Example 1

Making Sense of Studies Comparing Two Groups

Picture the Scenario

A smile is a universal greeting, a way to communicate with others even if we don't speak their language. Making our smiles as bright as possible with teeth as white as possible has become very desirable. We want our teeth and our smiles to be as perfect looking as those we see on TV and in magazines. Today, teeth-whitening products can be obtained at the dentist or at the drugstore. Numerous claims have been made about the ability of the products to whiten teeth. As with any other claims, some of which follow pseudoscience methods, it is difficult to sort out which products work best or if they work at all.

Studies that investigate claims like weight loss, teeth whiteners, or binge drinkers involve a comparison of two groups or two treatments (such as before and after weight, before and after teeth whiteness, or comparing males and females who binge drink).

Questions to Explore

- How can we use data from an experiment to summarize the evidence about the claims by tooth whitener manufacturers?
- How can we decide, based on the data, whether or not the claims are believable?

Thinking Ahead

This chapter shows how to compare two groups on a categorical or quantitative outcome. To do this, we'll use the inferential statistical methods that the previous two chapters introduced—confidence intervals and significance tests.

For categorical variables, the inferences compare proportions. In Examples 2–4 we'll look at aspirin and placebo treatments, studying their effects on proportions getting cancer. Exercise 10.40 examines a study investigating the "test of whiteness" between whitening gel and toothpaste.

For quantitative variables, the inferences compare means. In Examples 6–8 we'll examine the extent of nicotine addiction for female and male teenage smokers and we'll compare nicotine addiction for teenagers who smoke with teenagers who have smoked in the past but no longer do so.

Example 2

Aspirin, the Wonder Drug

Picture the Scenario

Here are two recent titles of newspaper articles about beneficial effects of aspirin:

> "Small doses of aspirin can lower the risk of heart attacks"
> "Aspirin could lower risk of colon cancer"

Most of us think of aspirin as a simple pill that helps relieve pain. In recent years, though, researchers have been on the lookout for new ways that aspirin may be helpful. Studies have shown that taking aspirin regularly may possibly forestall Alzheimer's disease and may increase the chance of survival for a person who has suffered a heart attack. Other studies have suggested that aspirin may protect against cancers of the pancreas, colon, and prostate. In the past decade, a growing number of studies have addressed the use of aspirin-like drugs to prevent cancer.[1]

Increasing attention has focused on aspirin since a landmark five-year study (Physicians Health Study Research Group, Harvard Medical School) about whether regular aspirin intake reduces deaths from heart disease. Studies have shown that treatment with daily aspirin for five years or longer reduces risk of colorectal cancer. These studies suggest that aspirin might reduce the risk of other cancers as well. Results of a recent meta-analysis combined the results of eight related studies with a minimum duration of treatment of four years or longer to determine the effects of aspirin on the risk of cancer death. A **meta-analysis** combines the results of several studies that address a set of related statistical questions. After analyzing the individual studies, the researchers assumed the different studies were measuring the same effect and pooled the results of the different studies. All experimental trials used were randomized and double-blind. The combined results provided evidence that daily aspirin reduced deaths due to several common cancers during and after the trials. We will explore some of these results.

Table 10.1 shows the study results. This is a **contingency table**, a data summary for categorical variables introduced in Section 3.1. Of the 25,570 individuals studied, 347 of those in the control group died of cancer, while 327 in the aspirin treatment died of cancer within 20 years following the study.

Table 10.1 Whether or Not Subject Died of Cancer, for Placebo and Aspirin Treatment Groups

	Death from Cancer		
Group	Yes	No	Total
Placebo	347	11,188	11,535
Aspirin	327	13,708	14,035

Questions to Explore

a. What is the response variable, and what are the groups to compare?

b. What are the two population parameters to compare? Estimate the difference between them using the data in Table 10.1.

Think It Through

a. In Table 10.1, the response variable is whether or not the subjects died of cancer, with categories yes and no. Group 1 is the subjects who took placebo and Group 2 is the subjects who took aspirin. They are the categories of the explanatory variable.

b. For the population from which this sample was taken, the proportion who died of cancer is represented by p_1 for taking placebo and p_2 for taking aspirin. The sample proportions of death from cancer were

$$\hat{p}_1 = 347/11535 = 0.030$$

for the $n_1 = 11{,}535$ in the placebo group and

$$\hat{p}_2 = 327/14035 = 0.023$$

for the $n_2 = 14{,}035$ in the aspirin group. Since $(\hat{p}_1 - \hat{p}_2) = 0.030 - 0.023 = 0.007$, the proportion of those who died of cancer was 0.007 higher for those who took placebo. In percentage terms, the difference was 3.0% − 2.3% = 0.7%, less than 1 percent.

Insight

The sample proportion of subjects who died of cancer was smaller for the aspirin group. But we really want to know if this result is true also for the population. To make an inference about the difference of population proportions, $(p_1 - p_2)$, we need to learn how much the difference $(\hat{p}_1 - \hat{p}_2)$ between the sample proportions would tend to vary from study to study. This is described by the standard error of the sampling distribution for the difference between the sample proportions.

Try Exercises 10.2 and 10.3, part a

Example 3

Cancer Death Rates for Aspirin and Placebo

Picture the Scenario

In Example 2, the sample proportions that died of cancer were $\hat{p}_1 = 347/11535 = 0.030$ for placebo (Group 1) and $\hat{p}_2 = 327/14035 = 0.023$ for aspirin (Group 2). The estimated difference was $\hat{p}_1 - \hat{p}_2 = 0.030 - 0.023 = 0.007$.

Questions to Explore

a. What is the standard error of this estimate?

b. How should we interpret this standard error?

Think It Through

a. Using the standard error formula given above with the values just recalled,

$$se = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

$$\sqrt{\frac{0.030(1-0.030)}{11535} + \frac{0.023(1-0.023)}{14035}} = 0.002.$$

b. Consider all the possible random samples of size about 11,535 for the placebo group and 14,035 for the aspirin group that could participate in this four-year study with follow-up. The difference $(\hat{p}_1 - \hat{p}_2)$ between the sample proportions of cancer deaths would not always equal 0.007 but would vary from sample to sample. The standard deviation of the $\hat{p}_1 - \hat{p}_2$ values for the different possible samples would equal about 0.002.

