaf area (cm^2)	n
1. Control, low light	245.3	13
2. Stress, low light	212.9	13
3. Control, moderate light	304.1	13
4. Stress, moderate light	268.8	13

(a) First, note that an ordinary pairwise difference is a contrast. For instance, to measure the effect of stress in low light we can consider the contrast

$$L = \bar{y}_1 - \bar{y}_2 = 245.3 - 212.9 = 32.4$$

For this contrast, the multipliers are $m_1 = 1, m_2 = -1, m_3 = 0, m_4 = 0$; note that they add to zero.

(b) To measure the effect of stress in moderate light we can consider the contrast

$$L = \bar{y}_3 - \bar{y}_4 = 304.1 - 268.8 = 35.3$$

For this contrast, the multipliers are $m_1 = 0, m_2 = 0, m_3 = 1, m_4 = -1$.

(c) To measure the overall effect of stress, we can average the contrasts in parts (a) and (b) to obtain the contrast

$$L = \frac{1}{2}(\bar{y}_1 - \bar{y}_2) + \frac{1}{2}(\bar{y}_3 - \bar{y}_4)$$
$$= \frac{1}{2}(32.4) + \frac{1}{2}(35.3) = 33.85$$

For this contrast, the multipliers are $m_1 = \frac{1}{2}, m_2 = -\frac{1}{2}, m_3 = \frac{1}{2}, m_4 = -\frac{1}{2}$. ■

Standard Error of a Linear Combination

Each linear combination L is an estimate, based on the $\bar{y}$'s, of the corresponding linear combination of the population means (μ's). As a basis for statistical inference, we need to consider the standard error of a linear combination, which is calculated as follows.

Standard Error of L

The standard error of the linear combination

$$L = m_1\bar{y}_1 + m_2\bar{y}_2 + \cdots + m_I\bar{y}_I$$

is

$$\text{SE}_L = s_{\text{pooled}}\sqrt{\sum_{i=1}^{I}\frac{m_i^2}{n_i}}$$

where $s_{\text{pooled}} = \sqrt{\text{MS(within)}}$ from the ANOVA.

The SE can be written explicitly as

$$\text{SE}_L = s_{\text{pooled}}\sqrt{\left(\frac{m_1^2}{n_1} + \frac{m_2^2}{n_2} + \cdots + \frac{m_I^2}{n_I}\right)}$$

If all the sample sizes (n_i) are equal, the SE can be written as

$$\text{SE}_L = s_{\text{pooled}}\sqrt{\frac{(m_1^2 + m_2^2 + \cdots + m_I^2)}{n}} = s_{\text{pooled}}\sqrt{\frac{1}{n}\sum_{i=1}^{I} m_i^2}$$

The following two examples illustrate the application of the standard error formula.

Example 11.8.3

Forced Vital Capacity For the linear combination L defined in Example 11.8.1, we find that

$$\sum_{i=1}^{I}\frac{m_i^2}{n_i} = \frac{0.23^2}{83} + \frac{0.22^2}{102} + \frac{0.24^2}{126} + \frac{0.22^2}{97} + \frac{0.09^2}{73}$$
$$= 0.0021789$$

The ANOVA for these data yields $s_{\text{pooled}} = \sqrt{0.59989} = 0.77453$. Thus, the standard error of L is

$$\text{SE}_L = 0.77453\sqrt{0.0021789} = 0.0362$$ ■

Example 11.8.4

Growth of Soybeans For the linear combination L defined in Example 11.8.2(a), we find that

$$\sum_{i=1}^{I} m_i^2 = (1)^2 + (-1)^2 + (0)^2 + (0)^2 = 2$$

so that

$$\text{SE}_L = s_{\text{pooled}}\sqrt{\frac{2}{13}}$$ ■

Confidence Intervals

Linear combinations of means can be used for testing hypotheses and for constructing confidence intervals. Critical values are obtained from Student's t distribution with

$$\text{df} = \text{df(within)}$$

from the ANOVA.* Confidence intervals are constructed using the familiar Student's t format. For instance, a 95% confidence interval is

$$L \pm t_{0.025}\text{SE}_L$$

The following example illustrates the construction of the confidence interval.

Example 11.8.5

Growth of Soybeans Consider the contrast defined in Example 11.8.2(c):

$$L = \frac{1}{2}(\bar{y}_1 - \bar{y}_2) + \frac{1}{2}(\bar{y}_3 - \bar{y}_4)$$

*This method of determining critical values does not take account of multiple comparisons. See Section 11.9.

This contrast is an estimate of the quantity

$$\lambda = \frac{1}{2}(\mu_1 - \mu_2) + \frac{1}{2}(\mu_3 - \mu_4)$$

which can be described as the true (population) effect of stress, averaged over the light conditions. Let us construct a 95% confidence interval for this true difference.

We found in Example 11.8.2 that the value of L is

$$L = 33.85$$

To calculate SE_L, we first calculate

$$\sum_{i=1}^{I} \frac{m_i^2}{n_i} = \frac{\left(\frac{1}{2}\right)^2}{13} + \frac{\left(-\frac{1}{2}\right)^2}{13} + \frac{\left(\frac{1}{2}\right)^2}{13} + \frac{\left(-\frac{1}{2}\right)^2}{13} = \frac{1}{13}$$

From the ANOVA, which is shown in Table 11.8.4, we find that $s_{\text{pooled}} = \sqrt{895.34} = 29.922$; thus,

$$\text{SE}_L = s_{\text{pooled}} \sqrt{\sum_{i=1}^{I} \frac{m_i^2}{n_i}} = 29.922 \sqrt{\frac{1}{13}} = 8.299$$

Table 11.8.4 ANOVA table for soybean growth experiment

Source	df	SS	MS	F ratio
Between stress depths	1	14858.5	14858.5	16.60
Between light levels	1	42751.6	42751.6	47.75
Interaction	1	26.3	26.3	0.029
Within groups	48	42976.3	895.34	
Total	51	100613		

From Table 4 with df $= 40 \approx 48$, we find $t_{40,0.025} = 2.021$. The confidence interval is

$$33.85 \pm (2.021)(8.299)$$

$$33.85 \pm 16.77$$

or (17.1, 50.6).

We are 95% confident that the effect of stress, averaged over the light conditions, is to reduce the leaf area by an amount whose mean value is between 17.1 cm^2 and 50.6 cm^2. ■

t Tests

To test the null hypothesis that the population value of a contrast is zero, the test statistic is calculated as

$$t_s = \frac{L}{\text{SE}_L}$$

and the t test is carried out in the usual way. The t test will be illustrated in Example 11.8.6.

Contrasts to Assess Interaction

Sometimes an investigator wishes to study the separate and joint effects of two or more factors on a response variable Y. In Section 11.7 the concept of interaction between two factors was introduced. Linear contrasts provide another way to study such interactions. The following is an example.

Example 11.8.6

Growth of Soybeans In the soybean growth experiment (Example 11.6.8 and Example 11.8.2), the two factors of interest are stress condition and light level. Table 11.8.5 shows the treatment means, arranged in a new format that permits us easily to consider the factors separately and together.

Table 11.8.5 Mean leaf areas for soybean experiment

		Light condition		
		Low light	Moderate light	Difference
Shaking condition	Control	245.3 (1)	304.1 (3)	58.8
	Stress	212.9 (2)	268.8 (4)	55.9
	Difference	−32.4	−35.3	

At each light level, the mean effect of stress can be measured by a contrast:

Effect of stress in low light: $\bar{y}_2 - \bar{y}_1 = 212.9 - 245.3 = -32.4$
Effect of stress in moderate light: $\bar{y}_4 - \bar{y}_3 = 268.8 - 304.1 = -35.3$

Now consider the question: Is the reduction in leaf area due to stress the same in both light conditions? One way to address this question is to compare $(\bar{y}_2 - \bar{y}_1)$ versus $(\bar{y}_4 - \bar{y}_3)$; the difference between these two values is a contrast:

$$\begin{aligned} L &= (\bar{y}_2 - \bar{y}_1) - (\bar{y}_4 - \bar{y}_3) \\ &= -32.4 - (-35.3) = 2.9 \end{aligned}$$

This contrast L can be used as the basis for a confidence interval or a test of hypothesis. We illustrate the test. The null hypothesis is

$$H_0\text{: } (\mu_2 - \mu_1) = (\mu_4 - \mu_3)$$

or, in words,

H_0: The effect of stress is the same in the two light conditions.

