Example 1

Investigating Customer Satisfaction

Picture the Scenario

In recent years, many companies have increased the attention paid to measuring and analyzing customer satisfaction. Here are examples of two recent studies of customer satisfaction:

- A company that makes personal computers has provided a toll-free telephone number for owners of their PCs to call and seek technical support. For years the company had two service centers for these calls: San Jose, California, and Toronto, Canada. Recently the company outsourced many of the customer service jobs to a new center in Bangalore, India, because employee salaries are much lower there. The company wanted to compare customer satisfaction at the three centers.
- An airline has a toll-free telephone number that potential customers can call to make flight reservations. Usually the call volume is heavy and callers are placed on hold until an agent is free to answer. Researchers working for the airline recently conducted a randomized experiment to analyze whether callers would remain on hold longer if they heard (a) an advertisement about the airline and its current promotions, (b) recorded Muzak ("elevator music"), or (c) recorded classical music. Currently, messages are five minutes long and then repeated; the researchers also wanted to find out if it would make a difference if messages were instead 10 minutes long before repeating.

Questions to Explore

In the second study, the company's CEO had some familiarity with statistical methods, based on a course he took in college. He asked the researchers:

- In this experiment, are the sample mean times that callers stayed on hold before hanging up significantly different for the three recorded message types?
- What conclusions can you make if you take into account both the type of recorded message and whether it was repeated every five minutes or every ten minutes?

Thinking Ahead

Chapter 10 showed how to compare two means. In practice, there may be *several* means to compare, such as in the second example. This chapter shows how to use statistical inference to compare several means. We'll see how to determine whether a set of sample means is significantly different and how to estimate the differences among corresponding population means. To illustrate, we'll analyze data from the second study in Examples 2–4 and 7.

Example 2

Tolerance of Being on Hold?

Picture the Scenario

Let's refer back to the second scenario in Example 1. An airline has a toll-free telephone number for reservations. Often the call volume is heavy, and callers are placed on hold until a reservation agent is free to answer. The airline hopes a caller remains on hold until the call is answered, so as not to lose a potential customer.

The airline recently conducted a randomized experiment to analyze whether callers would remain on hold longer, on average, if they heard (a) an advertisement about the airline and its current promotions, (b) Muzak, or (c) classical music (Vivaldi's *Four Seasons*). The company randomly selected one out of every 1000 calls in a particular week. For each call, they randomly selected one of the three recorded messages to play and then measured the number of minutes that the caller remained on hold before hanging up (these calls were purposely not answered). The total sample size was 15. The company kept the study small, hoping it could make conclusions without alienating too many potential customers! Table 14.1 shows the data. It also shows the mean and standard deviation for each recorded message type.

Table 14.1 Telephone Holding Times by Type of Recorded Message

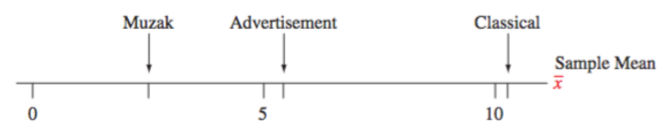
Each observation is the number of minutes a caller remained on hold before hanging up, rounded to the nearest minute.

Recorded Message	Holding Time Observations	Sample Size	Mean	Standard Deviation
Advertisement	5, 1, 11, 2, 8	5	5.4	4.2
Muzak	0, 1, 4, 6, 3	5	2.8	2.4
Classical	13, 9, 8, 15, 7	5	10.4	3.4

Questions to Explore

a. What are the hypotheses for the ANOVA test?

b. Figure 14.1 displays the sample means. Since these means are quite different, is there sufficient evidence to conclude that the population means differ?



▲ **Figure 14.1 Sample Means of Telephone Holding Times for Callers Who Hear One of Three Recorded Messages. Question** Since the sample means are quite different, can we conclude that the population means differ?