Insight

From the *se* formula, we see that *se* decreases as n_1 and n_2 increase. The standard error is very small for these data because the sample sizes were so large. This means that the $(\hat{p}_1 - \hat{p}_2)$ values would be very similar from study to study. It also implies that $(\hat{p}_1 - \hat{p}_2) = 0.007$ is quite precise as an estimate of the difference of population proportions.

Try Exercise 10.3, part b

Example 4

Comparing Cancer Death Rates for Aspirin and Placebo

Picture the Scenario

In the aspirin and cancer study, the estimated difference between placebo and aspirin in the proportions dying of cancer was $\hat{p}_1 - \hat{p}_2 = 0.003 - 0.023 = 0.007$. In Example 3 we found that this estimate has a standard error of 0.002.

Question to Explore

What can we say about the difference of population proportions of cancer deaths for those taking placebo versus those taking aspirin? Construct a 95% confidence interval for $(p_1 - p_2)$, and interpret.

Think It Through

From Table 10.1, shown again in the margin, the four outcome counts (347, 327, 11188, and 13708) were at least 10 for each group, so the large-samples confidence interval method is valid. A 95% confidence interval for $(p_1 - p_2)$ is

$$(\hat{p}_1 - \hat{p}_2) + 1.96(se), \text{ or } 0.007 + 1.96(0.002),$$
$$\text{which is } 0.007 + 0.004, \text{ or } (0.003, 0.011).$$

Suppose this experiment could be conducted with the entire population. The inference at the 95% confidence level that $(p_1 - p_2)$ is between 0.003 and 0.011 means that the population proportion p_1 of cancer deaths for those taking placebo would be between 0.003 higher and 0.011 higher than the population proportion p_2 of cancer deaths for those taking aspirin. Since both endpoints of the confidence interval (0.003, 0.011) for $(p_1 - p_2)$ are positive, we infer that $(p_1 - p_2)$ is positive. This means that p_1 is larger than p_2: The population proportion of cancer deaths is larger when subjects take the placebo than when they take aspirin. Table 10.2 shows how MINITAB reports this result and the margin shows screen shots from the TI-83+/84.

Table 10.2 MINITAB Output for Confidence Interval Comparing Proportions

```
     Number of Cancer Deaths          Observed p̂
             ↓                            ↓
Sample     X       N           Sample p
  1       347    11535         0.030082
  2       327    14035         0.023299
Difference = p(1) - p(2)
Estimate for difference: 0.00678346  ←  This is (p̂1 − p̂2)
95% CI for difference: (0.00279030, 0.0107766)
```

Insight

All the numbers in the confidence interval fall near 0. This suggests that the population difference $(p_1 - p_2)$ is small. However, this difference, small as it is, may be important in public health terms. For instance, projected over a population of 200 million adults (as in the United States or in Western Europe), a decrease over a twenty-year period of 0.01 in the proportion of people dying from cancer would mean two million fewer people dying from cancer.

This cancer study provided some of the first evidence that aspirin reduces deaths from several common cancers. Benefit was consistent across the different trial populations from the different randomized studies, suggesting that the findings have broader scope of generalization. However, it is important to replicate the study to see if results are similar or different for populations used in this study and other populations. Also, because there were fewer women than men in the study, findings about the effect of aspirin use and cancers related to women (such as breast cancer) were limited.

This example shows how the use of statistics can result in conclusions that benefit public health. An article[2] about proper and improper scientific methodology stated, "The most important discovery of modern medicine is not vaccines or antibiotics, it is the randomized double-blind study, by means of which we know what works and what doesn't."

Try Exercise 10.6

Example 5

TV Watching and Aggressive Behavior

Picture the Scenario

A study[3] considered whether greater levels of television watching by teenagers were associated with a greater likelihood of aggressive behavior. The researchers randomly sampled 707 families in two counties in northern New York state and made follow-up observations over 17 years. Table 10.3 shows results about whether a sampled teenager later conducted any aggressive act against another person, according to a self-report by that person or by his or her parent.

[3] By J.G. Johnson et al., *Science*, vol. 295, March 29, 2002, pp. 2468–2471.

	Aggressive Act		
TV Watching	Yes	No	Total
Less than 1 hour per day	5	83	88
At least 1 hour per day	154	465	619

We'll identify Group 1 as those who watched less than 1 hour of TV per day, on average, as teenagers. Group 2 consists of those who averaged at least 1 hour of TV per day, as teenagers. Denote the population proportion committing aggressive acts by p_1 for the lower level of TV watching and by p_2 for the higher level of TV watching.

Questions to Explore

a. Find and interpret the P-value for testing $H_0: p_1 = p_2$ against $H_a: p_1 \neq p_2$.

b. Make a decision about H_0 using the significance level of 0.05.

Think It Through

a. The study used random sampling and then classified teenagers by the level of TV watching, so the samples were independent random samples. Each count in Table 10.3 is at least five, so we can use a large-sample test. The sample proportions of aggressive acts were $\hat{p}_1 = 5/88 = 0.057$ for the lower level of TV watching and $\hat{p}_2 = 154/619 = 0.249$ for the higher level. Under the null hypothesis presumption that $p_1 = p_2$, the pooled estimate of the common value p is $\hat{p} = (5 + 154)/(88 + 619) = 159/707 = 0.225$.

The standard error for the test is

$$se_0 = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(0.225)(0.775)\left(\frac{1}{88} + \frac{1}{619}\right)} = 0.0476.$$

The test statistic for $H_0: p_1 = p_2$ is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{se_0} = \frac{(0.057 - 0.249) - 0}{0.0476} = \frac{-0.192}{0.0476} = -4.04.$$

For the two-sided alternative hypothesis, the P-value is the two-tail probability from the standard normal distribution. A z-score of -4.04 is far out in the left tail. See the margin figure. From tables (such as Table A) or software, it has a P-value $= 2(0.000027) = 0.000054$, or 0.0001 rounded to four decimal places. Extremely strong evidence exists against the null hypothesis that the population proportions committing aggressive acts are the same for the two levels of TV watching. The study provides strong evidence in support of H_a.

b. Since the P-value is less than 0.05, we can reject H_0. We support $H_a: p_1 \neq p_2$ and conclude that the population proportions of aggressive acts differ for the two groups. The sample values suggest that the population proportion is higher for the higher level of TV watching. The final row of Table 10.4 shows how MINITAB reports this result and the margin of the next page shows screen shots from the TI-83+/84.

