

# Chapter 13



case study

## Introduction

### Section 13.1 One-Way Analysis of Variance

# Analysis of Variance

## *Do Pets or Friends Help Reduce Stress?*

If you are a dog lover, perhaps having your dog along reduces the effect of stress. To examine how pets affect people in stressful situations, researchers recruited 45 women who said they were dog lovers. Fifteen of the subjects were randomly assigned to each of three groups: to do a stressful task alone (control, C), with a good friend present (F), or with their dog present (pet, P). The stressful task was to count backward by 13s or 17s. The subject's mean heart rate (in beats per minute) during the task is one measure of the effect of stress. The table below shows the data.<sup>1</sup>

Group	Rate	Group	Rate	Group	Rate	Group	Rate
P	69.169	P	68.862	C	84.738	C	75.477
F	99.692	C	87.231	C	84.877	C	62.646
P	70.169	P	64.169	P	58.692	P	70.077
C	80.369	C	91.754	P	79.662	F	88.015
C	87.446	C	87.785	P	69.231	F	81.600
P	75.985	F	91.354	C	73.277	F	86.985
F	83.400	F	100.877	C	84.523	F	92.492
F	102.154	C	77.800	C	70.877	P	72.262
P	86.446	P	97.538	F	89.815	P	65.446
F	80.277	P	85.000	F	98.200		
C	90.015	F	101.062	F	76.908		
C	99.046	F	97.046	P	69.538		

Here are some numerical summaries for the three groups:

Level	N	Mean	StDev
Control	15	82.524	9.242
Friend	15	91.325	8.341
Pet	15	73.483	9.970

Do the mean heart rates for the groups provide convincing evidence that the presence of a pet or friend affects heart rate during a stressful task? In this chapter, you'll develop the tools to help answer this question.

## Introduction

Which of three exercise routines helps overweight people lose the most weight? Who spends more time on the Internet—freshmen, sophomores, juniors, or seniors? Which of six brands of AAA batteries lasts longest? In each of these settings, we wish to compare the mean response for several populations or treatments. The statistical technique for comparing several means is called *analysis of variance*, or simply ANOVA. This brief chapter examines the big ideas and the computational details of ANOVA.

### • ACTIVITY *Come fly with me*

- **MATERIALS:** Three paper airplane pattern sheets (in the *Teacher's Resource Binder*), scissors, masking tape, tape measures, graphing calculator



The purpose of this Activity is to see which of three paper airplane models—A, B, or C—flies farthest. Specifically, the object is to determine whether there is a significant difference in the average distance flown for the three plane models. The null hypothesis is that there is no difference among the mean distances flown by Models A, B, and C:

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

The alternative hypothesis is that there is a difference in the average flight distances.

1. As a class, design an experiment to determine which of the three paper airplane models flies the farthest. Be sure to follow the principles of experimental design that you learned in Chapter 4.
2. Carry out your plan and collect the necessary data.
3. Compare the flight distances for the three models graphically and numerically. Does it appear that all the means are about the same, or is at least one mean different from the other means?

*Note:* Keep these data handy for further analysis later in the chapter.

## 13.1

# One-Way Analysis of Variance

### In Section 13.1, you'll learn about:

- Comparing means: the analysis of variance  $F$  test
- The idea of analysis of variance
- Conditions for ANOVA
- $F$  distributions and degrees of freedom
- ANOVA calculations
- Putting it all together: one-way ANOVA

The two-sample  $t$  procedures of Chapter 10 compare the means of two populations or the mean responses to two treatments in an experiment. Of course, studies don't always compare just two groups. In Chapter 11, we saw that a chi-square test for homogeneity could be used to compare the proportion of successes among several populations or treatments. We need a method for comparing any number of means.

Here's an example that introduces the basic idea of a significance test involving more than two means. As always, we begin with data analysis before proceeding to inference.

**EXAMPLE**

*Hummingbirds and Tropical Flowers*

Comparing several means

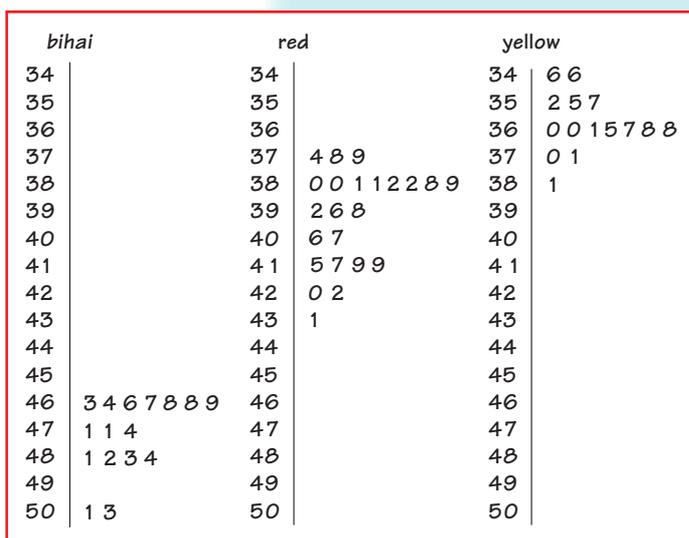


Ethan Temeles and W. John Kress of Amherst College studied the relationship between varieties of the tropical flower *Heliconia* on the island of Dominica and the different species of hummingbirds that pollinate the flowers.<sup>2</sup> Over time, the researchers believe, the lengths of the flowers and the forms of the hummingbirds' beaks have evolved to match each other. If that is true, flower varieties pollinated by different hummingbird species should have distinct distributions of length.

The table below gives length measurements (in millimeters) for random samples of three varieties of *Heliconia*, each pollinated by a different species of hummingbird. Do the three varieties display distinct distributions of length? In particular, do the mean lengths of their flowers differ?

<i>H. bihai</i>							
47.12	46.75	46.81	47.12	46.67	47.43	46.44	46.64
48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36
<i>H. caribaea red</i>							
41.90	42.01	41.93	43.09	41.47	41.69	39.78	40.57
39.63	42.18	40.66	37.87	39.16	37.40	38.20	38.07
38.10	37.97	38.79	38.23	38.87	37.78	38.01	
<i>H. caribaea yellow</i>							
36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.10
35.17	36.82	36.66	35.68	36.03	34.57	34.63	

Let's follow the strategy we learned way back in Chapter 1: use graphs and numerical summaries to compare the three distributions of flower length. Figure 13.1 is a side-by-side stemplot of the data. The lengths have been rounded to the nearest tenth of a millimeter.



Here are the summary statistics we will use in further analysis:

Sample	Variety	Sample size	Mean length	Standard deviation
1	<i>bihai</i>	16	47.60	1.213
2	Red	23	39.71	1.799
3	Yellow	15	36.18	0.975

What do we see? The three varieties differ so much in flower length that there is little overlap among them. In particular, the flowers of *bihai* are longer than either red or yellow. The mean lengths are 47.6 mm for *H. bihai*, 39.7 mm for *H. caribaea red*, and 36.2 mm for *H. caribaea yellow*. Are these observed differences in sample means statistically significant? We must develop a test for comparing more than two population means.

**FIGURE 13.1** Side-by-side stemplots comparing the lengths in millimeters of random samples of flowers from three varieties of *Heliconia*.



## Comparing Means: The Analysis of Variance $F$ Test

Call the mean lengths for the three populations of flowers  $\mu_1$  for *bihai*,  $\mu_2$  for red, and  $\mu_3$  for yellow. To compare these three population means, we might use the two-sample  $t$  test several times:

- Test  $H_0: \mu_1 = \mu_2$  to see if the mean length for *bihai* differs from the mean for red.
- Test  $H_0: \mu_1 = \mu_3$  to see if *bihai* differs from yellow.
- Test  $H_0: \mu_2 = \mu_3$  to see if red differs from yellow.

The weakness of doing three tests is that we get three  $P$ -values, one for each test alone. That doesn't tell us how likely it is that *three* sample means are spread apart as far as these are. This is the problem of *multiple comparisons* that we discussed in Chapter 11.

We want to test the null hypothesis that there are *no differences* among the mean lengths for the three populations of flowers:

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

The alternative hypothesis is that there is *some difference*. That is, not all three population means are equal:

$$H_a: \text{not all of } \mu_1, \mu_2, \text{ and } \mu_3 \text{ are equal}$$

As with the chi-square test for homogeneity, the alternative hypothesis is no longer one-sided or two-sided. It is “many-sided” because it allows any relationship other than “all three equal.” For example,  $H_a$  includes the case in which  $\mu_2 = \mu_3$  but  $\mu_1$  has a different value.

When the conditions for inference are met, the appropriate significance test for comparing means is the **analysis of variance  $F$  test**. Analysis of variance is usually abbreviated as ANOVA.

**B**e careful when defining parameters. For random sampling, each  $\mu_i$  refers to a corresponding population mean. In a completely randomized experiment,  $\mu_i$  refers to the actual mean response under treatment  $i$  for subjects like the ones in the study.

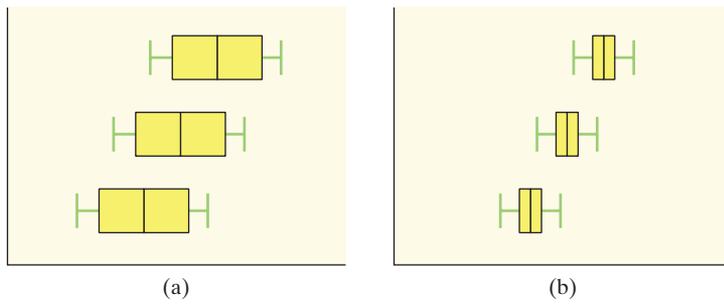
Analysis of variance  
 $F$  test

## The Idea of Analysis of Variance

The details of analysis of variance are a bit complicated, but the main idea of ANOVA is more straightforward and much more important to understand. Here it is: when we ask whether a set of sample means gives evidence for differences among the population means, what matters is not how far apart the sample means are but how far apart they are *relative to the variability of individual observations*.

Look at the two sets of boxplots in Figure 13.2. For simplicity, these distributions are all symmetric, so that the mean and median are the same. The center line in each boxplot is therefore the sample mean. Both sets of boxplots compare three samples. The sample means are the same in parts (a) and (b). Could differences this large easily arise due just to chance, or are they statistically significant?