For the preceding L, $\sum_{i=1}^{I} \frac{m_i^2}{n_i} = \frac{4}{13}$, and the standard error is

$$\text{SE}_L = s_{\text{pooled}} \sqrt{\sum_{i=1}^{I} \frac{m_i^2}{n_i}} = s_{\text{pooled}} \sqrt{\left(\frac{4}{13}\right)} = 29.922 \sqrt{\frac{4}{13}} = 16.6$$

The test statistic is

$$t_s = \frac{2.9}{16.6} = 0.2$$

From Table 4 with df = 40 we find $t_{40,0.20} = 1.303$. The data provide virtually no evidence that the effect of stress is different in the two light conditions. This is consistent with the F test for interactions conducted in Example 11.7.5. ■

The statistical definition of interaction introduced in Section 11.7 and viewed through the lens of contrasts here is rather specialized. It is defined in terms of the observed variable rather than in terms of a biological mechanism. Further, interaction as measured by a contrast is defined by *differences* between means. In some applications the biologist might feel that ratios of means are more meaningful or relevant than differences. The following example shows that the two points of view can lead to different answers.

Example 11.8.7

Chromosomal Aberrations A research team investigated the separate and joint effects in mice of exposure to high temperature (35 °C) and injection with the cancer drug cyclophosphamide (CTX). A completely randomized design was used, with eight mice in each treatment group. For each animal, the researchers measured the incidence of a certain chromosomal aberration in the bone marrow; the result is expressed as the number of abnormal cells per 1,000 cells. The treatment means are shown in Table 11.8.6.[24]

Table 11.8.6 Mean incidence of chromosomal aberrations following various treatments

		Injection	
		CTX	None
Temperature	Room	23.5	2.7
	High	75.4	20.9

Is the observed effect of CTX greater at room temperature or at high temperature? The answer depends on whether "effect" is measured absolutely or relatively.

Measured as a difference, the effect of CTX is

$$\textit{Room temperature:}\quad 23.5 - 2.7 = 20.8$$
$$\textit{High temperature:}\quad 75.4 - 20.9 = 54.5$$

Thus, the absolute effect of CTX is greater at the high temperature. However, this relationship is reversed if we express the effect of CTX as a ratio rather than as a difference:

$$\textit{Room temperature:}\quad \frac{23.5}{2.7} = 8.70$$
$$\textit{High temperature:}\quad \frac{75.4}{20.9} = 3.61$$

At room temperature CTX produces almost a ninefold increase in chromosomal aberrations, whereas at high temperature the increase is less than fourfold; thus, in relative terms, the effect of CTX is much greater at room temperature. ■

If the phenomenon under study is thought to be multiplicative rather than additive, so that relative rather than absolute change is of primary interest, then ordinary contrasts should not be used. One simple approach in this situation is to use a logarithmic transformation—that is, to compute $Y' = \log(Y)$, and then analyze Y' using contrasts. The motivation for this approach is that relations of constant *relative* magnitude in the Y scale become relations of constant *absolute* magnitude in the Y' scale.

Exercises 11.8.1–11.8.10

11.8.1 Refer to the FVC data of Example 11.8.1.

(a) Verify that the grand mean of all 481 FVC values is 4.56.

(b) Taking into account the age distribution among the 481 subjects and the age distribution in the U.S. population, explain intuitively why the grand mean (4.56 liters) is smaller than the age-adjusted mean (4.67 liters).

11.8.2 To see if there is any relationship between blood pressure and childbearing, researchers examined data from a large health survey. The following table shows the data on systolic blood pressure (mm Hg) for random samples from two populations of women: women who had borne no children and women who had borne five or more children. The pooled standard deviation from all eight groups was $s_{\text{pooled}} = 18$ mm Hg.[25]

	NO CHILDREN		FIVE OR MORE CHILDREN	
AGE	MEAN BLOOD PRESSURE	NO. OF WOMEN	MEAN BLOOD PRESSURE	NO. OF WOMEN
18–24	113	230	114	7
25–34	118	110	116	82
35–44	125	105	124	127
45–54	134	123	138	124
18–54	121	568	127	340

Carry out age adjustment, as directed, using the following reference distribution, which is the approximate distribution for U.S. women:[26]

AGE	RELATIVE FREQUENCY
18–24	0.17
25–34	0.29
35–44	0.31
45–54	0.23

(a) Calculate the age-adjusted mean blood pressure for women with no children.

(b) Calculate the age-adjusted mean blood pressure for women with five or more children.

(c) Calculate the difference between the values obtained in parts (a) and (b). Explain intuitively why the result is smaller than the unadjusted difference of 127 − 121 = 6 mg Hg.

(d) Calculate the standard error of the value calculated in part (a).

(e) Calculate the standard error of the value calculated in part (c).

11.8.3 Refer to the ATP data of Exercise 11.7.1. The sample means and standard deviations are as follows:

	RIVER BIRCH		EUROPEAN BIRCH	
	FLOODED	CONTROL	FLOODED	CONTROL
$\bar{y}$	1.19	1.78	0.29	1.20
s	0.18	0.24	0.20	0.16

Define linear combinations (that is, specify the multipliers) to measure each of the following:

(a) The effect of flooding in river birch

(b) The effect of flooding in European birch

(c) The difference between river birch and European birch with respect to the effect of flooding (that is, the interaction between flooding and species)

11.8.4 (*Continuation of Exercise 11.8.3*)

(a) Use a t test to investigate whether flooding has the same effect in river birch and in European birch. Use a nondirectional alternative and let $\alpha = 0.05$. (The pooled standard deviation is $s_{\text{pooled}} = 0.199$.)

(b) If the sample sizes were $n = 10$ rather than $n = 4$ for each group, but the means, standard deviations, and s_{pooled} remained the same, how would the result of part (a) change?

11.8.5 (*Continuation of Exercise 11.8.4*)

Consider the null hypothesis that flooding has no effect on ATP level in river birch. This hypothesis could be tested in two ways: as a contrast (using the method of Section 11.8), or with a two-sample t test (as in Exercise 7.2.11). Answer the following questions; do not actually carry out the tests.

(a) In what way or ways do the two test procedures differ?

(b) In what way or ways do the conditions for validity of the two procedures differ?

(c) One of the two procedures requires more conditions for its validity, but if the conditions are met, then this procedure has certain advantages over the other one. What are these advantages?

11.8.6 Consider the data from Exercise 11.7.3 in which the drugs ticrynafen (T) and hydrochlorothiazide (H) were compared. The data are summarized in the following table. The pooled standard deviation is $s_{\text{pooled}} = 11.83$ mm Hg.

	TICRYNAFEN (T)		HYDROCHLOROTHIAZIDE (H)	
	LOW DOSE	HIGH DOSE	LOW DOSE	HIGH DOSE
Mean	13.9	17.1	15.8	17.5
No. of patients	53	57	55	58

If the two drugs have equal effects on blood pressure, then T might be preferable because it has fewer side effects.

(a) Construct a 95% confidence interval for the difference between the drugs (with respect to mean blood pressure reduction), averaged over the two dosage levels.

(b) Interpret the confidence interval from part (a) in the context of this setting.

11.8.7 Consider the lettuce growth experiment described in Exercise 11.7.6. The accompanying table shows the mean leaf dry weight (gm) of the nine plants in each treatment group. MS(within) from the ANOVA was 0.3481.

	NUTRIENT SOLUTION	
	STANDARD	EXTRA NITROGEN
Low light	2.16	3.09
High light	3.26	4.48

Construct a 95% confidence interval for the effect of extra nitrogen, averaged over the two light conditions.