Think It Through

a. Denote the holding time means for the population that these three random samples represent by μ_1 for the advertisement, μ_2 for Muzak, and μ_3 for classical music. ANOVA tests whether these are equal. The null hypothesis is $H_0: \mu_1 = \mu_2 = \mu_3$. The alternative hypothesis is that at least two of the population means are different.

b. The sample means are quite different. But even if the population means are equal, we expect the sample means to differ because of sampling variability. So these differences alone are not sufficient evidence to enable us to reject H_0.

Insight

The strength of evidence against H_0 will also depend on the sample sizes and the variability of the data. We'll next study how to test these hypotheses.

Try Exercise 14.1, parts a and b

Example 3

Telephone Holding Times

Picture the Scenario

Examples 1 and 2 discussed a study of the length of time that 15 callers to an airline's toll-free telephone number remain on hold before hanging up. The study compared three recorded messages: an advertisement about the airline, Muzak, and classical music. Let μ_1, μ_2, and μ_3 denote the population mean telephone holding times for the three messages.

Questions to Explore

a. For testing H_0: $\mu_1 = \mu_2 = \mu_3$ based on this experiment, what value of the F test statistic would have a P-value of 0.05?

b. For the data in Table 14.1 on page 682, we'll see that software reports $F = 6.4$. Based on the answer to part a, will the P-value be larger, or smaller, than 0.05?

c. Can you reject H_0, using a significance level of 0.05? What can you conclude from this?

Think It Through

a. With $g = 3$ groups and a total sample size of $N = 15$ (5 in each group), the test statistic has

$$df_1 = g - 1 = 2 \text{ and } df_2 = N - g = 15 - 3 = 12.$$

From Table D (see the excerpt in the margin), with these df values an F test statistic value of 3.88 has a P-value of 0.05.

b. Since the F test statistic of 6.4 is farther out in the tail than 3.88 (see figure in margin), the right-tail probability above 6.4 is less than 0.05. So, the P-value is less than 0.05.

c. Since P-value < 0.05, there is sufficient evidence to reject H_0: $\mu_1 = \mu_2 = \mu_3$. We conclude that a difference exists among the three types of messages in the population mean amount of time that customers are willing to remain on hold.

Insight

We'll see that software reports a P-value $= 0.013$. This is quite strong evidence against H_0. If H_0 were true, there'd be only about a 1% chance of getting an F test statistic value larger than the observed F value of 6.4.

Try Exercise 14.2, parts a–c

Example 4

Telephone Holding Time Study

Picture the Scenario

Let's check the assumptions for the F test on telephone holding times (Example 3).

Question to Explore

Is it appropriate to apply ANOVA to the data in Table 14.1 to compare mean telephone holding times for three message types?

Think It Through

Subjects were selected randomly for the experiment and assigned randomly to the three recorded messages. From Table 14.1 (summarized in the margin), the largest sample standard deviation of 4.2 is less than twice the smallest standard deviation of 2.4 (In any case, the sample sizes are equal, so this assumption is not crucial). The sample sizes in Table 14.1 are small, so it is difficult to make judgments about shapes of population distributions. However, the dot plots in Figure 14.2a on page 683 did not show evidence of severe nonnormality. Thus, ANOVA is suitable for these data.

Insight

As in other statistical inferences, the method used to gather the data is the most crucial factor. Inferences have greater validity when the data are from an experimental study that randomly assigned subjects to groups or from a sample survey that used random sampling.

Try Exercise 14.11, part c

Example 5

Number of Good Friends and Happiness

Picture the Scenario

Chapter 11 investigated the association between happiness and several categorical variables, using data from the General Social Survey. The respondents indicated whether they were very happy, pretty happy, or not too happy. Is happiness associated with having lots of friends? A recent GSS asked, "About how many good friends do you have?" Here, we could treat either happiness (variable HAPPY in GSS) or number of good friends (NUMFREND in GSS) as the response variable. If we choose number of good friends, then we are in the ANOVA setting, having a quantitative response variable and a categorical explanatory variable (happiness).

For each happiness category, Table 14.3 shows the sample mean, standard deviation, and sample size for the number of good friends. It also shows the ANOVA table for the *F* test comparing the population means. The small P-value of 0.03 suggests that at least two of the three population means are different.