- The boxplots in Figure 13.2(a) have long boxes, which means that there is lots of variation among the individuals in each group. With this much variation among individuals, we would not be surprised if another set of samples gave quite different sample means. The observed differences among the sample means could easily happen just by chance if the three population means are equal.
- The boxplots in Figure 13.2(b) have the same centers as those in Figure 13.2(a), but the boxes are much shorter. That is, there is much less variation among the individuals in each group. It is unlikely that any sample from the first group



**FIGURE 13.2** Boxplots for two sets of three samples each. The sample means are the same in parts (a) and (b). Analysis of variance will find a more significant difference among the means in part (b) because there is less variation among the individuals within those samples.

would have a mean as small as the mean of the second group. Because means as far apart as those observed would rarely arise just by chance in repeated sampling, they are good evidence of real differences among the means of the three populations we are sampling from.

This comparison of the two parts of Figure 13.2 is too simple in one way. It ignores the effect of the sample sizes, an effect that boxplots do not show. *Small differences among sample means can be significant if the samples are large.*



*Large differences among sample means can fail to be significant if the samples are small.*

All we can be sure of is that for the same sample size, Figure 13.2(b) will give a much smaller  $P$ -value than Figure 13.2(a). Despite this qualification, the big idea remains: if sample means are far apart relative to the variation among individuals in the same groups, that's evidence that something other than chance is at work.

### DEFINITION: The Analysis of Variance Idea

**Analysis of variance (ANOVA)** tests whether several populations or treatments have the same mean by comparing how far apart the sample means are with how much variation there is within the samples.

It is one of the oddities of statistical language that methods for comparing means are named after the variance. The reason is that the test works by comparing two kinds of variation. Analysis of variance is a general method for studying sources of variation in responses. Comparing several means is the simplest form of ANOVA, called **one-way ANOVA**.

The **analysis of variance  $F$  statistic** for testing the equality of several means has this form:

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

We'll give more details about how to calculate this test statistic shortly. For now, note that the  $F$  statistic can take only values that are zero or positive. It is zero only when all the sample means are identical and gets larger as they move farther apart. Large values of  $F$  are evidence against the null hypothesis  $H_0$  that all population or treatment means are the same. Although the alternative hypothesis  $H_a$  is many-sided, the ANOVA  $F$  test is one-sided because any violation of  $H_0$  tends to produce a large value of  $F$ .

## Conditions for ANOVA

Like all inference procedures, ANOVA is valid only in some circumstances. Here are the conditions under which we can use ANOVA to compare population or treatment means.

### One-way ANOVA

#### Analysis of variance $F$ statistic

In Chapter 10, we saw that the two-sample  $t$  statistic usually has the form  $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ .

The numerator measures variation among the sample means, and the denominator measures variation among individuals within the same sample. Not surprisingly, the ANOVA  $F$  statistic for comparing the means of several groups or populations has a similar form.

### Conditions for ANOVA Inference

- **Random** The data are produced by random samples of size  $n_1$  from Population 1, size  $n_2$  from Population 2, . . . , and size  $n_k$  from Population  $k$  or by groups of size  $n_1, n_2, \dots, n_k$  in a randomized experiment.
- **Normal** All the population distributions (or the true distributions of responses to the treatments) are Normal, *or* the sample sizes are large and the distributions show no strong skewness or outliers.
- **Independent** Both the samples/groups themselves and the individual observations in each sample/group are independent. When sampling without replacement, check that all the populations are at least 10 times as large as the corresponding samples (the 10% condition).
- **Same SD** The response variable has the same standard deviation  $\sigma$  (with a value that is generally unknown) for all the populations or treatments.

The first three conditions should be somewhat familiar from our study of the two-sample  $t$  procedures for comparing two means. As usual, the design of the data production is the most important condition for inference. Biased sampling or confounding can make any inference meaningless. *If we do not actually draw separate random samples from each population or carry out a randomized comparative experiment, our scope of inference is very limited.*



### Robust

Because no real population has an exactly Normal distribution, the usefulness of inference procedures that assume Normality depends on how sensitive they are to departures from Normality. As long as there are no outliers or strong skewness, the ANOVA  $F$  test is fairly **robust** against non-Normality. This is especially true when the sample sizes are equal. In this setting, ANOVA becomes safer as the sample sizes get larger.

The fourth condition is annoying: ANOVA assumes that the variability of observations, measured by the standard deviation, is the same for all populations or treatments. The  $t$  test for comparing two means (Chapter 10) does not require equal standard deviations. Unfortunately, the ANOVA  $F$  test for comparing more than two means is less broadly valid. It is not easy to check the condition that the populations (or treatments) have equal standard deviations. You must either seek expert advice or rely on the robustness of ANOVA.

How serious are unequal standard deviations? ANOVA is not too sensitive to violations of the condition, especially when all samples have the same or similar sizes and no sample is very small. When designing a study, try to use samples or groups of about the same size. The sample standard deviations estimate the population standard deviations, so check before doing ANOVA that the sample standard deviations are similar to each other. We expect some variation among them due to chance. Here is a rule of thumb that is safe in almost all situations.

### Checking the Same SD Condition in ANOVA

The results of the ANOVA  $F$  test are approximately correct when the largest sample standard deviation is no more than twice as large as the smallest sample standard deviation.

Now we're ready to check conditions for performing an ANOVA  $F$  test using the *Heliconia* data.

## EXAMPLE

### *Hummingbirds and Tropical Flowers*

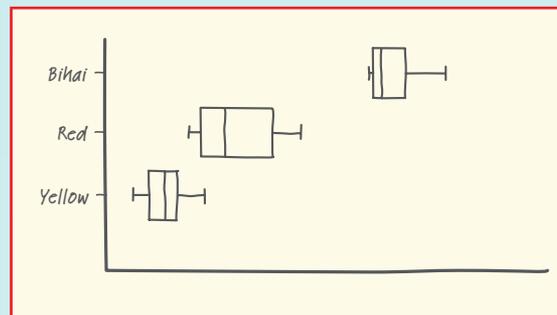
#### Checking conditions

Review the details of the *Heliconia* study from the previous example (page 5).

**PROBLEM:** Check conditions for carrying out an ANOVA  $F$  test in this setting.

**SOLUTION:**

- *Random* Researchers took separate random samples of 16 *bihai*, 23 *red*, and 15 *yellow Heliconia* flowers.
- *Normal* We entered the data into a TI-84 and made side-by-side boxplots of the flower lengths in the three samples. Although the distributions for the *bihai* and *red* varieties show some right-skewness, we don't see any strong skewness or outliers that would prevent use of one-way ANOVA.



- *Independent* Researchers took independent samples of *bihai*, *red*, and *yellow Heliconia*. Because sampling without replacement was used, there must be at least  $10(16) = 160$  *bihai*,  $10(23) = 230$  *red*, and  $10(15) = 150$  *yellow* flowers. This is pretty safe to assume.
- *Same SD* The sample standard deviations for the three varieties are

$$s_1 = 1.213 \quad s_2 = 1.799 \quad s_3 = 0.975$$

These standard deviations satisfy our rule of thumb:

$$\frac{\text{largest } s}{\text{smallest } s} = \frac{1.799}{0.975} = 1.85 \text{ (less than 2)}$$

We can safely use ANOVA to compare the mean lengths for the three populations.

**For Practice** Try Exercise 1



### CHECK YOUR UNDERSTANDING

Bromeliads are tropical flowering plants. Many are epiphytes that attach to trees and obtain moisture and nutrients from air and rain. Their leaf bases form cups that collect water and are home to the larvae of many insects. As a preliminary step to a study of changes in the nutrient cycle, Jacqueline Ngai and Diane Srivastava examined the effects of adding fertilizer in the form of nitrogen, phosphorus, or both to the cups. They randomly assigned 8 bromeliads growing in Costa Rica to each of four treatment groups, including an unfertilized control group. A monkey destroyed one of the plants in the control group, leaving 7 bromeliads in that group. Here are the numbers of new leaves on each plant over the 7 months following fertilization:<sup>3</sup>



Nitrogen	Phosphorus	Both	Neither
15	14	14	11
14	14	16	13
15	14	15	16
16	11	14	15
17	13	14	15
18	12	13	11
17	15	17	12
13	15	14	

1. State hypotheses for an ANOVA  $F$  test in this setting. Be sure to define your parameters.
2. Check conditions for carrying out the test in Question 1.

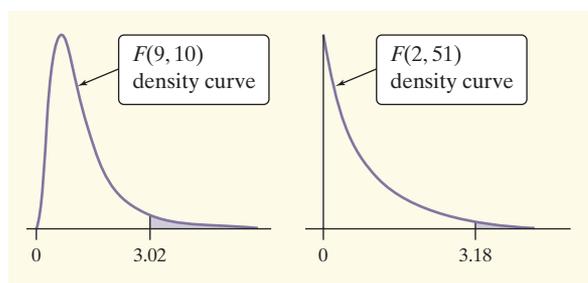
## $F$ Distributions and Degrees of Freedom

Recall that the ANOVA  $F$  statistic is

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

To find the  $P$ -value for this statistic, we must know the sampling distribution of  $F$  when the null hypothesis (all population or treatment means are equal) is true and the conditions are satisfied. This sampling distribution is an  **$F$  distribution**.

### $F$ distribution



**FIGURE 13.3** Density curves for two  $F$  distributions. Both are right-skewed and take only positive values. The upper 5% critical values are marked under the curves.

The  $F$  distributions are a family of right-skewed distributions that take only values greater than 0. The density curves in Figure 13.3 illustrate their shapes. A specific  $F$  distribution is determined by the *degrees of freedom* of the numerator and denominator of the  $F$  statistic. When describing an  $F$  distribution, always give the numerator degrees of freedom first. Our brief notation will be  $F$  (df1, df2) for the  $F$  distribution with df1 degrees of freedom in the numerator and df2 degrees of freedom in the denominator. *Interchanging the degrees of freedom changes the distribution, so the order is important.*



The degrees of freedom of the ANOVA  $F$  statistic depend on the number of means we are comparing and the number of observations in each sample. Here are the details.

**DEFINITION: Degrees of Freedom for the ANOVA  $F$  Test**

Suppose that we want to compare the means of  $k$  populations or treatments. The total number of observations in all samples or groups combined is

$$N = n_1 + n_2 + \cdots + n_k$$

If the null hypothesis that all the population or treatment means are equal is true and the conditions for inference are met, the ANOVA  $F$  statistic has the  $F$  distribution with  $k - 1$  degrees of freedom in the numerator and  $N - k$  degrees of freedom in the denominator.

Let's find the degrees of freedom for the ANOVA  $F$  test in the *Heliconia* study.

