Chapter 10

Introduction 610

Section 10.1 612
Comparing Two Proportions

Section 10.2 634
Comparing Two Means

Free Response AP® Problem, YAY! 662

Chapter 10 Review 662

Chapter 10 Review Exercises 664

Chapter 10 AP® Statistics Practice Test 666

Cumulative AP® Practice Test 3 669

Comparing Two Populations or Groups

case study

Fast-Food Frenzy!

More than $70 billion is spent each year in the drive-thru lanes of America's fast-food restaurants. Having quick, accurate, and friendly service at a drive-thru window translates directly into revenue for the restaurant. According to Jack Greenberg, former CEO of McDonald's, sales increase 1% for every six seconds saved at the drive-thru. So industry executives, stockholders, and analysts closely follow the ratings of fast-food drive-thru lanes that appear annually in *QSR*, a publication that reports on the quick-service restaurant industry.

The 2012 *QSR* magazine drive-thru study involved visits to a random sample of restaurants in the 20 largest fast-food chains in all 50 states. During each visit, the researcher ordered a modified main item (for example, a hamburger with no pickles), a side item, and a drink. If any item was not received as ordered, or if the restaurant failed to give the correct change or supply a straw and a napkin, then the order was considered "inaccurate." Service time, which is the time from when the car stopped at the speaker to when the entire order was received, was measured each visit. Researchers also recorded whether or not each restaurant had an order-confirmation board in its drive-thru.[1]

Here are some results from the 2012 *QSR* study:

- For restaurants with order-confirmation boards, 1169 of 1327 visits (88.1%) resulted in accurate orders. For restaurants with no order-confirmation board, 655 of 726 visits (90.2%) resulted in accurate orders.
- McDonald's average service time for 362 drive-thru visits was 188.83 seconds with a standard deviation of 17.38 seconds. Burger King's service time for 318 drive-thru visits had a mean of 201.33 seconds and a standard deviation of 18.85 seconds.

Was there a significant difference in accuracy at restaurants with and without order-confirmation boards? How much better was the average service time at McDonald's than at Burger King restaurants in 2012? By the end of the chapter, you should have acquired the tools to help answer questions like these.

Introduction

Which of two popular drugs—Lipitor or Pravachol—helps lower "bad cholesterol" more? Researchers designed an experiment, called the PROVE-IT Study, to find out. They used about 4000 people with heart disease as subjects. These individuals were randomly assigned to one of two treatment groups: Lipitor or Pravachol. At the end of the study, researchers compared the proportion of subjects in each group who died, had a heart attack, or suffered other serious consequences within two years. For those using Pravachol, the proportion was 0.263; for those using Lipitor, it was 0.224.[2] Could such a difference have occurred purely by the chance involved in the random assignment? This is a question about *comparing two proportions*.

Who studies more in college—men or women? Researchers asked separate random samples of 30 males and 30 females at a large university how many minutes they studied on a typical weeknight. The females reported studying an average of 165.17 minutes; the male average was 117.17 minutes. How large is the difference in the corresponding population means? This is a question about *comparing two means*.

Comparing two proportions or means based on random sampling or a randomized experiment is one of the most common situations encountered in statistical practice. In the PROVE-IT experiment, the goal of inference is to determine whether the treatments (Lipitor and Pravachol) *caused* the observed difference in the proportion of subjects who experienced serious consequences in the two groups. For the college studying survey, the goal of inference is to draw a conclusion about the actual mean study times for *all* women and *all* men at the university.

The following Activity gives you a taste of what lies ahead in this chapter.

ACTIVITY Is Yawning Contagious?

MATERIALS:

Set of 50 index cards or standard deck of playing cards for each pair of students

According to the popular TV show *Mythbusters*, the answer is "Yes." The *Mythbusters* team conducted an experiment involving 50 subjects. Each subject was placed in a booth for an extended period of time and monitored by hidden camera. Thirty-four subjects were given a "yawn seed" by one of the experimenters; that is, the experimenter yawned in the subject's presence before leaving the room. The remaining 16 subjects were given no yawn seed.

What happened in the experiment? The table below shows the results:[3]

	Yawn Seed?		
Subject Yawned?	**Yes**	**No**	**Total**
Yes	10	4	**14**
No	24	12	**36**
Total	**34**	**16**	**50**

Ten of the 34 subjects (29.4%) in the yawn-seed group yawned, compared to 4 of the 16 subjects (25.0%) in the no-yawn-seed group. The difference in the proportions who yawned for the two groups is 10/34 − 4/16 = 0.044. Adam Savage and Jamie Hyneman, the cohosts of *MythBusters*, used this difference as evidence that yawning is contagious. But is the evidence *convincing*?

In this Activity, your class will investigate whether the results of the experiment were really statistically significant. Let's see what would happen just by chance if we randomly reassign the 50 people in this experiment to the two groups (yawn seed and no yawn seed) many times, *assuming the treatment received doesn't affect whether or not a person yawns*.

1. We need 50 cards to represent the subjects in this study. In the *MythBusters* experiment, 14 people yawned and 36 didn't. Because we're assuming that the treatment received won't change whether each subject yawns, we use 14 cards to represent people who yawn and 36 cards to represent those who don't.

- *Using index cards:* Write "Yes" on 14 cards and "No" on 36 cards.
- *Using playing cards:* Remove the ace of spades and ace of clubs from the deck. All jacks, queens, kings, and aces represent subjects who yawn. All remaining cards represent subjects who don't yawn.

2. Shuffle and deal two piles of cards—one with 34 cards and one with 16 cards. The first pile represents the yawn-seed group and the second pile represents the no-yawn-seed group. The shuffling reflects our assumption that the outcome for each subject is not affected by the treatment.

Calculate the difference in the proportions who yawned for the two groups (yawn seed – no yawn seed). For example, if you get 9 yawners in the yawn-seed group and 5 yawners in the no-yawn-seed group, the resulting difference in proportions is

$$\frac{9}{34} - \frac{5}{16} = -0.048$$

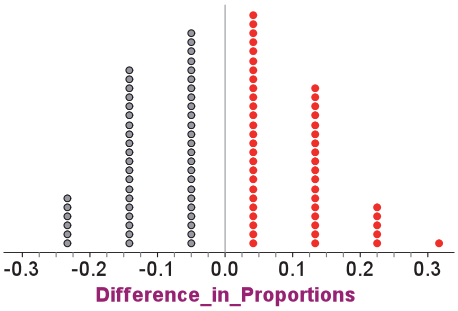
A negative difference would mean that a smaller proportion of people in the yawn-seed group yawned during the experiment than in the no-yawn-seed group.

3. Your teacher will draw and label axes for a class dotplot. Plot the result you got in Step 2 on the graph.

4. Repeat Steps 2 and 3 if needed to get a total of at least 40 repetitions of the simulation for your class.

5. Based on the class's simulation results, how surprising would it be to get a difference in proportions of 0.044 (what the *Mythbusters* got in their experiment) or larger simply due to the chance involved in the random assignment?

6. What conclusion would you draw about whether yawning is contagious? Explain.



Here is an example of what the class dotplot in the Activity might look like after 100 trials. In this simulation, 50 of the 100 trials (in red) produced a difference in proportions of at least 0.044, so the approximate *P*-value is 0.50. It is very likely that a difference this big could occur just due to the chance variation in random assignment! This result is not statistically significant and does not provide convincing evidence that yawning is contagious.

10.1 Comparing Two Proportions

WHAT YOU WILL LEARN **By the end of the section, you should be able to:**

- Describe the shape, center, and spread of the sampling distribution of $\hat{p}_1 - \hat{p}_2$.
- Determine whether the conditions are met for doing inference about $p_1 - p_2$.
- Construct and interpret a confidence interval to compare two proportions.
- Perform a significance test to compare two proportions.

Suppose we want to compare the proportions of individuals with a certain characteristic in Population 1 and Population 2. Let's call these parameters of interest p_1 and p_2. The ideal strategy is to take a separate random sample from each population and to compare the sample proportions $\hat{p}_1$ and $\hat{p}_2$ with that characteristic.

What if we want to compare the effectiveness of Treatment 1 and Treatment 2 in a completely randomized experiment? This time, the parameters p_1 and p_2 that we want to compare are the true proportions of successful outcomes for each treatment. We use the proportions of successes in the two treatment groups, $\hat{p}_1$ and $\hat{p}_2$, to make the comparison.

Here's a table that summarizes these two situations:

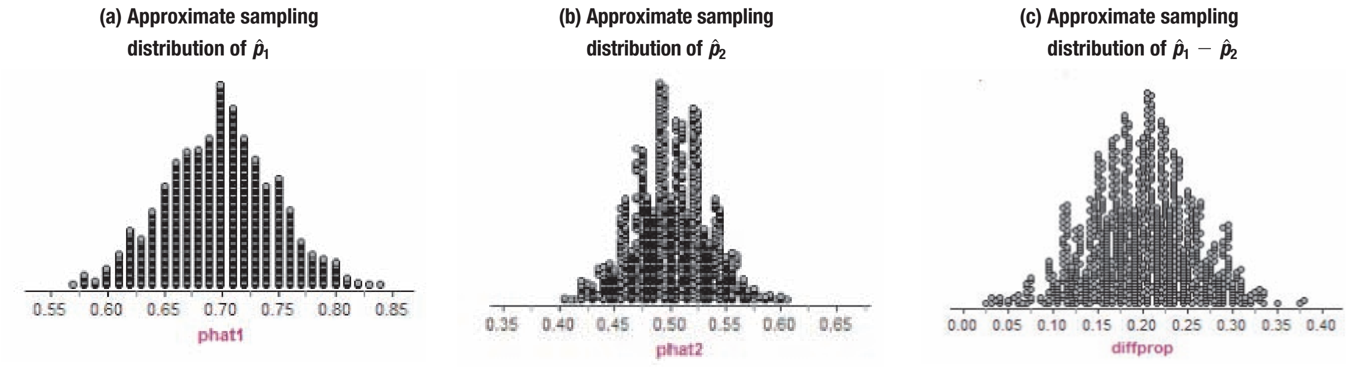
Population or treatment	Parameter	Statistic	Sample size
1	p_1	$\hat{p}_1$	n_1
2	p_2	$\hat{p}_2$	n_2

We compare the populations or treatments by doing inference about the difference $p_1 - p_2$ between the parameters. The statistic that estimates this difference is the difference between the two sample proportions, $\hat{p}_1 - \hat{p}_2$. To use $\hat{p}_1 - \hat{p}_2$ for inference, we must know its sampling distribution.

The Sampling Distribution of a Difference between Two Proportions

To explore the sampling distribution of $\hat{p}_1 - \hat{p}_2$, let's start with two populations having a known proportion of successes. Suppose that there are two large high schools, each with over 2000 students, in a certain town. At School 1, 70% of students did their homework last night. Only 50% of the students at School 2 did their homework last night. The counselor at School 1 takes an SRS of 100 students and records the proportion $\hat{p}_1$ that did the homework. School 2's counselor takes an SRS of 200 students and records the proportion $\hat{p}_2$ that did the homework. What can we say about the difference $\hat{p}_1 - \hat{p}_2$ in the sample proportions?

We used Fathom software to take an SRS of $n_1 = 100$ students from School 1 and a separate SRS of $n_2 = 200$ students from School 2 and to plot the values of $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_1 - \hat{p}_2$ from each sample. Our first set of simulated samples gave $\hat{p}_1 = 0.68$ and $\hat{p}_2 = 0.505$, so dots were placed above each of those values in Figure 10.1(a) and (b).


(a) Approximate sampling distribution of $\hat{p}_1$

(b) Approximate sampling distribution of $\hat{p}_2$

(c) Approximate sampling distribution of $\hat{p}_1 - \hat{p}_2$

FIGURE 10.1 Simulated sampling distributions of (a) the sample proportion $\hat{p}_1$ of successes in 1000 SRSs of size $n_1 = 100$ from a population with $p_1 = 0.70$, (b) the sample proportion $\hat{p}_2$ of successes in 1000 SRSs of size $n_2 = 200$ from a population with $p_2 = 0.50$, and (c) the difference in sample proportions $\hat{p}_1 - \hat{p}_2$ for each of the 1000 repetitions.

The difference in the sample proportions for this first set of samples is $\hat{p}_1 - \hat{p}_2 = 0.68 - 0.505 = 0.175$. A dot for this value appears in Figure 10.1(c). The three dotplots in Figure 10.1 show the results of repeating this process 1000 times. These are the approximate sampling distributions of $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}_1 - \hat{p}_2$.

In Chapter 7, we saw that the sampling distribution of a sample proportion $\hat{p}$ has the following properties:

Shape: Approximately Normal if $np \geq 10$ and $n(1 - p) \geq 10$

Center: $\mu_{\hat{p}} = p$

Spread: $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$ if $n \leq \dfrac{1}{10}N$

For the sampling distributions of $\hat{p}_1$ and $\hat{p}_2$ in this case:

	Sampling distribution of $\hat{p}_1$	Sampling distribution of $\hat{p}_2$
Shape	Approximately Normal; $n_1p_1 = 100(0.70) = 70 \geq 10$ and $n_1(1 - p_1) = 100(0.30) = 30 \geq 10$	Approximately Normal; $n_2p_2 = 200(0.50) = 100 \geq 10$ and $n_2(1 - p_2) = 200(0.50) = 100 \geq 10$
Center	$\mu_{\hat{p}_1} = p_1 = 0.70$	$\mu_{\hat{p}_2} = p_2 = 0.50$
Spread	$\sigma_{\hat{p}_1} = \sqrt{\dfrac{p_1(1-p_1)}{n_1}} = \sqrt{\dfrac{0.7(0.3)}{100}} = 0.0458$ because School 1 has a population of over 10(100) = 1000 students.	$\sigma_{\hat{p}_2} = \sqrt{\dfrac{p_2(1-p_2)}{n_2}} = \sqrt{\dfrac{0.5(0.5)}{200}} = 0.0354$ because School 2 has a population of over 10(200) = 2000 students.

The approximate sampling distributions in Figures 10.1(a) and (b) give similar results.

What about the sampling distribution of $\hat{p}_1 - \hat{p}_2$? Figure 10.1(c) suggests that it has an approximately Normal shape, is centered at about 0.198, and has standard deviation about 0.0572. The shape makes sense because we are combining two independent random variables, $\hat{p}_1$ and $\hat{p}_2$, that have approximately Normal distributions. How about the center? The true proportion of students who did last night's homework at School 1 is $p_1 = 0.70$ and at School 2 is $p_2 = 0.50$. We expect the difference $\hat{p}_1 - \hat{p}_2$ to center on the actual difference in the population proportions, $p_1 - p_2 = 0.70 - 0.50 = 0.20$. The spread, however, is a bit more complicated.

THINK ABOUT IT

How can we find formulas for the mean and standard deviation of the sampling distribution of $\hat{p}_1 - \hat{p}_2$? Both $\hat{p}_1$ and $\hat{p}_2$ are random variables. That is, their values would vary in repeated independent SRSs of size n_1 and n_2. Independent random samples yield independent random variables $\hat{p}_1$ and $\hat{p}_2$. The statistic $\hat{p}_1 - \hat{p}_2$ is the difference of these two independent random variables.