11.8.8 Refer to the MAO data of Exercise 11.4.1.

(a) Define a contrast to compare the MAO activity for schizophrenics without paranoid features versus the average of the two types with paranoid features.

(b) Calculate the value of the contrast in part (a) and its standard error.

(c) Apply a t test to the contrast in part (a). Let H_A be nondirectional and $\alpha = 0.05$.

11.8.9 Are the brains of left-handed people anatomically different? To investigate this question, a neuroscientist conducted postmortem brain examinations in 42 people. Each person had been evaluated before death for hand preference and categorized as consistently right-handed (CRH) or mixed-handed (MH). The table shows the results on the area of the anterior half of the corpus callosum (the structure that links the left and right hemispheres of the brain).[27] The MS(within) from the ANOVA was 2,498.

	AREA (MM^2)		
GROUP	MEAN	SD	n
1. Males: MH	423	48	5
2. Males: CRH	367	49	7
3. Females: MH	377	63	10
4. Females: CRH	345	43	20

(a) The difference between MH and CRH is 56 mm^2 for males and 32 mm^2 for females. Is this sufficient evidence to conclude that the corresponding population difference is greater for males than for females? Test an appropriate hypothesis. (Use a nondirectional alternative and let $\alpha = 0.10$.)

(b) As an overall measure of the difference between MH and CRH, one can consider the quantity $0.5(\mu_1 - \mu_2) + 0.5(\mu_3 - \mu_4)$. Construct a 95% confidence interval for this quantity. (This is a sex-adjusted comparison of MH and CRH, where the reference population is 50% male and 50% female.)

11.8.10 Consider the daffodil data of Exercise 11.4.5.

(a) Define a contrast to compare the stem length for daffodils from the open area versus the average of the north, south, east and west sides of the building.

(b) Calculate the value of the contrast in part (a) and its standard error.

(c) Apply a t test to the contrast in part (a). Let H_A be nondirectional and $\alpha = 0.05$.

11.9 Multiple Comparisons (Optional)

After conducting a global F test, we may find that there is significant evidence for a difference among the population means $\mu_1, \mu_2, \ldots, \mu_I$. In this situation, we are often interested in a detailed analysis of the sample means $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_I$ considering all pairwise comparisons. That is, we wish to test all possible pairwise hypotheses:

$$H_0: \mu_1 = \mu_2$$

$$H_0: \mu_1 = \mu_3$$

$$H_0: \mu_2 = \mu_3$$

and so on.

We saw in Section 11.1 that using repeated t tests leads to an increased overall risk of Type I error (e.g., finding evidence for a difference in population means when, in fact, there is no difference). In fact, it was this increased risk of Type I error that motivated the global F test in the first place. In this section we describe three multiple comparison methods to control the overall risk of Type I error: Bonferroni's method, Fisher's Least Significant Difference, and Tukey's Honest Significant Difference. First, however, we must examine the different types of Type I error that arise when considering multiple comparisons.

Experimentwise versus Comparisonwise Error

Consider a study involving the comparison of four population means: μ_1, μ_2, μ_3, and μ_4. As noted in Section 11.1, there are six possible comparisons:

$$H_0\colon \mu_1 = \mu_2 \quad H_0\colon \mu_1 = \mu_3 \quad H_0\colon \mu_1 = \mu_4 \quad H_0\colon \mu_2 = \mu_3 \quad H_0\colon \mu_2 = \mu_4 \quad H_0\colon \mu_3 = \mu_4$$

When considering these six comparisons we can speak of the chance of a Type I error for a particular comparison, say $H_0\colon \mu_1 = \mu_2$, called the **comparisonwise Type I error rate (α_{cw})**, or we can speak of the chance of making a Type I error among *any* of the six comparisons, called the **experimentwise Type I error rate (α_{ew})**.* For example, Table 11.1.2 displays the experimentwise Type I error rates for comparing different numbers of groups when the comparisonwise Type I error rate is $\alpha_{cw} = 0.05$.

While the relationship between α_{cw} and α_{ew} may be complex, it is always true that

$$\alpha_{ew} \leq k \times \alpha_{cw}$$

where k is the number of comparisons. Thus, if six independent comparisons were made at the $\alpha_{cw} = 0.05$ level, the experimentwise Type I error rate (α_{ew}) is at most $6 \times 0.05 = 0.30$.

Fisher's Least Significant Difference

In optional Section 11.8 we described a procedure for estimating linear contrasts. Fisher's Least Significant Difference (LSD) uses this procedure to produce all pairwise confidence intervals for differences of population means using $\alpha_{cw} = \alpha$, the Type I error rate used in the ANOVA. Intervals that do not contain zero provide evidence for a significant difference between the compared population means.

An example of the procedure follows.

Example 11.9.1

Oysters and Seagrass In a study to investigate the effect of oyster density on seagrass biomass, researchers introduced oysters to thirty 1-m^2 plots of healthy seagrass. At the beginning of the study the seagrass was clipped short in all plots. Next, 10 randomly chosen plots received a high density of oysters; 10, an intermediate density; and 10, a low density. As a control, an additional 10 randomly chosen clipped 1-m^2 plots received no oysters. After two weeks, the belowground

*Although the term *experimentwise* contains the word experiment, this terminology pertains to both experiments and observational studies.

seagrass biomass was measured in each plot (g/m^2). Data from some plots are missing. A summary of the data (Table 11.9.1) as well as the ANOVA table (Table 11.9.2) follow.[28]

Table 11.9.1 Belowground seagrass biomass (g/m^2)

	Oyster density			
	None (1)	Low (2)	Intermediate (3)	High (4)
Mean	34.81	33.13	28.33	15.00
SD	13.44	17.36	17.11	10.97
n	9	10	8	10

Table 11.9.2 ANOVA summary of belowground seagrass biomass (g/m^2)

	df	Sum of squares	Mean squares	F	P-value
Between	3	2365.5	788.51	3.5688	0.0243
Within	33	7291.1	220.94		
Total	36	9656.6			

The P-value for the ANOVA is 0.0243, indicating that there is significant evidence of a difference among the biomass means under these experimental conditions. Having evidence for a difference we proceed with comparisons.

Recall that for any linear contrast $L = m_1\overline{y}_1 + m_2\overline{y}_2 + \cdots + m_I\overline{y}_I$,

$$\mathrm{SE}_L = s_{\text{pooled}}\sqrt{\sum_{i=1}^{I}\frac{m_i^2}{n_i}}$$

where

$$s_{\text{pooled}} = \sqrt{\mathrm{MS(within)}}$$

Thus, to compare the no oyster condition (1) to the low oyster density condition (2) we define $D_{12} = \overline{Y}_1 - \overline{Y}_2$ so that as a linear contrast we have

$$\begin{aligned} d_{12} &= 1\overline{y}_1 + (-1)\overline{y}_2 + 0\overline{y}_3 + 0\overline{y}_4 \\ &= (1)(34.81) + (-1)(33.13) + (0)(28.33) + (0)(15.00) \\ &= 34.81 - 33.13 = 1.68 \end{aligned}$$

and, since $s_{\text{pooled}} = \sqrt{220.94} = 14.86$, we have

$$\begin{aligned} \mathrm{SE}_{D_{12}} &= 14.86 \times \sqrt{\frac{1^2}{9} + \frac{(-1)^2}{10} + \frac{0^2}{8} + \frac{0^2}{10}} \\ &= 14.86 \times \sqrt{\frac{1}{9} + \frac{1}{10}} \\ &= 6.82 \end{aligned}$$

A 95% confidence interval for the population mean difference in belowground biomass for the no oyster condition compared to the low oyster density condition, $\mu_1 - \mu_2$, is given by

$$\begin{aligned} d_{12} \pm t_{33,0.025} \times \text{SE}_{D_{12}} &= 1.68 \pm 2.0345 \times 6.82 \\ &= 1.68 \pm 13.89 \\ &= (-12.21, 15.57) \end{aligned}$$

We are 95% confident that the mean belowground biomass when there are no oysters is between 12.21 g/m^2 lower to 15.57 g/m^2 higher than when there is a low density of oysters. Since this interval contains zero, there is no evidence that the mean belowground biomass differs for these two conditions.