Table 14.3 Summary of ANOVA for Comparing Mean Number of Good Friends for Three Happiness Categories

The analysis is based on GSS data.

	Very happy	Pretty happy	Not too happy
Mean	10.4	7.4	8.3
Standard deviation	17.8	13.6	15.6
Sample size	276	468	87

Source	DF	SS	MS	F	P
Group	2	1626.8	813.4	3.47	0.032
Error	828	193900.9	234.2		
Total	830	195527.7			

Question to Explore

Use 95% confidence intervals to compare the population mean number of good friends for the three pairs of happiness categories—very happy with pretty happy, very happy with not too happy, and pretty happy with not too happy.

Think It Through

From Table 14.3, the MS error is 234.2. The residual standard deviation is $s = \sqrt{234.2} = 15.3$, with $df = 828$ (listed in the row for the MS error). For a 95% confidence interval with $df = 828$, software reports $t_{.025} = 1.963$. For comparing the very happy and pretty happy categories, the confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{.025}s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (10.4 - 7.4) \pm 1.963(15.3)\sqrt{\frac{1}{276} + \frac{1}{468}}$$

which is 3.0 ± 2.3, or $(0.7, 5.3)$.

We infer that the population mean number of good friends is between about 1 and 5 higher for those who are very happy than for those who are pretty happy. Since the confidence interval contains only positive numbers, this suggests that $\mu_1 - \mu_2 > 0$; that is, μ_1 exceeds μ_2. On the average, people who are very happy have more good friends than people who are pretty happy.

For the other comparisons, you can find

Very happy, not too happy: 95% CI for $\mu_1 - \mu_3$ is $(10.4 - 8.3) \pm 3.7$, or $(-1.6, 5.8)$.

Pretty happy, not too happy: 95% CI for $\mu_2 - \mu_3$ is $(7.4 - 8.3) \pm 3.5$, or $(-4.4, 2.6)$.

These two confidence intervals contain 0. So there's not enough evidence to conclude that a difference exists between μ_1 and μ_3 or between μ_2 and μ_3.

Insight

The confidence intervals are quite wide, even though the sample sizes are fairly large. This is because the sample standard deviations (and hence *s*) are large. Table 14.3 reports that the sample standard deviations are larger than the sample means, suggesting that the three distributions are skewed to the right. The margin figure shows a box plot of the overall sample data distribution of number of good friends, except for many large outliers. It would also be sensible to compare the median responses, but these are not available at the GSS website. Do non-normal population distributions invalidate this inferential analysis? We'll discuss this next.

Try Exercise 14.12

Example 6

Number of Good Friends

Picture the Scenario

Example 5 compared the population mean numbers of good friends, for three levels of reported happiness. There, we constructed a *separate* 95% confidence interval for the difference between each pair of means. Table

Table 14.4 Multiple Comparisons of Mean Good Friends for Three Happiness Categories

An asterisk * indicates a significant difference, with the confidence interval not containing 0.

Groups	Difference of means	Separate 95% CIs	Tukey 95% Multiple Comparison CIs
(Very happy, Pretty happy)	$\mu_1 - \mu_2$	(0.7, 5.3)*	(0.3, 5.7)*
(Very happy, Not too happy)	$\mu_1 - \mu_3$	(−1.6, 5.8)	(−2.6, 6.5)
(Pretty happy, Not too happy)	$\mu_2 - \mu_3$	(−4.4, 2.6)	(−5.1, 3.3)

Question to Explore

a. Explain how the Tukey multiple comparison confidence intervals differ from the separate confidence intervals in Table 14.4.

b. Summarize results shown for the Tukey multiple comparisons.