In Chapter 6, we learned that for any two random variables X and Y,

$$\mu_{X-Y} = \mu_X - \mu_Y$$

For the random variables $\hat{p}_1$ and $\hat{p}_2$, we have

$$\mu_{\hat{p}_1-\hat{p}_2} = \mu_{\hat{p}_1} - \mu_{\hat{p}_2} = p_1 - p_2$$

In the school homework survey,

$$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2 = 0.70 - 0.50 = 0.20$$

We also learned in Chapter 6 that for *independent* random variables X and Y,

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

For the random variables $\hat{p}_1$ and $\hat{p}_2$, we have

$$\sigma^2_{\hat{p}_1-\hat{p}_2} = \sigma^2_{\hat{p}_1} + \sigma^2_{\hat{p}_2} = \left(\sqrt{\frac{p_1(1-p_1)}{n_1}}\right)^2 + \left(\sqrt{\frac{p_2(1-p_2)}{n_2}}\right)^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

So $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$.

In the school homework survey,

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.7(0.3)}{100} + \frac{0.5(0.5)}{200}} = 0.058$$

This is similar to the result from the Fathom simulation.

Here are the facts we need.

THE SAMPLING DISTRIBUTION OF $\hat{p}_1 - \hat{p}_2$

Choose an SRS of size n_1 from Population 1 with proportion of successes p_1 and an independent SRS of size n_2 from Population 2 with proportion of successes p_2.

- **Shape:** When n_1p_1, $n_1(1-p_1)$, n_2p_2, and $n_2(1-p_2)$ are all at least 10, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal.
- **Center:** The mean of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2$.
- **Spread:** The standard deviation of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

as long as each sample is no more than 10% of its population.

When conditions are met, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ will be approximately Normal with mean $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$ and standard deviation $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$. Figure 10.2 displays this distribution.

The formula for the standard deviation of the sampling distribution involves the unknown parameters p_1 and p_2. Just as in Chapters 8 and 9, we must replace these by estimates to do inference. And just as before, we do this a bit differently for confidence intervals and for tests. We'll get to inference shortly. For now, let's focus on the sampling distribution of $\hat{p}_1 - \hat{p}_2$.

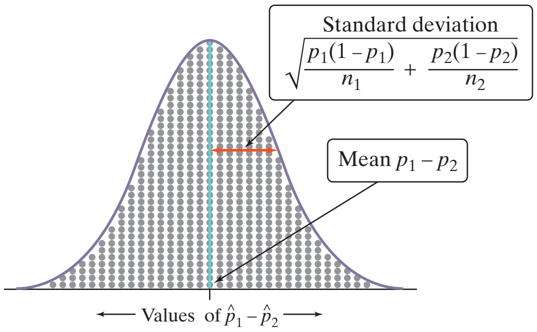

Standard deviation
$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$
Mean $p_1 - p_2$
Values of $\hat{p}_1 - \hat{p}_2$

FIGURE 10.2 Select independent SRSs from two populations having proportions of successes p_1 and p_2. The proportions of successes in the two samples are $\hat{p}_1$ and $\hat{p}_2$. When the samples are large, the sampling distribution of the difference $\hat{p}_1 - \hat{p}_2$ is approximately Normal.

EXAMPLE Yummy Goldfish!

Describing the sampling distribution of $\hat{p}_1 - \hat{p}_2$

Your teacher brings two bags of colored goldfish crackers to class. Bag 1 has 25% red crackers and Bag 2 has 35% red crackers. Each bag contains more than 1000 crackers. Using a paper cup, your teacher takes an SRS of 50 crackers from Bag 1 and a separate SRS of 40 crackers from Bag 2. Let $\hat{p}_1 - \hat{p}_2$ be the difference in the sample proportions of red crackers.

PROBLEM:

(a) What is the shape of the sampling distribution of $\hat{p}_1 - \hat{p}_2$? Why?

(b) Find the mean of the sampling distribution. Show your work.

(c) Find the standard deviation of the sampling distribution. Show your work.

SOLUTION:

(a) Because $n_1p_1 = 50(0.25) = 12.5$, $n_1(1 - p_1) = 50(0.75) = 37.5$, $n_2p_2 = 40(0.35) = 14$, and $n_2(1 - p_2) = 40(0.65) = 26$ are all at least 10, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal.

(b) The mean is $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2 = 0.25 - 0.35 = -0.10$.

(c) Because there are at least $10(50) = 500$ crackers in Bag 1 and $10(40) = 400$ crackers in Bag 2, the standard deviation is

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} = \sqrt{\frac{0.25(0.75)}{50} + \frac{0.35(0.65)}{40}} = 0.0971$$

For Practice *Try Exercise* 1

Confidence Intervals for $p_1 - p_2$

When data come from two independent random samples or two groups in a randomized experiment (the Random condition), the statistic $\hat{p}_1 - \hat{p}_2$ is our best guess for the value of $p_1 - p_2$. We can use our familiar formula to calculate a confidence interval for $p_1 - p_2$:

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

When the 10% condition is met, the standard deviation of the statistic $\hat{p}_1 - \hat{p}_2$ is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

If the Large Counts condition is met, we find the critical value z^* for the given confidence level from the standard Normal curve.

CONDITIONS FOR CONSTRUCTING A CONFIDENCE INTERVAL ABOUT A DIFFERENCE IN PROPORTIONS

- **Random:** The data come from two independent random samples or from two groups in a randomized experiment.
  - ○ **10%:** When sampling without replacement, check that $n_1 \leq \frac{1}{10}N_1$ and $n_2 \leq \frac{1}{10}N_2$.
- **Large Counts:** The counts of "successes" and "failures" in each sample or group—$n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, $n_2(1-\hat{p}_2)$—are all at least 10.

Because we don't know the values of the parameters p_1 and p_2, we replace them in the standard deviation formula with the sample proportions. The result is the **standard error** (also called the *estimated standard deviation*) of the statistic $\hat{p}_1 - \hat{p}_2$:

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

This value tells us how far the difference in sample proportions will typically be from the difference in population proportions if we repeat the random sampling or random assignment many times.

When the conditions are met, our confidence interval for $p_1 - p_2$ is therefore

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

$$(\hat{p}_1 - \hat{p}_2) \pm z^*\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

This is often called a **two-sample z interval for a difference between two proportions**.

TWO-SAMPLE *z* INTERVAL FOR A DIFFERENCE BETWEEN TWO PROPORTIONS

When the conditions are met, an approximate C% confidence interval for $\hat{p}_1 - \hat{p}_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where z^* is the critical value for the standard Normal curve with C% of its area between $-z^*$ and z^*.

The following example shows how to construct and interpret a confidence interval for a difference in proportions. As usual with inference problems, we follow the four-step process. Because you are expected to include these four steps whenever you construct a confidence interval or perform a significance test, *we will limit our use of the four-step icon to examples from this point forward*.

EXAMPLE Teens and Adults on Social Networking Sites

STEP 4

Confidence interval for $p_1 - p_2$

As part of the Pew Internet and American Life Project, researchers conducted two surveys in 2012. The first survey asked a random sample of 799 U.S. teens about their use of social media and the Internet. A second survey posed similar questions to a random sample of 2253 U.S. adults. In these two studies, 80% of teens and 69% of adults used social-networking sites.

PROBLEM:

(a) Calculate the standard error of the sampling distribution of the difference in the sample proportions (teens − adults). What information does this value provide?

(b) Construct and interpret a 95% confidence interval for the difference between the proportion of all U.S. teens and adults who use social-networking sites.

SOLUTION:

(a) The sample proportions of teens and adults who use social-networking sites are $\hat{p}_1 = 0.80$ and $\hat{p}_2 = 0.69$, respectively. The standard error of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.80(0.20)}{799} + \frac{0.69(0.31)}{2253}} = 0.0172$$

If we were to take many random samples of 799 teens and 2253 adults, the difference in the sample proportions of teens and adults who use social-networking sites will typically be 0.0172 from the true difference in proportions of all teens and adults who use social-networking sites.

(b) STATE: Our parameters of interest are p_1 = the proportion of all U.S. teens who use social-networking sites and p_2 = the proportion of all U.S. adults who use social-networking sites. We want to estimate the difference $p_1 - p_2$ at a 95% confidence level.

PLAN: We should use a two-sample z interval for $p_1 - p_2$ if the conditions are met.

- *Random:* The data come from independent random samples of 799 U.S. teens and 2253 U.S. adults.
  - 10%: The researchers are sampling without replacement, so we must check the 10% condition: there are at least 10(799) = 7990 U.S. teens and at least 10(2253) = 22,530 U.S. adults.
- *Large Counts:* We check the counts of "successes" and "failures":

$$n_1\hat{p}_1 = 799(0.80) = 639.2 \rightarrow 639 \qquad n_1(1-\hat{p}_1) = 799(1-0.80) = 159.8 \rightarrow 160$$
$$n_2\hat{p}_2 = 2253(0.69) = 1554.57 \rightarrow 1555 \qquad n_2(1-\hat{p}_2) = 2253(1-0.69) = 698.43 \rightarrow 698$$

Note that the observed counts have to be whole numbers! Because all four values are at least 10, this condition is met.

DO: We know that $n_1 = 799$, $\hat{p}_1 = 0.80$, $n_2 = 2253$, and $\hat{p}_2 = 0.69$. For a 95% confidence level, the critical value is $z^* = 1.96$. So our 95% confidence interval for $p_1 - p_2$ is

$$\begin{aligned}(\hat{p}_1 - \hat{p}_2) \pm z^*\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} &= (0.80 - 0.69) \pm 1.96\sqrt{\frac{0.80(0.20)}{799} + \frac{0.69(0.31)}{2253}} \\ &= 0.11 \pm 1.96(0.0172) \\ &= 0.11 \pm 0.034 \\ &= (0.076, 0.144)\end{aligned}$$

Using technology: Refer to the Technology Corner that follows the example. The calculator's `2-PropZInt` gives (0.07588, 0.14324).

This interval suggests that more teens than adults in the United States engage in social networking by between about 7.6 and 14.3 percentage points.

CONCLUDE: We are 95% confident that the interval from 0.07588 to 0.14324 captures the true difference in the proportion of all U.S. teens and adults who use social-networking sites.

For Practice *Try Exercise* **9**

The researchers in the previous example selected independent random samples from the two populations they wanted to compare. In practice, it's common to take one random sample that includes individuals from both populations of interest and then to separate the chosen individuals into two groups. The two-sample z procedures for comparing proportions are still valid in such situations, provided that the two groups can be viewed as independent samples from their respective populations of interest.

You can use technology to perform the calculations in the "Do" step. Remember that this comes with potential benefits and risks on the AP® exam.

21. TECHNOLOGY CORNER CONFIDENCE INTERVAL FOR A DIFFERENCE IN PROPORTIONS

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

The TI-83/84 and TI-89 can be used to construct a confidence interval for $p_1 - p_2$. We'll demonstrate using the previous example. Of $n_1 = 799$ teens surveyed, $X = 639$ said they used social-networking sites. Of $n_2 = 2253$ adults surveyed, $X = 1555$ said they engaged in social networking. To construct a confidence interval:

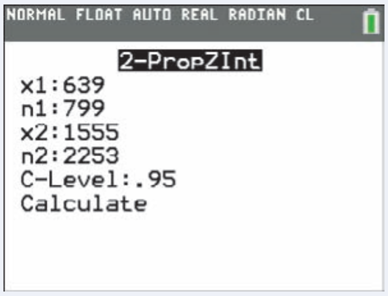
TI-83/84

- Press STAT, then choose TESTS and 2-PropZInt.

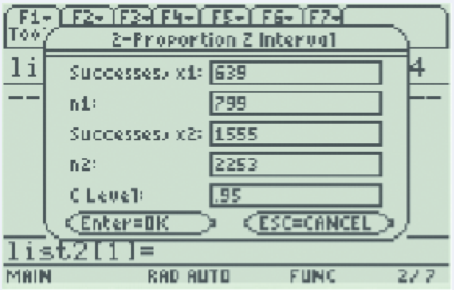
TI-89

- In the Statistics/List Editor, press 2nd F2 ([F7]) and choose 2-PropZInt.

- When the 2-PropZInt screen appears, enter the values shown.

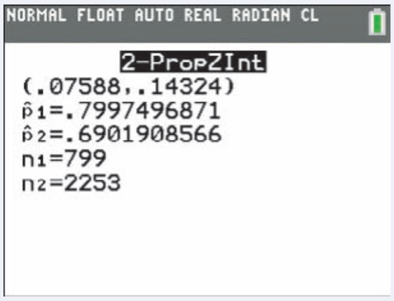


```
2-PropZInt
x1:639
n1:799
x2:1555
n2:2253
C-Level:.95
Calculate
```

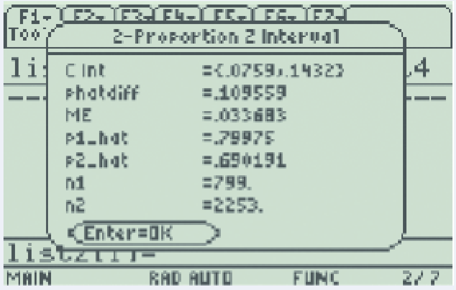


```
2-Proportion Z Interval
Successes, x1:  639
n1:             799
Successes, x2:  1555
n2:             2253
C Level:        .95
```

- Highlight "Calculate" and press ENTER.



```
2-PropZInt
(.07588,.14324)
p̂1=.7997496871
p̂2=.6901908566
n1=799
n2=2253
```



```
2-Proportion Z Interval
C Int      ={.0759,.14323}
phatdiff   =.109559
ME         =.033683
p1_hat     =.79975
p2_hat     =.690191
n1         =799.
n2         =2253.
```

AP® EXAM TIP The formula for the two-sample z interval for $p_1 - p_2$ often leads to calculation errors by students. As a result, we recommend using the calculator's 2-PropZInt feature to compute the confidence interval on the AP® exam. Be sure to name the procedure (two-proportion z interval) and to give the interval (0.076, 0.143) as part of the "Do" step.

CHECK YOUR UNDERSTANDING

Are teens or adults more likely to go online daily? The Pew Internet and American Life Project asked a random sample of 799 teens and a separate random sample of 2253 adults how often they use the Internet. In these two surveys, 63% of teens and 68% of adults said that they go online every day. Construct and interpret a 90% confidence interval for $p_1 - p_2$.

Significance Tests for $p_1 - p_2$

An observed difference between two sample proportions can reflect an actual difference in the parameters, or it may just be due to chance variation in random sampling or random assignment. Significance tests help us decide which explanation makes more sense.

The null hypothesis has the general form

$$H_0: p_1 - p_2 = \text{hypothesized value}$$

We'll restrict ourselves to situations in which the hypothesized difference is 0. Then the null hypothesis says that there is no difference between the two parameters:

$$H_0: p_1 - p_2 = 0 \text{ or, alternatively, } H_0: p_1 = p_2$$

The alternative hypothesis says what kind of difference we expect.

EXAMPLE Hungry Children

Stating hypotheses

Researchers designed a survey to compare the proportions of children who come to school without eating breakfast in two low-income elementary schools. An SRS of 80 students from School 1 found that 19 had not eaten breakfast. At School 2, an SRS of 150 students included 26 who had not had breakfast. More than 1500 students attend each school. Do these data give convincing evidence of a difference in the population proportions?

PROBLEM: State appropriate hypotheses for a significance test to answer this question. Define any parameters you use.

SOLUTION: We should carry out a test of

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 \neq 0$$

where p_1 = the true proportion of students at School 1 who did not eat breakfast and p_2 = the true proportion of students at School 2 who did not eat breakfast.

For Practice Try Exercise 13

The conditions for performing a significance test about $p_1 - p_2$ are the same as for constructing a confidence interval.