**EXAMPLE**

*Hummingbirds and Flower Lengths*

Degrees of freedom for  $F$

In the two earlier examples, we compared the mean lengths for three varieties of flowers, so  $k = 3$ . The three sample sizes are  $n_1 = 16$ ,  $n_2 = 23$ , and  $n_3 = 15$ . The total number of observations is therefore  $N = 16 + 23 + 15 = 54$ . The ANOVA  $F$  test has numerator degrees of freedom

$$k - 1 = 3 - 1 = 2$$

and denominator degrees of freedom

$$N - k = 54 - 3 = 51$$



**CHECK YOUR UNDERSTANDING**

Refer to the previous Check Your Understanding. Find the degrees of freedom for the ANOVA  $F$  test in the bromeliad study. Show your work.

## ANOVA Calculations

Now we will give the actual formula for the ANOVA  $F$  statistic. Suppose the Random condition is met. Subscripts from 1 to  $k$  tell us which sample or group a statistic refers to:

Population	Sample size	Sample mean	Sample std. dev.
1	$n_1$	$\bar{x}_1$	$s_1$
2	$n_2$	$\bar{x}_2$	$s_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$k$	$n_k$	$\bar{x}_k$	$s_k$

You can find the  $F$  statistic from just the sample sizes  $n_i$ , the sample means  $\bar{x}_i$ , and the sample standard deviations  $s_i$ . You don't have to go back to the individual observations.

- As we've already seen, the ANOVA  $F$  statistic has the form

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

### Mean squares

The measures of variation in the numerator and denominator of  $F$  are called **mean squares**. A mean square is a more general form of a sample variance. Recall that an ordinary sample variance  $s_x^2$  is a kind of average (or mean) of the squared deviations of observations from their mean, so it qualifies as a "mean square."

- The first step in calculating the  $F$  statistic is to determine the overall mean response  $\bar{x}$ . That is,  $\bar{x}$  is the mean of all  $N$  observations together. You can find  $\bar{x}$  from the  $k$  sample means by

$$\bar{x} = \frac{\text{sum of all sample observations}}{N} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{N}$$

This expression works because multiplying a group mean  $\bar{x}_i$  by the number of observations  $n_i$  it represents gives the sum of the observations in that group.

- The numerator of  $F$  is a mean square that measures variation among the  $k$  sample means  $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ . To measure this variation, look at the  $k$  deviations of the means of the samples from  $\bar{x}$ :

$$\bar{x}_1 - \bar{x}, \bar{x}_2 - \bar{x}, \dots, \bar{x}_k - \bar{x}$$

The mean square in the numerator of  $F$  is an average of the squares of these deviations. Each squared deviation is weighted by  $n_i$ , the number of observations it represents. We call this result the **mean square for groups** (MSG):

$$\text{MSG} = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_k(\bar{x}_k - \bar{x})^2}{k - 1}$$

- The mean square in the denominator of  $F$  measures variation among individual observations in the same sample. For any one sample, the sample variance  $s_i^2$  does this job. For all  $k$  samples together, we use an average of the individual sample variances. It is another weighted average, in which each  $s_i^2$  is weighted by its degrees of freedom  $n_i - 1$ . The resulting mean square is called the **mean square for error** (MSE):

$$\text{MSE} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{N - k}$$

"Error" doesn't mean a mistake has been made. It's a traditional term for chance variation.

### The ANOVA $F$ Test

To test the null hypothesis that all  $k$  populations or treatments have the same mean against the alternative hypothesis that not all the means are equal, calculate the ANOVA  $F$  statistic

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

The numerator of  $F$  is the **mean square for groups**

$$MSG = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_k(\bar{x}_k - \bar{x})^2}{k - 1}$$

The denominator of  $F$  is the **mean square for error**

$$MSE = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{N - k}$$

When  $H_0$  is true and the Random, Normal, Independent, and Same SD conditions are met, the  $F$  statistic has the  $F$  distribution with  $k - 1$  and  $N - k$  degrees of freedom.

Sums of squares  
ANOVA table

The denominators in the formulas for MSG and MSE are the two *degrees of freedom*  $k - 1$  and  $N - k$  of the  $F$  test. The numerators are called **sums of squares**, from their algebraic form. It is usual to present the results of ANOVA in an **ANOVA table**. Output from software usually includes an ANOVA table, like the one in Figure 13.4 for the *Heliconia* study. You can check that each mean square MS is the corresponding sum of squares SS divided by its degrees of freedom df, and that the  $F$  statistic is MSG divided by MSE.

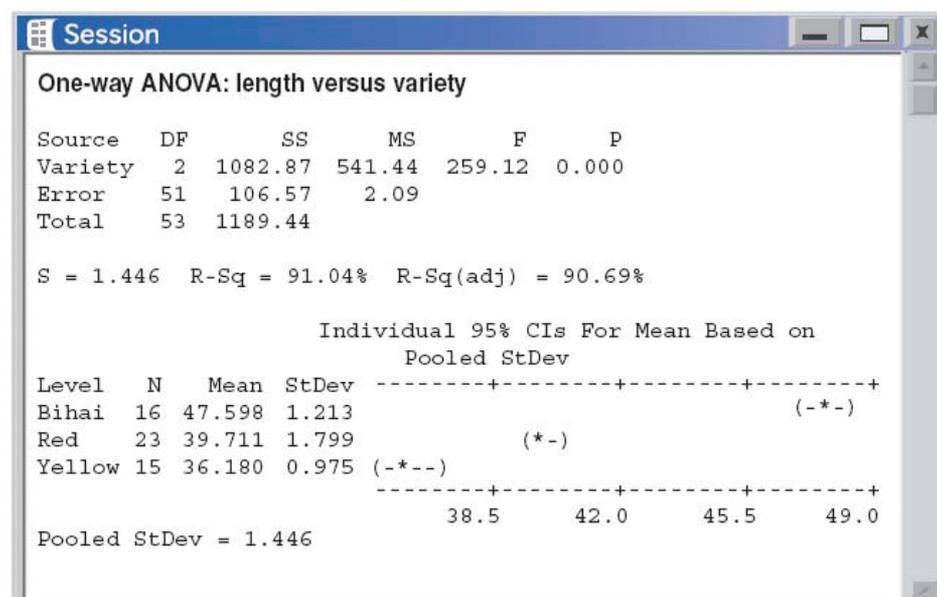


FIGURE 13.4 Minitab output from a one-way ANOVA for the flower length data.

We can use tables of  $F$  critical values to get the  $P$ -value for an ANOVA  $F$  test. Doing so is awkward, however, because we need a separate table for every pair of degrees of freedom  $df_1$  and  $df_2$ . Fortunately, software gives you  $P$ -values for the ANOVA  $F$  test without the need for a table. That leaves us free to interpret the results of the test.

## EXAMPLE

### *Hummingbirds and Flower Lengths*

#### Comparing tropical flowers: Conclusion

Figure 13.4 tells us that for the flower length data in the previous examples, the test statistic is  $F = 259.12$  with  $P$ -value  $P < 0.0001$ .

#### PROBLEM:

- Interpret the  $P$ -value in context.
- What conclusion would you draw? Justify your answer.

#### SOLUTION:

(a) If the null hypothesis  $H_0: \mu_1 = \mu_2 = \mu_3$  of no difference in the population mean flower lengths is true, the probability of getting a difference among the sample mean flower lengths as large as or larger than the one observed in the study just by the chance involved in the random sampling is less than 1 in 10,000.

(b) Because the  $P$ -value is so small (less than any of our usual  $\alpha$  levels), we would reject  $H_0$ . There is very strong evidence that the three varieties of flowers do not all have the same population mean length.

The Minitab ANOVA output in Figure 13.4 reports the  $P$ -value as 0.000 to three decimal places. This does *not* mean that the  $P$ -value is actually 0, just that it is very close to 0.

**For Practice** Try Exercise 13

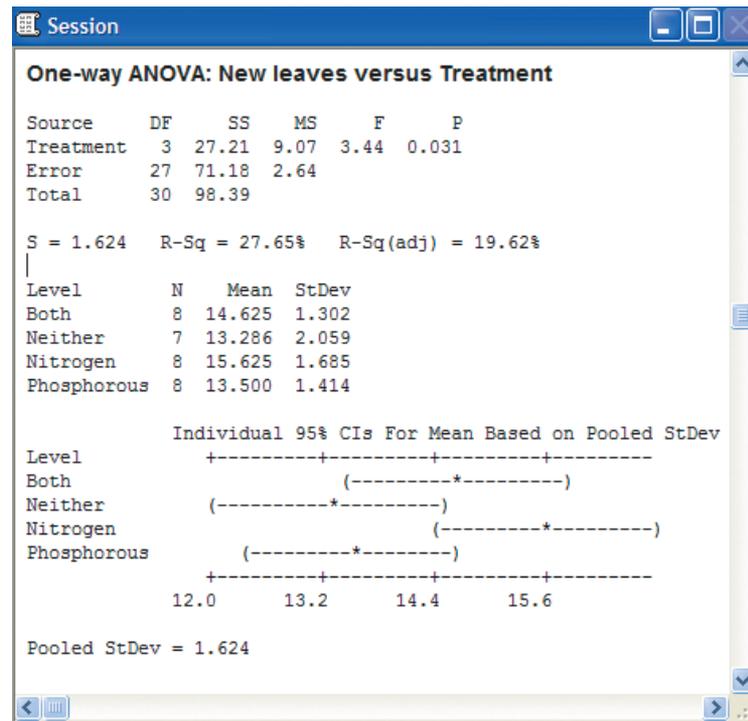
The  $F$  test doesn't say which of the three means are significantly different. It appears from our earlier data analysis that *bihai* flowers are distinctly longer than red or yellow. Red and yellow are closer in length, but the red flowers tend to be longer. The Minitab output in Figure 13.4 gives confidence intervals for all three means that help us see which means differ and by how much. None of the intervals overlap, and *bihai* is far above the other two. *Note: These are 95% confidence intervals for each mean separately. We are not 95% confident that all three intervals cover the three means. This is another example of the peril of multiple comparisons.*



#### CHECK YOUR UNDERSTANDING

The figure at the top of page 15 shows Minitab ANOVA output for the bromeliad study described in the Check Your Understanding on page 10.

- Interpret the  $P$ -value in context.
- What conclusion would you draw?