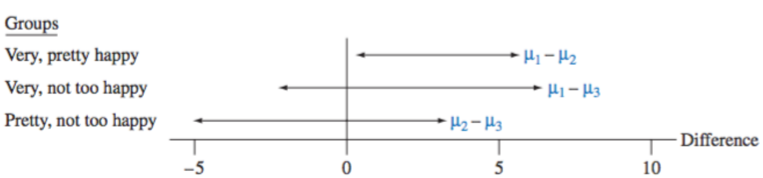
Think It Through

a. The Tukey confidence intervals hold with an *overall* confidence level of about 95%. This confidence applies to the entire set of three intervals. The Tukey confidence intervals are wider than the separate 95% confidence intervals because the multiple comparison approach uses a higher confidence level for each separate interval to ensure achieving the overall confidence level of 95% for the entire set of intervals.

b. The Tukey confidence interval for $\mu_1 - \mu_2$ contains only positive values, so we infer that $\mu_1 > \mu_2$. The mean number of good friends is higher, although perhaps barely so, for those who are very happy than for those who are pretty happy. The other two Tukey intervals contain 0, so we cannot infer that those pairs of means differ.

Insight

Figure 14.4 summarizes the three Tukey comparisons from Table 14.4. The intervals have different lengths because the group sample sizes are different.



▲ **Figure 14.4** Summary of Tukey Comparisons of Pairs of Means.

Try Exercise 14.15

Example 7

Telephone Holding Times

Picture the Scenario

Let's return to the data we analyzed in Examples 1–4 on telephone holding times for callers to an airline for which the recorded message is an advertisement, Muzak, or classical music.

Questions to Explore

a. Set up indicator variables to use regression to model the mean holding times with the type of recorded message as explanatory variable.

b. Table 14.6 shows a portion of a MINITAB printout for fitting this model. Use it to find the estimated mean holding time for the advertisement recorded message.

c. Use Table 14.6 to conduct the ANOVA F test (which Example 3 had shown).

Table 14.6 Printout for Regression Model $\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2$ for Telephone Holding Times and Type of Recorded Message

The indicator variables are x_1 for the advertisement and x_2 for Muzak.

```
Predictor        Coef    SE Coef      T        P
Constant       10.400     1.523     6.83    0.000
x1             -5.000     2.154    -2.32    0.039
x2             -7.600     2.154    -3.53    0.004
Analysis of Variance
Source           DF        SS       MS       F        P
Regression        2    149.20    74.60    6.43    0.013
Residual Error   12    139.20    11.60
Total            14    288.40
```

Think It Through

a. The factor (type of recorded message) has three categories—advertisement, Muzak, and classical music. We set up indicator variables x_1 and x_2 with

$$x_1 = 1 \text{ for the advertisement (and 0 otherwise)},$$
$$x_2 = 1 \text{ for Muzak (and 0 otherwise)},$$
$$\text{so } x_1 = x_2 = 0 \text{ for classical music.}$$

The regression model for the mean of y = telephone holding time is then

$$\mu_y = \alpha + \beta_1 x_1 + \beta_2 x_2.$$

b. From Table 14.6, the prediction equation is

$$\hat{y} = 10.4 - 5.0x_1 - 7.6x_2.$$

For the advertisement, $x_1 = 1$ and $x_2 = 0$, so the estimated mean is $\hat{y} = 10.4 - 5.0(1) - 7.6(0) = 5.4$. This is the sample mean for the five subjects in that group.

c. From Table 14.6, the F test statistic for testing

$$H_0: \beta_1 = \beta_2 = 0$$

is $F = 6.43$, with $df_1 = 2$ and $df_2 = 12$. This null hypothesis is equivalent to

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

Table 14.6 reports a P-value of 0.013. The regression approach provides the same F test statistic and P-value as the ANOVA did in Table 14.2 on page 686.

Insight

Testing that the beta coefficients equal zero is a way of testing that the population means are equal. Likewise, confidence intervals for those coefficients give us confidence intervals for differences between means. For instance, since $\beta_1 = \mu_1 - \mu_3$, a confidence interval for β_1 is also a confidence interval comparing μ_1 and μ_3. From Table 14.6, the estimate -5.0 of β_1 has $se = 2.154$. Since $df = 12$ for that interval, $t_{.025} = 2.179$, and a 95% confidence interval for $\mu_1 - \mu_3$ is

$$-5.0 \pm 2.179(2.154), \text{ or } -5.0 \pm 4.7, \text{ which is } (-9.7, -0.3).$$

This agrees with the 95% confidence interval you would obtain using the difference between the sample means and its standard error.