CONDITIONS FOR PERFORMING A SIGNIFICANCE TEST ABOUT A DIFFERENCE IN PROPORTIONS

- **Random:** The data come from two independent random samples or from two groups in a randomized experiment.
  - ○ **10%:** When sampling without replacement, check that $n_1 \leq \frac{1}{10}N_1$ and $n_2 \leq \frac{1}{10}N_2$.
- **Large Counts:** The counts of "successes" and "failures" in each sample or group—$n_1\hat{p}_1, n_1(1 - \hat{p}_1), n_2\hat{p}_2, n_2(1 - \hat{p}_2)$—are all at least 10.

If the conditions are met, we can proceed with calculations. To do a test, standardize $\hat{p}_1 - \hat{p}_2$ to get a z statistic:

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\text{standard deviation of statistic}}$$

If $H_0 : p_1 = p_2$ is true, the two parameters are the same. We call their common value p. But now we need a way to estimate p, so it makes sense to combine the data from the two samples as if they came from one larger sample. This **pooled** (or **combined**) **sample proportion** is

$$\hat{p}_C = \frac{\text{count of successes in both samples combined}}{\text{count of individuals in both samples combined}} = \frac{X_1 + X_2}{n_1 + n_2}$$

In other words, $\hat{p}_C$ gives the overall proportion of successes in the combined samples.

Let's look at how to calculate $\hat{p}_C$ in the hungry children example. The two-way table below summarizes the survey data. We have combined the independent SRSs from the two schools in the right-hand Total column.

	School		
Breakfast?	**1**	**2**	**Total**
No	19	26	**45**
Yes	61	124	**185**
Total	**80**	**150**	**230**

Because researchers want to compare the proportions of students at School 1 and School 2 who have not eaten breakfast, we treat the individuals in the "No" row as successes. It is easy to see from the table that the overall proportion of successes in the combined samples is $\hat{p}_C = \frac{45}{230} = 0.1957$. We can also get this result using the formula above:

$$\hat{p}_C = \frac{X_1 + X_2}{n_1 + n_2} = \frac{19 + 26}{80 + 150} = \frac{45}{230} = 0.1957$$

Recall that the standard deviation of $\hat{p}_1 - \hat{p}_2$ is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

Use $\hat{p}_C$ in place of both p_1 and p_2 in this expression for the denominator of the test statistic:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_C(1 - \hat{p}_C)}{n_1} + \frac{\hat{p}_C(1 - \hat{p}_C)}{n_2}}}$$

We can use a little algebra to rewrite the denominator of the test statistic:

$$\sqrt{\frac{\hat{p}_C(1 - \hat{p}_C)}{n_1} + \frac{\hat{p}_C(1 - \hat{p}_C)}{n_2}} =$$

$$\sqrt{\hat{p}_C(1 - \hat{p}_C)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} =$$

$$\sqrt{\hat{p}_C(1 - \hat{p}_C)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The final formula looks like the one given on the AP® exam formula sheet.

When the Large Counts condition is met, this will yield a z statistic that has approximately the standard Normal distribution when H_0 is true. Here are the details for the **two-sample z test for the difference between two proportions**.

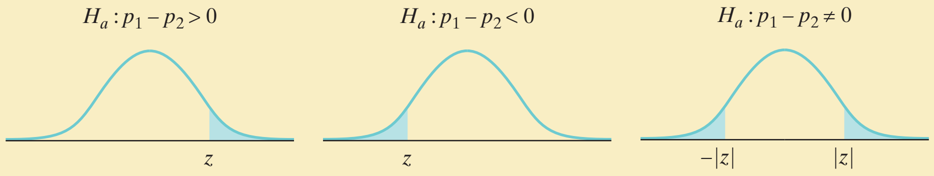
Some people prefer to use $\hat{p}_C$ to check the Large Counts condition. If the expected counts $n_1\hat{p}_C$, $n_1(1-\hat{p}_C)$, $n_2\hat{p}_C$, and $n_2(1-\hat{p}_C)$ are all at least 10, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal. Checking the observed counts of successes and failures is more conservative, as the expected counts will always be at least 10 if the observed counts are at least 10.

TWO-SAMPLE *z* TEST FOR THE DIFFERENCE BETWEEN TWO PROPORTIONS

Suppose the conditions are met. To test the hypothesis $H_0: p_1 - p_2 = 0$, first find the pooled proportion $\hat{p}_C$ of successes in both samples combined. Then compute the z statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\dfrac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \dfrac{\hat{p}_C(1-\hat{p}_C)}{n_2}}}$$

Find the *P*-value by calculating the probability of getting a z statistic this large or larger in the direction specified by the alternative hypothesis H_a:



Now we can finish the test we started earlier.

EXAMPLE Hungry Children

STEP 4

Significance test for $p_1 - p_2$

Researchers designed a survey to compare the proportions of children who come to school without eating breakfast in two low-income elementary schools. An SRS of 80 students from School 1 found that 19 had not eaten breakfast. At School 2, an SRS of 150 students included 26 who had not had breakfast. More than 1500 students attend each school. Do these data give convincing evidence at the $\alpha = 0.05$ level of a difference in the population proportions?

STATE: Our hypotheses are

$$H_0: p_1 - p_2 = 0$$
$$H_a: p_1 - p_2 \neq 0$$

where $p_1 =$ the true proportion of students at School 1 who did not eat breakfast and $p_2 =$ the true proportion of students at School 2 who did not eat breakfast.

PLAN: If conditions are met, we should perform a two-sample z test for $p_1 - p_2$.

- *Random*: The data were produced using two independent random samples—80 students from School 1 and 150 students from School 2.
  - 10%: The researchers are sampling without replacement, so we check the 10% condition: there are at least 10(80) = 800 students at School 1 and at least 10(150) = 1500 students at School 2.
- *Large Counts:* We check the counts of "successes" and "failures":

$$n_1\hat{p}_1 = 19,\ n_1(1-\hat{p}_1) = 61,\ n_2\hat{p}_2 = 26,\ n_2(1-\hat{p}_2) = 124$$

All four values are at least 10, so this condition is met.

	School		
Breakfast?	1	2	Total
No	19	26	45
Yes	61	124	185
Total	**80**	**150**	**230**

DO: We know that $n_1 = 80$, $\hat{p}_1 = \frac{19}{80} = 0.2375$, $n_2 = 150$, and $\hat{p}_2 = \frac{26}{150} = 0.1733$. Our point estimate for the difference in population proportions is $\hat{p}_1 - \hat{p}_2 = 0.2375 - 0.1733 = 0.0642$. The pooled proportion of students who didn't eat breakfast in the two samples is

$$\hat{p}_C = \frac{19 + 26}{80 + 150} = \frac{45}{230} = 0.1957$$

See the two-way table in the margin for confirmation.

- *Test statistic*

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}}} = \frac{0.0642 - 0}{\sqrt{\frac{0.1957(1-0.1957)}{80} + \frac{0.1957(1-0.1957)}{150}}} = 1.17$$

- *P-value* Figure 10.3 displays the *P*-value as an area under the standard Normal curve for this two-tailed test. Using Table A or `normalcdf`, the desired *P*-value is $2P(Z \geq 1.17) = 2(1 - 0.8790) = 0.2420$.

Using technology: Refer to the Technology Corner that follows the example. The calculator's `2-PropZTest` gives $z = 1.1683$ and *P*-value $= 0.2427$.

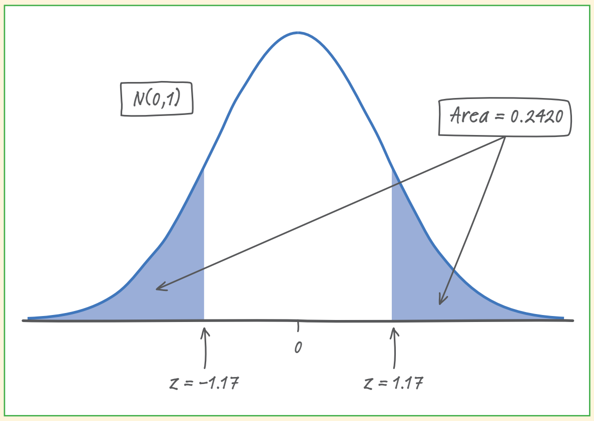


FIGURE 10.3 The *P*-value for the two-sided test.

CONCLUDE: Because our *P*-value, 0.2427, is greater than $\alpha = 0.05$, we fail to reject H_0. There is not convincing evidence that the true proportions of students at the two schools who didn't eat breakfast are different.

For Practice *Try Exercise* **15**

Exactly what does the *P*-value in the previous example tell us? If we repeated the random sampling process many times, we'd get a difference in sample proportions as large as or larger than 0.0642 in either direction about 24% of the time when H_0: $p_1 - p_2 = 0$ is true. With such a high probability of getting a result like this just by chance when the null hypothesis is true, we don't have enough evidence to reject H_0.

We can get additional information about the difference between the population proportions at School 1 and School 2 with a confidence interval. The TI-84's `2-PropZInt` gives the 95% confidence interval for $p_1 - p_2$ as $(-0.047, 0.175)$. That is, we are 95% confident that the difference in the true proportions of students who ate breakfast at the two schools is between 4.7 percentage points lower at School 1 and 17.5 percentage points higher at School 1. This is consistent with our "fail to reject H_0" conclusion in the example because 0 is included in the interval of plausible values for $p_1 - p_2$.

The two-sample *z* test and two-sample *z* interval for the difference between two proportions don't always give consistent results. That's because the "standard deviation of the statistic" used in calculating the test statistic is

$$\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}}$$

but for the confidence interval, it's

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

22. TECHNOLOGY CORNER

SIGNIFICANCE TEST FOR A DIFFERENCE IN PROPORTIONS

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

The TI-83/84 and TI-89 can be used to perform significance tests for comparing two proportions. Here, we use the data from the hungry children example. To perform a test of $H_0: p_1 - p_2 = 0$ versus $H_a: p_1 - p_2 \neq 0$:

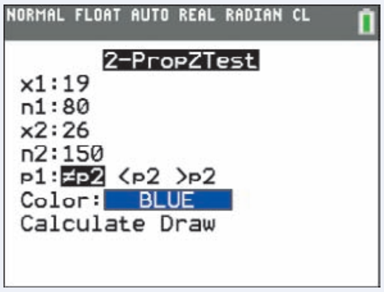
TI-83/84

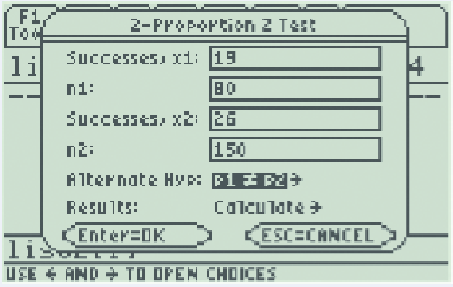
- Press `STAT`, then choose TESTS and `2-PropZTest`.

TI-89

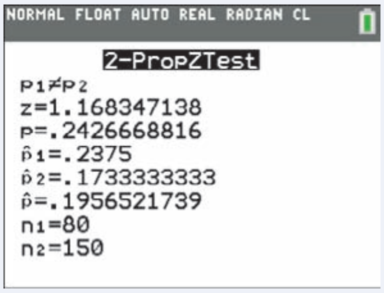
- In the Statistics/List Editor, press `2nd` `F1` ([F6]) and choose `2-PropZTest`.

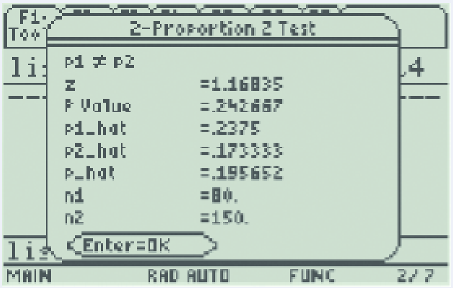
- When the `2-PropZTest` screen appears, enter the values $x_1 = 19$, $n_1 = 80$, $x_2 = 26$, $n_2 = 150$. Specify the alternative hypothesis $p_1 \neq p_2$, as shown.


```
NORMAL FLOAT AUTO REAL RADIAN CL
2-PropZTest
x1:19
n1:80
x2:26
n2:150
p1:≠p2 <p2 >p2
Color: BLUE
Calculate Draw
```

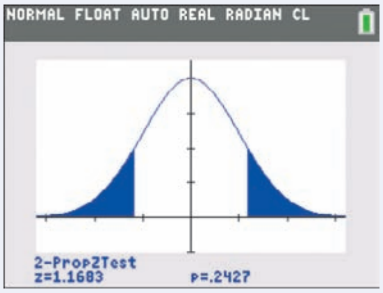

```
2-Proportion Z Test
Successes, x1:  19
n1:             80
Successes, x2:  26
n2:             150
Alternate Hyp:  p1 ≠ p2 →
Results:        Calculate →
<Enter=OK>      <ESC=CANCEL>
USE ← AND → TO OPEN CHOICES
```

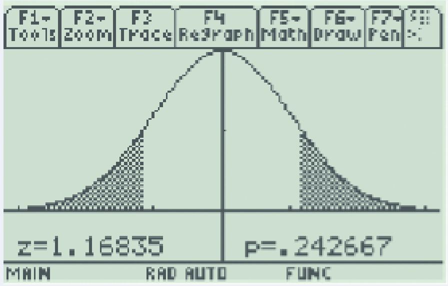
- If you select "Calculate" and press `ENTER`, you will see that the test statistic is $z = 1.168$ and the P-value is 0.2427. Do you see the combined proportion of students who didn't eat breakfast? It's the value labeled $\hat{p}$, 0.1957.


```
NORMAL FLOAT AUTO REAL RADIAN CL
2-PropZTest
p1≠p2
z=1.168347138
p=.2426668816
p̂1=.2375
p̂2=.1733333333
p̂=.1956521739
n1=80
n2=150
```


```
2-Proportion Z Test
p1 ≠ p2
z         =1.16835
P Value   =.242667
p1_hat    =.2375
p2_hat    =.173333
p_hat     =.195652
n1        =80.
n2        =150.
<Enter=OK>
MAIN      RAD AUTO      FUNC      2/7
```

- If you select the "Draw" option, you will see the screen shown here.


```
NORMAL FLOAT AUTO REAL RADIAN CL
2-PropZTest
z=1.1683        p=.2427
```



```
z=1.16835       p=.242667
MAIN      RAD AUTO      FUNC
```

AP® EXAM TIP The formula for the two-sample z statistic for a test about $p_1 - p_2$ often leads to calculation errors by students. As a result, we recommend using the calculator's `2-PropZTest` feature to perform calculations on the AP® exam. Be sure to name the procedure (two-proportion z test) and to report the test statistic ($z = 1.17$) and P-value (0.2427) as part of the "Do" step.

Inference for Experiments

Most of the examples in this section have involved doing inference about $p_1 - p_2$ using data that were produced by random sampling. In such cases, the parameters p_1 and p_2 are the true proportions of successes in the corresponding populations. However, many important statistical results come from randomized comparative experiments. Defining the parameters in experimental settings is more challenging.

The "Is Yawning Contagious?" Activity on page 610 describes an experiment that used 50 volunteer adults as subjects. Researchers randomly assigned 34 subjects to get a yawn seed and 16 subjects to get no yawn seed. Then researchers compared the proportions of people in the two groups who yawned. The parameters in this setting are:

p_1 = the true proportion of people like these who would yawn when given a yawn seed

p_2 = the true proportion of people like these who would yawn when no yawn seed is given

Most experiments on people use recruited volunteers as subjects. When subjects are not randomly selected, researchers cannot generalize the results of an experiment to some larger populations of interest. But researchers can draw cause-and-effect conclusions that apply to people like those who took part in the experiment. This same logic applies to experiments on animals or things. Also note that unless the experimental units are randomly selected, we don't need to check the 10% condition when performing inference about an experiment.