## THINK ABOUT IT

How does Minitab get those confidence intervals? Because MSE is an average of the individual variances, it is also called the *pooled sample variance*, written as  $s_p^2$ . When all  $k$  populations have the same population variance  $\sigma^2$ , as ANOVA assumes they do,  $s_p^2$  estimates  $\sigma^2$ . The square root of MSE is the **pooled standard deviation**  $s_p$ . It estimates the common standard deviation  $\sigma$  of each population or treatment. The Minitab output in Figure 13.4 (page 13) gives the value  $s_p = 1.446$ .

The pooled standard deviation  $s_p$  is a better estimator of the common  $\sigma$  than any of the individual sample standard deviations  $s_i$  because it combines (pools) the information in all  $k$  samples. We can get a confidence interval for any one of the means  $\mu_i$  from the usual formula:

$$\text{statistic} \pm (\text{critical value}) \cdot (\text{standard deviation of statistic})$$

The confidence interval for  $\mu_i$  is

$$\bar{x}_i \pm t^* \frac{s_p}{\sqrt{n_i}}$$

Use the critical value  $t^*$  from the  $t$  distribution with  $N - k$  degrees of freedom.

For the *bihai* variety, our 95% confidence interval for the population mean length is

$$47.598 \pm 2.008 \frac{1.446}{\sqrt{16}} = 47.598 \pm 0.726 = (46.872, 48.324)$$

using  $df = 54 - 3 = 51$ . This is the confidence interval for *bihai* that appears in the Minitab ANOVA output of Figure 13.4.

As the following example shows, it is possible to calculate the  $F$  statistic directly from the formulas. But it's pretty unpleasant to do so! We recommend using technology.

## EXAMPLE

### *Hummingbirds and Flower Lengths*

#### ANOVA calculations by hand

We can do the ANOVA test comparing the mean lengths of *bihai*, red, and yellow flower varieties using only the sample sizes, sample means, and sample standard deviations. There are  $k = 3$  groups with a total of  $N = 54$  flowers.

- The overall mean of the 54 lengths is

$$\begin{aligned}\bar{x} &= \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3}{N} \\ &= \frac{(16)(47.598) + (23)(39.711) + (15)(36.180)}{54} = 41.067\end{aligned}$$

- The mean square for groups is

$$\begin{aligned}\text{MSG} &= \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + n_3(\bar{x}_3 - \bar{x})^2}{k - 1} \\ &= \frac{1}{3 - 1} [(16)(47.598 - 41.067)^2 + (23)(39.711 - 41.067)^2 \\ &\quad + (15)(36.180 - 41.067)^2] \\ &= \frac{1082.996}{2} = 541.50\end{aligned}$$

- The mean square for error is

$$\begin{aligned}\text{MSE} &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{N - k} \\ &= \frac{(15)(1.213^2) + (22)(1.799^2) + (14)(0.975^2)}{51} = \frac{106.580}{51} = 2.09\end{aligned}$$

- Finally, the ANOVA  $F$  statistic is

$$F = \frac{\text{MSG}}{\text{MSE}} = \frac{541.50}{2.09} = 259.09$$

Our work differs slightly from the output in Figure 13.4 because of roundoff error.



You can carry out a one-way ANOVA test on the TI-83/84 and TI-89, as long as you are given the raw data (TI-83/84) and are comparing no more than six means. The following Technology Corner shows you how.

**TECHNOLOGY CORNER** One-way ANOVA on the calculator

We will repeat the one-way ANOVA procedure of the previous examples using the *Heliconia* data. Begin by entering the data into lists: *bihai*→L1/list1, *red*→L2/list2, and *yellow*→L3/list3.

## TI-83/84

- Press  $\boxed{\text{STAT}}$ , arrow to TESTS, and choose ANOVA (. Then enter the lists that contain the data. The completed command is shown below.

```
ANOVA(L1,L2,L3)
```

## TI-89

- In the Statistics/List Editor, press  $\boxed{2\text{nd}} \boxed{\text{F1}} (\boxed{\text{F6}})$ , and choose C:ANOVA. Set Data Input Method to “Data” and Number of Groups to “3.”

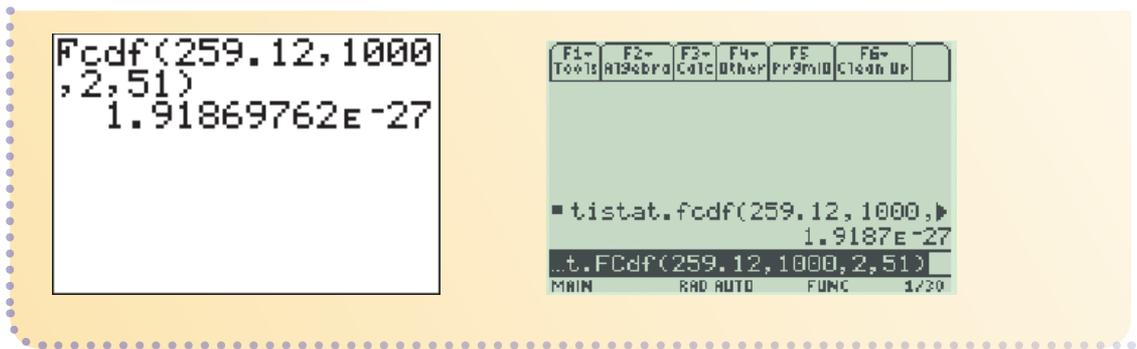
The results of the one-way ANOVA are shown in the following screens. The calculator reports that the  $F$  statistic is 259.12 and the  $P$ -value is  $1.92 \cdot 10^{-27}$ . The numerator degrees of freedom are  $k - 1 = 2$ , and by scrolling down, you see that the denominator degrees of freedom are  $N - k = 54 - 3 = 51$ .

What the calculator calls “Factor” is what we refer to as “Group.”

```
One-way ANOVA
F=259.1192995
P=1.918818E-27
Factor
df=2
SS=1082.87237
↓ MS=541.436183
```

```
One-way ANOVA
↑ MS=541.436183
Error
df=51
SS=106.565761
MS=2.08952472
SxP=1.44551884
```

If you know the  $F$  statistic and the numerator and denominator degrees of freedom, you can find the  $P$ -value with the command  $\text{Fcdf}$ , under the **DIS-TR** menu on the TI-83/84, and in the **CATALOG** under **FlashApps** on the TI-89. The syntax is  $\text{Fcdf}(\text{leftendpoint}, \text{rightendpoint}, \text{df numerator}, \text{df denominator})$ . See the following screen shots.



## Putting It All Together: One-Way ANOVA

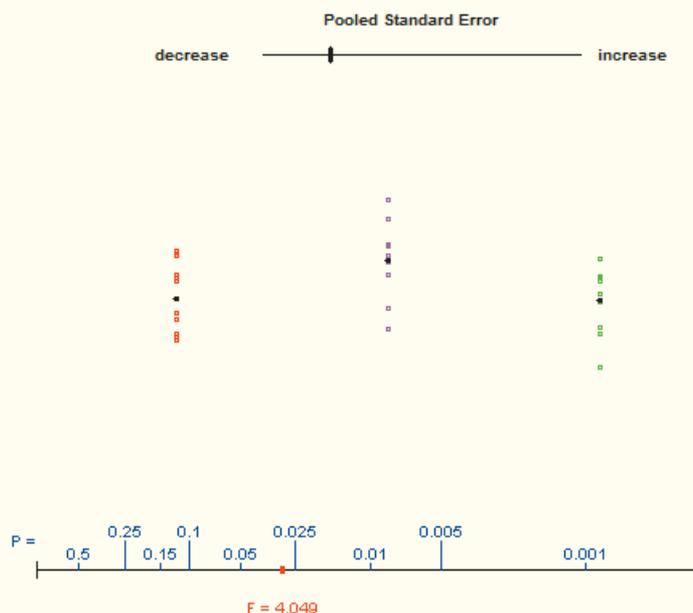
You can use the *One-Way ANOVA* applet to explore the analysis of variance idea for yourself. The applet allows you to change both the group means and the spread within groups. You can watch the ANOVA  $F$  statistic and its  $P$ -value change as you work.



### ACTIVITY *The One-Way ANOVA applet*

MATERIALS: Computer with Internet connection

The *One-Way ANOVA* applet displays the observations in three groups, with the group means highlighted by black dots. When you open or reset the applet, the scale at the bottom of the display shows that for these groups the ANOVA  $F$  statistic is  $F = 31.74$ , with  $P < 0.001$ . (The  $P$ -value is marked by a red dot that moves along the scale.)



*Part I: ANOVA compares several means*

1. The middle group has a larger mean than the other two. Grab its mean point with the mouse. How small can you make  $F$ ? What did you do to the mean to make  $F$  small? Roughly how significant is your small  $F$ ?
2. Starting with the three means aligned from your configuration at the end of step 1, drag any one of the group means either up or down. What happens to  $F$ ? What happens to the  $P$ -value? Convince yourself that the same thing happens if you move any one of the means, or if you move one slightly and then another slightly in the opposite direction.

*Part II: ANOVA uses within-group variation*

Reset the *One-Way ANOVA* applet to its original state. As in Figure 13.2(b), on page 7, the differences among the three means are highly significant (large  $F$ , small  $P$ -value) because the observations in each group cluster tightly about the group mean.

3. Use the mouse to slide the Pooled Standard Error at the top of the display to the right. You see that the group means do not change, but the spread of the observations in each group increases. What happens to  $F$  and  $P$  as the spread among the observations in each group increases? What are the values of  $F$  and  $P$  when the slider is all the way to the right? This is similar to Figure 13.2(a): variation within groups hides the differences among the group means.
4. Leave the Pooled Standard Error slider at the extreme right of its scale, so that spread within groups stays fixed. Use the mouse to move the group means apart. What happens to  $F$  and  $P$  as you do this?

For another look at the big ideas of ANOVA, try the *Simulating ANOVA Tables* applet at <http://statweb.calpoly.edu/chance/applets/anova/anova.html>.

Remember what you learned earlier about significance tests: a fail to reject  $H_0$  conclusion does not mean that  $H_0$  is true. All it says is that the data do not provide convincing evidence against the null hypothesis. In the case of the ANOVA  $F$  test, failing to reject  $H_0$  says that there is not enough evidence to conclude that the population/treatment means differ. As with other kinds of tests, we sometimes make a Type I or a Type II error as a result of an ANOVA  $F$  test.



Here is a final example that shows the entire one-way ANOVA process. As always, we follow the familiar four-step process.