Table 10.4 MINITAB Output for Example 5 on TV Watching and Aggression

```
Sample        X          N          Sample p
   1          5         88          0.056818

   2        154        619          0.248788

 Difference = p(1) - p(2)
 Estimate for difference: - 0.191970
 95% CI for difference: (-0.251124, -0.132816)
 Test for difference = 0(vs not = 0):z = - 4.04 P - Value = 0.000
```

Insight

This was an observational study. In practice, it would be impossible to conduct an experimental study by randomly assigning teenagers to watch little TV or watch much TV over several years. Also, just because a person watches more TV does not imply they watch more violence. We must be cautious of effects of lurking variables when we make conclusions. It is not proper to conclude that greater levels of TV watching *cause* later aggressive behavior. For instance, perhaps those who watched more TV had lower education levels, and perhaps lower education levels are associated with a greater likelihood of aggressive acts.

Try Exercise 10.8

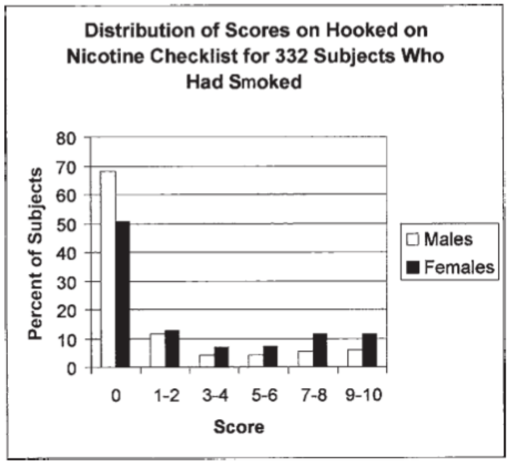
Example 6

Teenagers on Nicotine

Picture the Scenario

A recent study evaluated how addicted teenagers become to nicotine once they start smoking.[5] The 30-month study involved a random sample of 679 seventh-graders from two Massachusetts cities. Of this sample, 332 students who had ever smoked were the subjects. The response variable was constructed from a questionnaire called the Hooked on Nicotine Checklist (HONC). This is a list of ten questions such as "Have you ever tried to quit but couldn't?" and "Is it hard to keep from smoking in places where it is banned, like school?" The HONC score is the total number of questions to which a student answered yes during the study. Each student's HONC score falls between 0 and 10. The higher the score, the more hooked that student is on nicotine.

The study considered explanatory variables, such as gender, that might be associated with the HONC score. Figure 10.4, taken from the journal article about this study, shows the sample data distributions of the HONC scores for females and males who had ever smoked. Table 10.5 reports some descriptive statistics.



▲ **Figure 10.4** Sample Data Distribution of Hooked on Nicotine Checklist (HONC) Scores for Teenagers Who Have Smoked. Question Which group seems to have greater nicotine dependence, females or males? Explain your choice using the graph.

Table 10.5 Summary of Hooked on Nicotine Checklist (HONC) Scores, by Gender

		HONC Score	
Group	Sample Size	Mean	Standard Deviation
Females	150	2.8	3.6
Males	182	1.6	2.9

Question to Explore

How can we compare the sample HONC scores for females and males?

Think It Through

From Figure 10.4, a substantial proportion of teenagers who had ever smoked show no nicotine dependence. For both females and males the sample data distribution is highly skewed to the right. Because of the skew, we could use the median to summarize the sample HONC scores, but that has limited usefulness for such highly discrete data: The median is 0 both for boys and girls. We will estimate the population means and the difference between them. For Table 10.5, let's identify females as Group 1 and males as Group 2. Then $\bar{x}_1 = 2.8$ and $\bar{x}_2 = 1.6$. We estimate $(\mu_1 - \mu_2)$ by $(\bar{x}_1 - \bar{x}_2) = 2.8 - 1.6 = 1.2$. On average, females answered yes to about one more question on the HONC scale than males did.

Insight

This analysis uses descriptive statistics only. We'll next see how to make inferences about the difference between population means. You can conduct an inferential comparison of HONC scores by gender in Exercise 10.25. How do you think the nonnormality of HONC distributions affects inference about the population means?

Example 7

Nicotine Dependence

Picture the Scenario

Another explanatory variable in the teenage smoking study was whether a subject was still a smoker when the study ended. The study had 75 smokers and 257 ex-smokers at the end of the study. The HONC means describing nicotine addiction were 5.9 ($s = 3.3$) for the smokers and 1.0 ($s = 2.3$) for the ex-smokers.

Questions to Explore

a. Compare smokers and ex-smokers on their mean HONC scores.

b. What is the standard error of the difference in sample mean HONC scores? How do you interpret that *se*?

Think It Through

a. Let's regard the smokers as Group 1 and ex-smokers as Group 2. Then, $\bar{x}_1 = 5.9$ and $\bar{x}_2 = 1.0$. Since $(\bar{x}_1 - \bar{x}_2) = 5.9 - 1.0 = 4.9$, on average, smokers answered yes to nearly five more questions than ex-smokers did on the 10-question HONC scale. That's a large sample difference.

b. Applying the formula for the *se* of $(\bar{x}_1 - \bar{x}_2)$ to these data,

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(3.3)^2}{75} + \frac{(2.3)^2}{257}} = 0.41.$$

This describes the variability of the sampling distribution of $(\bar{x}_1 - \bar{x}_2)$, as shown in the margin figure. If other random samples of this size were taken from the study population, the difference between the sample means would vary from study to study. The standard deviation of the values for $(\bar{x}_1 - \bar{x}_2)$ would equal about 0.41.

Insight

This standard error will be used in confidence intervals and in significance tests for comparing means.

Try Exercise 10.16, part b

Example 8

Nicotine Addiction

Picture the Scenario

For the teenage smoking study, let's make an inference comparing the population mean nicotine addiction (as summarized by the HONC score) for smokers and ex-smokers. From Example 7, $\bar{x}_1 = 5.9$ with $s_1 = 3.3$ for $n_1 = 75$ smokers, and $\bar{x}_2 = 1.0$ with $s_2 = 2.3$ for $n_2 = 257$ ex-smokers.

Questions to Explore

a. Were the HONC sample data distributions for smokers and ex-smokers approximately normal? How does this affect inference?

b. Software reports a 95% confidence interval of (4.1, 5.7) for the difference between the population mean HONC score for smokers and ex-smokers. Show how it obtained this confidence interval and interpret it.