Here is an example that involves comparing two proportions.

EXAMPLE Cholesterol and Heart Attacks

STEP 4

Significance test in an experiment

High levels of cholesterol in the blood are associated with higher risk of heart attacks. Will using a drug to lower blood cholesterol reduce heart attacks? The Helsinki Heart Study recruited middle-aged men with high cholesterol but no history of other serious medical problems to investigate this question. The volunteer subjects were assigned at random to one of two treatments: 2051 men took the drug gemfibrozil to reduce their cholesterol levels, and a control group of 2030 men took a placebo. During the next five years, 56 men in the gemfibrozil group and 84 men in the placebo group had heart attacks. Is this difference statistically significant at the $\alpha = 0.01$ level?

STATE: We hope to show that gemfibrozil reduces heart attacks, so we have a one-sided alternative:

$$H_0: p_1 - p_2 = 0 \qquad \qquad H_0: p_1 = p_2$$
$$\text{or, equivalently,}$$
$$H_a: p_1 - p_2 < 0 \qquad \qquad H_a: p_1 < p_2$$

where p_1 is the actual heart attack rate for middle-aged men like the ones in this study who take gemfibrozil, and p_2 is the actual heart attack rate for middle-aged men like the ones in this study who take only a placebo. We'll use $\alpha = 0.01$.

Note that we did not need to check the 10% condition because the subjects in the experiment were not sampled without replacement from some larger population.

PLAN: If conditions are met, we will do a two-sample z test for $p_1 - p_2$.

- *Random:* The data come from two groups in a randomized experiment.
  - 10%: Don't need to check because there was no sampling.
- *Large Counts:* The number of successes (heart attacks!) and failures in the two groups are 56, 1995, 84, and 1946. These are all at least 10, so this condition is met.

DO: The proportions of men who had heart attacks in each group are

$$\hat{p}_1 = \frac{56}{2051} = 0.0273 \text{ (gemfibrozil group) and } \hat{p}_2 = \frac{84}{2030} = 0.0414 \text{ (placebo group)}$$

The pooled proportion of heart attacks for the two groups is

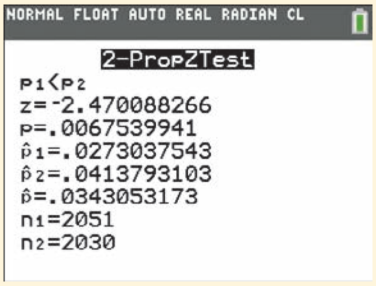
$$\hat{p}_C = \frac{\text{count of heart attacks in both samples combined}}{\text{count of subjects in both samples combined}} = \frac{56 + 84}{2051 + 2030} = \frac{140}{4081} = 0.0343$$

See the two-way table in the margin.

	Drug taken		
Heart attack?	Gemfibrozil	Placebo	**Total**
Yes	56	84	**140**
No	1995	1946	**3941**
Total	**2051**	**2030**	**4081**

We'll use the calculator's `2-PropZTest` to perform calculations.

- *Test statistic* $z = -2.47$
- *P-value* This is the area under the standard Normal curve to the left of $z = -2.47$, shown in Figure 10.4.



```
NORMAL FLOAT AUTO REAL RADIAN CL
2-PropZTest
p1<p2
z=-2.470088266
p=.0067539941
p̂1=.0273037543
p̂2=.0413793103
p̂=.0343053173
n1=2051
n2=2030
```

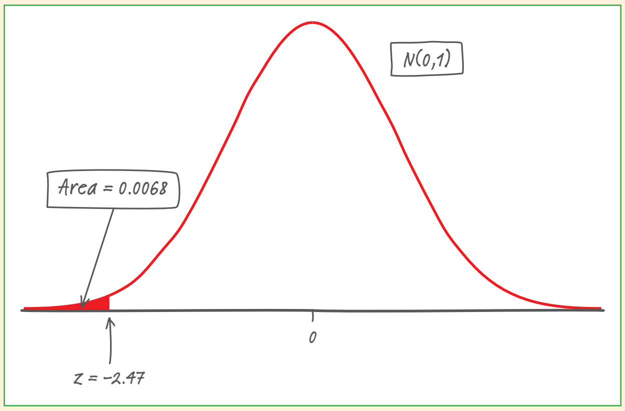


FIGURE 10.4 The *P*-value for the one-sided test.

CONCLUDE: Because the *P*-value, 0.0068, is less than 0.01, we can reject H_0. The results are statistically significant at the $\alpha = 0.01$ level. There is convincing evidence of a lower heart attack rate for middle-aged men like these who take gemfibrozil than for those who take only a placebo.

For Practice *Try Exercise* **21**

We chose $\alpha = 0.01$ in the example to reduce the chance of making a Type I error—finding convincing evidence that gemfibrozil reduces heart attack risk when it actually doesn't. This error could have serious consequences if an ineffective drug was given to lots of middle-aged men with high cholesterol!

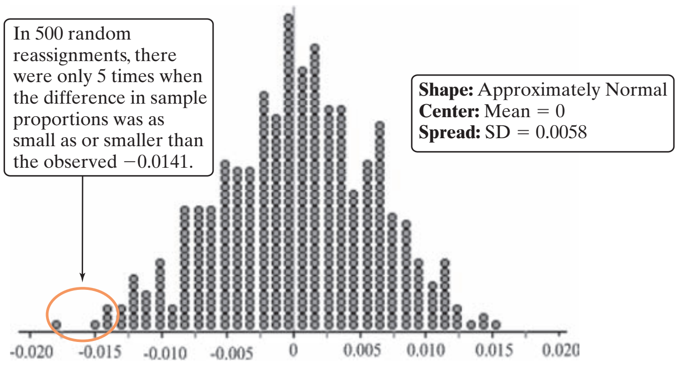
The random assignment in the Helsinki Heart Study allowed researchers to draw a cause-and-effect conclusion. They could say that gemfibrozil reduces the rate of heart attacks for middle-aged men like those who took part in the experiment. Because the subjects were not randomly selected from a larger population, researchers could not generalize the findings of this study any further. No conclusions could be drawn about the effectiveness of gemfibrozil at preventing heart attacks for all middle-aged men, for older men, or for women.

THINK ABOUT IT

Why do the inference methods for random sampling work for randomized experiments? Confidence intervals and tests for $p_1 - p_2$ are based on the sampling distribution of $\hat{p}_1 - \hat{p}_2$. But in experiments, we aren't sampling at random from any larger populations. We can think about what would happen if the random assignment were repeated many times under the assumption that $H_0: p_1 - p_2 = 0$ is true. That is, we assume that the specific treatment received doesn't affect an individual subject's response.

Let's see what would happen just by chance if we randomly reassign the 4081 subjects in the Helsinki Heart Study to the two groups many times, assuming the drug received *doesn't affect* whether or not each individual has a heart attack. We used Fathom software to redo the random assignment 500 times. The approximate **randomization distribution** of $\hat{p}_1 - \hat{p}_2$ is shown in Figure 10.5. It has an approximately Normal shape with mean 0 and standard deviation 0.0058. These are roughly the same as the shape, center, and spread of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ that we used to perform calculations in the previous example because

$$\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}} = \sqrt{\frac{0.0343(1-0.0343)}{2051} + \frac{0.0343(1-0.0343)}{2030}} = 0.0057$$


In 500 random reassignments, there were only 5 times when the difference in sample proportions was as small as or smaller than the observed −0.0141.

Shape: Approximately Normal
Center: Mean = 0
Spread: SD = 0.0058

FIGURE 10.5 Fathom simulation showing the approximate randomization distribution of $\hat{p}_1 - \hat{p}_2$ from 500 random reassignments of subjects to treatment groups in the Helsinki Heart Study.

In the Helsinki Heart Study, the difference in the proportions of subjects who had a heart attack in the gemfibrozil and placebo groups was $0.0273 - 0.0414 = -0.0141$. How likely is it that a difference this large or larger would happen just by chance when H_0 is true? Figure 10.5 provides a rough answer: 5 of the 500 random reassignments yielded a difference in proportions less than or equal to -0.0141. That is, our estimate of the *P*-value is 0.01. This is quite close to the 0.0068 *P*-value that we calculated in the previous example.

Figure 10.6 shows the value of the z test statistic for each of the 500 re-randomizations, calculated using our familiar formula

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \frac{\hat{p}_C(1-\hat{p}_C)}{n_2}}}$$

The standard Normal density curve is shown in blue. We can see that the z test statistic has approximately the standard Normal distribution in this case.

Whenever the conditions are met, the randomization distribution of $\hat{p}_1 - \hat{p}_2$ looks much like its sampling distribution. We are therefore safe using two-sample z procedures for comparing two proportions in a randomized experiment.

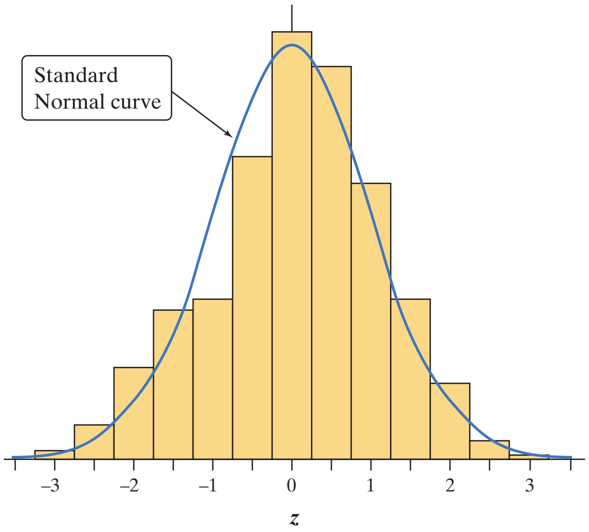

Standard Normal curve

FIGURE 10.6 The distribution of the z test statistic for the 500 random reassignments in Figure 10.5.

CHECK YOUR UNDERSTANDING

To study the long-term effects of preschool programs for poor children, researchers designed an experiment. They recruited 123 children who had never attended preschool from low-income families in Michigan. Researchers randomly assigned 62 of the children to attend preschool (paid for by the study budget) and the other 61 to serve as a control group who would not go to preschool. One response variable of interest was the need for social services as adults. Over a 10-year period, 38 children in the preschool group and 49 in the control group have needed social services.[4]

Does this study provide convincing evidence that preschool reduces the later need for social services? Justify your answer.

Section 10.1 Summary

- Choose independent SRSs of size n_1 from Population 1 with proportion of successes p_1 and of size n_2 from Population 2 with proportion of successes p_2. The sampling distribution of $\hat{p}_1 - \hat{p}_2$ has the following properties:
  - **Shape** Approximately Normal if the samples are large enough that n_1p_1, $n_1(1 - p_1)$, n_2p_2, and $n_2\,(1 - p_2)$ are all at least 10.
  - **Center** The mean is $p_1 - p_2$.
  - **Spread** As long as each sample is no more than 10% of its population, the standard deviation is $\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$.
- Confidence intervals and tests to compare the proportions p_1 and p_2 of successes for two populations or treatments are based on the difference $\hat{p}_1 - \hat{p}_2$ between the sample proportions.
- Before estimating or testing a claim about $p_1 - p_2$, check that these conditions are met:
  - **Random:** The data come from two independent random samples or from two groups in a randomized experiment.
    - ○ **10%:** When sampling without replacement, check that the two populations are at least 10 times as large as the corresponding samples.
  - **Large Counts:** The counts of "successes" and "failures" in each sample or group—$n_1\hat{p}_1$, $n_1(1 - \hat{p}_1)$, $n_2\hat{p}_2$, and $n_2\,(1 - \hat{p}_2)$—are all at least 10.
- When conditions are met, an approximate C% confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^*\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

  where z^* is the standard Normal critical value with C% of its area between $-z^*$ and z^*. This is called a **two-sample z interval for $p_1 - p_2$**.
- Significance tests of $H_0\colon p_1 - p_2 = 0$ use the **pooled (combined) sample proportion** in the standard error formula:

$$\hat{p}_C = \frac{\text{count of successes in both samples combined}}{\text{count of individuals in both samples combined}} = \frac{X_1 + X_2}{n_1 + n_2}$$

When conditions are met, the **two-sample z test for $p_1 - p_2$** uses the test statistic

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\dfrac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \dfrac{\hat{p}_C(1-\hat{p}_C)}{n_2}}}$$

with P-values calculated from the standard Normal distribution.

- Inference about the difference $p_1 - p_2$ in the effectiveness of two treatments in a completely randomized experiment is based on the **randomization distribution** of $\hat{p}_1 - \hat{p}_2$. When conditions are met, our usual inference procedures based on the sampling distribution of $\hat{p}_1 - \hat{p}_2$ will be approximately correct.
- Be sure to follow the four-step process whenever you construct a confidence interval or perform a significance test for comparing two proportions.

10.1 TECHNOLOGY CORNERS

TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site.

21. Confidence interval for a difference in proportions	page 618
22. Significance test for a difference in proportions	page 624

Section 10.1 Exercises

STEP 4 *Remember: We are no longer reminding you to use the four-step process in exercises that require you to perform inference.*

1. **Goldfish** Refer to the example on page 615. Suppose that your teacher decides to take SRSs of 100 crackers from both bags instead.

pg 615

(a) What is the shape of the sampling distribution of $\hat{p}_1 - \hat{p}_2$? Why?

(b) Find the mean of the sampling distribution. Show your work.

(c) Find the standard deviation of the sampling distribution. Show your work.

2. **Homework** Refer to page 612. Suppose that both school counselors decide to take SRSs of 150 students instead.

(a) What is the shape of the sampling distribution of $\hat{p}_1 - \hat{p}_2$? Why?

(b) Find the mean of the sampling distribution. Show your work.

(c) Find the standard deviation of the sampling distribution. Show your work.

3. **I want red!** A candy maker offers Child and Adult bags of jelly beans with different color mixes. The company claims that the Child mix has 30% red jelly beans, while the Adult mix contains 15% red jelly beans. Assume that the candy maker's claim is true. Suppose we take a random sample of 50 jelly beans from the Child mix and a separate random sample of 100 jelly beans from the Adult mix. Let $\hat{p}_C$ and $\hat{p}_A$ be the sample proportions of red jelly beans from the Child and Adult mixes, respectively.

(a) What is the shape of the sampling distribution of $\hat{p}_C - \hat{p}_A$? Why?

(b) Find the mean of the sampling distribution. Show your work.

(c) Find the standard deviation of the sampling distribution. Show your work.

4. **Literacy** A researcher reports that 80% of high school graduates, but only 40% of high school dropouts, would pass a basic literacy test.[5] Assume that the researcher's claim is true. Suppose we give

a basic literacy test to a random sample of 60 high school graduates and a separate random sample of 75 high school dropouts. Let $\hat{p}_G$ and $\hat{p}_D$ be the sample proportions of graduates and dropouts, respectively, who pass the test.

(a) What is the shape of the sampling distribution of $\hat{p}_G - \hat{p}_D$? Why?

(b) Find the mean of the sampling distribution. Show your work.

(c) Find the standard deviation of the sampling distribution. Show your work.

Explain why the conditions for constructing a two-sample z interval for $p_1 - p_2$ are not met in the settings of Exercises 5 through 8.