Repeating this process for the remaining five comparisons produces the intermediate computations and final intervals summarized in Table 11.9.3.

Table 11.9.3 Intermediate computations and 95% Fisher's LSD intervals comparing belowground biomass under different oyster density conditions*

Comparison	$d_{ab} = \bar{y}_a - \bar{y}_b$	$\sqrt{(1/n_a) + (1/n_b)}$	$\text{SE}_{D_{ab}} = s_{\text{pooled}} \times \sqrt{(1/n_a) + (1/n_b)}$	$t_{33,0.025} \times \text{SE}_{D_{ab}}$
None–low	1.68	0.459	6.828	13.891
None–intermediate	6.48	0.486	7.221	14.690
None–high	*19.81*	*0.459*	*6.828*	*13.891*
Low–intermediate	4.80	0.474	7.049	14.341
Low–high	*18.13*	*0.447*	*6.646*	*13.520*
Intermediate–high	13.33	0.474	7.049	14.341

Comparison	Lower 95%	Upper 95%
None–low	−12.2	15.6
None–intermediate	−8.2	21.2
None–high	*5.9*	*33.7*
Low–intermediate	−9.5	19.1
Low–high	*4.6*	*31.7*
Intermediate–high	−1.0	27.7

*Intervals not containing zero (i.e., there is a statistically significant difference between the group means) are in italics. Note that an interval will not contain zero whenever $|D_{ab}| > t \times \text{SE}_{D_{ab}}$. (The value of $t_{33,0.025} = 2.0345$ was determined using a computer. Using Table 4 we would obtain very similar results using the value listed for 30 degrees of freedom, $t_{30,0.025} = 2.042$.)

From Table 11.9.3 we observe that the only comparisons showing significant differences in mean biomass are the no- to high-oyster density and low- to high-oyster densities. ■

A general formula for computing a $100(1 - \alpha)\%$ Fisher LSD interval for $(\mu_a - \mu_b)$ is given in the following box.

100(1 − α)% Fisher LSD Interval for ($\mu_a - \mu_b$)

$$d_{ab} \pm t_{\mathrm{df},\alpha/2} \times \mathrm{SE}_{D_{ab}}$$

where

$$d_{ab} = \bar{y}_a - \bar{y}_b$$

$$\mathrm{SE}_{D_{ab}} = s_{\mathrm{pooled}}\sqrt{\frac{1}{n_a} + \frac{1}{n_b}}$$

$$s_{\mathrm{pooled}} = \sqrt{\mathrm{MS(within)}}$$

and

$$\mathrm{df} = \mathrm{df(within)}$$

How does Fisher's LSD control the experimentwise Type I error rate? One should use Fisher's LSD comparisons only after rejecting the ANOVA global null hypothesis that all population means are equal: H_0: $\mu_1 = \mu_2 = \cdots = \mu_I$. The ANOVA global F test acts as a screening procedure for the multiple comparisons and thus offers control over α_{ew}.

Displaying Results

The presentation of all six Fisher LSD intervals for the seagrass example in Table 11.9.3 is a useful working summary but is not suitable for effective communication of results. To organize the results for presentation in a simple table we take the following steps.

Step 1 *Array of group labels.* Arrange the group labels in increasing order of their means.

Step 2 *Systematic comparison of means, underlining nonsignificant differences.*

(a) Begin by examining the interval comparing the largest and smallest means. If the interval contains zero, the difference in means is not statistically significant and a line is drawn under the array of group labels to "connect" the groups with the largest and smallest means. If the interval does not contain zero, proceed to the next step.

(b) Ignore the group with the smallest mean and compare the remaining subarray of $I - 1$ means. As in step (2a), if the interval contains zero, the difference in means is not statistically significant and a line is drawn under the array of group labels being compared to "connect" the groups. Next consider the other subarray of $I - 1$ means—the means that remain if the group with the largest mean is ignored. Again, underline this subarray if the interval contains zero.

(c) Repeat step (2b) by successively comparing all subarrays of size $I - 2$, $I - 3$, and so on, until an interval is produced that contains zero or no more comparisons are possible.

Important Notes: During this procedure, never make a comparison within any subarray that has already been underlined; these group means are automatically declared not statistically significantly different. Also, when underlining, use a separate line for each step; never join a line to one that has already been drawn.

Step 3 *Translate the underlines to a tabular summary.* Create a summary table of the data using superscript letters to indicate which groups are not statistically significantly different.

Example 11.9.2

Oysters and Seagrass In this example we will follow the preceding procedure to display the oyster and seagrass Fisher's LSD comparisons displayed in Table 11.9.3.

Step 1 We first arrange the labels in order of the means (shown in Table 11.9.1).

High Intermediate Low None

Step 2 We compare the groups with the smallest (high oyster density) and largest (no oysters) means: $\mu_{\text{None}} - \mu_{\text{High}} = (5.9, 33.7)$. This interval does not contain zero, so these means are significantly different and no underline is made.* We now proceed to the next set (step 2b), the comparisons of subarrays of three means. First, we compare Intermediate to None:

$$\mu_{\text{None}} - \mu_{\text{Intermediate}} = (-8.2, 21.2)$$

This interval contains zero, so an underline is drawn as shown.

High Intermediate Low None

This underline indicates that these three groups do not have significantly different means. We now compare the next subarray of three means, High to Low: $\mu_{\text{High}} - \mu_{\text{Low}} = (4.6, 31.7)$. This interval does not contain zero, so no underlines are drawn. There is evidence for a difference in mean belowground biomass between the high and low oyster-density conditions.

Having compared all subarrays of three means, we continue with subarrays of two means. The only subarray of two means not already connected with an underline is the High–Intermediate comparison. This interval $\mu_{\text{Intermediate}} - \mu_{\text{High}} = (-1.0, 27.7)$ contains zero, so an underline is drawn as shown.

High Intermediate Low None

Step 3 Communicating these results, we give each line a letter and display these letters as superscripts in our table of group means as shown below and in Table 11.9.4. A graphical display is also possible and is displayed in Figure 11.9.1.

High	Intermediate	Low	None
b	a		

■

*Intuitively, this interval should not contain zero since we have rejected the global F test null hypothesis, though there are some instances where the results of our multiple comparison procedure and global F test may not agree.

Table 11.9.4 Belowground seagrass biomass (g/m^2) for different levels of oyster density

	Oyster density			
	None	Low	Intermediate	High
Mean	34.8^{a}	33.1^{a}	$28.3^{a,b}$	15.0^{b}
SD	13.4	17.4	17.1	11.0
n	9	10	8	10

*Groups sharing a common superscript have means that are not statistically significantly different based on Fisher's LSD comparisons with $\alpha_{cw} = 0.05$.

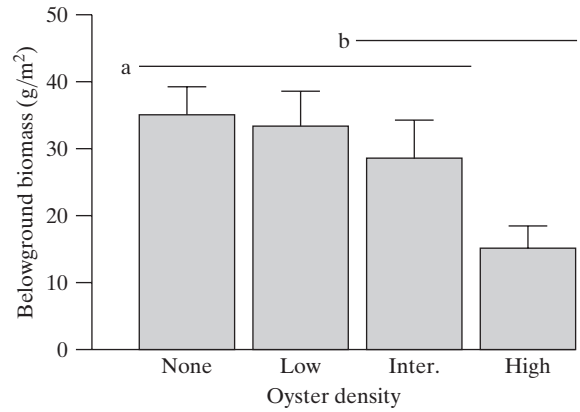


Figure 11.9.1 Belowground seagrass biomass (g/m^2) for different levels of oyster density. Bars display means plus one standard error. Groups sharing a common overbar are not statistically significantly different based on Fisher's LSD comparisons with $\alpha_{cw} = 0.05$

The Bonferroni Method

The **Bonferroni method** is based on a very simple and general relationship: The probability that at least one of several events will occur cannot exceed the sum of the individual probabilities. For instance, suppose we conduct six tests of hypotheses, each at $\alpha_{cw} = 0.01$. Then the overall risk of Type I error α_{ew}—that is, the chance of rejecting at least one of the six hypotheses when in fact all of them are true—cannot exceed

$$0.01 + 0.01 + 0.01 + 0.01 + 0.01 + 0.01 = (6)(0.01) = 0.06$$

Turning this logic around, suppose an investigator plans to conduct six tests of hypotheses and wants the overall risk of Type I error not to exceed $\alpha_{ew} = 0.05$. A conservative approach is to conduct each of the separate tests at the significance level $\alpha_{cw} = 0.05/6 = 0.0083$; this is called a **Bonferroni adjustment**.