Think It Through

a. For the ex-smokers (Group 2), the values $\bar{x}_2 = 1.0$ and $s_2 = 2.3$ suggest that the HONC distribution is far from bell-shaped because the lowest possible HONC score of 0 is less than 1 standard deviation below the mean. This is not problematic. With large samples, the confidence interval method does not require normal population distributions because the central limit theorem implies that the sampling distribution is approximately normal. (Recall that even for small samples, the method is robust when population distributions are not normal.)

b. For these sample sizes and s_1 and s_2 values, software reports that $df = 95$. The *t*-score for a 95% confidence interval is $t_{.025} = 1.985$. (As an approximation, Table B reports the value $t_{.025} = 1.984$ for $df = 100$. Also, the smaller of $(n_1 - 1)$ and $(n_2 - 1)$ is 74, so if you did not have software to find *df*, you could use the *t*-score with $df = 74$.)

Denote the population mean HONC score by μ_1 for smokers and μ_2 for ex-smokers. From Example 7, $(\bar{x}_1 - \bar{x}_2) = 4.9$ has a standard error of $se = 0.41$. The 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$(\bar{x}_1 - \bar{x}_2) + 1.985(se), \text{ or } 4.9 + 1.985(0.41),$$

$$\text{which equals } 4.9 + 0.8, \text{ or } (4.1, 5.7).$$

We can be 95% confident that plausible values for $(\mu_1 - \mu_2)$, the differex-smokers, falls between 4.1 and 5.7. We can infer that the population mean for the smokers is between 4.1 higher and 5.7 higher than for the ex-smokers. The margin shows screen shots from the TI-83+/84.

Insight

We are 95% confident that the smokers answer yes to about 4 to 6 more of the 10 questions, on average, than the ex-smokers. For the 10-point HONC scale, this is quite a substantial difference.

Try Exercise 10.23

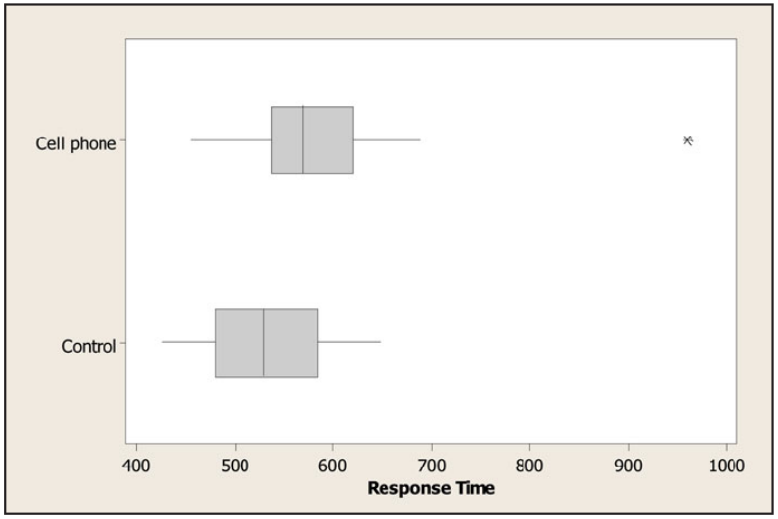
Example 9

Cell Phone Use While Driving Reaction Times

Picture the Scenario

An experiment[7] investigated whether cell phone use impairs drivers' reaction times, using a sample of 64 students from the University of Utah. Students were randomly assigned to a cell phone group or to a control group, 32 to each. On a simulation of driving situations, a target flashed red or green at irregular periods. Participants pressed a brake button as soon as they detected a red light. The control group listened to radio or books-on-tape while they performed the simulated driving. The cell phone group carried out a phone conversation about a political issue with someone in a separate room.

The experiment measured each subject's mean response time over many trials. Averaged over all trials and subjects, the mean response time was 585.2 milliseconds (a bit over half a second) for the cell phone group and 533.7 milliseconds for the control group. Figure 10.6 shows box plots of the responses for the two groups.



▲ **Figure 10.6** MINITAB Box Plots of Response Times for Cell Phone Study. Question Does either box plot show any irregularities that could affect the analysis?

Denote the population mean response time by μ_1 for the cell phone group and by μ_2 for the control group. Table 10.6 shows how MINITAB reports inferential comparisons of those two means. The margin on the next page contains screen shots from the TI-83+/84.

Questions to Explore

a. Show how MINITAB got the test statistic for testing $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 \neq \mu_2$.

b. Report and interpret the P-value, and state the decision you would make about the population using a 0.05 significance level.

```
                N          Mean        StDev
Cell
phone           32         585.2       89.6
Control         32         533.7       65.3
Difference = mu (Cell phone) − mu (Control)
Estimate for difference: 51.5172
95% CI for difference: (12.2393, 90.7951)
T-Test of difference = 0 (vs not = ):
T-Value = 2.63 P-Value = 0.011 DF = 56
```

c. What do the box plots tell us about the suitability of these analyses? What effect does the outlier for the cell phone group have on the analysis?

Think It Through

a. We estimate the difference $\mu_1 - \mu_2$ by $(\bar{x}_1 - \bar{x}_2) = 585.2 - 533.7 = 51.5$, shown in Table 10.6. The standard error of this estimate is

$$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{(89.6)^2}{32} + \frac{(65.3)^2}{32}} = 19.6.$$

The test statistic for H_0: $\mu_1 = \mu_2$ equals

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{se} = \frac{51.5}{19.6} = 2.63.$$

b. The P-value is the two-tail probability from a t distribution. Table 10.6 reports $df = 56$ and a P-value $= 0.01$. If H_0 were true, the probability would be 0.01 of getting a t test statistic this large or even larger in either tail. The P-value is less than 0.05, so we can reject H_0. Suppose the entire population used a cell phone in this experiment, or the entire population did not use a cell phone. A population mean applies in each case. Then we have enough evidence to conclude that the population mean response times would differ between the cell phone and control groups. The sample means suggest that the population mean is higher for the cell phone group.

c. The t inferences assume normal population distributions. The box plots do not show any substantial skew, but there is an extreme outlier for the cell phone group. One subject in that group had a very slow mean reaction time. Because this observation is so far from the others in that group, it's a good idea to make sure the results of the analysis aren't affected too strongly by that single observation.

If we delete the extreme outlier for the cell phone group, software reports

```
                    N         Mean        StDev
Cell phone          31        573.1       58.9
Control             32        533.7       65.3
Estimate for difference: 39.4265
95% CI for difference: (8.1007, 70.7522)
T-Test of difference = 0 (vs not = ):
T-Value = 2.52 P-Value = 0.015 DF = 60
```

The mean and standard deviation for the cell phone group now decrease substantially. However, the t test statistic is not much different, and the P-value is still small, 0.015, leading to the same conclusion.