5. **Don't drink the water!** The movie *A Civil Action* (Touchstone Pictures, 1998) tells the story of a major legal battle that took place in the small town of Woburn, Massachusetts. A town well that supplied water to eastern Woburn residents was contaminated by industrial chemicals. During the period that residents drank water from this well, 16 of the 414 babies born had birth defects. On the west side of Woburn, 3 of the 228 babies born during the same time period had birth defects.

6. **In-line skaters** A study of injuries to in-line skaters used data from the National Electronic Injury Surveillance System, which collects data from a random sample of hospital emergency rooms. The researchers interviewed 161 people who came to emergency rooms with injuries from in-line skating. Wrist injuries (mostly fractures) were the most common.[6] The interviews found that 53 people were wearing wrist guards and 6 of these had wrist injuries. Of the 108 who did not wear wrist guards, 45 had wrist injuries.

7. **Shrubs and fire** Fire is a serious threat to shrubs in dry climates. Some shrubs can resprout from their roots after their tops are destroyed. One study of resprouting took place in a dry area of Mexico.[7] The investigators randomly assigned shrubs to treatment and control groups. They clipped the tops of all the shrubs. They then applied a propane torch to the stumps of the treatment group to simulate a fire. All 12 of the shrubs in the treatment group resprouted. Only 8 of the 12 shrubs in the control group resprouted.

8. **Broken crackers** We don't like to find broken crackers when we open the package. How can makers reduce breaking? One idea is to microwave the crackers for 30 seconds right after baking them. Breaks start as hairline cracks called "checking." Randomly assign 65 newly baked crackers to the microwave and another 65 to a control group that is not microwaved. After one day, none of the microwave group and 16 of the control group show checking.[8]

9. **Who tweets?** Do younger people use Twitter more often than older people? In a random sample of 316 adult Internet users aged 18 to 29, 26% used Twitter. In a separate random sample of 532 adult Internet users aged 30 to 49, 14% used Twitter.[9]

pg 617

(a) Calculate the standard error of the sampling distribution of the difference in the sample proportions (younger adults − older adults). What information does this value provide?

(b) Construct and interpret a 90% confidence interval for the difference between the true proportions of adult Internet users in these age groups who use Twitter.

10. **Listening to rap** Is rap music more popular among young blacks than among young whites? A sample survey compared 634 randomly chosen blacks aged 15 to 25 with 567 randomly selected whites in the same age group. It found that 368 of the blacks and 130 of the whites listened to rap music every day.[10]

(a) Calculate the standard error of the sampling distribution of the difference in the sample proportions (blacks − whites). What information does this value provide?

(b) Construct and interpret a 95% confidence interval for the difference between the proportions of black and white young people who listen to rap every day.

11. **Young adults living at home** A surprising number of young adults (ages 19 to 25) still live in their parents' homes. A random sample by the National Institutes of Health included 2253 men and 2629 women in this age group.[11] The survey found that 986 of the men and 923 of the women lived with their parents.

(a) Construct and interpret a 99% confidence interval for the difference in the true proportions of men and women aged 19 to 25 who live in their parents' homes.

(b) Does your interval from part (a) give convincing evidence of a difference between the population proportions? Explain.

12. **Fear of crime** The elderly fear crime more than younger people, even though they are less likely to be victims of crime. One study recruited separate random samples of 56 black women and 63 black men over the age of 65 from Atlantic City, New Jersey. Of the women, 27 said they "felt vulnerable" to crime; 46 of the men said this.[12]

(a) Construct and interpret a 90% confidence interval for the difference in the true proportions of black women and black men in Atlantic City who would say they felt vulnerable to crime.

(b) Does your interval from part (a) give convincing evidence of a difference between the population proportions? Explain.

13. **Who owns iPods?** As part of the Pew Internet and American Life Project, researchers surveyed a random sample of 800 teens and a separate random sample of 400 young adults. For the teens, 79% said that they own an iPod or MP3 player. For the young adults, this figure was 67%. Do the data give convincing evidence of a difference in the proportions of all teens and young adults who would say that they own an iPod or MP3 player? State appropriate hypotheses for a test to answer this question. Define any parameters you use. pg 620

14. **Steroids in high school** A study by the National Athletic Trainers Association surveyed random samples of 1679 high school freshmen and 1366 high school seniors in Illinois. Results showed that 34 of the freshmen and 24 of the seniors had used anabolic steroids. Steroids, which are dangerous, are sometimes used in an attempt to improve athletic performance.[13] Do the data give convincing evidence of a difference in the proportion of all Illinois high school freshmen and seniors who have used anabolic steroids? State appropriate hypotheses for a test to answer this question. Define any parameters you use.

pg 622 15. **Who owns iPods?** Refer to Exercise 13. Carry out a significance test at the $\alpha = 0.05$ level.

16. **Steroids in high school** Refer to Exercise 14. Carry out a significance test at the $\alpha = 0.05$ level.

17. **Who owns iPods?** Refer to Exercise 13. Construct and interpret a 95% confidence interval for the difference between the population proportions. Explain how the confidence interval is consistent with the results of the test in Exercise 15.

18. **Steroids in high school** Refer to Exercise 14. Construct and interpret a 95% confidence interval for the difference between the population proportions. Explain how the confidence interval is consistent with the results of the test in Exercise 16.

19. **Children make choices** Many new products introduced into the market are targeted toward children. The choice behavior of children with regard to new products is of particular interest to companies that design marketing strategies for these products. As part of one study, randomly selected children in different age groups were compared on their ability to sort new products into the correct product category (milk or juice).[14] Here are some of the data:

Age group	*N*	Number who sorted correctly
4- to 5-year-olds	50	10
6- to 7-year-olds	53	28

Did a significantly higher proportion of the 6- to 7-year-olds than the 4- to 5-year-olds sort correctly? Give appropriate evidence to justify your answer.

20. **Marriage and status** "Would you marry a person from a lower social class than your own?" Researchers asked this question of a random sample of 385 black, never-married college students. Of the 149 men in the sample, 91 said "Yes." Among the 236 women, 117 said "Yes."[15] Did a significantly higher proportion of the men than the women who were surveyed say "Yes"? Give appropriate evidence to justify your answer.

21. **Driving school** A driving school owner believes that Instructor A is more effective than Instructor B at preparing students to pass the state's driver's license exam. An incoming class of 100 students is randomly assigned to two groups, each of size 50. One group is taught by Instructor A; the other is taught by Instructor B. At the end of the course, 30 of Instructor A's students and 22 of Instructor B's students pass the state exam. pg 625

(a) Do these results give convincing evidence at the $\alpha = 0.05$ level that Instructor A is more effective?

(b) Describe a Type I and a Type II error in this setting. Which error could you have made in part (a)?

22. **Preventing strokes** Aspirin prevents blood from clotting and so helps prevent strokes. The Second European Stroke Prevention Study asked whether adding another anticlotting drug, named dipyridamole, would be more effective for patients who had already had a stroke. Here are the data on strokes during the two years of the study:[16]

	Number of patients	Number of strokes
Aspirin alone	1649	206
Aspirin + dipyridamole	1650	157

The study was a randomized comparative experiment.

(a) Is there convincing evidence at the $\alpha = 0.05$ level that adding dipyridamole helps reduce the risk of stroke?

(b) Describe a Type I and a Type II error in this setting. Which is more serious? Explain.

Exercises 23 and 24 involve the following setting. Some women would like to have children but cannot do so for medical reasons. One option for these women is a procedure called in vitro fertilization (IVF), which involves injecting a fertilized egg into the woman's uterus.

23. **Prayer and pregnancy** Two hundred women who were about to undergo IVF served as subjects in an experiment. Each subject was randomly assigned to either a treatment group or a control group. Women in the treatment group were intentionally prayed for by several people (called *intercessors*) who did not know them, a process known as intercessory prayer. The praying continued for three weeks following IVF. The intercessors did not pray for the women in the control group. Here are the results: 44 of the 88 women in the treatment group got pregnant, compared to 21 out of 81 in the control group.[17]

    Is the pregnancy rate significantly higher for women who received intercessory prayer? To find out, researchers perform a test of $H_0: p_1 = p_2$ versus $H_a: p_1 > p_2$, where p_1 and p_2 are the actual pregnancy rates for women like those in the study who do and don't receive intercessory prayer, respectively.

(a) Name the appropriate test and check that the conditions for carrying out this test are met.

(b) The appropriate test from part (a) yields a *P*-value of 0.0007. Interpret this *P*-value in context.

(c) What conclusion should researchers draw at the $\alpha =$ 0.05 significance level? Explain.

(d) The women in the study did not know whether they were being prayed for. Explain why this is important.

24. **Acupuncture and pregnancy** A study reported in the medical journal *Fertility and Sterility* sought to determine whether the ancient Chinese art of acupuncture could help infertile women become pregnant.[18] One hundred sixty healthy women who planned to have IVF were recruited for the study. Half of the subjects (80) were randomly assigned to receive acupuncture 25 minutes before embryo transfer and again 25 minutes after the transfer. The remaining 80 women were assigned to a control group and instructed to lie still for 25 minutes after the embryo transfer. Results are shown in the table below.

	Acupuncture group	Control group
Pregnant	34	21
Not pregnant	46	59
Total	**80**	**80**

Is the pregnancy rate significantly higher for women who received acupuncture? To find out, researchers perform a test of $H_0: p_1 = p_2$ versus $H_a: p_1 > p_2$, where p_1 and p_2 are the actual pregnancy rates for women like those in the study who do and don't receive acupuncture, respectively.

(a) Name the appropriate test and check that the conditions for carrying out this test are met.

(b) The appropriate test from part (a) yields a *P*-value of 0.0152. Interpret this *P*-value in context.

(c) What conclusion should researchers draw at the $\alpha = 0.05$ significance level? Explain.

(d) The women in the study knew whether or not they received acupuncture. Explain why this is important.

Multiple choice: Select the best answer for Exercises 25 to 28.

Exercises 25 to 27 refer to the following setting. A sample survey interviews SRSs of 500 female college students and 550 male college students. Researchers want to determine whether there is a difference in the proportion of male and female college students who worked for pay last summer. In all, 410 of the females and 484 of the males say they worked for pay last summer.

25. Take p_M and p_F to be the proportions of all college males and females who worked last summer. The hypotheses to be tested are

(a) $H_0: p_M - p_F = 0$ versus $H_a: p_M - p_F \neq 0$.

(b) $H_0: p_M - p_F = 0$ versus $H_a: p_M - p_F > 0$.

(c) $H_0: p_M - p_F = 0$ versus $H_a: p_M - p_F < 0$.

(d) $H_0: p_M - p_F > 0$ versus $H_a: p_M - p_F = 0$.

(e) $H_0: p_M - p_F \neq 0$ versus $H_a: p_M - p_F = 0$.

26. The researchers report that the results were statistically significant at the 1% level. Which of the following is the most appropriate conclusion?

(a) Because the *P*-value is less than 1%, fail to reject H_0. There is not convincing evidence that the proportion of male college students in the study who worked for pay last summer is different from the proportion of female college students in the study who worked for pay last summer.

(b) Because the *P*-value is less than 1%, fail to reject H_0. There is not convincing evidence that the proportion of all male college students who worked for pay last summer is different from the proportion of all female college students who worked for pay last summer.

(c) Because the P-value is less than 1%, reject H_0. There is convincing evidence that the proportion of all male college students who worked for pay last summer is the same as the proportion of all female college students who worked for pay last summer.

(d) Because the P-value is less than 1%, reject H_0. There is convincing evidence that the proportion of all male college students in the study who worked for pay last summer is different from the proportion of all female college students in the study who worked for pay last summer.

(e) Because the P-value is less than 1%, reject H_0. There is convincing evidence that the proportion of all male college students who worked for pay last summer is different from the proportion of all female college students who worked for pay last summer.

27. Which of the following is the correct margin of error for a 99% confidence interval for the difference in the proportion of male and female college students who worked for pay last summer?

(a) $2.576\sqrt{\frac{0.851(0.149)}{550}+\frac{0.851(0.149)}{500}}$

(b) $2.576\sqrt{\frac{0.851(0.149)}{1050}}$

(c) $2.576\sqrt{\frac{0.880(0.120)}{550}+\frac{0.820(0.180)}{500}}$

(d) $1.960\sqrt{\frac{0.851(0.149)}{550}+\frac{0.851(0.149)}{500}}$

(e) $1.960\sqrt{\frac{0.880(0.120)}{550}+\frac{0.820(0.180)}{500}}$

28. In an experiment to learn whether Substance M can help restore memory, the brains of 20 rats were treated to damage their memories. First, the rats were trained to run a maze. After a day, 10 rats (determined at random) were given M and 7 of them succeeded in the maze. Only 2 of the 10 control rats were successful. The two-sample z test for "no difference" against "a significantly higher proportion of the M group succeeds"

(a) gives $z = 2.25$, $P < 0.02$.

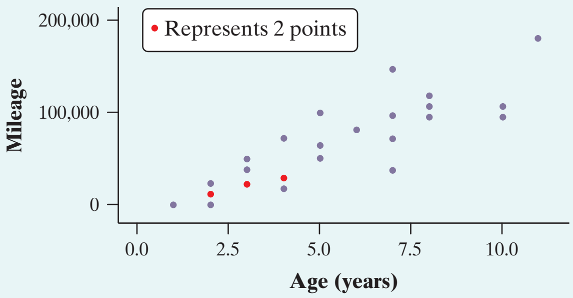
(b) gives $z = 2.60$, $P < 0.005$.

(c) gives $z = 2.25$, $P < 0.04$ but not < 0.02.

(d) should not be used because the Random condition is violated.

(e) should not be used because the Large Counts condition is violated.

Exercises 29 and 30 refer to the following setting. Thirty randomly selected seniors at Council High School were asked to report the age (in years) and mileage of their main vehicles. Here is a scatterplot of the data:



We used Minitab to perform a least-squares regression analysis for these data. Part of the computer output from this regression is shown below.

```
Predictor      Coef      Stdev     t-ratio       P
Constant     −13832      8773      −1.58       0.126
Age           14954      1546       9.67       0.000
s = 22723    R-sq = 77.0%    R-sq(adj) = 76.1%
```

29. **Drive my car** (3.2)

(a) What is the equation of the least-squares regression line? Be sure to define any symbols you use.

(b) Interpret the slope of the least-squares line in the context of this problem.

(c) One student reported that her 10-year-old car had 110,000 miles on it. Find and interpret the residual for this data value. Show your work.

30. **Drive my car** (3.2, 4.3)

(a) Explain what the value of r^2 tells you about how well the least-squares line fits the data.

(b) The mean age of the students' cars in the sample was $\bar{x} = 8$ years. Find the mean mileage of the cars in the sample. Show your work.

(c) Interpret the value of s in the context of this setting.

(d) Would it be reasonable to use the least-squares line to predict a car's mileage from its age for a Council High School teacher? Justify your answer.

10.2 Comparing Two Means

WHAT YOU WILL LEARN **By the end of the section, you should be able to:**

- Describe the shape, center, and spread of the sampling distribution of $\bar{x}_1 - \bar{x}_2$.
- Determine whether the conditions are met for doing inference about $\mu_1 - \mu_2$.
- Construct and interpret a confidence interval to compare two means.
- Perform a significance test to compare two means.
- Determine when it is appropriate to use two-sample *t* procedures versus paired *t* procedures.

In the previous section, we developed methods for comparing two proportions. What if we want to compare the mean of some quantitative variable for the individuals in Population 1 and Population 2? Our parameters of interest are the population means μ_1 and μ_2. Once again, the best approach is to take separate random samples from each population and to compare the sample means $\bar{x}_1$ and $\bar{x}_2$.