Note that the Bonferroni technique is very broadly applicable. The separate tests may relate to different response variables, different subsets, and so on; some may be t tests, some chi-square tests, and so on.

The Bonferroni approach can be used by a person reading a research report, if the author has included explicit P-values. For instance, if the report contains six P-values and the reader desires overall 5%-level protection against Type I error, then the reader will not regard a P-value as sufficient evidence of an effect unless it is smaller than $\alpha_{cw} = 0.0083$.

A Bonferroni adjustment can also be made for confidence intervals. For instance, suppose we wish to construct six confidence intervals and desire an overall probability of 95% that *all* the intervals contain their respective parameters ($\alpha_{ew} = 0.05$). Then this can be accomplished by constructing each interval at confidence level 99.17% (because $0.05/6 = 0.0083$ and $1 - 0.0083 = 0.9917$).

In general, to construct k Bonferonni-adjusted confidence intervals with an overall probability of $100(1 - \alpha_{ew})\%$ that *all* the intervals contain their respective parameters, we construct each interval at confidence level $100(1 - \alpha_{cw})\%$ where $\alpha_{cw} = \alpha_{ew}/k$. The mechanics of the computations are identical to those used for Fisher's LSD except the value of the t multiplier is modified: $t_{df,\alpha_{cw}/2}$. Note that the application of this idea requires unusual critical values, so that standard tables are not sufficient. Table 11 (at the end of this book) provides Bonferroni multipliers for confidence intervals that are based on a t distribution. Software can also be used to produce appropriate multipliers. Example 11.9.3 illustrates this idea.

Example 11.9.3

Oysters and Seagrass To compute the Bonferroni adjusted experimentwise 95% ($\alpha_{ew} = 0.05$) confidence intervals for our oyster and seagrass example, we first recall that a total of six comparisons are required so that $\alpha_{cw} = 0.05/6 = 0.0083$ and $t_{30,0.0083/2} = 2.825$ [because not all values of df are listed in Table 12, we use df = 30, the closest value to df(within) = 33]. Table 11.9.5 summarizes the collection of intervals in a manner similar to the Fisher LSD intervals in Table 11.9.3.

Table 11.9.5 Intermediate computations and experimentwise 95% (99.17% comparisonwise) Bonferroni intervals comparing belowground biomass under different oyster density conditions

Comparison	$d_{ab} = \bar{y}_a - \bar{y}_b$	$SE_{D_{ab}}$	$t_{30,0.025/6} \times SE_{D_{ab}}$	Lower 99.17%	Upper 99.17%
None–low	1.68	6.828	13.891	−17.6	21.0
None–intermediate	6.48	7.221	14.690	−13.9	26.9
None–high	*19.81*	*6.828*	*13.891*	*0.5*	*39.1*
Low–intermediate	4.80	7.049	14.341	−15.1	24.7
Low–high	18.13	6.646	13.520	−0.6	36.9
Intermediate–high	13.33	7.049	14.341	−6.6	33.2

*Intervals not containing zero (i.e., where there is a statistically significant difference between the group means) are in italics. Note the first two columns (d_{ab} and $SE_{D_{ab}}$) are identical to those presented in Table 11.9.3.

Using the method of underlining to visualize the comparisons, we have

High Intermediate Low None
a ———————————
b ———————————

The underlines indicate that the only significant difference in mean belowground seagrass biomass is between the high oyster density and no oyster conditions. A summary of the results is presented in Table 11.9.6. ■

Table 11.9.6 Belowground seagrass biomass (g/m^2) for different levels of oyster density

	Oyster density			
	None	Low	Intermediate	High
Mean	34.8^{a}	$33.1^{a,b}$	$28.3^{a,b}$	15.0^{b}
SD	13.4	17.4	17.1	11.0
n	9	10	8	10

*Groups sharing a common superscript have means that are not statistically significantly different based on Bonferroni comparisons with $\alpha_{ew} = 0.05$.

Note that the Fisher LSD intervals and the Bonferroni intervals are not identical (the Bonferroni are wider due to the smaller value of α_{cw}). Additionally, the conclusions differ as well. The Fisher LSD intervals indicate that there is evidence that the low and high oyster density conditions have different population means, while the Bonferroni intervals do not indicate a difference. This is because the Bonferroni intervals are less powerful and thus more conservative than the Fisher intervals. Unlike the Fisher intervals, the Bonferroni intervals are guaranteed to have α_{ew} less than or equal to the desired experimentwise Type I error rate.

Unfortunately, the Bonferroni intervals are often overly conservative so that the actual value of α_{ew} is much less than the desired experimentwise Type I error rate, and thus too much power is sacrificed for Type I error protection. A more complex procedure that (when sample sizes are equal) is able to achieve the desired experimentwise error exactly (and thus achieve higher power than Bonferroni) is Tukey's Honest Significant Difference.

Tukey's Honest Significant Differnece

Tukey's Honest Significant Difference (HSD) is very similar to the Fisher's LSD and Bonferonni adjusted intervals, but rather than using t multipliers in the confidence interval formulas, related values from a distribution known as the Studentized range distribution are used. Most computer packages will display all Tukey HSD pairwise intervals for any desired experimentwise Type I error rate, α_{ew}. As an example, Figure 11.9.2 displays the Tukey output from the statistical software package R using our oyster and seagrass data. Note that in addition to the intervals, most software also provides an "adjusted" P-value. Even though multiple comparisons are being made, if these "adjusted" P-values are compared to α_{ew}, an overall experimentwise Type I error rate of α_{ew} will still be maintained.

Figure 11.9.2 R software output presenting experimentwise 95% Tukey HSD intervals for the oyster and seagrass example

	diff	lwr	upr	p-adj
int-high	13.33	−5.74	32.40	0.2515
low-high	18.13	0.15	36.11	0.0475
no-high	19.81	1.34	38.28	0.0318
low-int	4.80	−14.27	23.87	0.9037
zero-int	6.48	−13.06	26.02	0.8063
zero-low	1.68	−16.79	20.15	0.9947

The intervals in Figure 11.9.2 show that the conclusions drawn from the Tukey HSD intervals match those from the Fisher LSD intervals: the high and low oyster density as well as the high and no oyster density means differ significantly. The endpoints of the experimentwise 95% Tukey HSD intervals are, however, different from both the Fisher LSD and Bonferroni intervals.

Conditions for Validity

All three multiple comparison procedures as described require the same standard ANOVA conditions given in Section 11.5. In addition, the validity conditions for Fisher's LSD intervals also require that the procedure not be used unless the global null hypothesis of all means being equal is rejected. In contrast, Tukey's HSD and Bonferroni intervals do not require that the global F test be performed a priori (though the computation of s_{pooled} is still needed). To exactly achieve the desired experimentwise Type I error rate, Tukey's HSD requires that all samples be the same

size. If the sample sizes are unequal, the actual error rate will be somewhat less than the nominal rate resulting in a loss of power.

An advantage of the Bonferroni method is that it is widely applicable and can easily be generalized to situations beyond ANOVA. One such example appears in the exercises.

Exercises 11.9.1–11.9.8

11.9.1 A botanist used a completely randomized design to allocate 45 individually potted eggplant plants to five different soil treatments. The observed variable was the total plant dry weight without roots (gm) after 31 days of growth. The treatment means were as shown in the following table.[29] The MS(within) was 0.2246. Use Fisher's LSD intervals to compare all pairs of means at $\alpha_{cw} = 0.05$. Present your results in a summary table similar to Table 11.9.4. (*Hint*: Take note that all sample sizes are equal; thus the calculated margin of error need only be calculated once for all comparisons. There is a total of 10 comparisons possible).

