CHAPTER

26

More about Analysis of Variance: Follow-up Tests and Two-Way ANOVA

NHPA/SuperStock

IN THIS CHAPTER WE COVER...

- Beyond one-way ANOVA
- Follow-up analysis: Tukey pairwise multiple comparisons
- Follow-up analysis: contrasts*
- Two-way ANOVA: conditions, main effects, and interaction
- Inference for two-way ANOVA
- Some details of two-way ANOVA*

Analysis of variance (ANOVA) is a statistical method for comparing the means of several populations based on independent random samples, or the mean responses to several treatments in a randomized comparative experiment. When we compare just two means, we use the two-sample t procedures described in Chapter 18. ANOVA allows comparison of any number of means. The basic form of ANOVA is one-way ANOVA, which treats the means being compared as mean responses to different levels of a single variable. For example, in Chapter 24 we used one-way ANOVA to compare the mean weights of adult male Wistar rats fed one of three types of diets. Figure 26.1 shows the Minitab ANOVA output for these data (displayed in Table 24.1, page 606).

BEYOND ONE-WAY ANOVA

You should recall or review the big ideas of **one-way ANOVA** from Chapter 24. One-way ANOVA compares the means $\mu_1, \mu_2, \ldots, \mu_k$ of k populations based on samples of sizes $n_1, n_2, \ldots, n_k$ from these populations.

- Using separate two-sample t procedures to compare many pairs of means is a bad idea, because we don't have a P-value or a confidence level for the complete set of comparisons together. This is the problem of **multiple comparisons.**
- One-way ANOVA gives a single test for the null hypothesis that all the population means are the same against the alternative hypothesis that not all are the same (H_0 simply is not true).

multiple comparisons

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

ANOVA *F* statistic

- ANOVA works by comparing how far apart the sample means are relative to the variation among individual observations in the same sample. The test statistic is the **ANOVA *F* statistic**

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

F distribution

The *P*-value comes from an ***F* distribution.**

ANOVA conditions

- The required **conditions for ANOVA** are *independent random samples* from each of the k populations (or a randomized comparative experiment with k treatments), *Normal distributions* for the response variable in each population, and a *common standard deviation* σ in all populations. Fortunately, ANOVA inference is quite robust against moderate violations of the Normality and common standard deviation conditions.
- In basic statistical practice, we combine the *F* test with descriptive data analysis to check the conditions for ANOVA and to see which means appear to differ and by how much.

Examples 24.1, 24.2, and 24.3 (pages 605, 608, and 609) showed all the steps required for a one-way ANOVA. This chapter moves beyond basic one-way ANOVA in two directions.

Follow-up analysis. The ANOVA *F* test in Figure 26.1 tells us only that the population means are not the same. We would like to say which means differ and by how much. For example, do the data allow us to say that the "extended diet" population does have a higher mean weight than the "chow diet" and the "restricted diet" populations of adult male Wistar lab rats? This is a *follow-up analysis* to the *F* test that goes beyond data analysis to confidence intervals and tests of significance for specific comparisons of means.

Two-way ANOVA. One-way ANOVA compares mean responses for several levels of just one explanatory variable. In Example 24.1, that variable is "the type of diet provided." Suppose that we have data on *two* explanatory variables, say,

Minitab

One-way ANOVA: Chow, Restricted, Extended

```
Source      DF        SS        MS        F         P
Factor       2     63400     31700    10.71     0.000
Error       47    139174      2961
Total       49    202573

S = 54.42    R-Sq = 31.30%    R-Sq(adj) = 28.37%

                                 Individual 95% CIs for Mean
                                 Based on Pooled StDev
Level        N    Mean   StDev   ----+------+------+------+-----
Chow        19  605.63   49.64   (------*------)
Restricted  16  657.31   50.68                (------*------)
Extended    15  691.13   63.41                       (------*------)
                                 ----+------+------+------+-----
                                     595    630    665    700

Pooled StDev = 54.42
```

▶ FIGURE 26.1
Minitab ANOVA output for the rat weight data of Example 24.3.

the type of diet provided and whether the lab rats are physically active. There are now 6 groups formed by combinations of diet type and physical activity, as follows:

		Variable 2	
		Active	Inactive
Variable 1	Chow	Group 1	Group 2
	Restricted	Group 3	Group 4
	Extended	Group 5	Group 6

One-way ANOVA will still tell us if there is evidence that mean body weight in these 6 experimental groups differs. But we want more: Does diet type matter? Does physical activity matter? And do these two variables *interact?* That is, does the effect of diet type change when the lab rats are physically active? Perhaps physical activity reduces the craving for cafeteria food, so that diet type has less effect when the rats are active than when they are inactive. To answer these questions we must extend ANOVA to take into account the fact that the 6 groups are formed from two explanatory variables. This is *two-way ANOVA*.

We will discuss follow-up analysis in ANOVA first, and then two-way ANOVA. Fortunately, the distinction between one-way and two-way doesn't affect the follow-up methods we will present. So once you have mastered these methods in the one-way setting, you can apply them immediately to two-way problems.

FOLLOW-UP ANALYSIS: TUKEY PAIRWISE MULTIPLE COMPARISONS

In Example 24.3 we saw that there is good evidence that the mean body weight of adult male Wistar rats is not the same when they are assigned to a diet consisting of chow only, chow plus restricted access to cafeteria food, and chow plus extended access to cafeteria food.[1] The sample means in Figure 26.1 suggest that (as we might expect) the mean body weight is highest in rats given extended access to cafeteria food and lowest in rats given chow only.

EXAMPLE 26.1 Comparing groups: individual *t* procedures

Cristian Ciobanu/Alamy

Let's use A, B, and C to refer to the chow, restricted, and extended groups, respectively. How much higher is the mean body weight of rats given restricted access to cafeteria food than that of rats given chow only? A 95% confidence interval comparing Groups A and B answers this question. Because the conditions for ANOVA require that the population standard deviation be the same in all three populations of rats, we can use a version of the two-sample t confidence interval that also assumes equal standard deviations.

The Minitab output in Figure 26.1 gives the pooled standard deviation (first defined in chapters 18 and 24, pages 462 and 624) as $s_p = 54.42$ grams (g). This is an estimate of the common standard deviation σ based on all three samples. It has 47 degrees of freedom, the degrees of freedom for "Error" in the ANOVA table. The standard error for the difference in sample means $\bar{x}_A - \bar{x}_B$ is (page 462)

$$s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$$

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

A 95% confidence interval for $\mu_A - \mu_B$ would therefore be

$$(\bar{x}_A - \bar{x}_B) \pm t^* s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

using $t^* = 2.012$ from technology (or approximately $t^* = 2.021$ from Table C for df = 40, conservatively, since there is no row for df = 47 in Table C).

pairwise difference

However, we really want to estimate all three **pairwise differences** among the population means,

$$\mu_A - \mu_B \qquad \mu_A - \mu_C \qquad \mu_B - \mu_C$$

Three 95% confidence intervals will not give us 95% confidence that all three *simultaneously capture their true parameter values*. This is the familiar problem of multiple comparisons that we discussed in Chapters 22 and 24.

overall confidence level

In general, we want to give confidence intervals for all pairwise differences among the population means $\mu_1, \mu_2, \ldots, \mu_k$ of k populations. We want an **overall confidence level** of (say) 95%. That is, in very many uses of the method, *all* the intervals will simultaneously capture the true differences 95% of the time. To do this, take the number of comparisons into account by replacing the t critical value t^* in Example 26.1 with another critical value based on the distribution of the difference between the largest and smallest of a set of k sample means. We will call this critical value m^*, for multiple comparisons. Values of m^* no longer come from a t table. They depend on the number of populations we are comparing and on the total number of observations in the samples, as well as on the confidence level we want. Software is very helpful for practical use. This method is named after its inventor, John Tukey (1915–2000), the same man who developed the ideas of modern data analysis. A short table of m^* values for a 95% confidence level (Table G) is provided for convenience at the end of this chapter.

TUKEY PAIRWISE MULTIPLE COMPARISONS

In the ANOVA setting, we have independent SRSs of size n_i from each of k populations having Normal distributions with means μ_i and a common standard deviation σ. **Tukey simultaneous confidence intervals** for all pairwise differences $\mu_i - \mu_j$ among the population means have the form

$$(\bar{x}_i - \bar{x}_j) \pm m^* s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Here $\bar{x}_i$ is the sample mean of the ith sample and s_p is the pooled estimate of σ. The critical value m^* depends on the confidence level C, the number of populations k, and the total number of observations N.

If all samples are the same size, the Tukey simultaneous confidence intervals provide an overall level C of confidence that *all* the intervals simultaneously capture the true pairwise differences. If the samples differ in size, the true confidence level is at least as large as C. That is, the conclusions are then conservative.

To carry out **simultaneous tests** of the hypotheses

$$H_0: \mu_i = \mu_j$$
$$H_a: \mu_i \neq \mu_j$$

for all pairs of population means, reject H_0 for any pair whose confidence interval does not contain 0. These tests have overall significance level no less than $1 - C$. That is, $1 - C$ is the probability that, when all the population means are equal, *any* of the tests incorrectly rejects its null hypothesis.

EXAMPLE 26.2 Rats and a cafeteria-style diet: multiple intervals

Figure 26.2 contains more Minitab output for the ANOVA comparing the mean body weights in 3 experimental groups of lab rats. We asked for Tukey multiple comparisons with an overall error rate of 5%. That is, the overall confidence level for the three intervals together is 95%.

The format of the Minitab output takes some study. Be sure you can see that the Tukey confidence intervals are

7.03 to 96.33	for	$\mu_B - \mu_A$
40.05 to 130.95	for	$\mu_C - \mu_A$
−13.47 to 81.12	for	$\mu_C - \mu_B$

If you do not have access to technology, these intervals can easily be computed by hand. Let's see how to obtain the first interval, for $\mu_A - \mu_B$. Table G at the end of this chapter gives values of m^* when using an overall confidence level of 95% and various combinations of N and k. Start by finding the right combination of k comparisons (top row) and $N - k$ degrees of freedom (left margin). In our example, $k = 3$ and $N - k = 47$, so $m^* = 2.434$, approximately (based on a

Minitab

```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons

Individual confidence level = 98.05%

Chow subtracted from:

            Lower  Center   Upper  ----+-------+-------+-------+-----
Restricted   7.03   51.68   96.33              (-----*-----)
Extended    40.05   85.50  130.95                    (-----*-----)
                                   ----+-------+-------+-------+-----
                                     -60       0       60      120

Restricted subtracted from:

             Lower  Center  Upper  ----+-------+-------+-------+-----
Extended    -13.47   33.82  81.12            (------*------)
                                   ----+-------+-------+-------+-----
                                     -60       0       60      120
```

◀ FIGURE 26.2
Additional Minitab ANOVA output showing Tukey pairwise multiple comparisons for the rat weight data, for Example 26.2.

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

conservative df = 40, since df = 47 is not available). The interval for $\mu_B - \mu_A$ is therefore

$$(\bar{x}_B - \bar{x}_A) \pm m^* s_p \sqrt{\frac{1}{n_B} + \frac{1}{n_A}} = (657.31 - 605.63) \pm (2.434)(54.42)\sqrt{\frac{1}{16} + \frac{1}{19}}$$
$$= 51.68 \pm 44.94$$
$$= 6.74 \text{ to } 96.62$$

Notice that the value of m^* we use here is larger than the value of t^* in Example 26.1. This is the price we pay for having 95% confidence not just in one interval but in all three simultaneously.

EXAMPLE 26.3 Rats and a cafeteria-style diet: multiple tests

The ANOVA null hypothesis is that all population means are equal,

$$H_0: \mu_A = \mu_B = \mu_C$$

We know from the output in Figure 26.1 that the ANOVA F test rejects this hypothesis ($F = 10.71$, $P < 0.0005$). So we have good evidence that *some* pairs of means are not the same. Which pairs? Look at the simultaneous 95% confidence intervals in Example 26.2. Which of these intervals do not contain 0? If an interval does not contain 0, we reject the hypothesis that this pair of population means are equal.

The conclusions are

We *can reject*	$H_0: \mu_B = \mu_A$
We *can reject*	$H_0: \mu_C = \mu_A$
We *cannot reject*	$H_0: \mu_C = \mu_B$

That is,

We *do* have enough evidence to conclude that $\mu_A = \mu_B$
We *do* have enough evidence to conclude that $\mu_A = \mu_C$
We *do not* have enough evidence to conclude that $\mu_B = \mu_C$

This Tukey simultaneous test of three null hypotheses has the property that when all three hypotheses are true, there is only a 5% probability that *any* of the three tests wrongly rejects its hypothesis.

Recall what a test at a fixed significance level such as 5% tells us: either we *do have enough evidence* to reject the null hypothesis, or the data *do not give enough evidence* to allow rejection.