Try Exercise 14.17

Example 8

Amounts of Fertilizer and Manure

Picture the Scenario

This example presents a typical ANOVA application, based on a study at Iowa State University.[2] A large field was portioned into 20 equal-size plots. Each plot was planted with the same amount of corn seed, using a fixed spacing pattern between the seeds. The goal was to study how the yield of corn later harvested from the plots (in metric tons) depended on the levels of use of nitrogen-based fertilizer and manure. Each factor was measured in a binary manner. The fertilizer level was low = 45-kg per hectare or high = 135 kg per hectare. The manure level was low = 84 kg per hectare or high = 168 kg per hectare.

Questions to Explore

a. What are four treatments you can compare with this experiment?

b. What comparisons are relevant when you control for (keep fixed) manure level?

Think It Through

a. Four treatments result from cross-classifying the two binary factors: fertilizer level and manure level. Table 14.7 shows the four treatments, defined for the $2 \times 2 = 4$ combinations of categories of the two factors (fertilizer level and manure level).

Table 14.7 Four Groups for Comparing Mean Corn Yield

These result from the two-way cross classification of fertilizer level with manure level.

	Fertilizer	
Manure	**Low**	**High**
Low		
High		

b. We can compare the mean corn yield for the two levels of fertilizer, controlling for manure level (that is, at a fixed level of manure use). For fields in which manure level was *low*, we can compare the mean yields for the two levels of fertilizer use. These refer to the first row of Table 14.7. Likewise, for fields in which manure level was *high*, we can compare the mean yields for the two levels of fertilizer use. These refer to the second row of Table 14.7.

Insight

Among the questions we'll learn how to answer in this section are: Does the mean corn yield depend significantly on the fertilizer level? Does it depend on the manure level? Does the effect of fertilizer depend on the manure level?

Try Exercise 14.22

Example 9

Corn Yield

Picture the Scenario

Let's analyze the relationship between corn yield and the two factors, fertilizer level and manure level. Table 14.9 shows the data and the sample mean and standard deviation for each group.

Table 14.9 Corn Yield by Fertilizer Level and Manure Level

Fertilizer Level	Manure Level	Plot 1	Plot 2	Plot 3	Plot 4	Plot 5	Sample Size	Mean	Std. Dev.
High	High	13.7	15.8	13.9	16.6	15.5	5	15.1	1.3
High	Low	16.4	12.5	14.1	14.4	12.2	5	13.9	1.7
Low	High	15.0	15.1	12.0	15.7	12.2	5	14.0	1.8
Low	Low	12.4	10.6	13.7	8.7	10.9	5	11.3	1.9

Questions to Explore

a. Summarize the factor effects as shown by the sample means.

b. Table 14.10 is an ANOVA table for two-way ANOVA. Specify the two hypotheses tested, give the test statistics and P-values, and interpret.

Table 14.10 Two-Way ANOVA for Corn Yield Data in Table 14.9

MS values for numerator of F statistics

Source	DF	SS	MS	F	P
Fertilizer	1	17.67	17.67	6.33	0.022
Manure	1	19.21	19.21	6.88	0.018
Error	17	47.44	2.79		
Total	19	84.32			

MS error is denominator of each F statistic

Think It Through

a. Table 14.9 (with means summarized in the margin) shows that for each manure level, the sample mean yield is higher for the plots using more fertilizer. For each fertilizer level, the sample mean yield is higher for the plots using more manure.

b. First, consider the hypothesis

H_0: Mean corn yield is equal for plots at the low and high levels of fertilizer, for each fixed level of manure.

For the fertilizer main effect, Table 14.10 reports that the between-groups estimate of the variance is 17.67. This is the mean square (MS) for fertilizer in Table 14.10. The within-groups estimate is the MS error, or 2.79. The F test statistic is the ratio,

$$F = 17.67 / 2.79 = 6.33.$$

From Table 14.10, the *df* values are 1 and 17 for the two estimates. From the F distribution with $df_1 = 1$ and $df_2 = 17$, the P-value is 0.022, also reported in Table 14.10. If the population means were equal at the two levels of fertilizer, the probability of an F test statistic value larger than 6.33 would be only 0.022. There is strong evidence that the mean corn yield depends on fertilizer level.