TREATMENT	A	B	C	D	E
Mean	4.37	4.76	3.70	5.41	5.38
n	9	9	9	9	9

11.9.2 Repeat Exercise 11.9.1, but use Bonferroni intervals with $\alpha_{ew} = 0.05$.

11.9.3 In a study of the dietary treatment of anemia in cattle, researchers randomly divided 144 cows into four treatment groups. Group A was a control group, and groups B, C, and D received different regimens of dietary supplementation with selenium. After a year of treatment, blood samples were drawn and assayed for selenium. The accompanying table shows the mean selenium concentrations (μg/dl).[30] The MS(within) from the ANOVA was 2.071.

GROUP	MEAN	n
A	0.8	36
B	5.4	36
C	6.2	36
D	5.0	36

(a) Compute three Bonferroni-adjusted intervals comparing diets B, C, and D to the control (diet A) using $\alpha_{ew} = 0.05$. (*Note*: This is an example of a situation for which the Bonferroni comparisons may be preferred over the Tukey HSD comparisons since not all comparisons are considered—we are only interested in comparing the control to each of the other three treatments.)

(b) In the context of the problem, interpret the Bonferroni interval computed in part (a) that compares the control (group A) to the group that is most different from it.

11.9.4 Consider the experiment and data in Exercise 11.9.3. The experimentwise 95% Tukey HSD intervals are displayed using the statistical software package R.

	diff	lwr	upr
B–A	4.6	3.72	5.48
C–A	5.4	4.52	6.28
D–A	4.2	3.32	5.08
C–B	0.8	−0.08	1.68
D–B	−0.4	−1.28	0.48
D–C	−1.2	−2.08	−0.32

(a) Using the preceding output to support your answer, is there evidence that each of the groups/diets B, C, and D, differs from the control, A?

(b) According to the preceding Tukey HSD intervals and summary of the data in Exercise 11.9.3, diet C yields the greatest mean selenium concentration and is significantly higher than the control. If the goal of the researchers is to find a diet that maximizes selenium concentration, is diet C the clear choice? That is, should we rule out diet B, diet D, or both? Refer to the Tukey HSD intervals to justify your answer.

11.9.5 Ten treatments were compared for their effect on the liver in mice. There were 13 animals in each treatment group. The ANOVA gave MS(within) = 0.5842. The mean liver weights are given in the table.[31]

TREATMENT	MEAN LIVER WEIGHT (GM)	TREATMENT	MEAN LIVER WEIGHT (GM)
1	2.59	6	2.84
2	2.28	7	2.29
3	2.34	8	2.45
4	2.07	9	2.76
5	2.40	10	2.37

(a) Use Fisher LSD intervals to compare all pairs of means with $\alpha_{cw} = 0.05$ and summarize the results in a table similar to Table 11.9.4. [*Time Saving Hints*: First note that the sample sizes are equal; hence the same margin of error ($t \times \text{SE}_{D_{ab}}$) can be used for all comparisons. Furthermore, since a summary table is desired, the actual intervals need not be computed: Simply check if $|d_{ab}| > t \times \text{SE}_{D_{ab}}$. If it is, then the computed interval would not contain zero, so the difference is significant. Finally, note that not all possible comparisons (there are 45) need to be checked: when using the method of underlining to summarize results, once a subarray of groups has been underlined all comparisons within the subarray are considered nonsignificant.]

(b) If Bonferroni's method is used with $\alpha_{ew} = 0.05$ instead of Fisher's LSD in part (a), are any pairs of means significantly different?

11.9.6 Consider the data from Example 11.2.1 on the weight gain of lambs. The MS(within) from the ANOVA for these data was 23.333. The sample mean of diet 2 was 15 and of diet 1 was 11.

(a) Use the Bonferroni method to construct a 95% confidence interval for the difference in population means of these two diets (assuming that intervals will also be computed for the other two possible comparisons as well).

(b) Suppose that the comparison in part (a) was the *only* comparison of interest (i.e., one comparison rather than three). How would the interval in part (a) change? Would it be wider, narrower, or stay the same? Explain.

11.9.7 As mentioned in this section, the Bonferroni procedure can be used in a variety of circumstances. Consider the plover nesting example from Section 10.5, which compares plover nest locations across three years. The percentage distribution appears in the following table.

	YEAR		
LOCATION	2004	2005	2006
Agricultural field (AF)	48.8	30.2	55.3
Prairie dog habitat (PD)	39.5	60.3	25.5
Grassland (G)	11.6	9.5	19.1
Total	99.9*	100.0	99.9*

*The sums of the 2004 and 2006 percentages differ from 100% due to rounding.

The *P*-value for the chi-square test of these data was found to be 0.007, indicating a significant difference in the distribution of nesting locations across the three years with $\alpha = 0.10$. Considering reduced tables and using chi-square tests to compare nesting distributions for pairs of years, we obtain the following *P*-values:

YEARS COMPARED	*P*-VALUE
2004 to 2005	0.100
2004 to 2006	0.307
2005 to 2006	0.001

Using a Bonferonni adjustment to achieve $\alpha_{ew} = 0.10$, for which pair(s) of years is there evidence of a significant difference in nesting location distributions? Indicate the value of α_{cw} used.

11.9.8 Exercise 10.5.1 presented the following problem: Patients with painful knee osteoarthritis were randomly assigned in a clinical trial to one of five treatments: glucosamine, chondroitin, both, placebo, or Celebrex, the standard therapy. One outcome recorded was whether or not each patient experienced substantial improvement in pain or in ability to function. The data are given in the following table.

	SUCCESSFUL OUTCOME		
TREATMENT	SAMPLE SIZE	NUMBER	PERCENT
Glucosamine	317	192	60.6
Chondroitin	318	202	63.5
Both	317	208	65.6
Placebo	313	178	56.9
Celebrex	318	214	67.3

(a) Suppose we wished to compare only the success rates of each of the treatments to the control (placebo) using four separate 2×2 chi-square tests. The *P*-values for these comparisons follow. Using a Bonferroni adjustment with $\alpha_{ew} = 0.05$, which treatments perform significantly different from the placebo? Indicate the value of α_{cw} used.

TREATMENTS COMPARED TO PLACEBO	*P*-VALUE
Glucosamine	0.346
Chondroitin	0.088
Both	0.024
Celebrex	0.007

(b) The *P*-value of the chi-square test that considers the entire 5×2 table is 0.054, which provides insufficient evidence to demonstrate any difference among the success rates of the five treatments using $\alpha = 0.05$. Explain why this result does not contradict the results of part (a). [*Hint*: How many comparisons are being considered by this chi-square test as compared to the number of comparisons in part (a)? To achieve $\alpha_{ew} = 0.05$ using a Bonferroni adjustment, how large would α_{cw} need to be? How large was it in part (a)? How does conducting many tests with a Bonferroni adjustment affect the power of each test?]

11.10 Perspective

In Chapter 11 we have introduced some statistical issues that arise when analyzing data from more than two samples and we have considered some classical methods of analysis. In this section we review these issues and briefly mention some alternative methods of analysis.

Advantages of Global Approach

Let us recapitulate the advantages of analyzing I independent samples by a global approach rather than by viewing each pairwise comparison separately.

1. *Multiple comparisons* In Section 11.1 we saw that the use of repeated t tests can greatly inflate the overall risk of Type I error. Some control of Type I error can be gained by the simple device of beginning the data analysis with a global F test. For more stringent control of Type I error, other multiple comparison methods are available (e.g., Bonferroni and Tukey HSD) and are described in optional Section 11.9. (Note that the problem of multiple comparisons is not confined to an ANOVA setting.)
2. *Use of structure in the treatments or groups* Analysis of suitable combinations of group means can be very useful in interpreting data. Many of the relevant techniques are beyond the scope of this book. The discussion in optional Sections 11.7 and 11.8 gave a hint of the possibilities. In Chapter 12 we will discuss some ideas that are applicable when the treatments themselves are quantitative (for instance, doses).
3. *Use of a pooled SD* We have seen that pooling all of the within-sample variability into a single pooled SD leads to a better estimate of the common population SD and thus to a more precise analysis. This is particularly advantageous if the individual sample sizes (n's) are small, in which case the individual SD estimates are quite imprecise. Of course, using a pooled SD is proper only if the population SDs are equal. It sometimes happens that one cannot take advantage of pooling the SDs because the assumption of equal population SDs is not tenable. One approach that can be helpful in this case is to analyze a transformed variable, such as $\log(Y)$; the SDs may be more nearly equal in the transformed scale.