The study found evidence that rats on a chow-only diet differ significantly in body weight from rats given restricted access and from rats given extended access to a cafeteria-style diet. However, the study did not find evidence that restricted and extended access to cafeteria food result in rats with significantly different body weights. That is, $\bar{x}_A = 605.63$ and $\bar{x}_B = 657.31$ are far enough apart to conclude that the population means μ_A and μ_B differ, $\bar{x}_A = 605.63$ and $\bar{x}_C = 691.13$ are far enough apart to conclude that the population means μ_A and μ_C differ, but $\bar{x}_B = 657.31$ is not far enough from $\bar{x}_C = 691.13$ to rule out the possibility that the population means μ_B and μ_C might be the same.

Notice that the Tukey method does not give a *P*-value for the three tests taken together. Rather, we have a set of "reject" or "fail to reject" conclusions with an overall significance level that we fixed in advance, 5% in this example. There are several other multiple-comparisons procedures that produce simultaneous confidence intervals with an overall confidence level or simultaneous tests with an overall probability of any false rejection. The Tukey procedures are arguably the most useful.[2] If you can interpret results from Tukey, you can understand output from other multiple-comparisons procedures.

APPLY YOUR KNOWLEDGE

26.1 Caffeine and sugar. Exercise 24.22 (page 628) describes a double-blind randomized experiment that assigned healthy undergraduate students to drink one of four beverages after fasting overnight: water, water with 75 mg of caffeine, water with 75 g of glucose, and water with 75 mg of caffeine and 75 g of glucose. The subjects performed a cognitive task, and their reaction times in the task are summarized below (SEM is the standard error of the mean):[3]

Beverage	*n*	$\bar{x}$	SEM
Water	18	389.35	18.50
Water and caffeine	18	320.16	17.98
Water and glucose	18	318.16	17.04
Water, caffeine, and glucose	18	336.44	14.02

An ANOVA *F* test gives a significant *P*-value of 0.0134, with MSE = 5186.0358.

(a) Because all four groups are the same size, the margin of error is the same for all 6 pairwise comparisons. Obtain this margin of error using Table G on page 26-40. Find the Tukey simultaneous 95% confidence intervals for all pairwise comparisons of population means.

(b) Explain in simple language what "95% confidence" means for these intervals.

(c) Which pairs of means differ significantly at the overall 5% significance level?

26.2 Logging in the rain forest. Exercise 24.5 (page 615) describes a study comparing forest plots in Borneo that had never been logged (Group 1) with similar plots nearby that had been logged 1 year earlier (Group 2) and 8 years earlier (Group 3). The three groups can be considered to be independent random samples. The data appear in Table 24.2 (page 616); the variable Trees is the number of trees in a plot.[4] The one-way ANOVA shown in Figure 24.4 compared the mean counts of trees in the 3 types of forest plots and was statistically significant, with $P = 0.0002$. It also gave MSE = 27.3574.

(a) Find the Tukey simultaneous 95% confidence intervals for all pairwise comparisons of population means. Use software or Table G on page 26-40.

(b) Explain in simple language what "95% confidence" means for these intervals.

(c) Which pairs of means differ significantly at the overall 5% significance level?

26.3 Which color attracts beetles best? To detect the presence of harmful insects in farm fields, we can put up boards covered with a sticky material and examine the insects trapped on the boards. Which colors attract insects best? Experimenters

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

placed six boards of each of four colors at random locations in a field of oats and measured the number of cereal leaf beetles trapped. Here are the data:[5]

Board color	Beetles trapped					
Blue	16	11	20	21	14	7
Green	37	32	20	29	37	32
White	21	12	14	17	13	20
Yellow	45	59	48	46	38	47

ANOVA gives very strong evidence ($P < 0.0005$, MSE = 32.167) that the colors differ in their ability to attract beetles.

(a) How many pairwise comparisons are there when we compare four colors?

(b) Use software or Table G on page 26-40 to obtain the Tukey simultaneous 95% confidence intervals for all pairwise comparisons of population means. Which pairs of colors are significantly different when we require a significance level of 5% for all comparisons as a group?

26.4 Dogs, friends, and stress. If you are a dog lover, perhaps having your dog along reduces the effect of stress. The EESEE story "Stress among Pets and Friends" describes a study designed to test this premise. The researchers recruited 45 women who said they were dog lovers, and randomly assigned them to three groups: alone (the control group, C), in the company of a good friend (F), or in the company of their dog (D). Each subject's mean heart rate while performing a stressful task was recorded. The study found very strong evidence (ANOVA: $F = 14.8$, $P = 0.00001$) that mean heart rates under stress do differ depending on whether a pet, a friend, or no one is present.

(a) We want to know whether the means for the two treatments (pet, friend) differ significantly from each other and from the mean for the control group. What are the corresponding three null hypotheses?

(b) We want to be 95% confident that we don't wrongly reject any of the three null hypotheses. What can we conclude from the following Tukey pairwise comparisons provided by the software R?

```
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = rate ~ group, data = Pet)
95% family-wise confidence level

            Estimate       lwr       upr
F - C == 0   8.8011   0.6296  16.9726
P - C == 0  -9.0410 -17.2125  -0.8695
P - F == 0 -17.8421 -26.0136  -9.6706
```

FOLLOW-UP ANALYSIS: CONTRASTS*

Multiple-comparisons methods give conclusions about *all* comparisons in some class with a measure of confidence that applies to all the comparisons taken together. For example, Tukey's method gives conclusions about all pairwise

*This material is optional.

comparisons among a set of population means. These methods are most useful when we did not have any specific comparison in mind before we produced the data.

Multiple-comparisons procedures sometimes give tests or confidence intervals for comparisons that don't interest us. And they may leave out comparisons that do interest us. If we have *specific questions in mind before we produce data*, it is more efficient to plan an analysis that asks these specific questions.

EXAMPLE 26.4 Which color attracts beetles best?

Holt Studios International/Alamy

What color should we use on sticky boards placed in a field of oats to attract cereal leaf beetles? Exercise 26.3 gives data from an experiment in which 24 boards (6 each of blue, green, white, and yellow) were placed at random locations in a field. ANOVA shows that there are significant differences among the mean numbers of beetles trapped by these colors. In Exercise 26.3 you followed ANOVA with Tukey pairwise comparisons.

But in fact we have specific questions in mind: We suspect that warm colors are generally more attractive than cold colors. That is, *before any data are gathered*, we suspect that blue and white boards will have similar properties, that green and yellow boards will give similar results, and also that the average beetle count for green and yellow will be greater than the average count for blue and white. Therefore, we want to test three hypotheses:

Hypothesis 1	Hypothesis 2	Hypothesis 3
H_0: $\mu_B = \mu_W$	H_0: $\mu_G = \mu_Y$	H_0: $(\mu_Y + \mu_G)/2 = (\mu_B + \mu_W)/2$
H_a: $\mu_B = \mu_W$	H_a: $\mu_G = \mu_Y$	H_a: $(\mu_Y + \mu_G)/2 > (\mu_B + \mu_W)/2$

Two of these hypotheses involve pairwise comparisons. The third does not and also has a one-sided alternative.

We can ask questions about population means by specifying *contrasts* among the means.

CONTRASTS

In the ANOVA setting comparing the means $\mu_1, \mu_2, \ldots, \mu_k$ of k populations, a **population contrast** is a combination of the means

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k$$

with numerical coefficients that add to 0, $c_1 + c_2 + \cdots + c_k = 0$.

EXAMPLE 26.5 Attracting beetles: contrasts

We can restate the three hypotheses in Example 26.4 in terms of three contrasts:

$$L_1 = (1)(\mu_B) + (0)(\mu_G) + (-1)(\mu_W) + (0)(\mu_Y)$$
$$L_2 = (0)(\mu_B) + (1)(\mu_G) + (0)(\mu_W) + (-1)(\mu_Y)$$
$$L_3 = (-1/2)(\mu_B) + (1/2)(\mu_G) + (-1/2)(\mu_W) + (1/2)(\mu_Y)$$

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

Check that the four coefficients in each line do add to 0. In terms of these contrasts, the hypotheses become

Hypothesis 1	Hypothesis 2	Hypothesis 3
$H_0: L_1 = 0$	$H_0: L_2 = 0$	$H_0: L_3 = 0$
$H_a: L_1 = 0$	$H_a: L_2 = 0$	$H_a: L_3 > 0$

Some statistical software will test hypotheses and give confidence intervals for any contrasts you specify. Because other software lacks this capability, here's how to proceed by hand, using information from basic ANOVA output.

To estimate a population contrast

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k$$

sample contrast

use the corresponding **sample contrast**

$$\hat{L} = c_1\bar{x}_1 + c_2\bar{x}_2 + \cdots + c_k\bar{x}_k$$

The sample contrast $\hat{L}$ has standard error (estimated standard deviation)

$$\text{SE}_{\hat{L}} = s_p\sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_k^2}{n_k}}$$

INFERENCE ABOUT A POPULATION CONTRAST

In the ANOVA setting, a level C **confidence interval** for a population contrast is

$$\hat{L} \pm t^*\text{SE}_{\hat{L}}$$

where $\hat{L}$ is the corresponding sample contrast and t^* is a critical value from the t distribution with the degrees of freedom for error in the ANOVA.

To test $H_0: L = 0$, use the t statistic

$$t = \frac{\hat{L}}{\text{SE}_{\hat{L}}}$$

with the same degrees of freedom.

For one-way ANOVA, the degrees of freedom for error are $N - k$, where N is the total number of observations and k is the number of populations compared (see Chapter 24, page 612). The box states the result more generally so that it applies to two-way ANOVA as well as one-way. If the contrast is a pairwise difference between means, the contrast confidence interval is exactly the individual confidence interval illustrated in Example 26.1.

EXAMPLE 26.6 Attracting beetles: inference for contrasts

Figure 26.3 displays the Minitab ANOVA output for the study on attracting cereal leaf beetles. The pooled estimate of σ is $s_p = 5.672$, and the degrees of freedom for error are 20. Minitab does not offer contrasts, so we must use a calculator.

The three sample contrasts and their standard errors are

Sample contrast	Standard error
$\hat{L}_1 = -1.334$	$\text{SE}_1 = 3.2747$
$\hat{L}_2 = -16.000$	$\text{SE}_2 = 3.2747$
$\hat{L}_3 = 23.667$	$\text{SE}_3 = 2.3153$

Here are the details for the third contrast:

$$L_3 = (-1/2)(\mu_B) + (1/2)(\mu_G) + (-1/2)(\mu_W) + (1/2)(\mu_Y)$$

$$\hat{L}_3 = (-1/2)(\bar{x}_B) + (1/2)(\bar{x}_G) + (-1/2)(\bar{x}_W) + (1/2)(\bar{x}_Y)$$

$$\hat{L}_3 = (-1/2)(14.833) + (1/2)(31.167) + (-1/2)(16.167) + (1/2)(47.167)$$

$$= 23.667$$

$$\text{SE}_3 = (5.672)\sqrt{\frac{(-1/2)^2}{6} + \frac{(1/2)^2}{6} + \frac{(-1/2)^2}{6} + \frac{(1/2)^2}{6}}$$

$$= (5.672)(0.4082)$$

$$= 2.3153$$

If you use software, your answers may differ slightly due to roundoff error in the hand calculations. A 95% confidence interval for L_3 uses $t^* = 2.086$ from Table C with df = 20:

$$\hat{L}_3 \pm t^*\text{SE}_3 = 23.667 \pm (2.086)(2.3153)$$

$$= 23.667 \pm 4.830$$

$$= 18.837 \text{ to } 28.497$$

We are 95% confident that the average number of beetles attracted by green and yellow boards exceeds the average for blue and white boards by between about 18.8 and 28.5 beetles per board.

Minitab

One-way ANOVA: Beetles versus Color

```
Source     DF         SS         MS         F         P
Color       3     4134.0     1378.0     42.84     0.000
Error      20      643.3       32.2
Total      23     4777.3

S = 5.672    R-Sq = 86.53%    R-Sq(adj) = 84.51%

                              Individual 95% CIs for Mean
                              Based on Pooled StDev
Level     N      Mean    StDev   -+-------+-------+-------+-------
Blue      6    14.833    5.345   (---*---)
Green     6    31.167    6.306             (---*---)
White     6    16.167    3.764   (---*---)
Yellow    6    47.167    6.795                       (---*---)
                              -+-------+-------+-------+-------
                              12       24       36       48

Pooled StDev = 5.672
```

◀ FIGURE 26.3
Minitab ANOVA output for the study on attracting cereal leaf beetles, for Example 26.6.

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

There is very strong evidence that the population contrast L_3 is greater than 0. The t statistic is

$$t = \frac{\hat{L}_3}{\text{SE}_3} = \frac{23.667}{2.3153} = 10.22$$

with 20 degrees of freedom, and $P < 0.0005$ from Table C. The other t tests conclude that μ_B and μ_W do not differ significantly, but that there is a significant difference between μ_G and μ_Y.

Our confidence intervals and tests for contrasts are individual procedures for each contrast. If we do inference about three contrasts, such as those in Examples 26.5 and 26.6, we face the problem of multiple comparisons again. That is, we do not have an overall confidence level for all three intervals together. There are more advanced multiple-comparisons methods that apply to contrasts just as Tukey's method applies to pairwise comparisons.

Do not use contrasts to blindly test any possible hypothesis combination. Contrasts are valid only if you have sound biological reasons to state particular hypotheses before you even look at the data.