Next, consider the hypothesis

H_0: Mean corn yield is equal for plots at the low and high levels of manure, for each fixed level of fertilizer.

For the manure main effect, the F test statistic is $F = 19.21 / 2.79 = 6.88$. From Table 14.10, $df_1 = 1$ and $df_2 = 17$, and the P-value is 0.018. There is strong evidence that the mean corn yield also depends on the manure level.

Insight

As with any significance test, the information gain is limited. We do not learn *how large* the fertilizer and manure effects are on the corn yield. We can use confidence intervals to investigate the sizes of the main effects. We'll now learn how to do this by regression modeling with indicator variables.

Try Exercise 14.27

Example 10

Estimate and Compare Mean Corn Yields

Picture the Scenario

Table 14.12 shows the result of fitting the regression model for predicting corn yield with indicator variables for fertilizer level and manure level.

Table 14.12 Estimates of Regression Parameters for Two-Way ANOVA of the Mean Corn Yield by Fertilizer Level and Manure Level

Predictor	Coef	SE Coef	T	P
Constant	11.6500	0.6470	18.01	0.000
fertilizer	1.8800	0.7471	2.52	0.022
manure	1.9600	0.7471	2.62	0.018

Questions to Explore

a. Find and use the prediction equation to estimate the mean corn yield for each group.

b. Use a parameter estimate from the prediction equation to compare mean corn yields for the high and low levels of fertilizer, at each manure level.

c. Find a 95% confidence interval comparing the mean corn yield at the high and low levels of fertilizer, controlling for manure level. Interpret it.

Think It Through

a. From Table 14.12, the prediction equation is (rounding to one decimal place)

$$\hat{y} = 11.6 + 1.9f + 2.0m.$$

The y-intercept equals 11.6. This is the estimated mean yield (in metric tons per hectare) when both indicator variables equal 0, that is, with fertilizer and manure at the low levels. The estimated means for the other cases result from substituting values for the indicator variables. For instance, at fertilizer level = high and manure level = low, $f = 1$ and $m = 0$, so the estimated mean yield is $\hat{y} = 11.6 + 1.9(1) + 2.0(0) = 13.5$. Doing this for all four groups, we get

	Fertilizer	
Manure	**Low**	**High**
Low	11.6	$11.6 + 1.9 = 13.5$
High	$11.6 + 2.0 = 13.6$	$11.6 + 1.9 + 2.0 = 15.5$

b. The coefficient of the fertilizer indicator variable f is 1.9. This is the estimated difference in mean corn yield between the high and low levels of fertilizer, for each level of manure (for instance, $13.5 - 11.6 = 1.9$ when manure level = low).

c. Again, The estimate of the fertilizer effect is 1.9. Its standard error, reported in Table 14.12, is 0.747. From Table 14.10 (see page 703), the df for the MS error is 17. From Table B, $t_{.025} = 2.11$ when $df = 17$. The 95% confidence interval is

$$1.9 \pm 2.11(0.747), \text{ which is } (0.3, 3.5).$$

At each manure level, we estimate that the mean corn yield is between 0.3 and 3.5 metric tons per hectare higher at the high fertilizer level than at the low fertilizer level. The confidence interval contains only positive values (does not contain 0), reflecting the conclusion that the mean yield is significantly higher at the higher level of fertilizer. This agrees with the P-value falling below 0.05 in the test for the fertilizer effect.

Insight

The *estimated* means are not the same as the *sample* means in Table 14.9. (Both sets are shown again in the margin.) The model *smooths* the sample means so that the difference between the estimated means for two categories of a factor is *exactly* the same at each category of the other factor. For example, from the estimated group means found previously, the fertilizer effect of $1.9 = 13.5 - 11.6 = 15.5 - 13.6$.