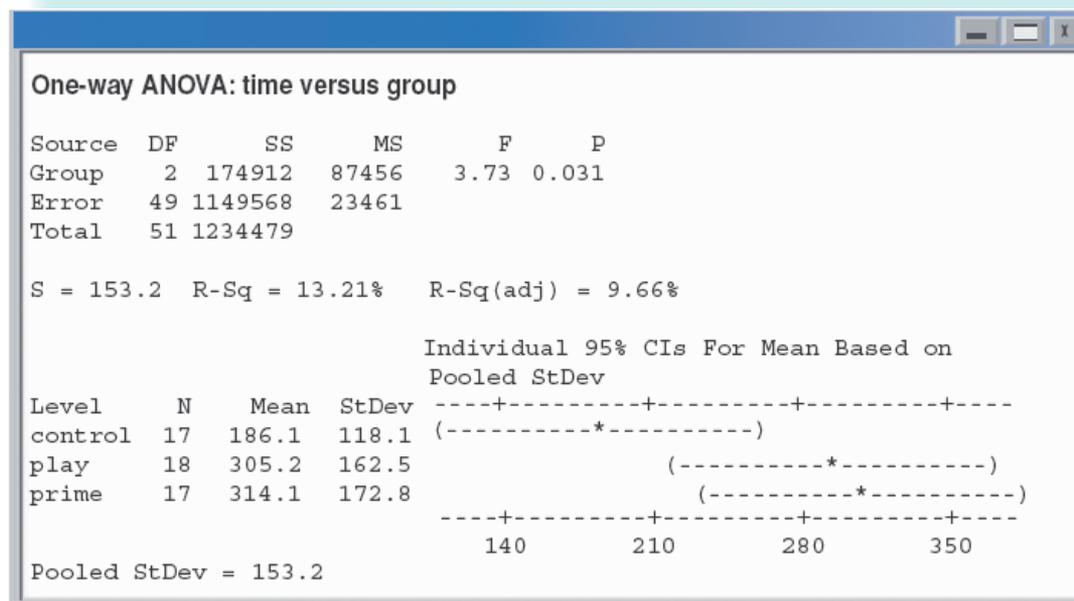
**EXAMPLE***Does Thinking about Money Change Behavior?***One-way ANOVA**

Kathleen Vohs of the University of Minnesota and her coworkers carried out several randomized comparative experiments on the effects of thinking about money. Here's an outline of one of the experiments. Ask student subjects to unscramble 30 sets of five words to make a meaningful phrase from four of the five. The control group unscrambled phrases like “cold it desk outside is” into “it is cold outside.” The “play money” group unscrambled similar sets of words, but a stack of Monopoly money was placed nearby. The “money prime” group unscrambled phrases that led to thinking about money, turning “high a salary desk paying” into “a high-paying salary.” Then each subject worked a hard puzzle, knowing that they could ask for help.

The table below shows the time in seconds that each subject worked on the puzzle before asking for help.<sup>4</sup> Psychologists think that money tends to make people self-sufficient. If so, the two groups that were encouraged in different ways to think about money should take longer on average to ask for help. Do the data support this idea?

Group	Time	Group	Time	Group	Time
Prime	609	Play	455	Control	118
Prime	444	Play	100	Control	272
Prime	242	Play	238	Control	413
Prime	199	Play	243	Control	291
Prime	174	Play	500	Control	140
Prime	55	Play	570	Control	104
Prime	251	Play	231	Control	55
Prime	466	Play	380	Control	189
Prime	443	Play	222	Control	126
Prime	531	Play	71	Control	400
Prime	135	Play	232	Control	92
Prime	241	Play	219	Control	64
Prime	476	Play	320	Control	88
Prime	482	Play	261	Control	142
Prime	362	Play	290	Control	141
Prime	69	Play	495	Control	373
Prime	160	Play	600	Control	156
		Play	67		

Minitab output from a one-way ANOVA on these data is shown below.



**STATE:** We want to perform a test of  $H_0: \mu_1 = \mu_2 = \mu_3$  versus  $H_a$ : the means are not all equal, where  $\mu_i$  = the actual mean time until subjects like the ones in this study ask for help;  $i = 1$ : money prime,  $i = 2$ : play money, and  $i = 3$ : control. Since no significance level was given, we'll use  $\alpha = 0.05$ .

**PLAN:** If conditions are met, we should perform an ANOVA  $F$  test to compare the means.

- *Random* Subjects were randomly assigned to the three treatment groups.
- *Normal* Figure 13.5 shows side-by-side stemplots of the data (rounded to the nearest ten seconds) for the three groups. We expect some irregularity in small samples, but there are no outliers or strong skewness that would prevent use of ANOVA.

	Prime		Play		Control
0	6 7	0	7 7	0	6 6 9 9
1	4 6 7	1	0	1	0 2 3 4 4 4 6 9
2	0 4 4 5	2	2 2 3 3 4 4 6 9	2	7 9
3	6	3	2 8	3	7
4	4 4 7 8 8	4	6	4	0 1
5	3	5	0 0 7	5	
6	1	6	0	6	

**FIGURE 13.5** Side-by-side stemplots comparing the time until subjects asked for help with a puzzle under each of the three treatments.

- *Independent* The random assignment yields independent groups. If the experiment was conducted properly, individual times to ask for help should also be independent: knowing one subject's time should give no additional information about another subject's time.
- *Same SD* The Minitab ANOVA output shows that the group standard deviations easily satisfy our rule of thumb: the largest (172.8) is no more than twice the smallest (118.1).

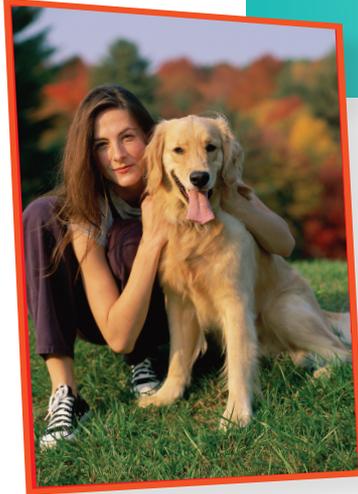
**DO:** From the computer output, we have

- *Test statistic*  $F = 3.73$
- *P-value* 0.031 using the  $F$  distribution with degrees of freedom  $k - 1 = 2$  and  $N - k = 49$ .

**CONCLUDE:** Since the  $P$ -value (0.031) is less than  $\alpha = 0.05$ , we reject  $H_0$ . The experiment gives convincing evidence that there is a difference in the true mean amount of time until people like the ones in this study ask for help under the three treatments.

 **For Practice** Try Exercise 17

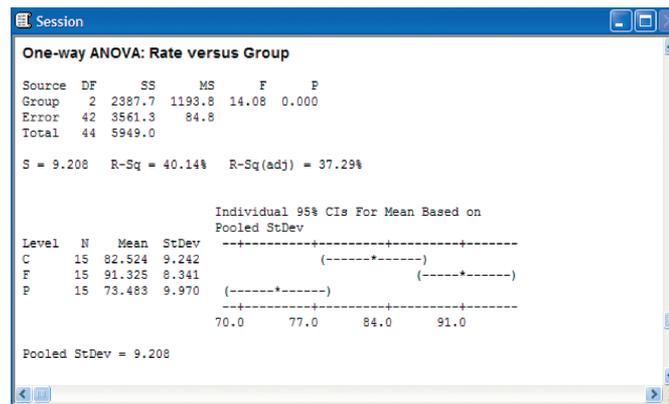
Although the ANOVA  $F$  test in the previous example doesn't tell us which means are different, the data suggest that reminding people of money in either of two ways does make them less likely to ask others for help. This is consistent with the idea that money makes people feel more self-sufficient.



## case closed

### *Do Pets or Friends Help Reduce Stress?*

In the chapter-opening Case Study (page 2), we described details of an experiment investigating whether the presence of a pet or a friend affects subjects' heart rates during a stressful task. Comparing the mean heart rates in the three groups calls for one-way analysis of variance. Figure 13.6 gives the Minitab ANOVA output for these data.



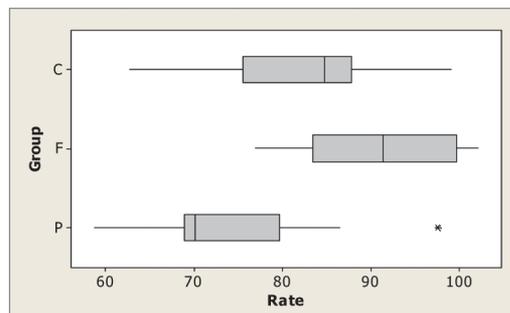
**FIGURE 13.6** Minitab output for the heart rate data (beats per minute) during stress. The control group worked alone, the “Friend” group had a friend present, and the “Pet” group had a pet dog present.



**STATE:** We want to perform a test of  $H_0: \mu_C = \mu_F = \mu_P$  versus  $H_a$ : the means are not all equal, where  $\mu_C$  = the actual mean heart rate of subjects like the ones in this study when performing a stressful task alone, and  $\mu_F$  and  $\mu_P$  are the corresponding parameters for performing the task with a friend present and a pet dog present. Since no significance level was given, we’ll use  $\alpha = 0.05$ .

**PLAN:** If conditions are met, we should perform an ANOVA  $F$  test to compare the means.

- *Random* Subjects were randomly assigned to the three treatment groups.
- *Normal* Side-by-side boxplots of the data are shown below. Unfortunately, there is one high outlier in the pet group. The ANOVA  $F$  test may not give accurate results in this case. To be safe, we’ll perform the analysis with and without the outlier.



- *Independent* The random assignment yields independent groups. If the experiment was conducted properly, individual heart rates should also be independent: knowing one subject's average heart rate should give no additional information about another subject's.
- *Same SD* The Minitab ANOVA output shows that the group standard deviations easily satisfy our rule of thumb: the largest (9.97) is no more than twice the smallest (8.341).

**DO:** From the Minitab output, we have

- *Test statistic*  $F = 14.08$
- *P-value* Approximately 0, using the  $F$  distribution with degrees of freedom 2 and 42.

When we ran the ANOVA  $F$  test again without the outlier, we obtained  $F = 19.34$  and a  $P$ -value of approximately 0.

**CONCLUDE:** With a  $P$ -value of approximately 0, we would reject  $H_0$  at the  $\alpha = 0.05$  level. The experiment gives strong evidence that the average mean heart rates of subjects like the ones in this study differ when performing a task alone, with a friend present, or with a pet present. From the Minitab output, the stress levels appear to be highest when a friend is present and lowest when a pet is present.