APPLY YOUR KNOWLEDGE

26.5 Green versus yellow. Using the Minitab output in Figure 26.3, verify the values for the sample contrast $\hat{L}_2$ and its standard error given in Example 26.6. Give a 95% confidence interval for the population contrast L_2. Carry out a test of the hypothesis H_0: $L_2 = 0$ against the two-sided alternative. Be sure to state your conclusions in the setting of the study.

26.6 Rats and a cafeteria-style diet: contrasts. Figure 26.1 gives basic ANOVA output for the study of the effects of diet type on rat body weights described in Example 26.1. We might describe the overall effect of access to cafeteria food by comparing the mean body weight for rats given chow only (Group A) with the average of the mean body weights for the two groups of rats with access (restricted or extended) to cafeteria food (Groups B and C).

(a) What population contrast L expresses this comparison?

(b) Starting from the output in Figure 26.1, give the sample contrast that estimates L and its standard error.

(c) Is there good evidence that the mean weight of rats given chow only is lower than the average for the two groups of rats allowed access to cafeteria food? State the hypotheses in terms of the population contrast L and carry out a test.

(d) How much smaller is the mean weight of rats given chow only than the average for the two groups of rats allowed access to cafeteria food? Give a 95% confidence interval.

TWO-WAY ANOVA: CONDITIONS, MAIN EFFECTS, AND INTERACTION

One-way analysis of variance compares the mean responses from any set of populations or experimental treatments when the responses satisfy the ANOVA conditions. Often, however, a sample or experiment has some design structure that leads to more specific questions than those answered by the one-way ANOVA F test or by Tukey pairwise comparisons. It is common, for example, to compare treatments

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

that are combinations of values of two explanatory (independent) variables, two **factors** in the language of experimental design. Here is an example.

factor

EXAMPLE 26.7 A two-way layout

The final stages of clinical trials for a new drug often involve finding the most effective dosage and delivery method. The answer may depend on a combination of the two. To investigate this question, randomly assign informed, consenting patients (the *subjects*) to receive the new drug in one of several forms. The drug will be taken either orally or intravenously. Some subjects will receive a low dose, others a medium dose, and others a high dose.

This experiment has two *factors:* delivery method, with 2 values; and dosage, with 3 values. The 6 combinations of one value of each factor form 6 *treatments:*

		Variable *C* Dosage		
		Low	**Medium**	**High**
Variable *R* Delivery	**Oral**	Group 1	Group 2	Group 3
	Intravenous	Group 4	Group 5	Group 6

This is a *two-way layout,* with values of one factor forming rows and values of the other forming columns. After treatment completion, each subject is evaluated and his or her condition rated—for example, with a score between 0 and 100 (100 for full recovery). This is the *response variable*.

To analyze data from such a study, we impose some additional conditions. Here are the **conditions for two-way ANOVA** that will govern our work in this chapter:

two-way ANOVA conditions

1. We have *responses for all combinations* of values of the two factors (all 6 cells in Example 26.7). No combinations are missing in our data. In general, call the two explanatory variables *R* and C (for Row and Column). Variable *R* has *r* different values, and variable C has *c* different values. The study compares all *rc* combinations of these values. Such designs are called **crossed,** or fully factorial.

crossed design

2. In an observational study, we have an *independent SRS* from each of the *rc* populations. If the study is an experiment, the available subjects are allocated at random among all *rc* treatments in a *completely randomized design*. In an intermediate design, subjects can represent *independent SRSs* from each of *r* populations and be separately assigned at random to *c* treatments in a *randomized block design*. We first saw these different designs in Chapter 8.
3. The response variable has a *Normal distribution* in each population. The population mean responses may differ, but all *rc* populations have a *common standard deviation* σ.
4. We have the *same number of individuals n* in each of the *rc* treatment groups or samples. Such designs are called **balanced.**

balanced design

The second and third conditions are just the usual conditions for ANOVA applied to the two-way layout. Study designs that satisfy the first and second conditions are very common. When you design a study, you can arrange to satisfy the

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

fourth condition; that is, you can choose to have equal numbers of individuals for each treatment. Balanced designs have several advantages in any ANOVA: *F* tests are most robust against violation of the "common standard deviation" condition when the subject counts are equal or close to equal, and Tukey's method then gives exact overall confidence or significance levels. In the two-way layout, there is an even stronger reason to prefer balanced designs. If the numbers of individuals differ among treatments (an unbalanced design), several alternative analyses of the data are possible. These analyses answer different sets of questions, and you must decide which questions you want to answer. All the sets of questions and all the analyses collapse to just one in the balanced case. This makes interpreting your data much simpler.

To understand the questions that two-way ANOVA answers, return to the drug trial in Example 26.7. In this section we will assume that we know the actual population mean responses for each treatment. That is, we deal with an ideal situation in which we don't have to worry about random variation in the mean responses.

EXAMPLE 26.8 First scenario: main effects with no interaction

Here is the table from Example 26.7 populated with made-up population mean responses to the 6 treatments. (It is unrealistic to expect to know the population means, but our objective in this section is simply to describe possible scenarios.)

		Variable *C* Dosage		
		Low	Medium	High
Variable *R* Delivery	Oral	30	45	50
	Intravenous	40	55	60

The mean patient condition scores increase with higher drug dosages and are also higher when the drug is delivered intravenously. The means increase *by the same amount* (10 points) when we move from oral to intravenous drug delivery, *no matter what drug dosage is given*. Turning to the other variable, the effect of moving from low to medium dosage is the same (15 points) for both oral and intravenous delivery, and the effect of moving from medium to high dosage is also the same (5 points) for both delivery methods. Because the result of changing the value of one variable is the same for all values of the other variable, we say that there is *no interaction* between the two variables.

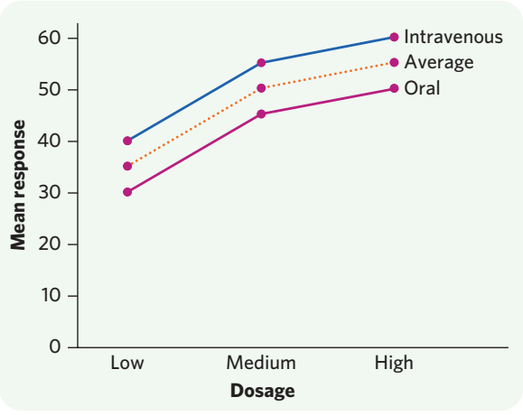
Now average the mean responses for oral and intravenous delivery. The average for oral delivery is

$$\frac{30 + 45 + 50}{3} = \frac{125}{3} = 41.7$$

and the average for intravenous delivery is

$$\frac{40 + 55 + 60}{3} = \frac{155}{3} = 51.7$$

Because the averages for the two delivery methods differ, we say that there is a *main effect* for delivery method. Similarly, average the mean responses for each



◀ FIGURE 26.4
Plot of the means for the made-up data from Example 26.8. In addition to the means for each of the six conditions, the plot displays the average for each of the low, medium, and high dosage groups as a dotted line. The parallel lines show that there is no interaction between the two factors.

drug dosage. They are 35 for the low dosage, 50 for the medium dosage, and 55 for the high dosage. So changing the drug dosage has an "on average" effect on the response (patient condition score). There is a *main effect* for drug dosage.

Figure 26.4 plots the cell means from Example 26.8. The two solid lines joining low, medium, and high dosages for oral and for intravenous delivery are *parallel*. This reflects the fact that the mean response always increases by 10 points when we move from oral to intravenous delivery, no matter what drug dosage is given. *Parallel lines in a plot of means show that there is no interaction*. It doesn't matter which variable you choose to place on the horizontal axis.

To see the main effect of drug dosage, look at the average response for oral and intravenous drug delivery at each dosage. This average is the dotted line in the plot. It changes as we move from low to medium to high dosages. *A variable has a main effect when the average response differs for different values of the variable*. "Average" here means averaged over all the values of the other variable. A main effect of drug dosage is present in Figure 26.4 because the dotted "average" line is not flat. A main effect for drug delivery method is also present but can't be seen directly in this plot.

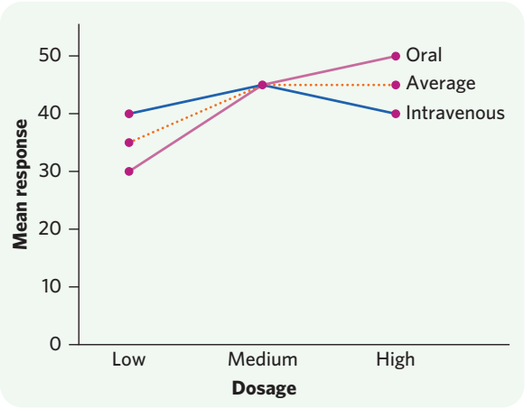
EXAMPLE 26.9 Second scenario: interactions and main effects

We continue exploring the hypothetical case of drug efficacy from Example 26.7. This time, however, let's imagine that the population means are as follows:

		Variable *C* Dosage		
		Low	**Medium**	**High**
Variable *R*	**Oral**	30	45	50
Delivery	**Intravenous**	40	45	40

Figure 26.5 plots these means, with the means for oral delivery connected by solid black lines and those for intravenous delivery by solid red lines.

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50



▶ FIGURE 26.5
Plot of the means for the made-up data from Example 26.9, along with the averages for each of the low, medium, and high dosage groups. The lines are not parallel, so there is an interaction between delivery method and dosage.

These means reflect a situation in which response to treatment decreases when the drug is given intravenously at a high dosage. This could happen, for instance, if treatment required the drug to stay in the system as long as possible but the intravenous delivery simply "dumped" the drug in the blood all at once. The mean responses for oral delivery are the same as in Example 26.8 and Figure 26.4. But when given intravenously, the medium dosage increases the mean response by only 5 points, and the high dosage drops the mean back to 40. There *is* an interaction between delivery and dosage: The difference between oral and intravenous delivery changes with the drug dosage given, so that the solid black and red lines in Figure 26.5 are *not parallel*.

There is still a main effect of dosage, because the average response (dotted line in Figure 26.5) changes as we move from low to medium to high dosages. What about the effect of delivery method? The average over all values of dosage for the oral delivery is

$$\frac{30+45+50}{3}=\frac{125}{3}=41.7$$

For the intravenous delivery this average is the same,

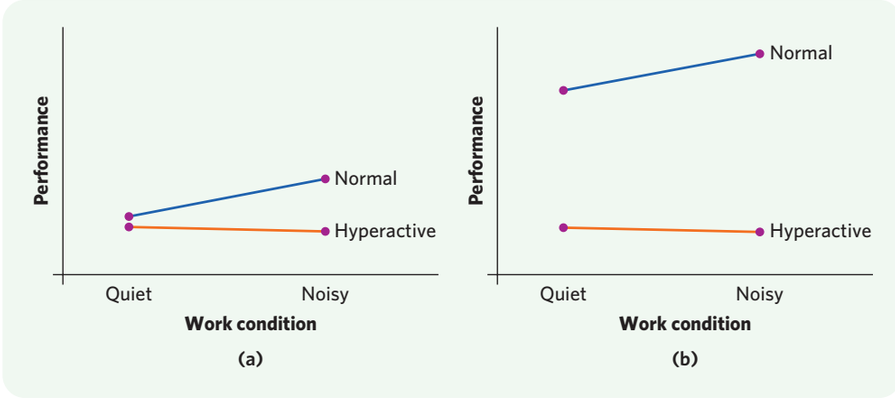
$$\frac{40+45+40}{3}=\frac{125}{3}=41.7$$

On average over all drug dosages, changing the delivery method has no effect. There is *no main effect* for delivery method. The interaction blurred the overall, "on average" effect of delivery method.

INTERACTIONS AND MAIN EFFECTS

An **interaction** is present between factors R and C in a two-way layout if the change in mean response when we move between two values of R is different for different values of C. (We can interchange the roles of R and C in this statement.)

A **main effect** for factor R is present if, when we average the responses for a fixed value of R over all values of C, we do not get the same result for all values of R.



◀ FIGURE 26.6
Plots of two sets of means for a made-up study comparing the performance of normal and hyperactive children under two conditions, for Example 26.10. The lines in (a) are not parallel, indicating an interaction between the two factors. In addition to the interaction effect, the large gap between the two lines in (b) indicates a very strong main effect of whether a child is hyperactive or not.

Main effects may have little meaning when interaction is present. After all, interaction says that the effect of changing one of the variables is different for different values of the other variable. The main effect, as an "on average" effect, may not tell us much. In Example 26.9, there is no main effect for the drug delivery method. But delivery method certainly matters—Figure 26.5 shows that there is little point giving the high drug dosage to patients if they will receive it intravenously. That's the interaction of dosage with delivery method.

EXAMPLE 26.10 Which effect is most important?

There are no simple rules for interpreting results from two-way ANOVA when strong interaction is present. You must look at plots of means and think. Figure 26.6 displays two different mean plots for a made-up study of the effects of classroom conditions on the performance of normal and hyperactive schoolchildren. The two conditions are "quiet" and "noisy," where the noisy condition is actually the usual environment in elementary school classrooms.