Suppose we want to compare the average effectiveness of two treatments in a completely randomized experiment. In this case, the parameters μ_1 and μ_2 are the true mean responses for Treatment 1 and Treatment 2, respectively. We use the mean response in the two groups, $\bar{x}_1$ and $\bar{x}_2$, to make the comparison. Here's a table that summarizes these two situations:

Population or treatment	Parameter	Statistic	Sample size
1	μ_1	$\bar{x}_1$	n_1
2	μ_2	$\bar{x}_2$	n_2

We compare the populations or treatments by doing inference about the difference $\mu_1 - \mu_2$ between the parameters. The statistic that estimates this difference is the difference between the two sample means, $\bar{x}_1 - \bar{x}_2$. To use $\bar{x}_1 - \bar{x}_2$ for inference, we must know its sampling distribution. Here is an Activity that gives you a preview of what lies ahead.

ACTIVITY | Does Polyester Decay?

MATERIALS:
10 small pieces of card stock (or index cards) per pair of students

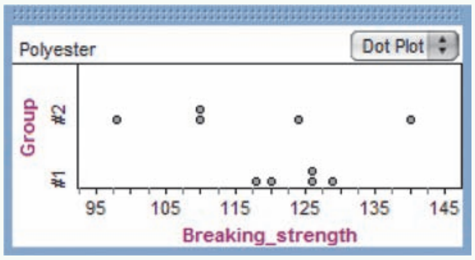
How quickly do synthetic fabrics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay. Lower strength means the fabric has decayed.

The researcher buried 10 strips of polyester fabric in well-drained soil in the summer. The strips were randomly assigned to two groups: 5 of them were buried for 2 weeks and the other 5 were buried for 16 weeks. Here are the breaking strengths in pounds:[19]

Group 1 (2 weeks):	118	126	126	120	129
Group 2 (16 weeks):	124	98	110	140	110

Do the data give convincing evidence that polyester decays more in 16 weeks than in 2 weeks?

1. The Fathom dotplot displays the data from the experiment. Discuss what this graph shows with your classmates.



For Group 1, the mean breaking strength was $\bar{x}_1 = 123.8$ pounds. For Group 2, the mean breaking strength was $\bar{x}_2 = 116.4$ pounds. The observed difference in average breaking strength for the two groups is $\bar{x}_1 - \bar{x}_2 = 123.8 - 116.4 = 7.4$ pounds. Is it plausible that this difference is due to the chance involved in the random assignment and not to the treatments themselves? To find out, your class will perform a simulation.

Suppose that the length of time in the ground has no effect on the breaking strength of the polyester specimens. Then each specimen would have the same breaking strength regardless of whether it was assigned to Group 1 or Group 2. In that case, we could examine the results of repeated random assignments of the specimens to the two groups.

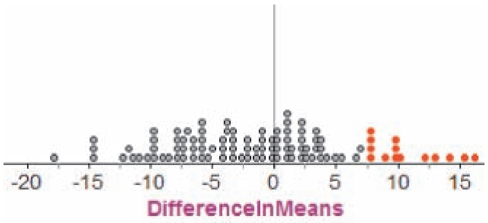
2. Write each of the 10 breaking-strength measurements on a separate card. Mix the cards well and deal them face down into two piles of 5 cards each. Be sure to decide which pile is Group 1 and which is Group 2 *before* you look at the cards. Calculate the difference in the mean breaking strength (Group 1 − Group 2). Record this value.

3. Your teacher will draw and label axes for a class dotplot. Plot the result you got in Step 2 on the graph.

4. Repeat Steps 2 and 3 if needed to get a total of at least 40 repetitions of the simulation for your class.

5. Based on the class's simulation results, how surprising would it be to get a difference in means of 7.4 or larger simply due to the chance involved in the random assignment?

6. What conclusion would you draw about whether polyester decays more when left in the ground for longer periods of time? Explain.



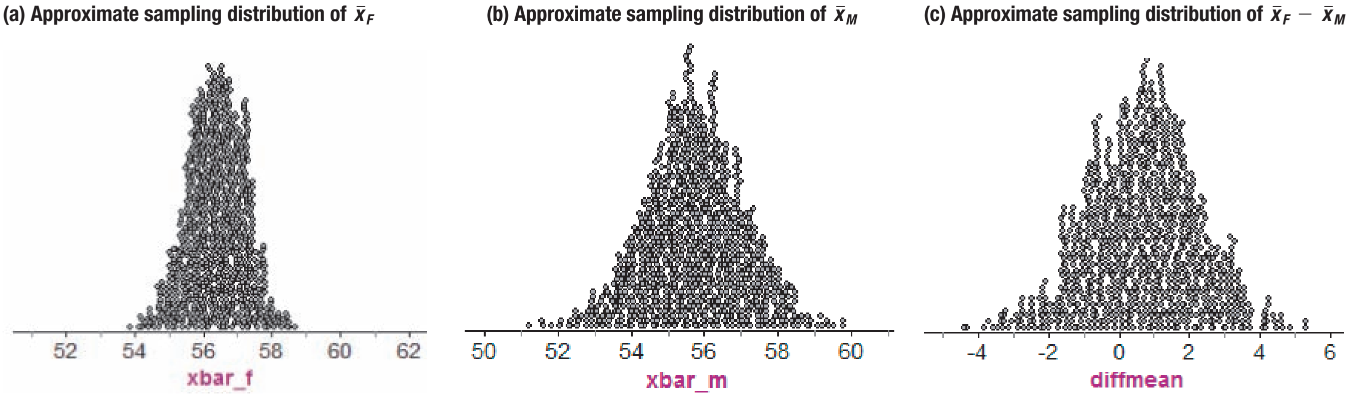
In this simulation, 14 of the 100 trials (in red) produced a difference in means of at least 7.4 pounds, so the approximate *P*-value is 0.14. It is likely that a difference this big could have happened just due to the chance variation in random assignment. The observed difference is not statistically significant and does not provide convincing evidence that polyester decays more in 16 weeks than in 2 weeks.

The Sampling Distribution of a Difference between Two Means

To explore the sampling distribution of $\bar{x}_1 - \bar{x}_2$, let's start with two Normally distributed populations having known means and standard deviations. Based on information from the U.S. National Health and Nutrition Examination Survey (NHANES), the heights of 10-year-old girls follow a Normal distribution with mean $\mu_F = 56.4$ inches and standard deviation $\sigma_F = 2.7$ inches. The heights of 10-year-old boys follow a Normal distribution with mean $\mu_M = 55.7$ inches and standard deviation $\sigma_M = 3.8$ inches.[20]

Suppose we take independent SRSs of 12 girls and 8 boys of this age and measure their heights. What can we say about the difference $\bar{x}_F - \bar{x}_M$ in the average heights of the sample of girls and the sample of boys?

We used Fathom software to take an SRS of 12 ten-year-old girls and 8 ten-year-old boys and to plot the values of $\bar{x}_F$, $\bar{x}_M$, and $\bar{x}_F - \bar{x}_M$ for each sample. Our first set of simulated samples gave $\bar{x}_F = 56.09$ inches and $\bar{x}_M = 54.68$ inches, so dots were placed above each of those values in Figure 10.7(a) and (b). The difference in the sample means is $\bar{x}_F - \bar{x}_M = 56.09 - 54.68 = 1.41$ inches. A dot for this value appears in Figure 10.7(c). The three dotplots in Figure 10.7 show the results of repeating this process 1000 times. These are the approximate sampling distributions of $\bar{x}_F$, $\bar{x}_M$, and $\bar{x}_F - \bar{x}_M$.



(a) Approximate sampling distribution of $\bar{x}_F$

Shape: Approximately Normal
Center: Mean = 56.40 inches
Spread: SD = 0.80 inches

(b) Approximate sampling distribution of $\bar{x}_M$

Shape: Approximately Normal
Center: Mean = 55.73 inches
Spread: SD = 1.35 inches

(c) Approximate sampling distribution of $\bar{x}_F - \bar{x}_M$

Shape: Approximately Normal
Center: Mean = 0.67 inches
Spread: SD = 1.56 inches

FIGURE 10.7 Simulated sampling distributions of (a) the sample mean height $\bar{x}_F$ in 1000 SRSs of size $n_F = 12$ from the population of 10-year-old girls, (b) the sample mean height $\bar{x}_M$ in 1000 SRSs of size $n_M = 8$ from the population of 10-year-old boys, and (c) the difference in sample means $\bar{x}_F - \bar{x}_M$ for each of the 1000 repetitions.

In Chapter 7, we saw that the sampling distribution of a sample mean $\bar{x}$ has the following properties:

Shape: (1) If the population distribution is Normal, then so is the sampling distribution of $\bar{x}$; (2) if the population distribution isn't Normal, the sampling distribution of $\bar{x}$ will be approximately Normal if the sample size is large enough (say, $n \geq 30$) by the central limit theorem (CLT).

Center: $\mu_{\bar{x}} = \mu$

Spread: $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$ if the sample is no more than 10% of the population

For the sampling distributions of $\bar{x}_F$ and $\bar{x}_M$ in this case:

	Sampling distribution of $\bar{x}_F$	**Sampling distribution of $\bar{x}_M$**
Shape	Normal, because the population distribution is Normal	Normal, because the population distribution is Normal
Center	$\mu_{\bar{x}_F} = \mu_F = 56.4$ inches	$\mu_{\bar{x}_M} = \mu_M = 55.7$ inches
Spread	$\sigma_{\bar{x}_F} = \dfrac{\sigma_F}{\sqrt{n_F}} = \dfrac{2.7}{\sqrt{12}} = 0.78$ inches because there are way more than 10(12) = 120 ten-year-old girls in the United States.	$\sigma_{\bar{x}_M} = \dfrac{\sigma_M}{\sqrt{n_M}} = \dfrac{3.8}{\sqrt{8}} = 1.34$ inches because there are way more than 10(8) = 80 ten-year-old boys in the United States.

The approximate sampling distributions in Figures 10.7(a) and (b) give similar results.

What about the sampling distribution of $\bar{x}_F - \bar{x}_M$? Figure 10.7(c) suggests that it has a roughly Normal shape, is centered at about 0.67 inches, and has standard deviation about 1.56 inches. The shape makes sense because we are combining two independent Normal random variables, $\bar{x}_F$ and $\bar{x}_M$. How about the center? The actual mean height of 10-year-old girls is $\mu_F = 56.4$ inches. For 10-year-old boys, the actual mean height is $\mu_M = 55.7$ inches. We'd expect the difference $\bar{x}_F - \bar{x}_M$ to center on the actual difference in the population means, $\mu_F - \mu_M = 56.4 - 55.7 = 0.7$ inches. The spread, however, is a bit more complicated.

THINK ABOUT IT

How can we find formulas for the mean and standard deviation of the sampling distribution of $\bar{x}_1 - \bar{x}_2$? Both $\bar{x}_1$ and $\bar{x}_2$ are random variables. That is, their values would vary in repeated independent SRSs of size n_1 and n_2. Independent random samples yield independent random variables $\bar{x}_1$ and $\bar{x}_2$. The statistic $\bar{x}_1 - \bar{x}_2$ is the difference of these two independent random variables.

In Chapter 6, we learned that for any two random variables X and Y,

$$\mu_{X-Y} = \mu_X - \mu_Y$$

For the random variables $\bar{x}_1$ and $\bar{x}_2$, we have

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_{\bar{x}_1} - \mu_{\bar{x}_2} = \mu_1 - \mu_2$$

In the observational study of the heights of 10-year-olds,

$$\mu_{\bar{x}_F - \bar{x}_M} = \mu_F - \mu_M = 56.4 - 55.7 = 0.70 \text{ inches}$$

We also learned in Chapter 6 that for *independent* random variables X and Y,

$$\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

For the random variables $\bar{x}_1$ and $\bar{x}_2$, we have

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

So $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$.

In the observational study of the heights of 10-year-olds,

$$\sigma_{\bar{x}_F-\bar{x}_M}=\sqrt{\frac{\sigma_F^2}{n_F}+\frac{\sigma_M^2}{n_M}}=\sqrt{\frac{2.7^2}{12}+\frac{3.8^2}{8}}=1.55$$

This is similar to the result from the Fathom simulation.

Here are the facts we need.

THE SAMPLING DISTRIBUTION OF $\bar{x}_1-\bar{x}_2$

Choose an SRS of size n_1 from Population 1 with mean μ_1 and standard deviation σ_1 and an independent SRS of size n_2 from Population 2 with mean μ_2 and standard deviation σ_2.

- **Shape:** When the population distributions are Normal, the sampling distribution of $\bar{x}_1-\bar{x}_2$ is Normal. In other cases, the sampling distribution of $\bar{x}_1-\bar{x}_2$ will be approximately Normal if the sample sizes are large enough ($n_1 \geq 30$ and $n_2 \geq 30$).
- **Center:** The mean of the sampling distribution of $\bar{x}_1-\bar{x}_2$ is $\mu_1-\mu_2$.
- **Spread:** The standard deviation of the sampling distribution of $\bar{x}_1-\bar{x}_2$ is

$$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

as long as each sample is no more than 10% of its population.

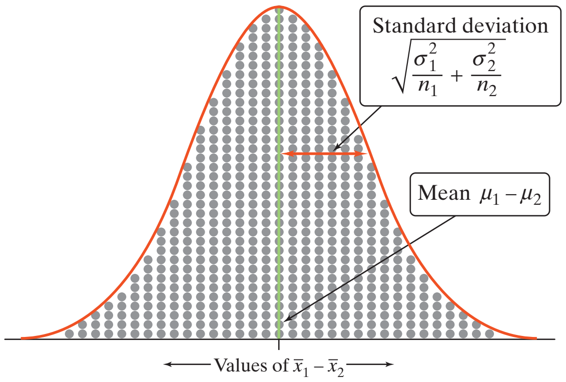
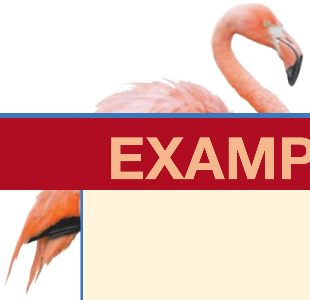

Standard deviation $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
Mean $\mu_1-\mu_2$
Values of $\bar{x}_1-\bar{x}_2$

FIGURE 10.8 Select independent SRSs from two populations having means μ_1 and μ_2 and standard deviations σ_1 and σ_2. The two sample means are $\bar{x}_1$ and $\bar{x}_2$. If the population distributions are both Normal, the sampling distribution of the difference $\bar{x}_1-\bar{x}_2$ is Normal. The sampling distribution will be approximately Normal in other cases if both samples are large enough ($n_1 \geq 30$ and $n_2 \geq 30$).

When conditions are met, the sampling distribution of $\bar{x}_1-\bar{x}_2$ will be approximately Normal with mean $\mu_{\bar{x}_1-\bar{x}_2}=\mu_1-\mu_2$ and standard deviation $\sigma_{\bar{x}_1-\bar{x}_2}=\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$. Figure 10.8 displays this distribution.

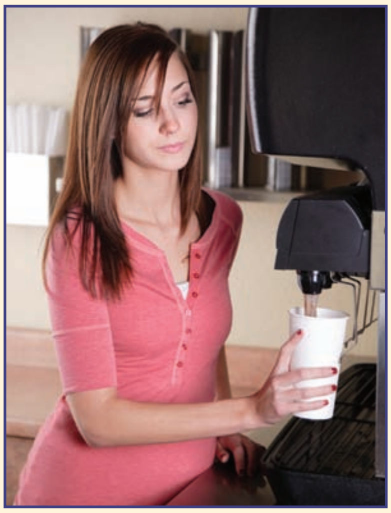
The formula for the standard deviation of the sampling distribution involves the parameters σ_1 and σ_2, which are usually unknown. Just as in Chapters 8 and 9, we must replace these by estimates to do inference. We'll get to confidence intervals and significance tests shortly. For now, let's focus on the sampling distribution of $\bar{x}_1-\bar{x}_2$.