Try Exercise 14.28

Example 11

Corn Yield Data

Picture the Scenario

Let's return to our analysis of the corn yield data, summarized in the margin table. In Example 10 we analyzed these data assuming no interaction. Let's see if that analysis is valid. Table 14.14 is an ANOVA table for a model that allows interaction in assessing the effects of fertilizer level and manure level on the mean corn yield.

Table 14.14 Two-Way ANOVA of Mean Corn Yield by Fertilizer Level and Manure Level, Allowing Interaction

Source	DF	SS	MS	F	P
Fertilizer	1	17.67	17.67	6.37	0.023
Manure	1	19.21	19.21	6.92	0.018
Interaction	1	3.04	3.04	1.10	0.311
Error	16	44.40	2.78		
Total	19	84.32			

F statistic for test of no interaction

Question to Explore

Give the result of the test of H_0: no interaction, and interpret.

Think It Through

The test statistic for H_0: no interaction is

$$F = (\text{MS for interaction})/(\text{MS error}) = 3.04/2.78 = 1.10.$$

Based on the F distribution with $df_1 = 1$ and $df_2 = 16$ for these two mean squares, the ANOVA table reports P-value = 0.31. This is not much evidence of interaction. We would not reject H_0 at the usual significance levels, such as 0.05.

Insight

Because there is not much evidence of interaction, we are justified in conducting the simpler two-way ANOVA about main effects. The tests presented previously in Table 14.10 for effects of fertilizer and manure on mean corn yield are valid.

Try Exercise 14.29

Example 12

Political Ideology by Gender and Race

Picture the Scenario

In most years, the General Social Survey asks subjects to report their political ideology, measured with seven categories in which 1 = extremely liberal, 4 = moderate, 7 = extremely conservative. Table 14.15 shows results from the 2008 General Social Survey on mean political ideology classified by gender and by race.

Table 14.15 Mean Political Ideology by Gender and by Race

	Race	
Gender	**Black**	**White**
Female	4.164 (n = 165)	4.268 (n = 840)
Male	3.819 (n = 116)	4.444 (n = 719)

For the test of H_0: no interaction, software reports an F test statistic of 6.19 with $df_1 = 1$ and $df_2 = 1836$, for a P-value of 0.013. So, in comparing females and males on their mean political ideology, we should do it separately by race. The MS error for the model allowing interaction equals 2.534, so the residual standard deviation is $s = \sqrt{2.534} = 1.592$.

Analysis of Variance for POLVIEWS, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
SEX	1	5.138	1.652	1.652	0.65	0.420
RACE	1	24.786	30.780	30.780	12.15	0.001
SEX*RACE	1	15.693	15.693	15.693	6.19	0.013
Error	1836	4651.981	4651.981	2.534		
Total	1839	4697.598				

Questions to Explore

a. Interpret the significant interaction by comparing sample mean political ideology of females and males for each race descriptively and using 95% confidence intervals.

b. Interpret the confidence intervals derived in part a.

Think It Through

a. The sample means show that for black subjects, females are more conservative (have the higher mean). By contrast, for white subjects, males are more conservative. For a confidence interval comparing mean political ideology for females and males who are black, the standard error (using the sample sizes reported in the table) is

$$se = s\sqrt{\frac{1}{n \text{ for black females}} + \frac{1}{n \text{ for black males}}} =$$

$$1.592\sqrt{\frac{1}{165} + \frac{1}{116}} = 0.193.$$

The 95% confidence interval is

$$(4.164 - 3.819) \pm 1.96(0.193), \text{ which is } 0.345 \pm 0.378, \text{ or } (-0.03, 0.72).$$

Likewise, you can find that the 95% confidence interval comparing mean political ideology for females and males who are white is -0.1768 ± 0.159, or $(-0.33, -0.02)$.

b. Since the confidence interval for black subjects contains zero we cannot infer that there is a difference in the populations. For white subjects, however, all values in the interval are negative. We infer that white females are *less* conservative than white males in the population.

Insight

The confidence interval for white subjects has an endpoint that is close to 0. So, the true gender effect in the population could be small.

Try Exercise 14.33