There is an interaction: Typical ("normal") children perform a bit better under noisy conditions, but hyperactive children perform slightly less well under noisy conditions. The interaction is exactly the same size in the two plots of Figure 26.6. To see this, look at the slopes of the "Hyperactive" lines in the two plots: They are the same. The slopes of the two "Normal" lines are also the same. So the size of the gap between normal and hyperactive changes by the same amount when we move from quiet to noisy in both plots even though the gap is much larger in Figure 26.6(b).

In Figure 26.6(a), this interaction is the most important conclusion of the study. Both main effects are small: Normal children do a bit better than hyperactive children, for example, but not a great deal better on the average.

In Figure 26.6(b), the main effect of "hyperactive or not" is the big story. Normal children perform much better than hyperactive children in both environments. The interaction is still there, but it is not very important in the face of the large difference in average performance between hyperactive and normal children.

In this section we pretended that we knew the population means so that we could discuss patterns without needing statistical inference. In practice, we simply don't know the population means. However, plotting the sample means for all groups is an essential part of data analysis for a two-way layout. Examine the graph of sample means for interaction and main effects just as we did in this section. Of

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

course, you will almost never find exactly parallel lines representing exactly no interaction in real data. Two-way ANOVA inference helps guide you because it assesses whether the interaction in the data is statistically significant. We are now ready to introduce two-way ANOVA inference.

APPLY YOUR KNOWLEDGE

Figure 26.7 shows two made-up plots of means for a two-way study that compares the yields of two varieties of soybeans (V1 and V2) when three different herbicides (H1, H2, and H3) are applied to the fields. Exercises 26.7 and 26.8 ask you to interpret these plots.

26.7 Recognizing effects. Consider the mean responses plotted in Figure 26.7(a).

(a) Is there an interaction between soybean variety and herbicide type? Why or why not?

(b) Is there a main effect of herbicide type? Why or why not?

(c) Is there a main effect of soybean variety? Why or why not?

26.8 Recognizing effects. Consider the mean responses plotted in Figure 26.7(b).

(a) Is there an interaction between soybean variety and herbicide type? Why or why not?

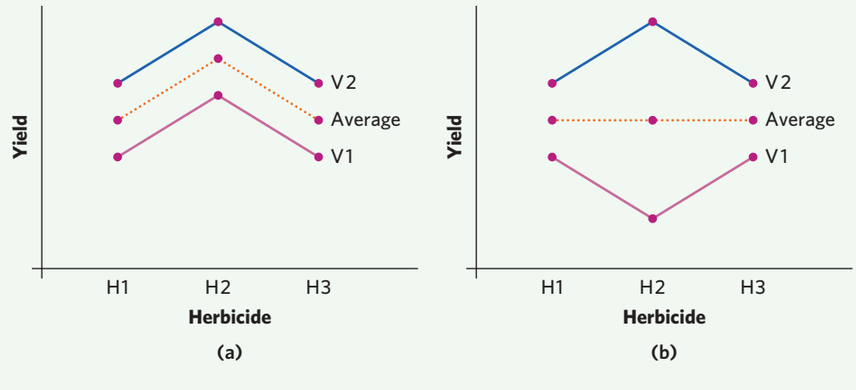
(b) Is there a main effect of herbicide type? Why or why not?

(c) Is there a main effect of soybean variety? Why or why not?

26.9 Mycorrhizal symbiosis and salt stress. Mycorrhizal symbiosis can help plants alleviate salt stress. Researchers compared the growth of mycorrhizal and nonmycorrhizal lettuce plants grown from seeds under 3 saline watering conditions (0, 40, and 80 millimoles NaCl/liter, mmol/l). The mean shoot dry weights (in grams, g) after 7 weeks are displayed below, based on 5 lettuce plants grown in each condition:[6]

	Salinity (in mmol/l)		
	0	40	80
Nonmycorrhizal	1.34	0.57	0.42
Mycorrhizal	2.54	1.77	1.51

Plot the means, and discuss the interaction and the two main effects.



▶ FIGURE 26.7
Plots of two sets (a and b) of possible means (solid lines) for a study in which the two factors are soybean variety and type of herbicide. The plots also show the average for each herbicide (dotted line).

INFERENCE FOR TWO-WAY ANOVA

Inference for two-way ANOVA is in many ways similar to inference for one-way ANOVA. Here is a brief outline:

1. Find and plot the group sample means. Study the plot to understand the interaction and main effects. Do data analysis to check the conditions for ANOVA.
2. Use software for basic ANOVA inference. There are now three F tests with three P-values, which answer the questions

   Is the interaction statistically significant?
   Is the main effect for variable R statistically significant?
   Is the main effect for variable C statistically significant?
3. You may wish to carry out a follow-up analysis. For example, Tukey's method makes pairwise comparisons among the means of all treatment groups.

We will illustrate two-way ANOVA inference with several examples. In the first example, the interaction is both small and insignificant, so that the message is in the main effects.

EXAMPLE 26.11 Dietary manipulations in fruit flies

STATE: Reproduction has a high physiological cost. A diet rich in proteins can trigger increased reproductive output in fruit flies, which we would expect to lead to the depletion of reserves such as body fat. An experiment assessed the percent of body fat in female fruit flies fed one of four diets, three of which were enriched with yeast (a high-protein food). The experiment used both wild-type fruit flies and mutants with a longer reproductive cycle. There are 8 groups in a two-way layout:

	Amount of yeast in diet (mg)			
	0	1	3	7
Wild-type	Group 1	Group 2	Group 3	Group 4
Mutant	Group 5	Group 6	Group 7	Group 8

Martin Shields/Science Source

The two factors are the genotype (wild or mutant) and the amount of yeast in the diet (in milligrams per day). Table 26.1 displays the data.[7] The response variable is the percent of body fat (lipid) after two weeks on the diet.

PLAN: Plot the sample means and discuss interaction and main effects. Check the conditions for ANOVA inference. Use two-way ANOVA F tests to determine the significance of interaction and main effects.

SOLVE: The experiment has a randomized block design, with fruit flies randomly assigned to diets separately for each genotype. We consider the blocks (genotypes) as one of the factors in analyzing the data because we are interested in comparing them.

Figure 26.8 displays side-by-side dotplots of the percent of body fat for the 5 observations in each group. The dotplots show no departures from Normality. The group standard deviations do not satisfy our rule of thumb that the largest (3.17) be no more than twice the smallest (0.69). *As the number of treatment groups increases, even samples from populations with exactly the same standard deviation are more likely to produce sample standard deviations that violate our "twice as large" rule*

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

TABLE 26.1 Percent lipid in female fruit flies by genotype and diet

Yeast amount (mg)	Fruit fly genotype				
	Mutant				
0	25.87	26.33	22.75	22.90	23.09
1	23.87	23.35	20.71	27.15	21.45
3	18.59	17.16	11.09	13.69	12.46
7	10.57	16.59	12.50	14.90	10.42
	Wild-type				
0	23.99	27.54	26.23	24.18	26.04
1	23.01	20.35	22.66	24.70	21.02
3	15.17	15.47	15.98	14.83	14.14
7	11.25	12.87	11.31	12.29	10.11

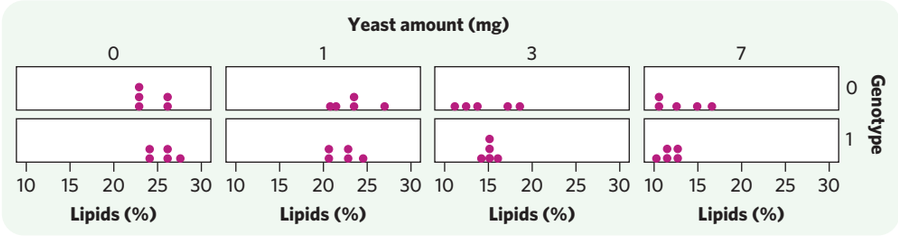
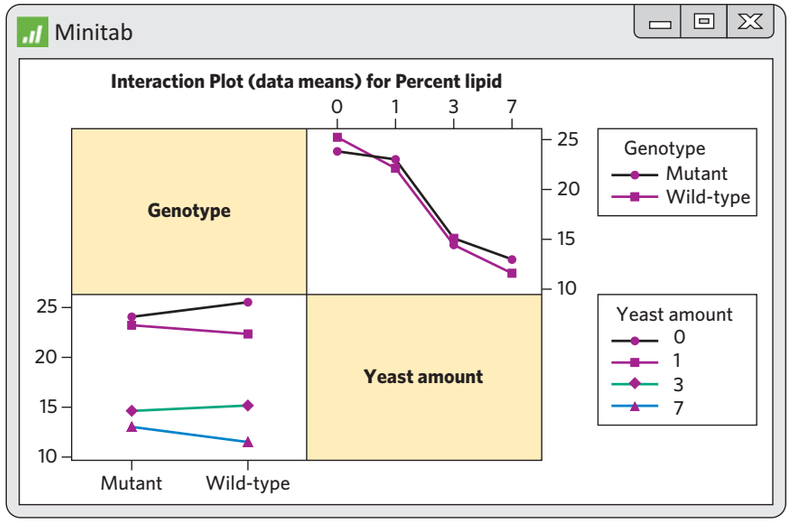


FIGURE 26.8 Side-by-side dotplots comparing the percent of body fat of wild-type and mutant female fruit flies for four different amounts of yeast in the diet, for Example 26.11. Genotype 0 = mutant; Genotype 1 = wild-type.

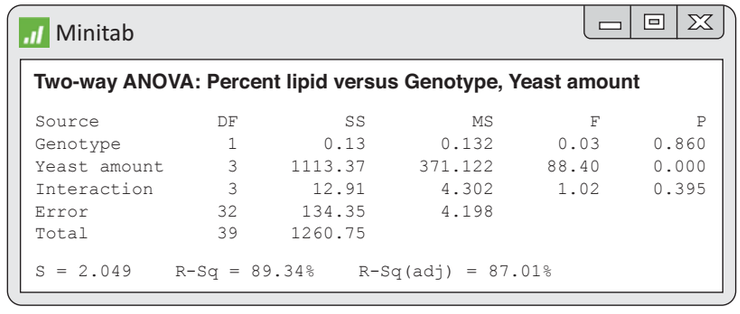
of thumb. (Think of comparing the shortest and tallest person among more and more people.) So our rule of thumb is often conservative for two-way ANOVA. Looking back at the dotplots, we can see that the data all cover a similar range and do not give the clear impression that they might have come from populations with different standard deviations. In fact, the smallest sample standard deviation arose from a sample with closely clustered observations. This chance event can easily happen with very small sample sizes, as in this experiment. We will proceed with ANOVA inference assuming that the population variances are similar enough, but we should keep in mind that our calculations could be somewhat inaccurate if this assumption is unfounded.

Figure 26.9 shows the plots of means produced by Minitab. The two plots display the same 8 sample means, with each variable marked on the horizontal axis. The plots are easy to interpret: The interaction and the main effect of genotype are both small, and the main effect of yeast amount is quite large. Figure 26.10 shows the two-way ANOVA output from Minitab. The three F tests in the Minitab output substantiate what the plots of means show: Interaction ($P = 0.395$) and genotype ($P = 0.860$) are not significant, but yeast amount ($P < 0.001$) is highly significant.

CONCLUDE: Higher amounts of protein in the diet, in the form of yeast supplements, lead to a depletion of fat reserves in fruit flies (other results in the study link this effect to an increased reproductive output). This is the only significant effect that appears in these data. In particular, the genotype of the fruit flies has very little effect on body fat at any amount of dietary yeast.



◀ FIGURE 26.9
Plots of the group means from the fruit fly study, from Minitab. The two plots use the same eight means. They differ only in the choice of which variable to mark on the horizontal axis.



◀ FIGURE 26.10
Partial two-way ANOVA output from Minitab for the fruit fly study data in Table 26.1.

The next example illustrates the situation in which there is significant interaction, but main effects are larger and more important. Think of Figure 26.6(b).