EXAMPLE Medium or Large Drink?

Describing the sampling distribution of $\bar{x}_1-\bar{x}_2$

A fast-food restaurant uses an automated filling machine to pour its soft drinks. The machine has different settings for small, medium, and large drink cups. According to the machine's manufacturer, when the large setting is chosen, the amount of liquid *L* dispensed by the machine follows a Normal distribution with mean 27 ounces and



standard deviation 0.8 ounces. When the medium setting is chosen, the amount of liquid M dispensed follows a Normal distribution with mean 17 ounces and standard deviation 0.5 ounces. To test the manufacturer's claim, the restaurant manager measures the amount of liquid in each of 20 cups filled with the large setting and 25 cups filled with the medium setting. Let $\bar{x}_L - \bar{x}_M$ be the difference in the sample mean amount of liquid under the two settings.

PROBLEM:

(a) What is the shape of the sampling distribution of $\bar{x}_L - \bar{x}_M$? Why?

(b) Find the mean of the sampling distribution. Show your work.

(c) Find the standard deviation of the sampling distribution. Show your work.

SOLUTION:

(a) The sampling distribution of $\bar{x}_L - \bar{x}_M$ is Normal because both population distributions are Normal.

(b) The mean is $\mu_{\bar{x}_L - \bar{x}_M} = \mu_L - \mu_M = 27 - 17 = 10$ ounces.

(c) The standard deviation is $\sigma_{\bar{x}_L - \bar{x}_M} = \sqrt{\frac{\sigma_L^2}{n_L} + \frac{\sigma_M^2}{n_M}} = \sqrt{\frac{(0.80)^2}{20} + \frac{(0.50)^2}{25}} = 0.205$ ounces.

Note that we do not need to check the 10% condition because we are not sampling without replacement from a finite population.

For Practice *Try Exercise* **31**

The Two-Sample *t* Statistic

When data come from two independent random samples or two groups in a randomized experiment (the Random condition), the statistic $\bar{x}_1 - \bar{x}_2$ is our best guess for the value of $\mu_1 - \mu_2$. If the 10% condition is met, the standard deviation of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If the Normal/Large Sample condition is met, we can standardize the observed difference $\bar{x}_1 - \bar{x}_2$ to obtain a z statistic that is modeled well by a standard Normal distribution:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

In the unlikely event that both population standard deviations are known, this *two-sample z statistic* is the basis for inference about $\mu_1 - \mu_2$.

Suppose now that the population standard deviations σ_1 and σ_2 are not known. We estimate them by the standard deviations s_1 and s_2 from our two samples. The result is the **standard error** (also called the *estimated standard deviation*) of $\bar{x}_1 - \bar{x}_2$:

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Now when we standardize the point estimate $\bar{x}_1 - \bar{x}_2$, the result is the **two-sample *t* statistic:**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

The statistic t has the same interpretation as any z or t statistic: it says how far $\bar{x}_1 - \bar{x}_2$ is from its mean in standard deviation units. When the Normal/Large Sample condition is met, the two-sample t statistic has approximately a t distribution. It does not have exactly a t distribution even if the populations are both exactly Normal. In practice, however, the approximation is very accurate.

CONDITIONS FOR PERFORMING INFERENCE ABOUT $\mu_1 - \mu_2$

- **Random:** The data come from two independent random samples or from two groups in a randomized experiment.
  - ○ **10%:** When sampling without replacement, check that $n_1 \le \frac{1}{10}N_1$ and $n_2 \le \frac{1}{10}N_2$.
- **Normal/Large Sample:** Both population distributions (or the true distributions of responses to the two treatments) are Normal or both sample sizes are large ($n_1 \ge 30$ and $n_2 \ge 30$). If either population (treatment) distribution has unknown shape and the corresponding sample size is less than 30, use a graph of the sample data to assess the Normality of the population (treatment) distribution. Do not use two-sample t procedures if the graph shows strong skewness or outliers.

There are two practical options for using the two-sample t procedures when the conditions are met. The two options are exactly the same except for the degrees of freedom used for t critical values and P-values.

Option 1 (Technology): Use the t distribution with degrees of freedom calculated from the data by the formula below. Note that the df given by this formula is usually not a whole number.

$$\text{df} = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{1}{n_1 - 1}\left(\dfrac{s_1^2}{n_1}\right)^2 + \dfrac{1}{n_2 - 1}\left(\dfrac{s_2^2}{n_2}\right)^2}$$

You can thank statisticians B. L. Welch and F. E. Satterthwaite for discovering this fairly remarkable formula.

Option 2 (Conservative): Use the t distribution with degrees of freedom equal to the *smaller* of $n_1 - 1$ and $n_2 - 1$. With this option, the resulting confidence interval has a margin of error *as large as or larger than* is needed for the desired confidence level. The significance test using this option gives a P-value *equal to or* *greater than* the true *P*-value. As the sample sizes increase, confidence levels and *P*-values from Option 2 become more accurate.[21]

Confidence Intervals for $\mu_1 - \mu_2$

If the Random, 10%, and Normal/Large Sample conditions are met, we can use our standard formula to construct a confidence interval for $\mu_1 - \mu_2$:

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

$$= (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

We can use either technology or the conservative approach with Table B to find the critical value t^* for the given confidence level. This method is called a **two-sample *t* interval for a difference between two means**.

TWO-SAMPLE *t* INTERVAL FOR A DIFFERENCE BETWEEN TWO MEANS

When the conditions are met, an approximate *C*% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here, t^* is the critical value with *C*% of its area between $-t^*$ and t^* for the *t* distribution with degrees of freedom using either Option 1 (technology) or Option 2 (the smaller of $n_1 - 1$ and $n_2 - 1$).

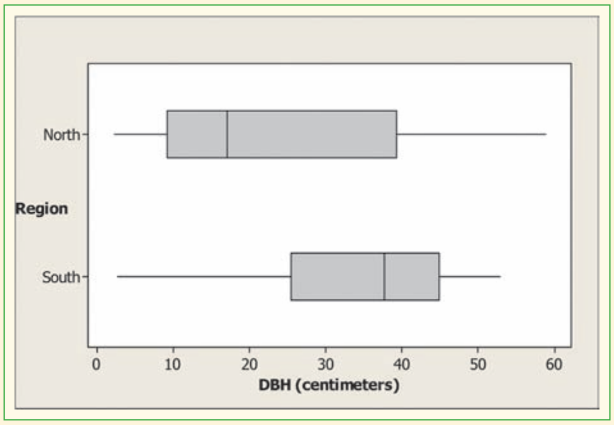
The following example shows how to construct and interpret a confidence interval for a difference in means. As usual with inference problems, we follow the four-step process.

EXAMPLE Big Trees, Small Trees, Short Trees, Tall Trees

STEP 4

Confidence interval for $\mu_1 - \mu_2$

The Wade Tract Preserve in Georgia is an old-growth forest of longleaf pines that has survived in a relatively undisturbed state for hundreds of years. One question of interest to foresters who study the area is "How do the sizes of longleaf pine trees in the northern and southern halves of the forest compare?" To find out, researchers took random samples of 30 trees from each half and measured the diameter at breast height (DBH) in centimeters.[22] Here are comparative boxplots of the data and summary statistics from Minitab.



Descriptive Statistics: North, South

```
Variable     N       Mean       StDev
North       30      23.70       17.50
South       30      34.53       14.26
```

PROBLEM:

(a) Based on the graph and numerical summaries, write a few sentences comparing the sizes of longleaf pine trees in the two halves of the forest.

(b) Construct and interpret a 90% confidence interval for the difference in the mean DBH of longleaf pines in the northern and southern halves of the Wade Tract Preserve.

SOLUTION:

(a) The distribution of DBH measurements in the northern sample is skewed to the right, while the distribution of DBH measurements in the southern sample is skewed to the left. It appears that trees in the southern half of the forest have larger diameters. The mean and median DBH for the southern sample are both much larger than the corresponding measures of center for the northern sample. Furthermore, the boxplots show that more than 75% of the southern trees have diameters that are above the northern sample's median. There is more variability in the diameters of the northern longleaf pines, as we can see from the larger range, *IQR*, and standard deviation for this sample. No outliers are present in either sample.

(b) **STATE:** Our parameters of interest are μ_1 = the true mean DBH of all trees in the southern half of the forest and μ_2 = the true mean DBH of all trees in the northern half of the forest. We want to estimate the difference $\mu_1 - \mu_2$ at a 90% confidence level.

PLAN: If conditions are met, we'll construct a two-sample *t* interval for $\mu_1 - \mu_2$.

- *Random:* The data came from independent random samples of 30 trees each from the northern and southern halves of the forest.
  - *10%:* Because sampling without replacement was used, there have to be at least 10(30) = 300 trees in each half of the forest. This is fairly safe to assume.
- *Normal/Large Sample:* The boxplots give us reason to believe that the population distributions of DBH measurements may not be Normal. However, because both sample sizes are at least 30, we are safe using two-sample *t* procedures.

DO: From the Minitab output, $\bar{x}_1 = 34.53$, $s_1 = 14.26$, $n_1 = 30$, $\bar{x}_2 = 23.70$, $s_2 = 17.50$, and $n_2 = 30$. We'll use the conservative df = the smaller of $n_1 - 1$ and $n_2 - 1$, which is 29. For a 90% confidence level the critical value from Table B is $t^* = 1.699$. So a 90% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_1) \pm t^*\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (34.53 - 23.70) \pm 1.699\sqrt{\frac{14.26^2}{30} + \frac{17.50^2}{30}}$$

$$= 10.83 \pm 7.00 = (3.83, 17.83)$$

	Upper-tail probability *p*		
df	**.10**	**.05**	**.025**
28	1.313	1.701	2.048
29	1.311	1.699	2.045
30	1.310	1.697	2.042
	80%	90%	95%
	Confidence level *C*		

Using technology: Refer to the Technology Corner that follows the example. The calculator's `2-SampTInt` gives (3.9362, 17.724) using df = 55.728.

CONCLUDE: We are 90% confident that the interval from 3.9362 to 17.724 centimeters captures the difference in the actual mean DBH of the southern trees and the actual mean DBH of the northern trees.

For Practice *Try Exercise* **37**

The 90% confidence interval in the example does not include 0. This gives convincing evidence that the difference in the mean diameter of northern and southern trees in the Wade Tract Preserve isn't 0. However, the confidence interval provides more information than a simple reject or fail to reject H_0 conclusion. It gives a set of plausible values for $\mu_1 - \mu_2$. The interval suggests that the mean diameter of the southern trees is between 3.83 and 17.83 cm larger than the mean diameter of the northern trees.

We chose the parameters in the DBH example so that $\bar{x}_1 - \bar{x}_2$ would be positive. What if we had defined μ_1 as the mean DBH of the northern trees and μ_2 as the mean DBH of the southern trees? The 90% confidence interval for $\mu_1 - \mu_2$ would be

$$(23.70 - 34.53) \pm 1.699\sqrt{\frac{17.50^2}{30} + \frac{14.26^2}{30}} = -10.83 \pm 7.00 = (-17.83, -3.83)$$

This interval suggests that the mean diameter of the northern trees is between 3.83 and 17.83 cm smaller than the mean diameter of the southern trees. Changing the order of subtraction doesn't change the result.

As with other inference procedures, you can use technology to perform the calculations in the "Do" step. Remember that technology comes with potential benefits and risks on the AP® exam.

23. TECHNOLOGY CORNER — TWO-SAMPLE *t* INTERVALS ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

You can use the two-sample *t* interval command on the TI-83/84 or TI-89 to construct a confidence interval for the difference between two means. We'll show you the steps using the summary statistics from the pine trees example.

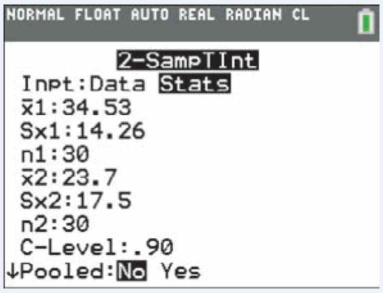
TI-83/84

- Press STAT, then choose TESTS and `2-SampTInt...`.

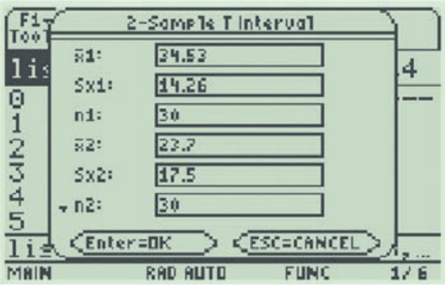
TI-89

- Press 2nd F2 ([F7]) Ints and choose `2-SampTInt...`.

- Choose Stats as the input method and enter the summary statistics as shown.

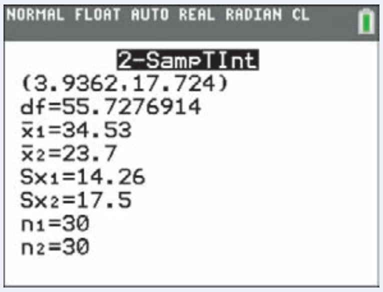


```
NORMAL FLOAT AUTO REAL RADIAN CL
        2-SampTInt
 Inpt:Data Stats
 x̄1:34.53
 Sx1:14.26
 n1:30
 x̄2:23.7
 Sx2:17.5
 n2:30
 C-Level:.90
↓Pooled:No Yes
```

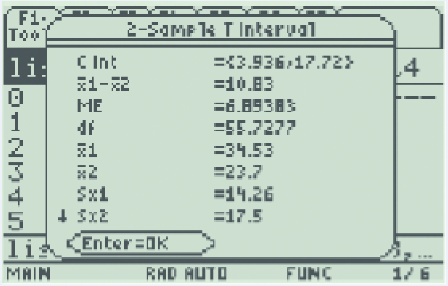


```
        2-Sample T Interval
   x̄1:      34.53
   Sx1:     14.26
   n1:      30
   x̄2:      23.7
   Sx2:     17.5
 ▾ n2:      30
 Enter=OK     ESC=CANCEL
MAIN        RAD AUTO      FUNC      1/6
```

- Enter the confidence level: `C-level: .90`. For Pooled: choose "No." We'll discuss pooling later.
- Highlight Calculate and press ENTER.



```
NORMAL FLOAT AUTO REAL RADIAN CL
        2-SampTInt
 (3.9362,17.724)
 df=55.7276914
 x̄1=34.53
 x̄2=23.7
 Sx1=14.26
 Sx2=17.5
 n1=30
 n2=30
```



```
        2-Sample T Interval
   C Int       ={3.936,17.72}
   x̄1-x̄2       =10.83
   ME          =6.89383
   df          =55.7277
   x̄1          =34.53
   x̄2          =23.7
   Sx1         =14.26
 ↓ Sx2         =17.5
 Enter=OK
MAIN        RAD AUTO      FUNC      1/6
```

AP® EXAM TIP The formula for the two-sample *t* interval for $\mu_1 - \mu_2$ often leads to calculation errors by students. As a result, we recommend using the calculator's `2-SampTInt` feature to compute the confidence interval on the AP® exam. Be sure to name the procedure (two-sample *t* interval) and to give the interval (3.9362, 17.724) and df (55.728) as part of the "Do" step.