Insight

Even though the difference between the sample means decreased from 51.5 to 39.4 when we deleted the outlier, the standard error also got smaller (you can check that it equals 15.7) because of the smaller standard deviation for the cell phone group after removing the outlier. That's why the t test statistic did not change much. In practice, you should not delete outliers from a data set without sufficient cause (for example, if it seems the observation was incorrectly recorded). However, it's a good idea to check for *sensitivity* of an analysis to an outlier, as we did here, by repeating the analysis without it. If the results change much, it means that the inference including the outlier is on shaky ground.

Try Exercise 10.25

Example 10

Arthroscopic Surgery

Picture the Scenario

A random trial study assessed the usefulness of arthroscopic surgery.[8] Over a three-year period, patients suffering from osteoarthritis who had at least moderate knee pain were recruited from a medical center in Houston. Patients were randomly assigned to one of three groups, to receive one of two types of arthroscopic surgeries or a surgery that was actually a placebo procedure. In the arthroscopic surgeries, the lavage group had the joint flushed with fluid but instruments were not used to remove tissue, whereas the debridement group had tissue removal as well. In the placebo procedure, the same incisions were made in the knee as with surgery and the surgeon manipulated the knee as if surgery was being performed, but none was actually done. The study was double-blind.

A knee-specific pain scale was created for the study. Administered two years after the surgery, it ranged from 0 to 100, with higher scores indicating more severe pain. Table 10.7 shows summary statistics. The pain scale was the response variable, and the treatment group was the explanatory variable.

Table 10.7 Summary of Knee Pain Scores

The descriptive statistics compare lavage and debridement arthroscopic surgery to a placebo (fake surgery) treatment.

	Knee Pain Score		
Group	**Sample Size**	**Mean**	**Standard Deviation**
1. Placebo	60	51.6	23.7
2. Arthroscopic—lavage	61	53.7	23.7
3. Arthroscopic—debridement	59	51.4	23.2

Denote the population mean of the pain scores by μ_1 for the placebo group and by μ_2 for the lavage arthroscopic group. Most software gives you the option of assuming equal population standard deviations. Table 10.8 is the MINITAB output for the two-sample t inferences. (We'll consider the debridement arthroscopic group in an exercise.)

```
Sample     N      Mean    StDev     SE Mean

   1       60     51.6     23.7        3.1

   2       61     53.7     23.7        3.0

Difference = mu(1)   mu(2)

Estimate for difference:  − 2.10000

95% CI for difference: (−10.63272,6.43272)

T-Test of difference = 0 (vs not = ):

T-Value =  − 0.49  P-Value = 0.627DF = 119

Both use Pooled StDev = 23.7000
```

Questions to Explore

a. Does the t test inference in Table 10.8 seem appropriate for these data?

b. Show how to find the pooled standard deviation estimate of σ, the standard error, the test statistic, and its *df*.

c. Identify the P-value for testing $H_0: \mu_1 = \mu_2$ against $H_a: \mu_1 \neq \mu_2$. With the 0.05 significance level, can you reject H_0? Interpret.

Think It Through

a. If the arthroscopic surgery has no real effect, its population distribution of pain score should be the same as for the placebo. Not only will the population means be equal, but so will the population standard deviations. So, in testing whether this surgery has no effect, we will use a method that assumes equal population standard deviations. Table 10.8 shows that the sample standard deviations are identical. This is mere coincidence, and we would not typically expect this even if the population standard deviations were identical. But it suggests that we can use the method assuming equal population standard deviations.

b. The pooled standard deviation estimate of the common value σ of $\sigma_1 = \sigma_2$ is

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(60 - 1)23.7^2 + (61 - 1)23.7^2}{60 + 61 - 2}} = 23.7,$$

shown at the bottom of Table 10.8. We estimate the difference $\mu_1 - \mu_2$ by $(\bar{x}_1 - \bar{x}_2) = 51.6 - 53.7 = -2.1$. The standard error of $(\bar{x}_1 - \bar{x}_2)$ equals

$$se = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 23.7\sqrt{\frac{1}{60} + \frac{1}{61}} = 4.31.$$

The test statistic for $H_0: \mu_1 = \mu_2$ equals

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{se} = \frac{51.6 - 53.7}{4.31} = -0.49.$$

Its $df = (n_1 + n_2 - 2) = (60 + 61 - 2) = 119$. When H_0 is true, the test statistic has the t distribution with $df = 119$.

c. The two-sided P-value equals 0.63. So, if the true population means were equal, by random variation, it would not be surprising to observe a test statistic of this size. The P-value is larger than 0.05, so we cannot reject H_0. Consider the population of people who suffer from osteoarthritis and could conceivably receive one of these treatments. In summary, we don't have enough evidence to conclude that the mean pain level differs for the placebo treatment and the arthroscopic surgery treatment.

Insight

Table 10.8 reports a 95% confidence interval for $(\mu_1 - \mu_2)$ of $(-10.6, 6.4)$. Because the interval contains 0, it is plausible that there is no difference between the population means. This is the same conclusion as that for the significance test. We infer that the population mean for the knee pain score could be as much as 10.6 lower or as much as 6.4 higher for the placebo treatment than for the arthroscopic surgery. On the pain scale with range 100, this is a relatively small difference in practical terms.

Try Exercise 10.36

Example 11

Alcohol Consumption and Risk of Stroke

Picture the Scenario

A recent article in a medical journal[10] stated, "Compared with participants who had less than one drink per week, those who drank more had a reduced overall risk of stroke (relative risk, 0.79; 95% confidence interval, 0.66 to 0.94)."

Question to Explore

How do you interpret the relative risk value and the reported confidence interval?

Think It Through

Let $\hat{p}_1$ = sample proportion of the regular alcohol drinkers who had strokes and $\hat{p}_2$ = sample proportion of the light or nondrinkers who had strokes. Then the sample relative risk of 0.79 means that $\hat{p}_1/\hat{p}_2 = 0.79$. So $\hat{p}_1 = 0.79\hat{p}_2$. The proportion of the regular alcohol drinkers who had strokes was 0.79 times the proportion of the light or nondrinkers who had strokes. Since the relative risk was less than 1.0, the proportion of strokes was smaller for the regular alcohol drinkers. (How can you determine which is Group 1 and which is Group 2 for the ratio $\hat{p}_1/\hat{p}_2$? The quoted sentence says that those who drank more had a *reduced* risk of stroke. Since the relative risk of 0.79 is less then 1.0, this means that the group that drank more is in the numerator for this calculation, so it is Group 1.)