## SECTION 13.1

## Summary

- **One-way analysis of variance (ANOVA)** compares the means of several populations or treatments. The **ANOVA  $F$  test** tests the null hypothesis that all the populations or treatments have the same mean. If the  $F$  test shows significant differences, examine the data to see where the differences lie and whether they are large enough to be important.
- The **conditions for ANOVA** are
  - **Random** The data are produced by random samples of size  $n_1$  from Population 1, size  $n_2$  from Population 2, . . . , and size  $n_k$  from Population  $k$  or by groups of size  $n_1, n_2, \dots, n_k$  in a randomized experiment.
  - **Normal** All the population distributions (or the true distributions of responses to the treatments) are Normal, *or* the sample sizes are large and the distributions show no strong skewness or outliers.
  - **Independent** Both the samples/groups themselves and the individual observations in each sample/group are independent. When sampling without replacement, check that all the populations are at least 10 times as large as the corresponding samples (the 10% condition).
  - **Same SD** The response variable has the same standard deviation  $\sigma$  (with a value that is generally unknown) for all the populations or treatments.
- In practice, ANOVA inference is relatively **robust** when the populations are non-Normal, especially when the samples are large. Before doing the  $F$  test, check the observations in each sample for outliers or strong skewness. Also verify that the largest sample standard deviation is no more than twice as large as the smallest standard deviation.
- When the null hypothesis is true, the **ANOVA  $F$  statistic** for comparing  $k$  means from a total of  $N$  observations in all samples combined has the  **$F$  distribution** with  $k - 1$  and  $N - k$  degrees of freedom.
- ANOVA calculations are reported in an **ANOVA table** that gives sums of squares, mean squares, and degrees of freedom for variation among groups and for variation within groups, along with the  $F$  test statistic and  $P$ -value. In practice, we use software to do the calculations.

**13.1 TECHNOLOGY CORNER**

One-way ANOVA on the calculator ..... page 17

## SECTION 13.1 Exercises

- pg 9
- 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 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:<sup>5</sup>

Board Color	Insects 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

- State hypotheses for an ANOVA  $F$  test in this setting. Be sure to define your parameters.
- Check conditions for carrying out the test in part (a).

- More rain for California?** The changing climate will probably bring more rain to California, but we don't know whether the additional rain will come during the winter wet season or extend into the long dry season in spring and summer. Kenwyn Suttle of the University of California at Berkeley and his coworkers randomly assigned plots of open grassland to three treatments: added water equal to 20% of annual rainfall either during January to March (winter) or during April to June (spring), and no added water (control). Here are some of the data, for plant biomass (in grams per square meter) produced by each plot in a single year.<sup>6</sup>

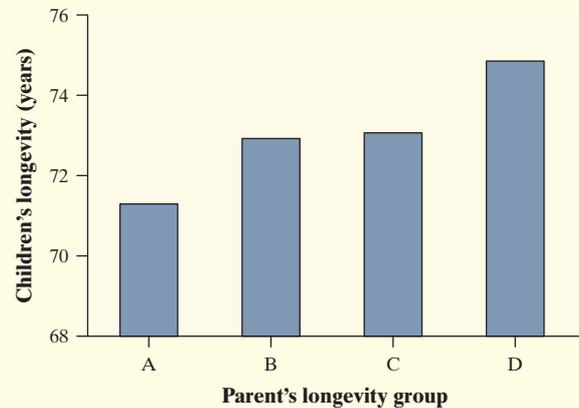
Winter	Spring	Control
264.1514	318.4182	129.0538
187.7312	281.6830	144.6578
291.1431	288.8433	172.7772
176.2879	382.6673	113.2813
141.7525	326.8877	142.1562
169.9737	293.8502	117.9808

- State hypotheses for an ANOVA  $F$  test in this setting. Be sure to define your parameters.
- Check conditions for carrying out the test in part (a).

- Choose your parents wisely** To live long, it helps to have long-lived parents. One study of this "inheritance of longevity" looked at records of families in a small mountain valley in France where the style of life changed little over time.<sup>7</sup> All married couples in which both husband and wife were born between 1745 and 1849 were divided into four groups as follows:

	Short-Lived Wife	Long-Lived Wife
Short-lived husband	Group A	Group C
Long-lived husband	Group B	Group D

"Long-lived" parents were in the top 70% of the distribution of age at death, "short-lived" parents were in the bottom 30%. To examine extended life, the investigators looked at the ages at death of the children of these parents who lived to at least age 55. The report includes a bar graph like the one below.



Explain why it is not wise to carry out an ANOVA  $F$  test in this setting.

- Marital status and salary** Married men tend to earn more than single men. An investigation of the relationship between marital status and income collected data on all 8235 men employed as managers or professionals by a large manufacturing firm. Here are descriptive statistics for the salaries of these men:<sup>8</sup>

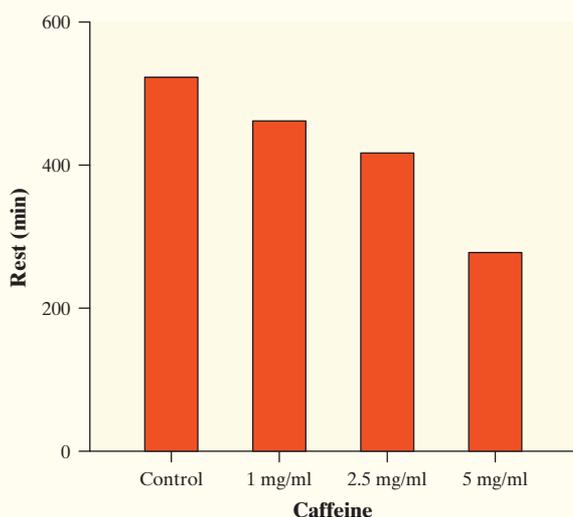
	Single	Married	Divorce	Widowed
$n_i$	337	7,730	126	42
$\bar{x}_i$	\$21,384	\$26,873	\$25,594	\$26,936
$s_i$	\$5,731	\$7,159	\$6,347	\$8,119

Explain why it is not wise to carry out an ANOVA  $F$  test in this setting.

5. **Road rage** “The phenomenon of road rage has been frequently discussed but infrequently examined.” So begins a report based on interviews with 1382 randomly selected drivers.<sup>9</sup> The respondents’ answers to interview questions produced scores on an “angry/threatening driving scale” with values between 0 and 19. What driver characteristics go with road rage? There were no significant differences among races or levels of education. What about the effect of the driver’s age? Here are the means and standard deviations for drivers in three age groups:

Age group	$n$	$\bar{x}$	$s$
Less than 30 years	244	2.22	3.11
30 to 55 years	734	1.33	2.21
Over 55 years	364	0.66	1.60

- (a) State hypotheses for an ANOVA  $F$  test in this setting. Be sure to define your parameters.
- (b) Show that the Random, Independent, and Same SD conditions are met.
- (c) Data from the study confirms that the Normal condition is met. Explain what this means.
6. **Do fruit flies sleep?** Mammals and birds sleep. Insects such as fruit flies rest, but can this rest be considered sleep? Biologists now think that insects do sleep. One experiment gave caffeine to fruit flies to see if it affected their rest. We know that caffeine reduces sleep in mammals, so if it reduces rest in fruit flies, that’s another hint that the rest is really sleep. The paper reporting the study contains a bar graph similar to the one below.<sup>10</sup>



- (a) The explanatory variable is amount of caffeine, in milligrams per milliliter (mg/ml) of blood. The response variable is minutes of rest (measured by an infrared motion sensor) during a 12-hour dark period.

Outline a completely randomized design using 36 flies per group for this experiment.

- (b) State hypotheses for an ANOVA  $F$  test in this setting. Be sure to define your parameters.
- (c) Show that the Random and Independent conditions are met.
- (d) Data from the study confirm that the Normal and Same SD conditions are met. Explain what this means in each case.

*Exercises 7 to 10 describe situations in which we want to compare the mean responses in several populations. For each setting, identify the parameters of interest. Then give  $k$ , the  $n_i$ , and  $N$ . Finally, give the degrees of freedom of the ANOVA  $F$  statistic.*

7. **Morning or evening?** Are you a morning person, an evening person, or neither? Does this personality trait affect how well you perform? A random sample of 100 students took a psychological test that found 16 morning people, 30 evening people, and 54 who were neither. All the students then took a test of their ability to memorize at 8:00 A.M. and again at 9:00 P.M. The response variable is the score at 8:00 A.M. minus the score at 9:00 P.M.
8. **Writing essays** Do strategies such as preparing a written outline help students write better essays? College students were divided at random into four groups of 20 students each, then asked to write an essay on an assigned topic. Group A (the control group) received no additional instruction. Group B was required to prepare a written outline. Group C was given 15 ideas that might be relevant to the essay topic. Group D was given the ideas and also required to prepare an outline. An expert scored the quality of the essays on a scale of 1 to 7.
9. **Test accommodations** Many states require schoolchildren to take regular statewide tests to assess their progress. Children with learning disabilities who read poorly may not do well on mathematics tests because they can’t read the problems. Most states allow “accommodations” for learning-disabled children. Randomly assign 100 learning-disabled children in equal numbers to three types of accommodation and a control group: math problems are read by a teacher; by a computer; by a computer that also shows a video; and standard test conditions. Compare the mean scores on the state mathematics assessment.
10. **A medical study** The Quebec (Canada) Cardiovascular Study recruited men aged 34 to 64 at random from towns in the Quebec City metropolitan area. Of these, 1824 met the criteria (no diabetes, free of heart disease, and so on) for a study of the relationship

between being overweight and medical risks. The 719 normal-weight men had a mean triglyceride level of 1.5 millimoles per liter (mmol/l) of blood; the 885 overweight men had mean 1.7 mmol/l; and the 220 obese men had mean 1.9 mmol/l.<sup>11</sup>

11. **Road rage** The report in Exercise 5 says that  $F = 34.96$ , with  $P < 0.01$ .
- Interpret the  $P$ -value in context.
  - What conclusion would you draw? Justify your answer.
12. **Do fruit flies sleep?** The paper described in Exercise 6 also states, “Flies given caffeine obtained less rest during the dark period in a dose-dependent fashion ( $n = 36$  per group,  $P < 0.0001$ ).” The  $P$ -value in the report comes from the ANOVA  $F$  test.
- Interpret the  $P$ -value in context.
  - What conclusion would you draw? Justify your answer.