EXAMPLE 26.12 Mycorrhizal colonies and plant nutrition

STATE: Mycorrhizal fungi are present in the roots of many plants. This is a symbiotic relationship, in which the plant supplies nutrition to the fungus and the fungus helps the plant absorb nutrients from the soil. An experiment compared the effects of adding nitrogen fertilizer to two genotypes of tomato plants, a wild-type variety susceptible to mycorrhizal colonies and a mutant variety that is not. Nitrogen was added at rates of 0, 28, or 160 kilograms per hectare (kg/ha). Here is the two-way layout for the 6 treatment combinations:

		Tomato genotype	
		Mutant	**Wild**
Nitrogen	**0 kg/ha**	Group 1	Group 4
	28 kg/ha	Group 2	Group 5
	160 kg/ha	Group 3	Group 6

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

TABLE 26.2 Percent of phosphorus in tomato plants

Group	Genotype	Nitrogen amount (kg/ha)	Percent phosphorus	Group	Genotype	Nitrogen amount (kg/ha)	Percent phosphorus
1	mutant	0	0.29	4	wild-type	0	0.64
1	mutant	0	0.25	4	wild-type	0	0.54
1	mutant	0	0.27	4	wild-type	0	0.53
1	mutant	0	0.24	4	wild-type	0	0.52
1	mutant	0	0.24	4	wild-type	0	0.41
1	mutant	0	0.20	4	wild-type	0	0.43
2	mutant	28	0.21	5	wild-type	28	0.41
2	mutant	28	0.24	5	wild-type	28	0.37
2	mutant	28	0.21	5	wild-type	28	0.50
2	mutant	28	0.22	5	wild-type	28	0.43
2	mutant	28	0.19	5	wild-type	28	0.39
2	mutant	28	0.17	5	wild-type	28	0.44
3	mutant	160	0.18	6	wild-type	160	0.34
3	mutant	160	0.20	6	wild-type	160	0.31
3	mutant	160	0.19	6	wild-type	160	0.36
3	mutant	160	0.19	6	wild-type	160	0.37
3	mutant	160	0.16	6	wild-type	160	0.26
3	mutant	160	0.17	6	wild-type	160	0.27

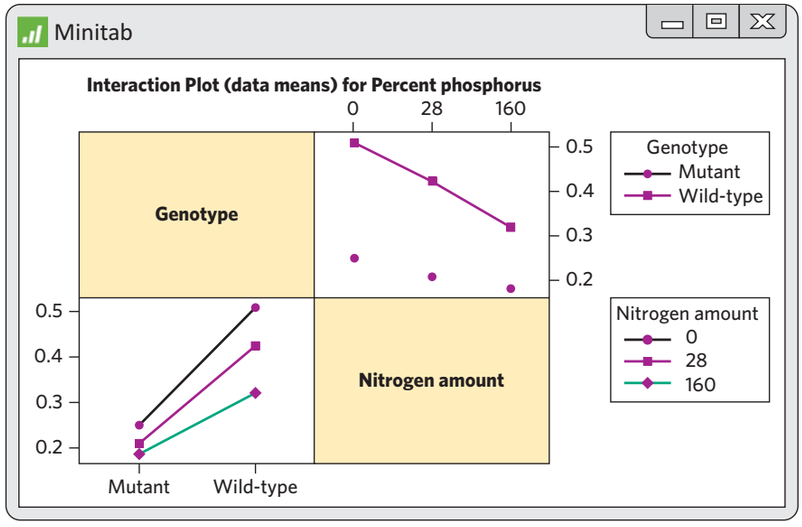
Six plants of each type were assigned at random to each level of fertilizer. The response variables describe the level of nutrients in a plant after 19 weeks, when the tomatoes are fully ripe. We will look at one response, the percent of phosphorus in the plant. Table 26.2 contains the data.[8]

PLAN: Plot the sample means and discuss interaction and main effects. Check the conditions for ANOVA inference. Use two-way ANOVA F tests to determine the significance of interaction and main effects.

SOLVE: The design is a randomized block design, with tomatoes randomly assigned to nitrogen level separately for each genotype. We consider the blocks (genotypes) as one of the factors in analyzing the data because we are interested in comparing the two genotypes.

Figure 26.11 displays Minitab's plots of the 6 sample means. The lines are not parallel, so interaction is present. The interaction is rather small compared with the main effects. The main effect of genotype is expected: Wild-type tomato plants with mycorrhizal colonies have higher phosphorus levels than the mutants at all levels of fertilization, because they benefit from symbiosis with the fungus. The main effect of fertilizer is a bit surprising: Phosphorus level goes down as the level of nitrogen fertilizer increases.

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50



◀ FIGURE 26.11
Plots from Minitab of the group means for the study of phosphorus levels in tomatoes, for Example 26.12.

Examination of the data (we don't show the details) finds no outliers or strong skewness. But the largest sample standard deviation (0.083 in Group 4) is much larger than twice the smallest (0.014 in Group 3). As in Example 26.11, it is not entirely unexpected to find relatively large differences in sample standard deviations when the sample sizes are small and the number of treatment groups is large. Nonetheless, ANOVA inference may not give correct P-values for these data. The P-values for the three two-way ANOVA F tests are $P = 0.008$ for interaction and $P < 0.001$ for both main effects. These agree with the mean plots and are so small that even if not accurate, they strongly suggest significance.

CONCLUDE: Wild-type plants, with their mycorrhizal colonies, have higher phosphorus levels than mutants that lack such colonies. Nitrogen fertilizer actually reduces phosphorus levels in both types of plants. The reduction is stronger for wild-type plants, but this interaction is not very large in practical terms.

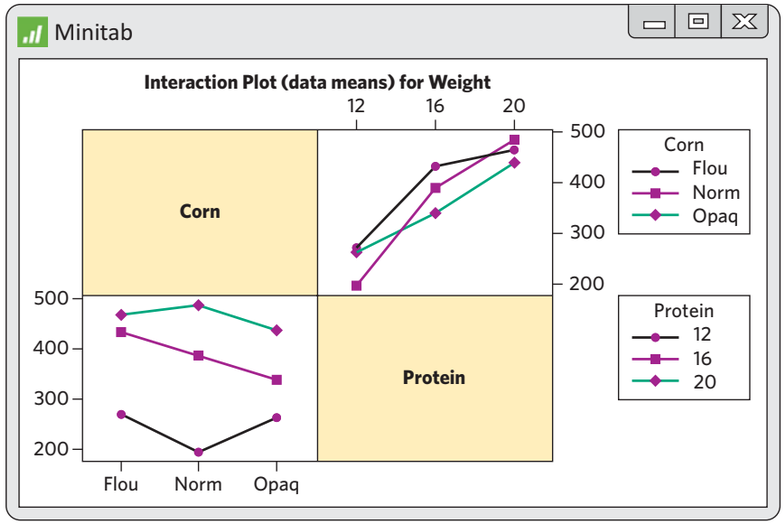
Finally, here is an example in which strong interaction makes one of the main effects meaningless. Two-way ANOVA with strong interaction is often difficult to interpret simply, as the following example illustrates.

EXAMPLE 26.13 Better corn for heavier chicks?

STATE: Corn varieties with altered amino acid content might be advantageous for feeding animals. Here is an excerpt from a study that compared normal corn ("norm" in the data file *eg26-13.dat*) with two altered varieties called opaque-2 ("opaq") and floury-2 ("flou").

> Nine treatments were arranged in a 3 × 3 factorial experiment to compare opaque-2, floury-2, and normal corn at dietary protein levels of 20, 16, and 12%. Corn-soybean meal diets containing either opaque-2, floury-2, or normal corn were formulated so that, for a given protein level, an equivalent amount of corn protein was supplied by each type of corn. Male broiler-type chicks were randomly allotted to treatments at 1 day of age. Feed and water were provided *ad libitum*. Chicks were weighed at weekly intervals until termination of the experiment at 21 days.[9]

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50



▶ **FIGURE 26.12**
Plots from Minitab of the mean weights of 21-day-old chicks fed 9 different diets, for Example 26.13.

There are 10 chicks in each group. The response variable is the weight in grams after 21 days. Which combinations of corn type and protein content lead to the most growth?

PLAN: Plot the sample means and discuss interaction and main effects. Check the conditions for ANOVA inference. Use two-way ANOVA F tests to determine the significance of interaction and main effects. If necessary, use Tukey pairwise comparisons to identify significant differences among treatments.

SOLVE: The experiment has a balanced, completely randomized design with 9 treatments. Figure 26.12 shows Minitab's plots of the sample means. The mean weight of the chicks increases with the percent of protein in the diet, as expected. We are primarily interested in comparing the three types of corn. There are important interaction effects. Normal corn does poorest of the three types at 12% protein and best at 20%. Floury is best at both 12% and 16%. Opaque is always inferior to floury and beats normal corn only at 12% protein.

Although we don't show the details, ANOVA is justified. There are no outliers or strong skewness. The largest sample standard deviation (61.16 for Group 4) is a little bit more than twice the smallest (25.99 for Group 6), but this is common when we have 9 groups even when all populations really have the same σ. Figure 26.13 contains Minitab's ANOVA output. We included the means for the 9 groups, for the three types of corn, and for the three levels of protein. The main effects can be seen in the different mean weights for the corn types and for the protein levels. The three F tests are all highly significant. Although there is a substantial main effect for corn (the means range from 346.84 g for opaque to 390.27 g for floury), this has little meaning in light of the interaction that we just described.

When strong interaction makes one or both main effects hard to interpret, it is often useful to find the Tukey pairwise multiple comparisons for all the treatment groups. One way to do this is to do a one-way ANOVA with just "group" as the explanatory variable. There are 36 pairwise comparisons among 9 groups. Minitab's Tukey output is both long and hard to understand in such cases. Here

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

Minitab

ANOVA: Weight versus Corn, Protein

Analysis of Variance for Weight

Source	DF	SS	MS	F	P
Corn	2	31137	15568	7.13	0.001
Protein	2	745621	372810	170.66	0.000
Corn*Protein	4	60894	15223	6.97	0.000
Error	81	176951	2185		
Total	89	1014602			

S = 46.7395 R-Sq = 82.56% R-Sq(adj) = 80.84%

Means

Corn	N	Weight
flou	30	390.27
norm	30	356.63
opaq	30	346.84

Protein	N	Weight
12	30	243.49
16	30	387.33
20	30	462.93

Corn	Protein	N	Weight
flou	12	10	271.39
flou	16	10	434.22
flou	20	10	465.21
norm	12	10	195.34
norm	16	10	389.03
norm	20	10	485.53
opaq	12	10	263.73
opaq	16	10	338.75
opaq	20	10	438.05

◀ FIGURE 26.13
Two-way ANOVA output from Minitab for the study of the effect of diet on chick growth, for Example 26.13.

is a condensed version using an idea that some software (though not Minitab) implements. Arrange the 9 groups in the order of their sample means, from smallest to largest. We have identified the groups both by their group number in the data file and by the treatment. Connect all pairs that do *not* differ significantly at the overall 5% level with an underline:

```
Treatment:  N12  O12  F12  O16  N16  F16  O20  F20  N20
 Group:       4    7    1    8    5    2    9    3    6
              ------    ------
                             ----------
                                  ----------------
```

Group 4 has a significantly smaller mean weight than any other group. Groups 7 and 1 do not differ significantly, but they are higher than Group 4 and lower than all other groups. Group 8 is not significantly different from 5 but is higher than 4, 7, and 1 and lower than 2, 9, 3, and 6. And so on. The most interesting finding is that at the high end, Groups 2, 9, 3, and 6 do not have significantly different mean weights. Three of these are the three 20% protein groups, but floury corn with 16% protein belongs with these three.

CONCLUDE: More protein clearly helps chicks grow faster. The three types of corn do not differ significantly when the diet has 20% protein. Floury corn is superior to both opaque and normal corn at middle (16%) and low (12%) protein levels, though not all differences at these levels are statistically significant.

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

APPLY YOUR KNOWLEDGE

NHPA/SuperStock

26.10 Hooded rats: social play times. How does social isolation during a critical development period affect the behavior of hooded rats? Psychology students assigned 24 young female rats at random to either isolated or group housing, then similarly assigned 24 young male rats. This is a randomized block design with the gender of the 48 rats as the blocking variable and housing type as the treatment. Later, the students observed the rats at play in a group setting and recorded data on three types of behavior (object play, locomotor play, and social play).[10] The file *ex26-10.dat* records the time (in seconds) that each rat devoted to social play during the observation period.

(a) Make a plot of the 4 group means. Is there a large interaction between gender and housing type? Which main effect appears to be more important?

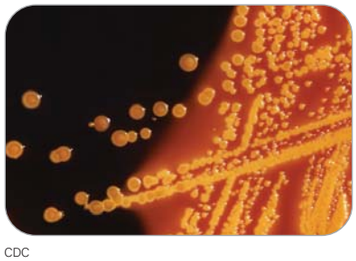
(b) Verify that the conditions for ANOVA inference are satisfied.

(c) Here is the ANOVA table from Minitab:

Source	DF	SS	MS	F	P
Sex	1	11193.5	11193.5	12.85	0.001
Housing	1	623.5	623.5	0.72	0.402
Sex*Housing	1	72.5	72.5	0.08	0.774

What are the F statistics and P-values for interaction and the two main effects? Do these values support your tentative interpretation of the graph?

26.11 Evolution in bacteria. Exercise 24.31 (page 630) described an experiment on *E. coli* bacteria to test the theory of evolution. Lines of *E. coli* bacteria grown and kept at a neutral pH of 7.2 were "evolved," that is, grown for 2000 generations (about 300 days) at a stressful acidic pH of 5.5 and then tested against their ancestor at various pH values to determine their relative fitness. A control was also performed in which bacteria were "evolved" at the neutral pH and tested against their ancestors at various pH values. The objective was to determine whether any change in relative fitness could be attributed to directional evolution (as a result of an environmental change) or simply a chance event. Here are the relative fitness scores obtained for test environments of pH 5.5 (acid), 7.2 (neutral), or 8.0 (basic). There were six replicates using different original bacterial lines for each test group:[11]



CDC

Bacteria "evolved" at pH 5.5						
Test in acid pH	1.24	1.22	1.23	1.24	1.18	1.09
Test in neutral pH	0.99	0.99	0.98	0.94	0.95	0.95
Test in basic pH	0.56	0.83	0.82	0.72	0.86	0.84
Bacteria "evolved" at pH 7.2						
Test in acid pH	1.02	1.04	0.99	1.10	1.12	1.08
Test in neutral pH	1.15	1.06	1.04	0.93	1.02	1.03
Test in basic pH	0.80	1.06	1.07	1.11	0.96	1.04

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

A score of 1 indicates that both bacteria types (evolved and ancestral) are equally fit. A score larger than 1 indicates that the evolved line is more fit than the ancestral line.