Other Experimental Designs

The techniques of this chapter are valid only for independent samples. But the basic idea—partitioning variability within and between treatments into interpretable components—can be applied in many experimental designs. For instance, all the techniques discussed in this chapter can be adapted (by suitable modification of the SE calculation) to analysis of data from an experiment with more than two experimental factors or situations for which all or some experimental factors are numeric rather than categorical. These and related techniques belong to the large subject called *analysis of variance*, of which we have discussed only a small part.

Nonparametric Approaches

There are k-sample analogs of the Wilcoxon-Mann-Whitney test and other nonparametric tests (e.g., the Kruskal–Wallis test). These tests have the advantage of not assuming underlying normal distributions. However, many of the advantages of the parametric techniques—such as the use of linear combinations—do not easily carry over to the nonparametric setting.

Ranking and Selection

In some investigations the primary aim of the investigator is not to answer research questions about the populations but simply to *select* one or several "best" populations. For instance, suppose 10 populations (stocks) of laying hens are available and it is desired to select the one population with the highest egg-laying potential. The investigator will select a random sample of n chickens from each stock and will observe for each chicken Y = total number of eggs laid in 500 days.[32] One relevant question is: How large should n be so that the stock that is *actually* best (has the highest μ) is likely to also *appear* best (have the highest $\overline{Y}$)? This and similar questions are addressed by a branch of statistics called *ranking and selection theory*.

Supplementary Exercises 11.S.1–11.S.19

(*Note*: Exercises preceded by an asterisk refer to optional sections.)

11.S.1 Consider the research described in Exercise 11.4.6 in which 10 women in an aerobic exercise class, 10 women in a modern dance class, and a control group of 9 women were studied. One measurement made on each woman was change in fat-free mass over the course of the 16-week training period. Summary statistics are given in the following table.[8] The ANOVA SS(between) is 2.465 and the SS(within) is 50.133.

	AEROBICS	MODERN DANCE	CONTROL
Mean	0.00	0.44	0.71
SD	1.31	1.17	1.68
n	10	10	9

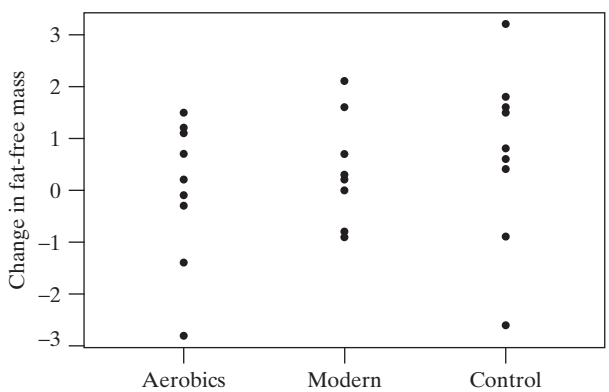
(a) State in words, in the context of this problem, the null hypothesis that is tested by the analysis of variance.

(b) Construct the ANOVA table and test the null hypothesis. Let $\alpha = 0.05$.

11.S.2 Refer to Exercise 11.S.1. The F test is based on certain conditions concerning the population distributions.

(a) State the conditions.

(b) The following dotplots show the raw data. Based on these plots and on the information given in Exercise 11.S.1, does it appear that the F test conditions are met? Why or why not?



11.S.3 In a study of the eye disease retinitis pigmentosa (RP), 211 patients were classified into four groups according to the pattern of inheritance of their disease. Visual acuity (spherical refractive error, in diopters) was measured for each eye, and the two values were then averaged to give one observation per person. The accompanying table shows the number of persons in each group and the group mean refractive error.[33] The ANOVA of the 211 observations yields SS(between) = 129.49 and SS(within) = 2,506.8. Construct the ANOVA table and carry out the F test at $\alpha = 0.05$.

GROUP	NUMBER OF PERSONS	MEAN REFRACTIVE ERROR
Autosomal dominant RP	27	+0.07
Autosomal recessive RP	20	−0.83
Sex-linked RP	18	−3.30
Isolate RP	146	−0.84
Total	211	

11.S.4 (*Continuation of Exercise 11.S.3*) Another approach to the data analysis is to use the eye, rather than the person, as the observational unit. For the 211 persons there were 422 measurements of refractive error; the accompanying table summarizes these measurements. The ANOVA of the 422 observations yields SS(between) = 258.97 and SS(within) = 5,143.9.

GROUP	NUMBER OF EYES	MEAN REFRACTIVE ERROR
Autosomal dominant RP	54	+0.07
Autosomal recessive RP	40	−0.83
Sex-linked RP	36	−3.30
Isolate RP	292	−0.84
Total	422	

(a) Construct the ANOVA table and bracket the *P*-value for the *F* test. Compare with the *P*-value obtained in Exercise 11.S.3. Which of the two *P*-values is of doubtful validity, and why?

(b) The mean refractive error for the sex-linked RP patients was −3.30. Calculate the standard error of this mean two ways: (i) regarding the person as the observational unit and using s_{pooled} from the ANOVA of Exercise 11.S.3; (ii) regarding the eye as the observational unit and using s_{pooled} from the ANOVA of this exercise. Which of these standard errors is of doubtful validity, and why?

***11.S.5** In a study of the mutual effects of the air pollutants ozone and sulfur dioxide, Blue Lake snap beans were grown in open-top field chambers. Some chambers were fumigated repeatedly with sulfur dioxide. The air in some chambers was carbon filtered to remove ambient ozone. There were three chambers per treatment combination, allocated at random. After one month of treatment, total yield (kg) of bean pods was recorded for each chamber, with results shown in the accompanying table.[34] For these data, SS(between) = 1.3538 and SS(within) = 0.27513. Complete the ANOVA table and carry out the *F* test at $\alpha = 0.05$.

	OZONE ABSENT		OZONE PRESENT	
	SULFUR DIOXIDE		SULFUR DIOXIDE	
	ABSENT	PRESENT	ABSENT	PRESENT
	1.52	1.49	1.15	0.65
	1.85	1.55	1.30	0.76
	1.39	1.21	1.57	0.69
Mean	1.587	1.417	1.340	0.700
SD	0.237	0.181	0.213	0.056

Prepare an interaction graph (like Figure 11.7.3).

***11.S.6** Consider the data from Exercise 11.S.5. For these data, SS(ozone) = 0.696, SS(sulfur) = 0.492, SS(interaction) = 0.166, and SS(within) = 0.275.

(a) Construct the ANOVA table.

(b) Carry out an *F* test for interactions; use $\alpha = 0.05$.

(c) Test the null hypothesis that ozone has no effect on yield. Use $\alpha = 0.05$.

***11.S.7** Refer to Exercise 11.S.5. Define contrasts to measure each effect specified, and calculate the value of each contrast.

(a) The effect of sulfur dioxide in the absence of ozone

(b) The effect of sulfur dioxide in the presence of ozone

(c) The interaction between sulfur dioxide and ozone

***11.S.8** (*Continuation of Exercises 11.S.6 and 11.S.7*) For the snap-bean data, use a *t* test to test the null hypothesis of no interaction against the alternative that sulfur dioxide is more harmful in the presence of ozone than in its absence. Let $\alpha = 0.05$. How does this compare with the *F* test of Exercise 11.S.6(b) (which has a nondirectional alternative)?