*Exercises 13 and 14 refer to the following setting.* People often match their behavior to their social environment. One study of this idea first established that the type of music most preferred by black college students is R&B and that whites most preferred rock music. Will students hosting a small group of other students choose music that matches the makeup of the people attending? Assign 90 black business students at random to three equal-sized groups. Do the same for 96 white students. Each student sees a picture of the people he or she will host. Group 1 sees 6 blacks, Group 2 sees 3 whites and 3 blacks, and Group 3 sees 6 whites. Ask how likely the host is to play the type of music preferred by the other race. Use ANOVA to compare the three groups to see whether the racial mix of the gathering affects the choice of music.<sup>12</sup>

13. **What music will you play?** For the black subjects,  $F = 2.47$ .
- What are the degrees of freedom?
  - Find the  $P$ -value using technology. Interpret this result in context.
  - What conclusion would you draw? Explain.
14. **What music will you play?** For the white subjects,  $F = 16.48$ .
- What are the degrees of freedom?
  - Find the  $P$ -value using technology. Interpret this result in context.
  - What conclusion would you draw? Explain.
15. **Attitudes toward math** Do high school students from different racial/ethnic groups have different attitudes toward mathematics? Measure the level of interest in mathematics on a five-point scale for a national random sample of students. Here are

summaries for students who were taking math at the time of the survey:<sup>13</sup>

Racial/ethnic group	$n$	$\bar{x}$	$s_x$
African American	809	2.57	1.40
White	1860	2.32	1.36
Asian/Pacific Islander	654	2.63	1.32
Hispanic	883	2.51	1.31
Native American	207	2.51	1.28

- Show that the Random, Independent, and Same SD conditions for ANOVA are met.
  - Data from the study confirm that the Normal condition is met. Explain what this means.
  - Calculate the overall mean response  $\bar{x}$ , the mean squares MSG and MSE, and the  $F$  statistic.
  - Which  $F$  distribution would you use to find the  $P$ -value of the ANOVA  $F$  test? Software gives  $P < 0.001$ . What do you conclude from this study?
16. **Exercise and weight loss** What conditions help overweight people exercise regularly? Subjects were randomly assigned to three treatments: a single long exercise period 5 days per week; several 10-minute exercise periods 5 days per week; and several 10-minute periods 5 days per week on a home treadmill that was provided to the subjects. The study report contains the following information about weight loss (in kilograms) after six months of treatment:<sup>14</sup>

Treatment	$n$	$\bar{x}$	$s_x$
Long exercise periods	37	10.2	4.2
Short exercise periods	36	9.3	4.5
Short periods with equipment	42	10.2	5.2

- Show that the Random, Independent, and Same SD conditions for ANOVA are met.
- Data from the study confirm that the Normal condition is met. Explain what this means.
- Calculate the overall mean response  $\bar{x}$ , the mean squares MSG and MSE, and the  $F$  statistic.
- Which  $F$  distribution would you use to find the  $P$ -value of the ANOVA  $F$  test? Software says that  $P = 0.634$ . What do you conclude from this study?

*Exercises 17 and 18 refer to the following setting.* How does logging in a tropical rain forest affect the forest in later years? Researchers compared 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). Although the study was



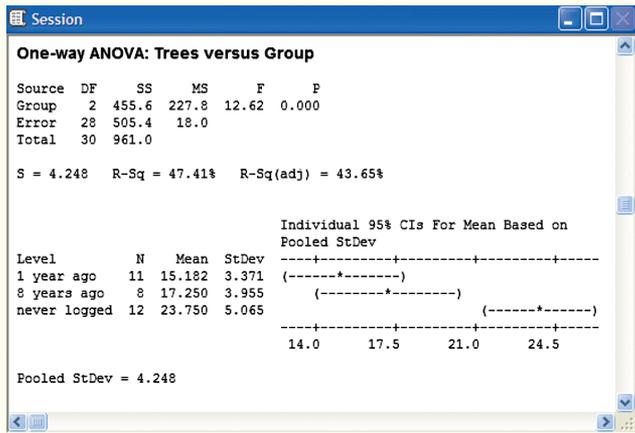
pg 14

not an experiment, the authors explain why we can consider the plots to be randomly selected. The data appear in the table below. The variable Trees is the count of trees in a plot; Species is the count of tree species in a plot.<sup>15</sup>

Never Logged		Logged 1 Year Ago		Logged 8 Years Ago	
Trees	Species	Trees	Species	Trees	Species
27	22	12	11	18	17
22	18	12	11	22	18
29	22	15	14	15	14
21	20	9	7	18	18
19	15	20	18	19	15
33	21	18	15	22	15
16	13	17	15	12	10
20	13	14	12	12	12
24	19	14	13		
27	13	17	15		
28	19	19	8		
19	15				

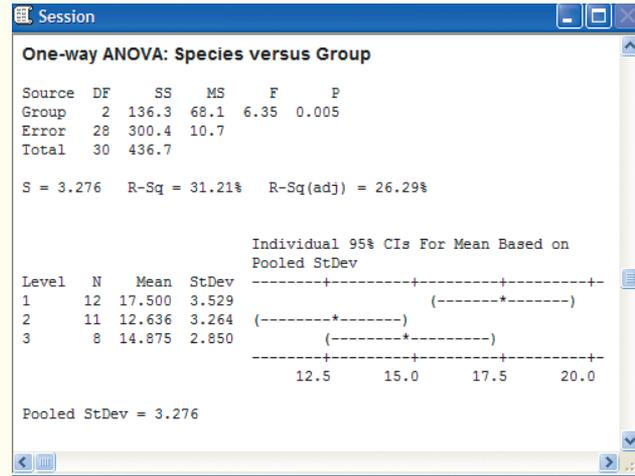
17. Logging in the rain forest

- pg 19 (a) Make a graph suitable for comparing the number of trees per plot for the three groups. Write a few sentences describing what you see.  
 (b) The figure below shows ANOVA output for these data. Carry out the ANOVA  $F$  test.



18. Logging in the rain forest

- (a) Make a graph suitable for comparing the number of tree species per plot for the three groups. Write a few sentences describing what you see.  
 (b) The figure below shows ANOVA output for these data. Carry out the ANOVA  $F$  test.



19. Can you hear these words? To test whether a hearing aid is right for a patient, audiologists play a tape on which words are pronounced at low volume. The patient tries to repeat the words. There are several different lists of words that are supposed to be equally difficult. Are the lists equally difficult when there is background noise? To find out, an experimenter had subjects with normal hearing listen to four lists with a noisy background. The response variable was the percent of the 50 words in a list that the subject repeated correctly. The data set contains 96 responses.<sup>16</sup> Here are two study designs that could produce these data:

- Design A:** The experimenter assigns 96 subjects to four groups at random. Each group of 24 subjects listens to one of the lists. All individuals listen and respond separately.  
**Design B:** The experimenter has 24 subjects. Each subject listens to all four lists in random order. All individuals listen and respond separately.

Does Design A allow use of one-way ANOVA to compare the lists? Does Design B allow use of one-way ANOVA to compare the lists? Briefly explain your answers.

20. Which test? The example on page 19 describes one of the experiments done by Kathleen Vohs and her coworkers to demonstrate that even being reminded of money makes people more self-sufficient and less involved with other people. Here are three more of these experiments. For each experiment, which inference method from Chapters 8 to 13 would you use, and why?

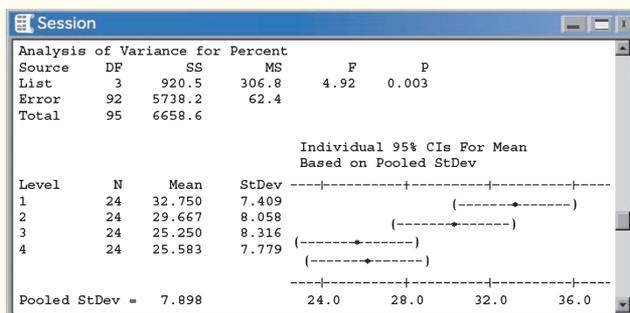
- (a) Randomly assign student subjects to money and control groups. The control group unscrambles neutral phrases, and the money group unscrambles money-oriented phrases, as described in the earlier example. Then ask the subjects to volunteer to help the experimenter by coding data sheets, which takes

about five minutes per sheet. Subjects said how many sheets they would volunteer to code. “Participants in the money condition volunteered to help code fewer data sheets than did participants in the control condition.”

(b) Randomly assign student subjects to high-money, low-money, and control groups. After playing Monopoly for a short time, the high-money group is left with \$4000 in Monopoly money, the low-money group with \$200, and the control group with no money. Each subject is asked to imagine a future with lots of money (high-money group), a little money (low-money group), or just their future plans (control group). Another student walks in and spills a box of 27 pencils. How many pencils does the subject pick up? “Participants in the high-money condition gathered fewer pencils” than subjects in the other two groups.

(c) Randomly assign student subjects to three groups. All do paperwork while a computer on the desk shows a screensaver of currency floating underwater (Group 1), a screensaver of fish swimming underwater (Group 2), or a blank screen (Group 3). Each subject must now develop an advertisement and can choose whether to work alone or with a partner. Count how many in each group make each choice. “Choosing to perform the task with a coworker was reduced among money condition participants.”

21. **Can you hear these words?** The figure below displays the Minitab output for one-way ANOVA applied to the hearing data described in Exercise 19. The response variable is “Percent,” and “List” identifies the four lists of words. Based on this analysis, is there good reason to think that the four lists are not all equally difficult? Perform an appropriate test to help answer this question. Graphs of the data (not shown) confirm that the Normal condition is met.

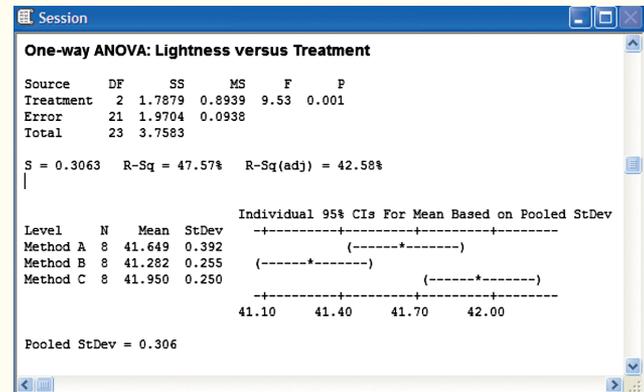


22. **Which blue is most blue?** The color of a fabric depends on the dye used and also on how the dye is applied. This matters to clothing manufacturers, who want the color of the fabric to be just right. A manufacturer dyes fabric made of ramie with the same “procion blue” die but applies the dye in three

different ways. Then he uses a colorimeter to measure the lightness of the color on a scale in which black is 0 and white is 100. Here are the data for 24 pieces of fabric divided into three groups of 8 that were randomly assigned to be dyed in each way:<sup>17</sup>

<b>Method A:</b>	41.72	41.83	42.05	41.44
	41.27	42.27	41.12	41.49
<b>Method B:</b>	40.98	40.88	41.30	41.28
	41.66	41.50	41.39	41.27
<b>Method C:</b>	41.68	41.65	42.30	42.04
	42.25	41.99	41.72	41.97

The figure below displays the Minitab output for one-way ANOVA applied to these data. Based on this analysis, is there good reason to think that the three methods do not all yield the same shade of blue? Perform an appropriate test to help answer this question.