(a) Plot the means on a single graph and describe the main effects and interaction.

(b) Software gives the following information:

```
Evolution pH     F=4.38     P=0.045
Test pH          F=27.85    P<0.001
Interaction      F=17.23    P<0.001
```

Do these values support your interpretation of the graph?

(c) Check that the conditions for ANOVA are met. If you find deviations from the recommendations, describe them and explain how they might affect your interpretation in (b).

SOME DETAILS OF TWO-WAY ANOVA*

All ANOVA F statistics work on the same principle: Compare the variation due to the effect being tested with a benchmark level of variation that would be present even if that effect were absent. The three F tests for two-way ANOVA use the same benchmark as the one-way ANOVA F test, namely, the variation among individual responses within each treatment group.

In two-way ANOVA, we have two factors (also called explanatory or independent variables) that form treatments in a two-way layout. Factor R has r values and Factor C has c values, so that there are rc treatments. If the design is balanced, then the same number n of subjects are assigned to each treatment. The two-way layout that results is as follows:

		Column factor *C*			
		1	**2**	···	***c***
Row factor *R*	**1**	n subjects	n subjects	···	n subjects
	2	n subjects	n subjects	···	n subjects
	⋮	⋮	⋮		⋮
	r	n subjects	n subjects	···	n subjects

The number of treatments is $k = rc$

The total number of observations is $N = rcn = kn$

Figure 26.14 presents both one-way and two-way ANOVA output for the same data, from the fruit fly study in Example 26.11. The one-way analysis just compares the means of rc treatments, ignoring the two-way layout. In discussing one-way ANOVA, we called the number of treatments k. Now $k = rc$. The two-way analysis takes into account that each treatment is formed by combining a value of R with a value of C.

*This optional material requires the optional section on some details of ANOVA from Chapter 24, page 621.

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

Minitab

One-way ANOVA: Percent lipid versus Groups

Source	DF	SS	MS	F	P
Groups	7	1126.41	160.92	38.33	0.000
Error	32	134.35	4.20		
Total	39	1260.75			

S = 2.049 R-Sq = 89.34% R-Sq(adj) = 87.01%

Two-way ANOVA: Percent lipid versus Genotype, Yeast

Source	DF	SS	MS	F	P
Genotype	1	0.13	0.132	0.03	0.860
Yeast amount	3	1113.37	371.122	88.40	0.000
Interaction	3	12.91	4.302	1.02	0.395
Error	32	134.35	4.198		
Total	39	1260.75			

S = 2.049 R-Sq = 89.34% R-Sq(adj) = 87.01%

▶ FIGURE 26.14
Compare the sums of squares in these one-way and two-way ANOVA outputs for the fruit fly data from Table 26.1.

In both settings, analysis of variance breaks down the overall variation in the observations into several pieces. The overall variation is expressed numerically by the **total sum of squares**

total sum of squares

$$\text{SST} = \sum (\text{individual observation} - \text{mean of all observations})^2$$

where the sum runs over all N individual observations. If we divide SST by $N - 1$, we get the variance of the observations. So SST is closely related to a familiar measure of variability. Because SST uses just the N individual observations, it is the same for both one-way and two-way analyses. You can see in Figure 26.14 that SST = 1260.75 for these data.

One-way ANOVA. We saw in Chapter 24 that the one-way ANOVA F test compares the variation among the k treatment means with the variation among responses to each treatment. If the means vary more than we would expect based on the variation among subjects who receive the same treatment, that's evidence of a difference among the mean responses in the k populations.

Let's give a bit more detail. One-way ANOVA breaks down the total variation into the sum of two parts:

$$\begin{aligned}\text{total variation among responses} &= \text{variation among treatment means} \\ &\quad + \text{variation among responses to each treatment} \\ \text{total sum of squares} &= \text{sum of squares for groups} \\ &\quad + \text{sum of squares for error} \\ \text{SST} &= \text{SSG} + \text{SSE}\end{aligned}$$

Formulas for the sum of squares for groups (SSG) and the sum of squares for error (SSE) appear in Chapter 24 as the numerators of the mean square for groups (MSG) and the mean square for error (MSE), respectively, but we won't concern ourselves with the algebra. Remember that "error" is the traditional term in ANOVA for variation among observations. It doesn't imply that some mistake has been made. In the one-way output in Figure 26.14, you see that the breakdown

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

for these data is, aside from rounding error,

$$\begin{aligned} \text{SST} &= \text{SSG} + \text{SSE} \\ 1260.75 &= 1126.41 + 134.35 \end{aligned}$$

The one-way ANOVA F test is performed in two stages:

1. Divide each sum of squares by its *degrees of freedom* to get the *mean squares* MSG for groups and MSE for error:

$$\text{MSG} = \frac{\text{SSG}}{k-1} \qquad \text{MSE} = \frac{\text{SSE}}{N-k}$$

2. The one-way ANOVA F statistic compares MSG with MSE:

$$F = \frac{\text{MSG}}{\text{MSE}}$$

Find the P-value from the F distribution with $k-1$ and $N-k$ degrees of freedom.

ANOVA table

The **analysis of variance table** in the output reports sums of squares, their degrees of freedom, mean squares, and the F statistic with its P-value.

Two-way ANOVA. Now look at the two-way analysis of variance table in Figure 26.14:

- The total sum of squares and the error sum of squares are the same as in the one-way analysis.
- The sum of squares for groups in one-way is the sum of the three sums of squares for main effects and interaction in two-way.

This is the heart of two-way analysis of variance: Break down the variation among the rc groups into three parts—variation due to the main effect of Factor R, variation due to the main effect of Factor C, and variation due to interaction between the two factors. Each type of variation is measured by a sum of squares. The formulas for the two main effects sums of squares are similar to that for the one-way sum of squares for groups, but we will again ignore the algebraic details. The interaction sum of squares is best thought of as what's left over, the variation among treatments that isn't explained by the two main effects. In symbols,

$$\begin{aligned} \text{total sum of squares} &= \text{sum of squares for main effect of Factor } R \\ &\quad + \text{sum of squares for main effect of Factor C} \\ &\quad + \text{sum of squares for interaction between } R \text{ and C} \\ &\quad + \text{sum of squares for error} \\ \text{SST} &= \text{SSR} + \text{SSC} + \text{SSRC} + \text{SSE} \end{aligned}$$

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

Each of these sums of squares has its own degrees of freedom, and these break down in the same way:

$$\begin{aligned}\text{total df} &= \text{df for main effect of Factor } R \\ &\quad + \text{df for main effect of Factor C} \\ &\quad + \text{df for interaction between } R \text{ and C} \\ &\quad + \text{df for error} \\ rcn - 1 &= (r-1) + (c-1) + (r-1)(c-1) + rc(n-1) \\ &= N - 1\end{aligned}$$

You can check that the total degrees of freedom are $N - 1$ and the degrees of freedom for error are $N - k$, the same as for one-way ANOVA. The one-way degrees of freedom for groups are equal to the sum of the degrees of freedom for the three two-way effects.

EXAMPLE 26.14 Comparing one-way and two-way ANOVA

The data behind Figure 26.14 appear in Table 26.1. The two factors are R = genotype and C = yeast amount. Factor R has $r = 2$ values: mutant and wild-type. Factor C has $c = 4$ values: 0 mg, 1 mg, 3 mg, and 7 mg. There are $k = 8$ treatments and $n = 5$ observations for each treatment, resulting in $N = 40$ observations.

The total degrees of freedom are $N - 1 = 39$. The degrees of freedom for error is $N - k = 40 - 8 = 32$. In the one-way analysis, the degrees of freedom for "Groups" are $k - 1 = 7$. The two-way analysis breaks this into degrees of freedom $r - 1 = 1$ for Factor R, $c - 1 = 3$ for Factor C, and $(r - 1)(c - 1) = 1 \times 3 = 3$ for interaction.

Here are the two breakdowns of the total variation and the degrees of freedom that appear in Figure 26.14:

One-way			Two-way		
Sums of squares		**df**	**Sums of squares**		**df**
SSG	1126.41	7	SSR	0.13	1
			SSC	1113.37	3
			SSRC	12.91	3
SSE	134.35	32	SSE	134.35	32
SST	1260.75	39	SST	1260.75	39

The neat breakdown of SSG into three effects depends on the balance of the two-way layout. It doesn't hold when the counts of observations are not the same for all treatments. That's why two-way ANOVA is more complicated and harder to interpret for unbalanced data than it is for balanced data.

Two-way ANOVA *F* tests. Finally, form three F statistics exactly as in the one-way setting.

1. Divide each sum of squares by its *degrees of freedom* to get the *mean squares* for the three effects and for error:

$$\text{MSR} = \frac{\text{SSR}}{r-1} \qquad \text{MSC} = \frac{\text{SSC}}{c-1} \qquad \text{MSRC} = \frac{\text{SSRC}}{(r-1)(c-1)}$$

$$\text{MSE} = \frac{\text{SSE}}{N-k}$$

2. The three F statistics compare the mean squares for the three effects with MSE.

TWO-WAY ANOVA *F* TESTS

The F statistics for the three types of treatment effects in two-way ANOVA are

For the main effect of Factor R,	$F = \frac{\text{MSR}}{\text{MSE}}$	with dfs $r-1$ and $N-k$
For the main effect of Factor C,	$F = \frac{\text{MSC}}{\text{MSE}}$	with dfs $c-1$ and $N-k$
For the interaction of R and C,	$F = \frac{\text{MSRC}}{\text{MSE}}$	with dfs $(r-1)(c-1)$ and $N-k$

In all cases, large values of F are evidence against the null hypothesis that the effect is not present in the populations.

APPLY YOUR KNOWLEDGE

26.12 Hooded rats: social play times, continued. Exercise 26.10 gave the two-way ANOVA table for a study of the effect of social isolation on hooded rats. The response variable is the time (in seconds) that a rat devoted to social play during an observation period. Start your work in this exercise with the two-way ANOVA table.

(a) Explain how the sums of squares from the two-way ANOVA table can be combined to obtain the one-way ANOVA sum of squares for the 4 groups (SSG). What is the value of SSG?

(b) Give the degrees of freedom, mean square (MSG), and F statistic for testing for the effect of groups in the one-way ANOVA setting.

(c) Is there a significant effect of group on the amount of time spent in play? Give and interpret the P-value in the context of this experiment.

(d) Use software to carry out one-way ANOVA of time on group. Verify that your results in parts (b), (c), and (d) agree with the software output.

26.13 Evolution in bacteria, continued. Exercise 26.11 gives the F statistics and P-values for the main effects and interaction of a two-way ANOVA of bacterial relative fitness as a function of evolution pH and test pH.

(a) What are the degrees of freedom for each of the three F values?

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

(b) The mean squares (MS) for evolution pH, test pH, and interaction are 0.027225, 0.172978, and 0.107033, respectively. Use this information to fill in the complete ANOVA table for the two-way ANOVA test (you can refer to Figure 26.10 for a model).

CHAPTER 26 SUMMARY

- **Two-way analysis of variance (ANOVA)** compares the means of several populations formed by combinations of two factors *R* and C in a **two-way layout.**
- The **conditions for ANOVA** state that we have an **independent SRS** from each population (or a completely randomized experimental design), that each population has a **Normal distribution,** and that all populations have the **same standard deviation.** In this chapter we consider only examples that satisfy the additional conditions that the design producing the data be **crossed** (all combinations of the factors are present) and **balanced** (all factor combinations are represented by the same number of individuals).
- A factor has a **main effect** if the mean responses for each value of that factor, averaged over all values of the other factor, are not the same. The two factors **interact** if the effect of moving between two values of one factor is different for different values of the other factor. **Plot the treatment mean responses** to examine main effects and interaction.
- There are three **ANOVA *F* tests:** for the null hypotheses of no main effect for Factor *R*, no main effect for Factor C, and no interaction between the two factors.
- **Follow-up analysis** is often helpful in both one-way and two-way ANOVA settings. **Tukey pairwise multiple comparisons** give confidence intervals for all differences between pairs of treatment means with an **overall level of confidence.** That is, we can be (say) 95% confident that *all* the intervals simultaneously capture the true population differences between means. When the data lend themselves to more specific hypotheses, contrasts of combinations of means can be created and used for inference.

STATISTICS IN SUMMARY

Here are the most important skills you should have acquired from reading this chapter.

A. Recognition

1. Recognize the two-way layout, in which we have a quantitative response to treatments formed by combinations of values of two factors.
2. Recognize when comparing mean responses to the treatments in a two-way layout is helpful in understanding data.
3. Recognize when you can use two-way ANOVA to compare means. Check the data production, the presence of outliers, and the sample standard deviations for the groups you want to compare. Look for data production designs that are crossed and balanced.