The 95% confidence interval for the population relative risk of stroke was (0.66, 0.94). Since all numbers in the interval are less than 1.0, we can infer that $p_1/p_2 < 1.0$ (see the margin figure). That is, it seems that $p_1 < p_2$. We can infer that the population proportion of strokes is smaller for those who drink alcohol regularly.

Insight

Regular alcohol drinking seems to correspond to a reduction in the incidence of stroke. Since the upper endpoint of the confidence interval is 0.94, however, the population relative risk could be close to 1.0. So the effect of drinking might be weak. Another reason to be cautious is that this was an observational study, rather than a randomized clinical trial like in the study comparing aspirin and placebo. In any case, the conclusion does not imply that it's good to drink a lot of alcohol every day.

Try Exercise 10.41

Example 12

Cell Phones and Driving

Picture the Scenario

In this chapter, Example 9 analyzed whether the use of cell phones impairs reaction times in a driving skills test. The analysis used independent samples—one group used cell phones and a separate control group did not use them. An alternative design uses the same subjects for both groups. Reaction times are measured when subjects performed the driving task without using cell phones and then again while the same subjects used cell phones.

Table 10.9 shows the mean of the reaction times (in milliseconds) for each subject under each condition. Figure 10.8 shows box plots of the data for the two conditions.

Table 10.9 Reaction Times on Driving Skills Before and While Using Cell Phone

The difference score is the reaction time using the cell phone minus the reaction time not using it, such as $636 - 604 = 32$ milliseconds.

	Using Cell Phone?				Using Cell Phone?		
Student	No	Yes	Difference	Student	No	Yes	Difference
1	604	636	32	17	525	626	101
2	556	623	67	18	508	501	−7
3	540	615	75	19	529	574	45
4	522	672	150	20	470	468	−2
5	459	601	142	21	512	578	66
6	544	600	56	22	487	560	73
7	513	542	29	23	515	525	10
8	470	554	84	24	499	647	148
9	556	543	−13	25	448	456	8
10	531	520	−11	26	558	688	130
11	599	609	10	27	589	679	90
12	537	559	22	28	814	960	146
13	619	595	−24	29	519	558	39
14	536	565	29	30	462	482	20
15	554	573	19	31	521	527	6
16	467	554	87	32	543	536	−7

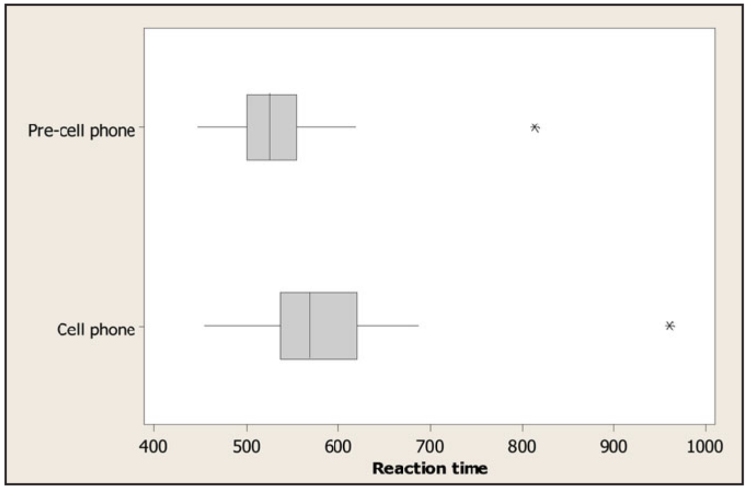
Questions to Explore

a. Summarize the sample data distributions for the two conditions.

b. To compare the mean response times using statistical inference, should we treat the samples as independent or dependent?

Think It Through

a. The box plots show that the reaction times tend to be larger for the cell phone condition. But each sample data distribution has one extremely large outlier.



▲ **Figure 10.8** MINITAB Box Plots of Observations on Reaction Times. Question From the data file or Table 10.9, what do the outliers for the two distributions have in common?

b. The pre–cell phone (no cell phone) responses in Table 10.9 were made by the same subjects as the responses using cell phones. These are matched-pairs data because each control observation (Sample 1) pairs with a cell phone observation (Sample 2). Because this part of the study used the same subjects for each sample, the samples are dependent.

Insight

The data file and Table 10.9 show that subject number 28 had a large reaction time in each case. This one subject had a very slow reaction time, regardless of the condition.

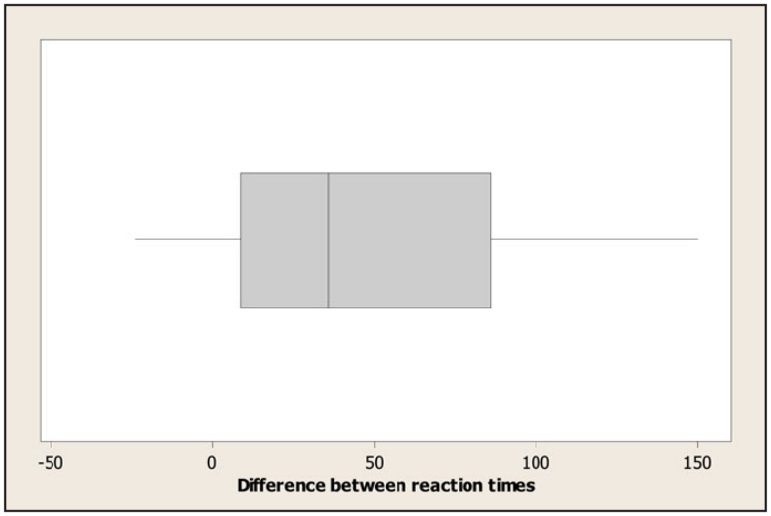
Try Exercise 10.47, part a

Example 13

Cell Phones and Driver Reaction Time

Picture the Scenario

The matched-pairs data in Table 10.9 showed the reaction times for the sampled subjects using and not using cell phones. Figure 10.9 is a box plot of the $n = 32$ difference scores. Table 10.10 shows MINITAB output for these data. The first part of Table 10.10 shows the mean, standard deviation, and standard error for each sample and for the differences. The second part shows inference about the mean of the differences.