23. **Python weights** A study of the effect of nest temperature on the development of water pythons separated python eggs at random into nests at three temperatures: cold, neutral, and hot. Exercise 33 in Chapter 11 (page 725) shows that the proportions of eggs that hatched at each temperature did not differ significantly. Now we will examine the weights of the little pythons. In all, 16 eggs hatched at the cold temperature, 38 at the neutral temperature, and 75 at the hot temperature. The report of the study summarizes the data in the common form “mean ± standard error of the mean” as follows:<sup>18</sup>

Temperature	n	Weight (grams) at Hatching	Propensity to Strike
Cold	18	28.89 ± 0.08	6.40 ± 5.67
Neutral	38	32.93 ± 5.61	5.82 ± 4.24
Hot	75	32.27 ± 4.10	4.30 ± 2.70

Graphs of the data (not shown) confirm that the Normal condition is met.

(a) We will compare the mean weights at hatching. Recall that the standard error of the mean is  $s_x/\sqrt{n}$ . Find the standard deviations of the weights in the three groups, and verify that they satisfy our rule of thumb for using ANOVA.

(b) Starting from the sample sizes  $n_i$ , the means  $\bar{x}_i$ , and the standard deviations  $s_i$ , carry out an ANOVA. That is, find MSG, MSE, the  $F$  statistic, and the  $P$ -value. Is there evidence that nest temperature affects the mean weight of newly hatched pythons? Justify your answer.

24. **Python strikes** The data in the previous exercise also describe the “propensity to strike” of the hatched pythons at 30 days of age. This is the number of taps on the head with a small brush until the python launches a strike. (Don’t try this at home!) The data are again summarized in the form “sample mean  $\pm$  standard error of the mean.” Graphs of the data (not shown) confirm that the Normal condition is met. Follow the outline in parts (a) and (b) of the previous exercise for propensity to strike. Does nest temperature appear to influence propensity to strike?
25.  **$F$  versus  $t$**  We have two methods to compare the means of two groups: the two-sample  $t$  test of Section 10.2 and the ANOVA  $F$  test with  $k = 2$ . We prefer the  $t$  test because it allows one-sided alternatives and does not assume that both populations have the same standard deviation. Let’s apply both tests to the same data.

There are two types of life insurance companies. “Stock” companies have shareholders, and “mutual” companies are owned by their policyholders. Take an SRS of each type of company from those listed in a directory of the industry. Then ask the annual cost per \$1000 of insurance for a \$50,000 policy insuring the life of a 35-year-old man who does not smoke. Here are the data summaries:<sup>19</sup>

	Stock companies	Mutual companies
$n_i$	13	17
$\bar{x}_i$	\$2.31	\$2.37
$s_i$	\$0.38	\$0.58

Graphs of the data (not shown) confirm that the Normal condition is met.

- (a) Calculate the two-sample  $t$  statistic for testing  $H_0: \mu_1 = \mu_2$  against the two-sided alternative. Use the conservative method to find the  $P$ -value.
- (b) Calculate MSG, MSE, and the ANOVA  $F$  statistic for the same hypotheses. What is the  $P$ -value of  $F$ ?

(c) How close are the two  $P$ -values? (The square root of the  $F$  statistic is a  $t$  statistic with  $N - k = n_1 + n_2 - 2$  degrees of freedom. This is the “pooled two-sample  $t$ ” mentioned on page 645. So  $F$  for  $k = 2$  is exactly equivalent to a  $t$  statistic, but it is a slightly different  $t$  from the one we use.)

**Multiple choice: Select the best answer for Exercises 26 to 29.**

26. A study of the effects of smoking classifies subjects as nonsmokers, moderate smokers, or heavy smokers. The investigators interview a random sample of 200 people in each group. Among the questions is “How many hours do you sleep on a typical night?” The degrees of freedom for the ANOVA  $F$  statistic comparing mean hours of sleep are
- (a) 2 and 197. (d) 3 and 597.  
 (b) 2 and 199. (e) 3 and 599.  
 (c) 2 and 597.
27. The alternative hypothesis for the ANOVA  $F$  test in the previous exercise is
- (a) the mean hours of sleep in the groups are all the same.  
 (b) the mean hours of sleep in the groups are all different.  
 (c) the mean hours of sleep in the groups are not all the same.  
 (d) the mean hours of sleep is highest for the nonsmokers.  
 (e) there is an association between sleep and smoking status.

*Exercises 28 and 29 refer to the following setting.* The air in poultry-processing plants often contains fungus spores. Large spore concentrations can affect the health of the workers. To measure the presence of spores, air samples are pumped to an agar plate and “colony-forming units (CFUs)” are counted after an incubation period. Here are data from the “kill room” of a plant that slaughters 37,000 turkeys a day, taken at four seasons of the year. Each observation was made on a different day. The units are CFUs per cubic meter of air.<sup>20</sup>

	Fall	Winter	Spring	Summer
	1231	384	2105	3175
	1254	104	701	2526
	752	251	2947	1763
	1088	97	842	1090

Here is Minitab output for ANOVA to compare mean CFUs in the four seasons:

Source	DF	SS	MS	F	P
Season	3	8236359	2745453	5.38	0.014
Error	12	6124211	510351		
Total	15	14360570			

Individual 95% CIs For Mean Based on Pooled StDev			
Level	N	Mean	StDev
Fall	4	1081.3	231.5
Spring	4	1648.8	1071.2
Summer	4	2138.5	906.4
Winter	4	209.0	136.6

0      1000      2000      3000

28. The most striking conclusion from the numerical summaries for the turkey-processing plant is that
- there appears to be little difference among the seasons.
  - on average, CFUs are much lower in winter than in other seasons.
  - the air in the plant is clearly unhealthy.
  - there is more variability in CFUs during the summer than in other seasons.
  - the graphs are all roughly symmetrical.
29. The  $P$ -value 0.014 in the output may not be accurate because the conditions for ANOVA are not satisfied. Which condition is clearly *not* satisfied based solely on the Minitab output?
- Random
  - Normal
  - Independent
  - Same SD
  - None of these

## SECTION 13.1 Notes

1. These are some of the data from the EESEE story “Stress among Pets and Friends.” The study results appear in K. Allen, J. Blascovich, J. Tomaka, and R. M. Kelsey, “Presence of human friends and pet dogs as moderators of autonomic responses to stress in women,” *Journal of Personality and Social Psychology*, 83 (1988), pp. 582–589.
2. Ethan J. Temeles and W. John Kress, “Adaptation in a plant-hummingbird association,” *Science*, 300 (2003), pp. 630–633. I thank Ethan J. Temeles for providing the data.
3. Jacqueline T. Ngai and Diane S. Srivastava, “Predators accelerate nutrient cycling in a bromeliad ecosystem,” *Science*, 314 (2006), p. 963. I thank Jacqueline Ngai for providing the data.
4. Kathleen D. Vohs, Nicole L. Mead, and Miranda R. Goode, “The psychological consequences of money,” *Science*, 314 (2006), pp. 1154–1156. I thank Kathleen Vohs for providing the data.
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. K. B. Suttle, Meredith A. Thomsen, and Mary E. Power, “Species interactions reverse grassland responses to changing climate,” *Science*, 315 (2007), pp. 640–642.
7. Amandine Cournil, Jean-Marie Legay, and François Schächter, “Evidence of sex-linked effects on the inheritance of human longevity: a population-based study in the Valserine Valley (French Jura), 18–20th centuries,” *Proceedings of the Royal Society of London: Biological Sciences*, 267 (2000), pp. 1021–1025.
8. Sanders Korenman and David Neumark, “Does marriage really make men more productive?” *Journal of Human Resources*, 26 (1991), pp. 282–307.
9. Elisabeth Wells-Parker et al., “An exploratory study of the relationship between road rage and crash experience in a representative sample of U.S. drivers,” *Accident Analysis and Prevention*, 34 (2002), pp. 271–278.
10. Based on the online supplement to Paul J. Shaw et al., “Correlates of sleep and waking in *Drosophila melanogaster*,” *Science*, 287 (2000), pp. 1834–1837.
11. Annie C. St. Pierre et al., “Insulin resistance syndrome, body mass index and the risk of ischemic heart disease,” *Canadian Medical Association Journal*, 172 (2005), pp. 1301–1305.
12. David B. Wooten, “One-of-a-kind in a full house: some consequences of ethnic and gender distinctiveness,” *Journal of Consumer Psychology*, 4 (1995), pp. 205–224.
13. John P. Thomas, “Influences on mathematics learning and attitudes among African American high school students,” *Journal of Negro Education*, 69 (2000), pp. 165–183.
14. John M. Jakicic et al., “Effects of intermittent exercise and use of home exercise equipment on adherence, weight loss, and fitness in overweight women,” *Journal of the American Medical Association*, 282 (1999), pp. 1554–1560.
15. C. H. Cannon, D. R. Peart, and M. Leighton, “Tree species diversity in commercially logged Bornean rainforest,” *Science*, 281 (1998), pp. 1366–1367. I thank Charles Cannon for providing the data.
16. The data and the full story can be found in the Data and Story Library at lib.stat.cmu.edu. The original study is by Faith Loven, “A study of interlist equivalency of the CID W-22 word list presented in quiet and in noise,” MS thesis, University of Iowa, 1981.
17. Yvan R. Germain, “The dyeing of ramie with fiber reactive dyes using the cold pad-batch method,” MS thesis, Purdue University, 1988.
18. R. Shine, T. R. L. Madden, M. J. Elphick, and P. S. Harlow, “The influence of nest temperatures and maternal brooding on hatchling phenotypes in water pythons,” *Ecology*, 78 (1997), pp. 1713–1721.
19. Mark Kroll, Peter Wright, and Pochera Theerathorn, “Whose interests do hired managers pursue? An examination of select mutual and stock life insurers,” *Journal of Business Research*, 26 (1993), pp. 133–148.
20. Michael W. Peugh, “Field investigation of ventilation and air quality in duck and turkey slaughter plants,” MS thesis, Purdue University, 1996.