B. Interpreting Two-Way ANOVA

1. Plot the sample means for the treatments or groups. Based on your plot, describe the main effects and interaction that appear to be present.
2. Decide which effects are most important in practice. Pay particular attention to whether a large interaction makes one or both main effects less meaningful.
3. Use software to carry out two-way ANOVA inference. From the *P*-values of the three *F* tests, learn which effects are statistically significant.

C. Follow-up Analysis

1. Decide when it is helpful to know which differences among treatment means are significant.
2. Use software to carry out Tukey pairwise multiple comparisons among all the means you want to compare.
3. Understand the meaning of the overall confidence level and the overall level of significance provided by Tukey's method for a set of confidence intervals or a set of significance tests.

THIS CHAPTER IN CONTEXT

In Chapter 24 we described the one-way analysis of variance F test, a method used as a first step in comparing the means of several populations. One-way ANOVA allows us to avoid the problems associated with multiple comparisons by first asking if there is evidence that these population means are not all equal. When the ANOVA F test is statistically significant, we have evidence that at least one population mean in the set is significantly different from at least one other population mean. But ANOVA does not say which ones.

In this chapter we examine techniques, such as Tukey comparisons and contrasts, that allow us to compare a set of population means two at a time while maintaining a reasonable overall probability of committing a Type I error. These follow-up tests are not fundamentally different from the two-sample t procedures that we saw in Chapter 18 and used to perform statistical inference for the difference $\mu_1 - \mu_2$ between the means of two distinct populations.

We also expand the use of analysis of variance to more complex settings. Specifically, we describe how the two-way ANOVA procedure allows comparisons of population means obtained from designs involving two different factors, or explanatory variables. The advantage of a two-way design is that it allows us to consider possible interactions between the factors. This is a concept that we will revisit in companion Chapter 28 on multiple regression.

CHECK YOUR SKILLS

26.14 To avoid the problem of multiple comparisons in an ANOVA follow-up analysis, one should

(a) use contrasts for all pairwise comparisons of means.

(b) use tests that guarantee a given significance level for all pairwise comparisons of means (such as Tukey tests).

(c) use a series of two-sample t tests for all pairwise comparisons of means.

26.15 An ANOVA comparing three population means finds a significant P-value. With a 5% significance level, the pairwise Tukey tests find that μ_1 is significantly different from μ_2 but the other tests are not significant. These results

(a) say that we have significant evidence that $\mu_1 = \mu_2$ but not enough evidence to reject the other two null hypotheses.

(b) must be wrong because if $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$, then $\mu_1 = \mu_3$ is not possible.

(c) are not very convincing because they are based on multiple comparisons.

26.16 Inference about a population contrast tests

(a) whether any linear combination of population means we are interested in is equal to zero.

(b) whether a linear combination of population means is equal to zero if that hypothesis made sense before seeing the sample data.

(c) whether one population mean is equal or not to a linear combination of the other population means in a series of experiments.

26.17 The purpose of a two-way ANOVA is to examine

(a) the combined effect of two factors on a quantitative variable.

(b) the variances of two populations.

(c) the means of two populations.

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

Minitab

Two-way ANOVA: Number versus Housing, Sex

Source	DF	SS	MS	F	P
Housing	1	697.8	697.840	4.27	0.041
Sex	1	742.6	742.563	4.55	0.035
Interaction	1	184.5	184.507	1.13	0.290
Error	140	22870.7	163.362		
Total	143	24495.7			

S = 12.78 R-Sq = 6.63% R-Sq(adj) = 4.63%

▲ **FIGURE 26.15** Two-way ANOVA table for Exercises 26.20 to 26.23.

26.18 In a two-way ANOVA, an interaction effect is present when

(a) the values of both factors are the same.

(b) the values of both factors are different.

(c) the effect of changing one of the variables is different for different values of the other variable.

26.19 In a two-way ANOVA test, when the interaction effect is significant,

(a) one should interpret the main effects with caution, especially when the interaction effect is large.

(b) one should simply ignore the main effects.

(c) one should describe the two main effects separately.

A study examined the effects of social isolation versus social housing on the development of rats. Two-way ANOVA was used to investigate the effects of housing condition and the rats' gender on the number of play behaviors. Use the ANOVA table in Figure 26.15 to answer the questions below.

26.20 How many treatment groups were there for this experiment?

(a) 1 (b) 2 (c) 4

26.21 What is the value of the test statistic for interaction between housing condition and gender?

(a) 0.041 (b) 0.290 (c) 1.13

26.22 Which statement below provides the best summary of the housing effect?

(a) Housing is significant at the 1% level but not at the 5% level.

(b) Housing is significant at the 5% level but not at the 1% level.

(c) Housing is not significant at either the 1% or the 5% level.

26.23 You conclude from this experiment that the effect of sex on number of play behaviors

(a) is significant at the 5% level and this effect does not depend on housing type.

(b) is significant at the 5% level but this effect depends on housing type.

(c) should be ignored because of the interaction effect.

CHAPTER 26 EXERCISES

26.24 Logging in the rain forest, continued. In Exercise 26.2 you compared the mean numbers of trees in forest plots in Borneo with different logging histories, using the data in Table 24.2 (page 616). This table also provides data on the variable Species, the number of tree species in a plot. The one-way ANOVA on this variable was carried out in Exercise 24.28 (page 629) and was found statistically significant ($P = 0.006$). Here are the Tukey 95% simultaneous confidence intervals given by software:

1.599	to	9.901	for	$\mu_A - \mu_B$
−0.650	to	8.317	for	$\mu_A - \mu_C$
−6.400	to	2.567	for	$\mu_B - \mu_C$

(a) Write a short summary of your interpretation of the ANOVA test and follow-up analysis.

(b) Use software or Table G on page 26-40 to verify the software output provided here.

26.25 Does nature heal best? The body has a natural electrical field that helps wounds heal. Might higher or lower levels speed healing? An experiment on skin healing rate in anesthetized newts compared the natural electrical field of the skin to several imposed electrical field intensities (0, 0.5, 1.25, 1.5, and the control 1). The response variable is the difference in healing rate (in micrometers per hour) of cuts made in the experimental limb and the undisturbed limb of each newt. Negative values mean that the experimental limb healed more slowly. A one-way ANOVA gives a P-value of 0.005. Here

are the 10 pairwise Tukey 95% simultaneous confidence intervals given by software:

−12.24	to	13.65	for	$\mu_{0.5} - \mu_0$
−6.01	to	18.51	for	$\mu_1 - \mu_0$
−4.86	to	20.63	for	$\mu_{1.25} - \mu_0$
−20.13	to	5.36	for	$\mu_{1.5} - \mu_0$
−6.18	to	17.27	for	$\mu_1 - \mu_{0.5}$
−5.05	to	19.41	for	$\mu_{1.25} - \mu_{0.5}$
−20.31	to	4.14	for	$\mu_{1.5} - \mu_{0.5}$
−9.87	to	13.14	for	$\mu_{1.25} - \mu_1$
−25.14	to	−2.13	for	$\mu_{1.5} - \mu_1$
−27.28	to	−3.25	for	$\mu_{1.5} - \mu_{1.25}$

(a) Explain why doing a Tukey pairwise follow-up analysis is legitimate here.

(b) What can you conclude from the ANOVA test and follow-up analysis?

26.26 Does nature heal best? Continued. Interpreting many pairwise Tukey comparisons can be challenging. It is always preferable to start with a specific hypothesis based on biological reasoning. Refer to Exercise 26.25, with raw data available in Table 26.3.

▶ TABLE 26.3 Effect of electrical field on healing rate in newts

Level 0	Level 0.5	Level 1	Level 1.25	Level 1.5
−10	−1	−7	1	−13
−12	10	15	8	−49
−9	3	−4	−15	−16
−11	−3	−16	14	−8
−1	−31	−2	−7	−2
6	4	−13	−1	−35
−31	−12	5	11	−11
−5	−3	−4	8	−46
13	−7	−2	11	−22
−2	−10	−14	−4	2
−7	−22	5	7	10
−8	−4	11	−14	−4
	−1	10	0	−10
	−3	3	5	2
		6	−2	−5
		−1		
		13		
		−8		

(a) Run a contrast test between nature's way (control group, "1") and the average of all four treatment groups. State the hypotheses, compute the statistic, and find the *P*-value. What is your conclusion?

(b) Explain why this test is easier to interpret biologically than the 10 Tukey tests of Exercise 26.25.

(c) Plot the means for the 5 groups on one graph. It might be tempting based on this graph to run a contrast for Groups 1 and 1.25 against all 3 other groups. Explain why this would not be appropriate.

26.27 Toxicology of lead acetate. Lead acetate is a widely used chemical often involved in dyeing textiles or hair. The National Toxicology Program (NTP) reports a study of the impact of exposure to lead acetate on the hematology of female B6C3F1 mice. The mice were randomly assigned to receive in their diet a given amount of lead acetate for 15 days. Various aspects of their hematology were examined at the end of the experiment. Here are the data for their blood hemoglobin levels after the 15-day exposure to lead acetate:[12]

Lead acetate	Hemoglobin (g/dl)							
0 (control)	16.4	16.4	16.8	16.9	17.0	16.7	17.0	16.4
500 mg/kg	15.7	16.4	16.6	16.7	17.2	17.1	15.7	16.8
1000 mg/kg	16.6	16.4	16.0	16.2	16.6	17.0	16.4	16.5
2000 mg/kg	16.0	13.5	15.1	14.4	14.9	15.0	14.0	15.3

(a) Does the consumption of lead acetate influence the hemoglobin level of female B6C3F1 mice? Run the ANOVA test, following the four-step process outlined in Chapter 24.

(b) Here are the 6 pairwise Tukey 95% simultaneous confidence intervals given by software for this study:

−0.9060	to	0.5560	for	$\mu_{500} - \mu_0$
−0.9995	to	0.5138	for	$\mu_{1000} - \mu_0$
−2.6560	to	−1.1940	for	$\mu_{2000} - \mu_0$
−0.8245	to	0.6888	for	$\mu_{1000} - \mu_{500}$
−2.4810	to	−1.0190	for	$\mu_{2000} - \mu_{500}$
−2.4388	to	−0.9255	for	$\mu_{2000} - \mu_{1000}$

Interpret the ANOVA results in light of this follow-up analysis.

26.28 Toxicology of lead acetate, continued. To analyze the study in Exercise 26.27, you could legitimately run a contrast test comparing the control (0 mg/kg lead acetate) to the three other treatment groups. Compute a 95% confidence interval for this contrast and conclude.

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

26.29 Neural basis of 3D vision. Exercise 24.32 (page 631) discussed the neural coding of stereopsis, the process of 3D vision based on the disparity of visual signals reaching both retinas. The exercise provided the response (in spikes per second) of a neuron to stereograms of 6 different disparities.

(a) How many pairwise comparisons are there among the means of 6 populations?

(b) Use Tukey's method to compare these means at the overall 5% significance level.

26.30 Hooded rats: object play times. Exercise 26.10 describes an experiment to study the effects of social isolation on the behavior of hooded rats. You have analyzed the effects on social play time. Now look at another response variable, the time that a rat spends in object play during an observation period. The file *ex26-30.dat* records the time (in seconds) that each rat devoted to object play.

(a) Make a plot of the 4 group means. Is there a large interaction between gender and housing type? Which main effect appears to be more important?

(b) Verify that the conditions for ANOVA inference are satisfied.

(c) Give the complete two-way ANOVA table. What are the F statistics and P-values for interaction and the two main effects? Explain why the test results confirm the tentative conclusions you drew from the plot of means.

26.31 Hooded rats: social play counts. The researchers who conducted the study in Exercise 26.30 also recorded the number of times each of three types of behavior (object play, locomotor play, and social play) occurred. The file *ex26-31.dat* contains the counts of social play episodes by each rat during the observation period. Use two-way ANOVA to analyze the effects of gender and housing.

26.32 Hooded rats: object play counts. The researchers who conducted the study in Exercise 26.30 also recorded the number of times each of three types of behavior (object play, locomotor play, and social play) occurred. The file *ex26-32.dat* contains the counts of object play episodes for each rat during the observation period. Carry out a complete analysis of the effects of gender and housing type.

26.33 Temperature response in goldfish. Goldfish are cold-blooded and have strong physiological responses to sudden temperature changes ("acute response"). However, they can also adapt to long-term changes in temperature over the course of weeks ("acclimation"). An experiment compared the acute response to a given test temperature (10, 12, 15, 22, or 25 degrees Celsius) in goldfish acclimated for two months to a cold (12 degrees Celsius) or warm (22 degrees Celsius) environment. Eighteen goldfish were used for each of the 10 treatment groups. Table 26.4 gives their ventilation rate in number of opercular beats per minute.[13] Here is software output for the two-way ANOVA test of these data:

```
Test temperature  F=68.13, P<0.001
Acclimation  F=12.85, P<0.001
Test temperature*Acclimation F=1.84, P=0.124
```

What are the long-term and short-term effects of temperature change on goldfish ventilation rate? Write a short conclusion based on your analysis. Follow the four-step process used in Examples 26.11 through 26.13.