***11.S.9** (*Computer exercise*) Refer to the snap-bean data of Exercise 11.S.5. Apply a reciprocal transformation to the data. That is, for each yield value *Y*, calculate $Y' = 1/Y$.

(a) Calculate the ANOVA table for Y' and carry out the *F* test.

(b) It often happens that the SDs are more nearly equal for transformed data than for the original data. Is this true for the snap-bean data when a reciprocal transformation is used?

(c) Make a normal probability plot of the residuals, $(y'_{ij} - \bar{y}'_i)$. Does this plot support the condition that the populations are normal?

***11.S.10** (*Computer exercise—continuation of Exercises 11.S.8 and 11.S.9*) Repeat the test in Exercise 11.S.7 using Y' instead of *Y*, and compare with the results of Exercise 11.S.7.

11.S.11 Suppose a drug for treating high blood pressure is to be compared to a standard blood pressure drug in a study of humans.

(a) Describe an experimental design for a study that makes use of blocking. Be careful to note which parts of the design involve randomness and which parts do not.

(b) Can the experiment you described in part (a) involve blinding? If so, explain how blinding could be used.

11.S.12 In a study of balloon angioplasty, patients with coronary artery disease were randomly assigned to one of four treatment groups: placebo, probucol (an experimental drug), multivitamins (a combination of beta carotene, vitamin E, and vitamin C), or probucol combined with multivitamins. Balloon angioplasty was performed on each of the patients. Later, "minimal luminal diameter" (a measurement of how well the angioplasty did in dilating the artery) was recorded for each of the patients. Summary statistics are given in the following table.[35]

	PLACEBO	PROBUCOL	MULTI-VITAMINS	PROBUCOL AND MULTIVITAMINS
n	62	58	54	56
Mean	1.43	1.79	1.40	1.54
SD	0.58	0.45	0.55	0.61

(a) Complete the ANOVA table and bracket the *P*-value for the *F* test.

SOURCE	DF	SS	MS	*F*
Between treatments	____	5.4336	____	____
Within treatments	____	____	____	____
Total	229	73.9945	____	____

(b) If $\alpha = 0.01$, do you reject the null hypothesis of equal population means? Why or why not?

***11.S.13** Refer to Exercise 11.S.12. Define contrasts to measure each effect specified, and calculate the value of each contrast.

(a) The effect of probucol in the absence of multivitamins
(b) The effect of probucol in the presence of multivitamins
(c) The interaction between probucol and multivitamins

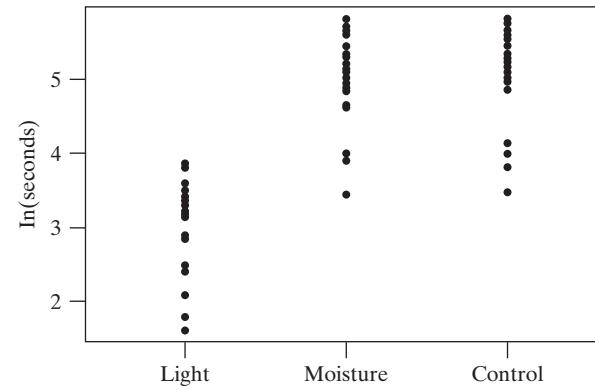
***11.S.14** Refer to Exercise 11.S.12. Construct a 95% confidence interval ($\alpha_{cw} = 0.05$) for the effect of probucol in the absence of multivitamins. That is, construct a 95% confidence interval for $\mu_{\text{probucol}} - \mu_{\text{placebo}}$.

***11.S.15** Refer to Exercise 11.S.12. Assuming all possible comparisons of group means will be computed, use the Bonferroni method to construct a 95% confidence interval for the effect of probucol in the absence of multivitamins. That is, construct a Bonferroni-adjusted 95% ($\alpha_{ew} = 0.05$) confidence interval for $\mu_{\text{probucol}} - \mu_{\text{placebo}}$.

***11.S.16** Three college students collected several pillbugs from a woodpile and used them in an experiment in which they measured the time, in seconds, that it took for a bug to move 6 inches within an apparatus they had created. There were three groups of bugs: one group was exposured to strong light, for one group the stimulus was moisture, and a third group served as a control. The data are shown in the following table.[36]

	LIGHT	MOISTURE	CONTROL
	23	170	229
	12	182	126
	29	286	140
	12	103	260
	5	330	330
	47	55	310
	18	49	45
	30	31	248
	8	132	280
	45	150	140
	36	165	160
	27	206	192
	29	200	159
	33	270	62
	24	298	180
	17	100	32
	11	162	54
	25	126	149
	6	229	201
	34	140	173
Mean	23.6	169.2	173.5
SD	12.3	83.5	86.0
n	20	20	20

Clearly the SDs show that the variability is not constant between groups, so a transformation is needed. Taking the natural logarithm of each observation results in the following dotplots and summary statistics.



	LIGHT	MOISTURE	CONTROL
Mean	2.99	4.98	4.99
SD	0.65	0.62	0.66

For the transformed data, the ANOVA SS(between) is 53.1103 and the SS(within) is 23.5669.

(a) State the null hypothesis in symbols.
(b) Construct the ANOVA table and test the null hypothesis. Let $\alpha = 0.05$.
(c) Calculate the pooled standard deviation, s_{pooled}.

***11.S.17** Mountain climbers often experience several symptoms when they reach high altitudes during their climbs. Researchers studied the effects of exposure to high altitude on human skeletal muscle tissue. They set up a 2×2 factorial experiment in which subjects trained for six weeks on a bicycle. The first factor was whether subjects trained under hypoxic conditions (corresponding to an altitude of 3,850m) or normal conditions. The second factor was whether subjects trained at a high level of energy expenditure or at a low level (25% less than the high level). There were either 7 or 8 subjects at each combination of factor levels. The accompanying table shows the results for the response variable "percentage change in vascular endothelial growth factor mRNA."[37]

	HYPOXIC		NORMAL	
ENERGY	LOW LEVEL	HIGH LEVEL	LOW LEVEL	HIGH LEVEL
Mean	117.7	173.2	95.1	114.6
No. of patients	7	7	8	8

Prepare an interaction graph (like Figure 11.7.3).

***11.S.18** Consider the data from Exercise 11.S.17.

(a) Complete the following ANOVA table.

SOURCE	DF	SS	MS	*F* RATIO
Between hypoxic and normal	1	12126.5	____	____
Between energy level	1	10035.7	____	____
Interaction	1	____	____	____
within groups	26	56076.0	____	____
Total	29	80738.7	____	____

(b) Conduct a test for interactions. Use $\alpha = 0.05$.

(c) Based on your conclusions in part (b), is it sensible to examine the main effects of condition and of energy level?

(d) Test the null hypothesis that energy level has no effect on the response. Use $\alpha = 0.05$.

(e) Test the null hypothesis that the effect on the response of hypoxic training is the same as the effect on the response of normal training. Use $\alpha = 0.05$.

***11.S.19** In a study to examine the utility of using ammonia gas to sanitize animal feeds, researchers inoculated corn silage with a strain of *Salmonella.* Next, two petri dishes of 5 g of contaminated feed were exposed to concentrated anhydrous ammonia gas and two control petri dishes of 5 g of contaminated feed were not treated with the gas. This experiment was repeated twice, for a total of three trials, as only two petri dishes could be placed in the pressurized gas chamber at any given time. Twenty-four hours after inoculation and gassing, the number of bacterial colonies (colony forming units or cfu) on each dish were counted. Because the data were highly skewed, the log(cfu) was analyzed.[38]

(a) Identify the blocking, treatment, and response variables in this problem.

(b) Complete the following ANOVA table for this blocked analysis.

	DF	SS	MS	*F* RATIO
Between treatments	1	1.141	1.141	7.107
Between trials	2	3.611	____	____
Within groups	8	____	____	
Total	11	6.036		

(c) Using the complete table from part (b), is there evidence that the ammonia gas treatment affects the contamination level (i.e., mean log cfu)? Use $\alpha = 0.05$.

(d) Do the preceding analysis and information allow you to infer that ammonia reduces contamination? If not, what other information would be necessary to make such a claim?