▲ **Figure 10.9** MINITAB Box Plot of Difference Scores from Table 10.9. Question How is it that some of the scores plotted here are negative?

Table 10.10 Software Output for Matched-Pairs Analysis With Table 10.9

The next page shows screen shots from the TI-83+/84.

```
Paired T for Cell phone - Pre-cell phone

                   N     Mean     StDev    SE Mean

Cell phone        32   585.188   89.646    15.847

Pre-cell phone    32   534.563   66.448    11.746

Difference        32   50.6250  52.4858    9.2783

95% CI for mean difference: (31.7019, 69.5481)

T-Test of mean difference = 0 (vs not = 0):

T-Value = 5.46 P-Value = 0.000
```

Questions to Explore

a. How can you conduct and interpret the significance test reported in Table 10.10?

b. How can you construct and interpret the confidence interval reported in Table 10.10?

Think It Through

a. The box plot shows skew to the right for the difference scores. Two-sided inference is robust to violations of the assumption of a normal population distribution. The box plot does not show any severe outliers that would raise questions about the validity of using the mean to summarize the difference scores.

The sample mean difference is $\bar{x}_d = 50.6$, and the standard deviation of the difference scores is $s_d = 52.5$. The standard error is $se = s_d/\sqrt{n} = 52.5/\sqrt{32} = 9.28$. The t test statistic for the significance test of H_0: $\mu_d = 0$ (and hence equal population means for the two conditions) against H_a: $\mu_d \neq 0$ is

$$t = \bar{x}_d/se = 50.6/9.28 = 5.46.$$

With 32 difference scores, $df = n - 1 = 31$. Table 10.10 reports the two-sided P-value of 0.000. There is extremely strong evidence that the population mean reaction times are different.

b. For a 95% confidence interval for $\mu_d = \mu_1 - \mu_2$, with $df = 31$, $t_{.025} = 2.040$. We can use $se = 9.28$ from part a. The confidence interval equals

$$\bar{x}_d + t_{.025}(se), \text{ or } 50.6 + 2.040(9.28),$$
$$\text{which equals } 50.6 + 18.9, \text{ or } (31.7, 69.5).$$

At the 95% confidence level, we infer that the population mean when using cell phones is between about 32 and 70 milliseconds higher than when not using cell phones. The confidence interval does not contain 0, so we can infer that the population mean reaction time is greater when using a cell phone. The confidence interval *is* more informative than the significance test because it predicts just how large the difference might be.

The article about the study did not indicate whether the subjects were randomly selected, so inferential conclusions are tentative.

Insight

The study also showed that reaction times were similar with hands-free versus hand-held cell phones. It also showed that the probability of missing a simulated traffic signal doubled when students used cell phones.

A later related study[13] showed that the mean reaction time of students using cell phones was similar to that of elderly drivers not using cell phones. The study concluded, "The net effect of having younger drivers converse on a cell phone was to make their average reactions equivalent to those of older drivers who were not using a cell phone."

Try Exercises 10.47 and 10.48

Example 14

Beliefs in Heaven and Hell

Picture the Scenario

A recent General Social Survey asked subjects whether they believed in heaven and whether they believed in hell. For the 1314 subjects who responded, Table 10.11 shows the data in contingency table form. The rows of Table 10.11 are the response categories for belief in heaven. The columns are the same categories for belief in hell. In the U.S. adult population, let p_1 denote the proportion who believe in heaven and let p_2 denote the proportion who believe in hell.

Table 10.11 Beliefs in Heaven and Hell

	Belief in Hell		
Belief in Heaven	**Yes**	**No**	**Total**
Yes	955	162	**1117**
No	9	188	**197**
Total	**964**	**350**	**1314**

Questions to Explore

a. How can we estimate p_1, p_2, and their difference?

b. Are the samples used to estimate p_1 and p_2 independent or dependent, samples?

c. How can we use a sample mean of difference scores to estimate $(p_1 - p_2)$?

Think It Through

a. The counts in the two margins of Table 10.11 summarize the responses. Of the 1314 subjects, 1117 said they believed in heaven, so $\hat{p}_1 = 1117/1314 = 0.85$. Of the same 1314 subjects, 964 said they believed in hell, so $\hat{p}_2 = 964/1314 = 0.73$. Since $(\hat{p}_1 - \hat{p}_2) = .85 - .73 = 0.12$, we estimate that 12% more people believe in heaven than in hell.

b. Each sample of 1314 responses refers to the same 1314 subjects. Any given subject's response on belief in heaven can be matched with that subject's response on belief in hell. So the samples for these two proportions are dependent.

c. Recall that a proportion is a mean when we code the responses by 1 and 0. For belief in heaven or belief in hell, let 1 = yes and 0 = no. Then, for belief in heaven, 1117 responses were 1 and 197 responses were 0. The mean of the 1314 observations was $[1117(1) + 197(0)]/1314 = 1117/1314 = 0.85$, which is $\hat{p}_1$. For belief in hell, 964 responses were 1 and 350 responses were 0, and the mean was $\hat{p}_2 = 964/1314 = 0.73$.

Table 10.12 shows the possible paired binary responses, for the 0 and 1 coding. The difference scores for the heaven and hell responses are $d = 1, 0$, or -1, as shown. The sample mean of the 1314 difference scores equals

$$[0(955) + 1(162) - 1(9) + 0(188)]/1314 = 153/1314 = 0.12.$$

This equals the difference of proportions, $\hat{p}_1 - \hat{p}_2 = 0.85 - 0.73 = 0.12$. In summary, the mean of the difference scores equals the difference between the sample proportions.

Table 10.12 Using 0 and 1 Responses for Binary Matched-Pairs Data

Heaven	Hell	Interpretation	Difference d	Frequency
1	1	Believe in heaven and in hell	$1 - 1 = 0$	955
1	0	Believe in heaven but not in hell	$1 - 0 = 1$	162
0	1	Believe in hell but not in heaven	$0 - 1 = -1$	9
0	0	Do not believe in heaven or hell	$0 - 0 = 0$	188

Insight

We've converted two samples with 1314 binary observations each to a single sample of 1314 difference scores. We can use single-sample methods with the differences, as we just did for matched-pairs analysis of means.