26.34 Neural basis of 3D vision, continued. Exercise 26.29 discussed the neural coding of 3D vision based on image stereopsis. However, 3D vision also involves the appreciation of distances. How does the brain combine information about volume (from binocular disparity) with information about distance (from lens focus and gaze angle)? Figure 26.16 shows Minitab's interaction plot for the activity (in spikes per second) of the primary visual cortex neuron of Exercise 24.32 when visual patterns with 1 of 6 disparity values were presented on a screen located at a distance of either 20 cm, 40 cm, or 80 cm. A negative disparity corresponds to the impression of a near object. There were 12 recordings for each treatment group, collected in random order.[14] The file *ex26-34.dat* contains the raw data. Here is software output for the two-way ANOVA test of these data:

```
Distance  F=142.46, P<0.001
Disparity  F=39.79, P<0.001
Distance*Disparity interaction F=6.82, P<0.001
```

How does this neuron respond to both disparity and distance cues? Write a short conclusion based on your analysis. Follow the four-step process used in Examples 26.11 through 26.13.

26.35 Temperature response in goldfish, continued. Go back to Exercise 26.33 on goldfish ventilation rate.

(a) What are the degrees of freedom for the main effects and for the interaction?

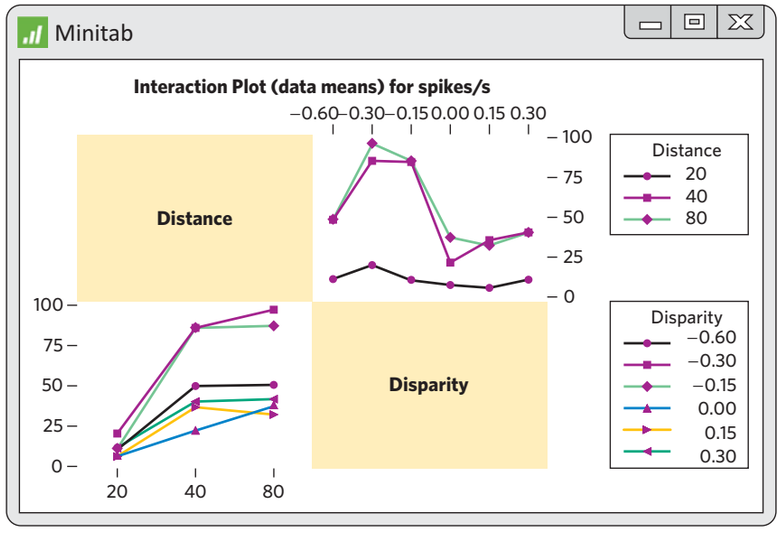
(b) Software gives mean squares of 39,289.4, 7411.3, and 1058.7 for test temperature, acclimation, and interaction, respectively. What is the value of MSE?

(c) Compute the sums of squares. What is the value of SST? Write all your answers into a two-way ANOVA table (you can refer to Figure 26.10 for a model).

(d) If you have access to software, verify that your answers match the software output (aside from rounding errors).

▶ TABLE 26.4 Goldfish ventilation rates (number of opercular beats per minute)

Cold-acclimated fish					Warm-acclimated fish				
10 °C	12 °C	15 °C	22 °C	25 °C	10 °C	12 °C	15 °C	22 °C	25 °C
60	61	70	111	113	43	25	64	114	104
88	36	98	85	124	39	30	69	86	158
36	32	94	84	192	8	68	76	120	160
50	51	86	53	146	30	54	82	75	120
27	92	67	99	100	44	63	78	40	141
37	88	80	58	110	64	35	49	96	79
65	48	67	122	102	12	80	96	90	122
40	26	52	67	144	15	80	43	65	172
26	40	93	136	160	47	71	62	54	110
68	60	77	110	216	14	12	48	86	153
74	76	88	96	132	28	27	46	87	96
63	60	69	72	75	40	8	54	20	110
80	48	88	52	148	70	15	30	78	128
60	60	80	77	87	21	40	70	80	158
70	51	75	125	100	47	30	48	87	117
100	38	84	77	105	20	4	66	112	85
40	66	62	78	102	34	45	48	67	189
76	72	53	92	91	35	30	46	52	120



▲ FIGURE 26.16 Plots from Minitab of the mean activity of a primary visual cortex neuron under six conditions of disparity and three visual distances, for Exercise 26.34.

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

26.36 Neural basis of 3D vision, continued. Go back to Exercise 26.34 on neural activity.

(a) What are the degrees of freedom for the main effects and for the interaction?

(b) Software gives mean squares of 47,449.7, 13,253.6, and 2272.8 for distance, disparity, and interaction, respectively. What is the value of MSE?

(c) Compute the sums of squares. What is the value of SST? Write all your answers into a two-way ANOVA table (you can refer to Figure 26.10 for a model).

(d) The file *ex26-36.dat* contains the raw data for this experiment. If you have access to software, verify that your answers match the software output (aside from rounding errors).

26.37 Herbicide and corn hybrids. Genetic engineering has produced new corn hybrids that resist the effects of herbicides. This allows more effective control of weeds, because herbicides don't damage the corn. A study compared the effects of the herbicide glusofinate on a number of corn hybrids. The percents of necrosis (leaf burn) 10 days after application of glusofinate for several application rates (kilograms per hectare) and three corn hybrids, two resistant and one not, are provided in the file *ex26-37.dat*.[15]

(a) Construct a plot of means to examine the effects of application rate and hybrid and their interaction.

(b) Are the conditions for ANOVA inference satisfied? Explain.

26.38 More on dietary manipulations in fruit flies. The experiment described in Example 26.11 also examined the number of eggs produced per female. Here are the means for each of the 8 test groups:

	Amount of yeast in diet (mg)			
	0	1	3	7
Wild-type	6.4	15.1	33.8	56.3
Mutant	7.7	20.8	50.8	78.3

(a) Plot the means on a single graph and describe the main effects and interaction.

(b) A two-way ANOVA analysis gives significant *P*-values for both main effects and the interaction effect. However, although there are no outliers and no major deviation from Normality in the data, the 8 standard deviations range from 2.1 to 12.0 eggs per female. What do you make of this information? How might it affect your interpretation of the ANOVA results?

26.39 Comparisons among means for one factor in a two-way analysis. We have illustrated the Tukey pairwise comparisons among all treatment means in both one-way and two-way settings. The method can also compare the mean responses to just one of the two factors in a two-way setting. Just do a one-way ANOVA on the two-way data with only one factor listed. This combines data for all levels of the other factor, so it is useful only when interactions are small. Return to the data on phosphorus in tomatoes, Table 26.2. Do a one-way ANOVA that uses all 36 observations, with fertilizer type (nitrogen level) as the only factor. Ask for Tukey pairwise comparisons among the three levels of nitrogen, with overall confidence level 95%.

(a) How many observations per group does your analysis use?

(b) What do you conclude from the *F* statistic and its *P*-value?

(c) Which pairwise differences of means for the 3 nitrogen levels are significant at the overall 5% level?

(d) Do you think these pairwise comparisons are useful for these data? (*Hint:* What population does each of the three samples represent?)

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

NOTES AND DATA SOURCES

1. P. M. Johnson and P. J. Kenny, "Dopamine D2 receptors in addiction-like reward dysfunction and compulsive eating in obese rats, *Nature Neuroscience*, 13 (2010), pp. 635–641, doi:10.1038/nn.2519.
2. If your software doesn't provide automated tests for multiple comparisons and you don't have access to a table of Tukey critical values like Table G, you can perform the Bonferroni procedures by hand for a simple, though somewhat conservative, approximation. For any c multiple-comparisons tests, the Bonferroni approach uses traditional t tests but requires a significance level reduced by a factor of c for each test (that is, $\alpha =$ overall significance level divided by c). For example, if you want an overall significance level of 5% and run 10 multiple-comparisons tests, then each test will be declared significant only if its P-value is $0.05/10 = 0.005$ or less.
3. A. Adan and J. M. Serra-Grabulosa, "Effects of caffeine and glucose, alone and combined, on cognitive performance," *Human Psychopharmacology: Clinical and Experimental*, 25 (2010), pp. 310–317, doi:10.1002/hup.1115.
4. We thank Charles Cannon of Duke University for providing the data. The study report is C. H. Cannon, D. R. Peart, and M. Leighton, "Tree species diversity in commercially logged Bornean rainforest," *Science*, 281 (1998), pp. 1366–1367.
5. Modified from M. C. Wilson and R. E. Shade, "Relative attractiveness of various luminescent colors to the cereal leaf beetle and the meadow spittlebug," *Journal of Economic Entomology*, 60 (1967), pp. 578–580.
6. R. Aroca et al., "Arbuscular mycorrhizal symbiosis influences strigolactone production under salinity and alleviates salt stress in lettuce plants," *Journal of Plant Physiology*, 170 (2013), pp. 47–55, doi:10.1016/j.jplph.2012.08.020.
7. F. H. Simmons, "Physiology of the trade-off between fecundity and survival in *Drosophila melanogaster*, as revealed through dietary manipulation," MS thesis, University of California, Irvine, 1996.
8. Data courtesy of David LeBauer, University of California, Irvine.
9. Simulated data, based on data summaries in G. L. Cromwell et al., "A comparison of the nutritive value of *opaque-2*, *floury-2* and normal corn for the chick," *Poultry Science*, 57 (1968), pp. 840–847.
10. We thank Andy Niemiec and Robbie Molden for data from a summer science project at Kenyon College. Provided by Brad Hartlaub.
11. Data courtesy of Brad Hughes, University of California, Irvine.
12. Hematology data for Lead(2+) Acetate from National Toxicology Program Study No. IMM92043, `http://ntp-server.niehs.nih.gov`.
13. We thank Rudi Berkelhamer of the University of California at Irvine for the data. The data are part of a larger set collected for an undergraduate lab exercise in scientific methods.
14. B. Stricanne, "Etudes d'intégration multisensorielles dans la voie visuelle occipito-pariétale du primate," PhD thesis, Université Paris VI, 1996.
15. C. E. Mowen, "Use of glusofinate in glusofinate resistant corn hybrids," MS thesis, Purdue University, 1999.

Baldi-16-3176 02_BAL_01337_Book November 11, 2017 15:50

▶ TABLE G Critical values m^* for Tukey pairwise multiple comparisons with 95% confidence level, k comparisons, and $N - k$ degrees of freedom (df)

	Number of comparisons (k)							
df	3	4	5	6	7	8	9	10
5	3.254	3.690	4.011	4.266	4.476	4.654	4.809	4.946
6	3.068	3.462	3.751	3.980	4.168	4.329	4.468	4.591
7	2.945	3.310	3.578	3.789	3.964	4.112	4.241	4.354
8	2.857	3.202	3.455	3.654	3.818	3.957	4.078	4.185
9	2.792	3.122	3.362	3.553	3.708	3.841	3.956	4.057
10	2.741	3.060	3.291	3.473	3.623	3.750	3.861	3.958
11	2.701	3.009	3.234	3.410	3.555	3.678	3.785	3.879
12	2.668	2.969	3.188	3.359	3.500	3.620	3.723	3.815
13	2.640	2.935	3.149	3.316	3.454	3.570	3.671	3.760
14	2.617	2.907	3.116	3.280	3.415	3.528	3.627	3.714
15	2.597	2.882	3.088	3.249	3.381	3.493	3.590	3.676
16	2.580	2.861	3.064	3.222	3.352	3.462	3.557	3.642
17	2.565	2.843	3.043	3.199	3.327	3.435	3.529	3.612
18	2.552	2.826	3.024	3.178	3.304	3.411	3.504	3.586
19	2.541	2.812	3.007	3.159	3.285	3.390	3.482	3.562
20	2.530	2.799	2.992	3.143	3.267	3.371	3.461	3.541
21	2.521	2.787	2.979	3.128	3.251	3.354	3.444	3.522
22	2.512	2.777	2.967	3.115	3.236	3.339	3.427	3.505
23	2.505	2.768	2.956	3.103	3.223	3.325	3.412	3.490
24	2.498	2.758	2.946	3.092	3.211	3.312	3.399	3.475
25	2.491	2.751	2.937	3.082	3.200	3.300	3.386	3.463
26	2.485	2.744	2.928	3.072	3.190	3.289	3.375	3.451
27	2.479	2.737	2.920	3.064	3.181	3.280	3.364	3.439
28	2.474	2.730	2.913	3.056	3.172	3.270	3.355	3.429
29	2.470	2.724	2.907	3.048	3.164	3.262	3.346	3.420
30	2.465	2.719	2.901	3.041	3.157	3.253	3.338	3.411
40	2.434	2.681	2.856	2.992	3.103	3.197	3.277	3.348
50	2.415	2.657	2.830	2.963	3.072	3.163	3.241	3.310
60	2.403	2.642	2.812	2.944	3.050	3.140	3.217	3.285
80	2.388	2.624	2.791	2.920	3.024	3.113	3.188	3.255
100	2.379	2.613	2.778	2.906	3.009	3.096	3.171	3.236
120	2.373	2.606	2.770	2.896	2.999	3.085	3.159	3.224
150	2.367	2.598	2.761	2.887	2.989	3.075	3.147	3.212
200	2.361	2.591	2.753	2.877	2.978	3.063	3.136	3.200