Try Exercise 10.58

Example 15

Beliefs in Heaven and Hell

Picture the Scenario

We continue our analysis of the data on belief in heaven and/or hell. When software analyzes the 1314 paired difference scores, we get the results in Table 10.13:

Table 10.13 Software Output for Analyzing Difference Scores from Table 10.12 to Compare Beliefs in Heaven and Hell

```
                N        Mean       StDev      SE Mean
Difference    1314      0.1218     0.5706      0.0157
95% CI for mean difference: (0.091, 0.153)
```

Questions to Explore

a. Explain how software got the confidence interval from the other results shown.

b. Interpret the reported 95% confidence interval.

Think It Through

a. We've already seen that the sample mean of the 1314 difference scores is $(\hat{p}_1 - \hat{p}_2) = 0.12$. Table 10.13 reports that the standard error of the sample mean difference is $se = 0.0157$. For $n = 1314$, $df = 1313$, and $t_{.025} = 1.962$. The population mean difference is $(p_1 - p_2)$, the difference between the population proportion p_1 who believe in heaven and the population proportion p_2 who believe in hell. A 95% confidence interval for $(p_1 - p_2)$ equals

Sample mean difference $\pm\ t_{.025}(se)$, which is
$0.12 \pm 1.962(0.0157)$, or $(0.09, 0.15)$.

b. We can be 95% confident that the population proportion p_1 believing in heaven is between 0.09 higher and 0.15 higher than the population proportion p_2 believing in hell. Since the interval contains only positive values, we infer that $p_1 > p_2$.

Insight

As in the matched-pairs inferences comparing two means, we conducted the inference comparing two proportions by using inference for a single parameter—namely, the population mean of the differences.

Try Exercise 10.59

Example 16

Speech Recognition Systems

Picture the Scenario

Research in comparing the quality of different speech recognition systems uses a series of isolated words as a benchmark test, finding for each system the proportion of words for which an error of recognition occurs. Table 10.14 shows data from one of the first articles[14] that showed how to conduct such a test. The article compared speech recognition systems called generalized minimal distortion segmentation (GMDS) and continuous density hidden Markov model (CDHMM). Table 10.14 shows the counts of the four possible sequences to test outcomes for the two systems with a given word, for a test using 2000 words.

Table 10.14 Results of Test Using 2000 Words to Compare Two Speech Recognition Systems (GMDS and CDHMM)

	CDHMM		
GMDS	**Correct**	**Incorrect**	**Total**
Correct	1921	58	**1979**
Incorrect	16	5	**21**
Total	**1937**	**63**	**2000**

Question to Explore

Conduct McNemar's test of the null hypothesis that the probability of a correct outcome is the same for each system.

Think It Through

The article about this test did not indicate how the words were chosen. Inferences are valid if the 2000 words were a random sample of the possible words on which the systems could have been tested. Let p_1 denote the population proportion of correct results for GMDS and let p_2 denote the population proportion correct for CDHMM. The test statistic for McNemar's test of H_0: $p_1 = p_2$ against H_a: $p_1 \neq p_2$ is

$$z = \frac{58 - 16}{\sqrt{58 + 16}} = 4.88.$$

The two-sided P-value is 0.000001. There is extremely strong evidence against the null hypothesis that the correct detection rates are the same for the two systems.

Insight

We learn more by estimating parameters. A confidence interval for $p_1 - p_2$ indicates that the GMDS system is better, but the difference in correct detection rates is small. See Exercise 10.56.

Try Exercise 10.57

Example 17

Death Penalty and Race

Picture the Scenario

The United States is one of only a few Western nations that still imposes the death penalty. Are those convicted of murder more likely to get the death penalty if they are black than if they are white?

Table 10.16 comes from one of the first studies on racial inequities of the death penalty.[15] The 326 subjects were defendants convicted of homicide in Florida murder trials. The variables are the defendant's race and whether the defendant received the death penalty. The contingency table shows that about 12% of white defendants and about 10% of black defendants received the death penalty.

	Death Penalty			
Defendant's Race	**Yes**	**No**	**Total**	**Percentage Yes**
White	19	141	160	11.9
Black	17	149	166	10.2

In this study, the difference between the percentages of white defendants and black defendants who received the death penalty is small (1.7%), but the percentage was lower for black defendants.

Is there some explanation for these results? Does a control variable lurk that explains why relatively fewer black defendants got the death penalty in Florida? Researchers who study the death penalty stress that the *victim's* race is often an important factor. So, let's control for victim's race. We'll construct a table like Table 10.16 *separately* for cases in which the victim was white and for cases in which the victim was black. Table 10.17 shows this three-variable table.

Table 10.17 Defendant's Race and Death Penalty Verdict, Controlling for Victim's Race

		Death Penalty			
Victim's Race	**Defendant's Race**	**Yes**	**No**	**Total**	**Percentage Yes**
White	White	19	132	151	12.6
	Black	11	52	63	17.5
Black	White	0	9	9	0.0
	Black	6	97	103	5.8

is 5.8 − 0.0 = 5.8. In summary, controlling for victim's race, more black defendants than white defendants received the death penalty in Florida.

b. Table 10.16 showed that a *larger* percentage of white defendants than black defendants got the death penalty. By contrast, controlling for victim's race, Table 10.17 showed that a *smaller* percentage of white defendants than black defendants got the death penalty. This was true for each victim's race category.

c. In Table 10.17, look at the percentages who got the death penalty. When the victim was white, they are quite a bit larger (12.6 and 17.5) than when the victim was black (0.0 and 5.8). That is, defendants who killed a white person were more likely to get the death penalty. Now look at the totals for the four combinations of victim's race and defendant's race. The most common cases are white defendants having white victims (151 times) and black defendants having black victims (103 times). In summary, white defendants usually had white victims and black defendants usually had black victims. Killing a white person was more likely to result in the death penalty than killing a black person. These two factors operating together produce an overall association that shows (in Table 10.16) that a higher percentage of white defendants than black defendants got the death penalty.

Insight

Overall, relatively more white defendants in Florida got the death penalty than black defendants. Controlling for victim's race, however, relatively more black defendants got the death penalty. The effect changes direction after we control for victim's race. This shows that the association at each level of a control variable can have a different direction than overall when that third variable is ignored instead of controlled. This phenomenon is called **Simpson's paradox**.

In Table 10.17, the death penalty was imposed most often when a black defendant had a white victim. By contrast, it was never imposed when a white defendant had a black victim. Similar results have occurred in other studies of the death penalty. See Exercises 10.64 and 3.58.

Try Exercise 10.64