The calculator's 90% confidence interval for $\mu_1 - \mu_2$ is (3.936, 17.724). This interval is narrower than the one we found by hand earlier: (3.83, 17.83). Why the difference? We used the conservative df = 29, but the calculator used df = 55.73. With more degrees of freedom, the calculator's critical value is smaller than our $t^* = 1.699$, which results in a smaller margin of error and a narrower interval.

CHECK YOUR UNDERSTANDING

The U.S. Department of Agriculture (USDA) conducted a survey to estimate the average price of wheat in July and in September of the same year. Independent random samples of wheat producers were selected for each of the two months. Here are summary statistics on the reported price of wheat from the selected producers, in dollars per bushel:[23]

Month	n	$\bar{x}$	s_x
July	90	\$2.95	\$0.22
September	45	\$3.61	\$0.19

Construct and interpret a 99% confidence interval for the difference in the true mean wheat price in July and in September.

Significance Tests for $\mu_1 - \mu_2$

An observed difference between two sample means can reflect an actual difference in the parameters μ_1 and μ_2, or it may just be due to chance variation in random sampling or random assignment. Significance tests help us decide which explanation makes more sense. The null hypothesis has the general form

$$H_0: \mu_1 - \mu_2 = \text{hypothesized value}$$

We're often interested in situations in which the hypothesized difference is 0. Then the null hypothesis says that there is no difference between the two parameters:

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{or, alternatively,} \quad H_0: \mu_1 = \mu_2$$

The alternative hypothesis says what kind of difference we expect.

If the Random, 10%, and Normal/Large Sample conditions are met, we can proceed with calculations. To do a test, standardize $\bar{x}_1 - \bar{x}_2$ to get a two-sample t statistic:

$$\text{test statistic} = \frac{\text{statistic} - \text{parameter}}{\text{standard deviation of statistic}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

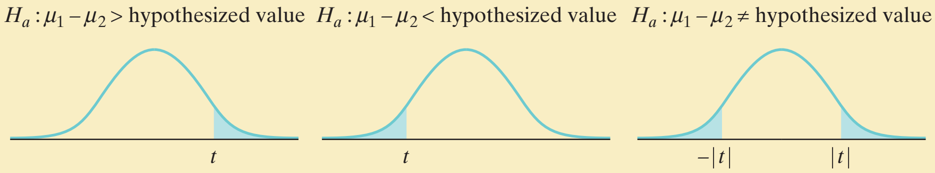
To find the P-value, use the t distribution with degrees of freedom given by Option 1 (technology) or Option 2 (df = smaller of $n_1 - 1$ and $n_2 - 1$). Here are the details for the **two-sample *t* test for the difference between two means**.

TWO-SAMPLE *t* TEST FOR THE DIFFERENCE BETWEEN TWO MEANS

Suppose the conditions are met. To test the hypothesis $H_0: \mu_1 - \mu_2 =$ hypothesized value, compute the two-sample *t* statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

Find the *P*-value by calculating the probability of getting a *t* statistic this large or larger in the direction specified by the alternative hypothesis H_a. Use the *t* distribution with degrees of freedom approximated by technology or the smaller of $n_1 - 1$ and $n_2 - 1$.



Here's an example that shows how to perform a two-sample *t* test for a randomized experiment.

EXAMPLE Calcium and Blood Pressure

Comparing two means

STEP 4

Does increasing the amount of calcium in our diet reduce blood pressure? Examination of a large sample of people revealed a relationship between calcium intake and blood pressure. The relationship was strongest for black men. Such observational studies do not establish causation. Researchers therefore designed a randomized comparative experiment.

The subjects were 21 healthy black men who volunteered to take part in the experiment. They were randomly assigned to two groups: 10 of the men received a calcium supplement for 12 weeks, while the control group of 11 men received a placebo pill that looked identical. The experiment was double-blind. The response variable is the decrease in systolic (top number) blood pressure for a subject after 12 weeks, in millimeters of mercury. An increase appears as a negative number.[24] Here are the data:

Group 1 (calcium):	7	−4	18	17	−3	−5	1	10	11	−2	
Group 2 (placebo):	−1	12	−1	−3	3	−5	5	2	−11	−1	−3

PROBLEM:

(a) Do the data provide convincing evidence that a calcium supplement reduces blood pressure more than a placebo? Carry out an appropriate test to support your answer.

(b) Interpret the *P*-value you got in part (a) in the context of this experiment.

SOLUTION:

(a) STATE: We want to perform a test of

$$H_0: \mu_1 - \mu_2 = 0 \qquad \qquad H_0: \mu_1 = \mu_2$$
$$\text{or, equivalently,}$$
$$H_a: \mu_1 - \mu_2 > 0 \qquad \qquad H_a: \mu_1 > \mu_2$$

where μ_1 is the true mean decrease in systolic blood pressure for healthy black men like the ones in this study who take a calcium supplement and μ_2 is the true mean decrease in systolic blood pressure for healthy black men like the ones in this study who take a placebo. No significance level was specified, so we'll use $\alpha = 0.05$.

PLAN: If conditions are met, we will carry out a two sample *t* test for $\mu_1 - \mu_2$.

- *Random:* The 21 subjects were randomly assigned to the two treatments.
  - 10%: Don't need to check because there was no sampling.
- *Normal/Large Sample:* With such small sample sizes, we need to graph the data to see if it's reasonable to believe that the actual distributions of differences in blood pressure when taking calcium or placebo are Normal. Figure 10.9 shows hand sketches of calculator boxplots for these data. The graphs show no strong skewness and no outliers. So we are safe using two-sample *t* procedures.

AP® EXAM TIP When checking the Normal condition on an AP® exam question involving inference about means, be sure to include graphs. Don't expect to receive credit for describing graphs that you made on your calculator but didn't put on paper.

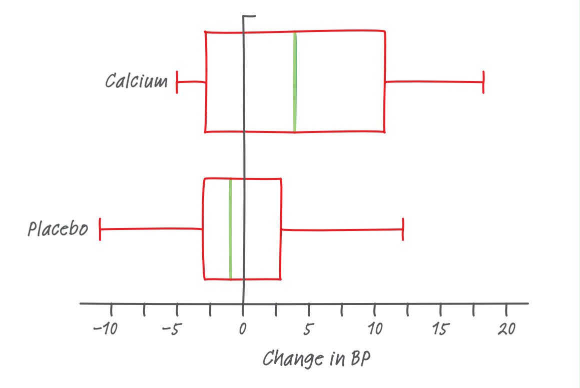


FIGURE 10.9 Sketches of boxplots of the changes in blood pressure for the two groups of subjects in the calcium and blood pressure experiment.

DO: From the data, we calculated summary statistics:

Group	Treatment	*n*	$\bar{x}$	s_x
1	Calcium	10	5.000	8.743
2	Placebo	11	−0.273	5.901

- *Test statistic*

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{[5.000 - (-0.273)] - 0}{\sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}}} = \frac{5.273}{3.2878} = 1.604$$

- *P-value* By the conservative method, the smaller of $n_1 - 1$ and $n_2 - 1$ gives df = 9. Because H_a counts only positive values of *t* as evidence against H_0, the *P*-value is the area to the right of $t = 1.604$ under the *t* distribution curve with df = 9. Figure 10.10 illustrates this *P*-value. Table B shows that the *P*-value lies between 0.05 and 0.10.

	Upper-tail probability *p*		
df	.10	.05	.025
8	1.397	1.860	2.306
9	1.383	1.833	2.262
10	1.372	1.812	2.228

Using technology: Refer to the Technology Corner that follows the example. The calculator's `2-SampTTest` gives $t = 1.60$ and *P*-value = 0.0644 using df = 15.59.

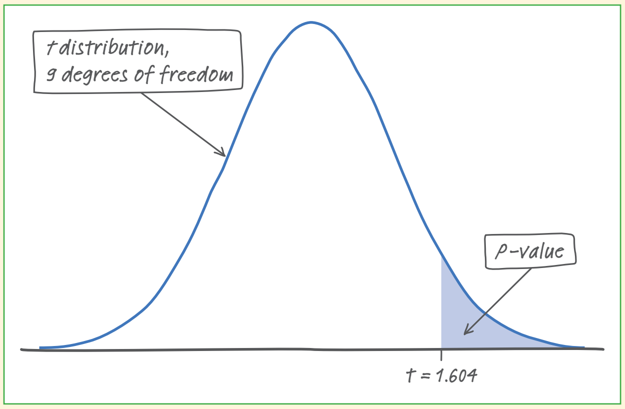


FIGURE 10.10 The *P*-value for the one-sided test using the conservative method, which leads to the *t* distribution with 9 degrees of freedom.

CONCLUDE: Because the *P*-value is greater than $\alpha = 0.05$, we fail to reject H_0. The experiment does not provide convincing evidence that the true mean decrease in systolic blood pressure is higher for men like these who take calcium than for men like these who take a placebo.

(b) Assuming H_0: $\mu_1 - \mu_2 = 0$ is true, there is a 0.0644 probability of getting a difference in mean blood pressure reduction for the two groups (calcium − placebo) of 5.273 or greater just by the chance involved in the random assignment.

For Practice *Try Exercise* **41**

When a significance test leads to a fail to reject H_0 decision, as in the previous example, be sure to interpret the results as "We don't have convincing evidence to conclude H_a." Saying anything that sounds like you believe H_0 is (or might be) true is incorrect.

THINK ABOUT IT

Why didn't researchers find a significant difference in the calcium and blood pressure experiment? The difference in mean systolic blood pressures for the two groups was 5.273 millimeters of mercury. This seems like a fairly large difference. With the small group sizes, however, this difference wasn't large enough to reject H_0: $\mu_1 - \mu_2 = 0$ in favor of the one-sided alternative. We suspect that larger groups might show a similar difference in mean blood pressure reduction, which would indicate that calcium has a significant effect. If so, then the researchers in this experiment made a Type II error—failing to reject a false H_0. In fact, later analysis of data from an experiment with more subjects resulted in a *P*-value of 0.008. *Sample size strongly affects the power of a test.* It is easier to detect an actual difference in the effectiveness of two treatments if both are applied to large numbers of subjects.

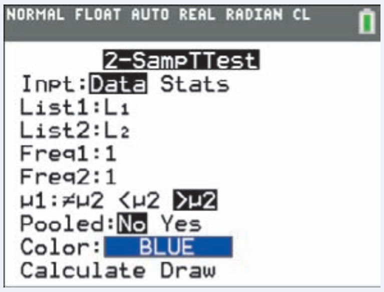
24. TECHNOLOGY CORNER TWO-SAMPLE *t* TESTS ON THE CALCULATOR

TI-Nspire instructions in Appendix B; HP Prime instructions on the book's Web site.

Technology gives smaller *P*-values for two-sample *t* tests than the conservative method. That's because calculators and software use the more complicated formula on page 640 to obtain a larger number of degrees of freedom.

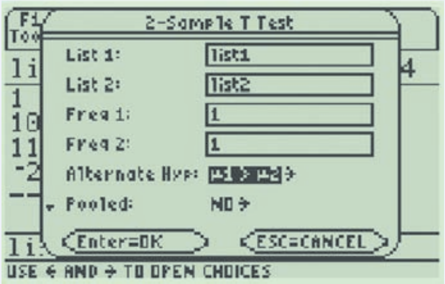
- Enter the Group 1 (calcium) data in L1/list1 and the Group 2 (placebo) data in L2/list2.
- To perform the significance test, go to STAT/TESTS (Tests menu in the Statistics/List Editor on the TI-89) and choose `2-SampTTest`.
- In the 2-SampTTest screen, specify "Data" and adjust your other settings as shown.

TI-83/84



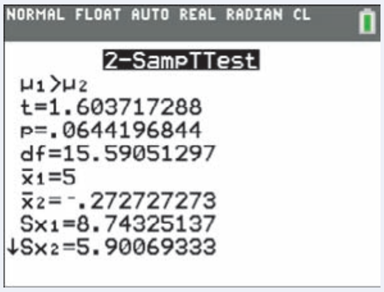
```
NORMAL FLOAT AUTO REAL RADIAN CL
     2-SampTTest
Inpt:Data Stats
List1:L1
List2:L2
Freq1:1
Freq2:1
μ1:≠μ2 <μ2 >μ2
Pooled:No Yes
Color: BLUE
Calculate Draw
```

TI-89

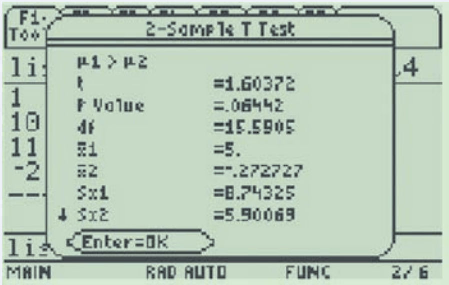


```
2-Sample T Test
List 1:            list1
List 2:            list2
Freq 1:            1
Freq 2:            1
Alternate Hyp:     μ1 > μ2 →
Pooled:            NO →
Enter=OK           ESC=CANCEL
USE ← AND → TO OPEN CHOICES
```

- Highlight "Calculate" and press ENTER. (The Pooled option will be discussed shortly.)

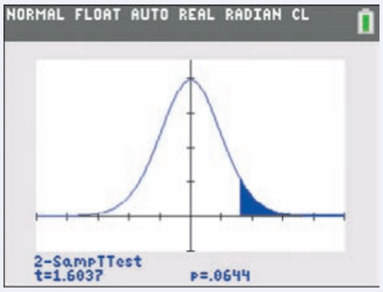


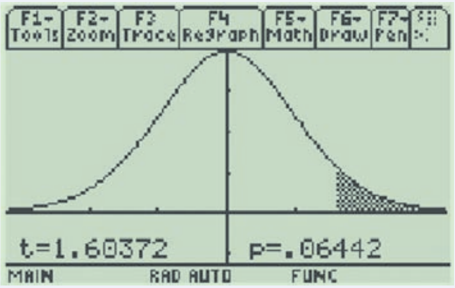
```
NORMAL FLOAT AUTO REAL RADIAN CL
     2-SampTTest
μ1>μ2
t=1.603717288
p=.0644196844
df=15.59051297
x̄1=5
x̄2=⁻.272727273
Sx1=8.74325137
↓Sx2=5.90069333
```



```
2-Sample T Test
μ1 > μ2
t                  =1.60372
P Value            =.06442
df                 =15.5905
x̄1                 =5.
x̄2                 =⁻.272727
Sx1                =8.74325
↓ Sx2              =5.90069
Enter=OK
MAIN         RAD AUTO        FUNC        2/6
```

If you select "Draw" instead of "Calculate," the appropriate *t* distribution will be displayed, showing the test statistic and the shaded area corresponding to the *P*-value.





AP® EXAM TIP The formula for the two-sample *t* statistic for a test about $\mu_1 - \mu_2$ often leads to calculation errors by students. As a result, we recommend using the calculator's `2-SampTTest` feature to perform calculations on the AP® exam. Be sure to name the procedure (two-sample *t* test) and to report the test statistic ($t = 1.60$), *P*-value (0.0644), and df (15.59) as part of the "Do" step.

Inference for Experiments Confidence intervals and tests for $\mu_1 - \mu_2$ are based on the sampling distribution of $\bar{x}_1 - \bar{x}_2$. But in experiments, we aren't sampling at random from any larger populations. We can think about what would happen if the random assignment were repeated many times under the assumption that $H_0\colon \mu_1 - \mu_2 = 0$ is true. That is, we assume that the specific treatment received doesn't affect an individual subject's response.

Let's see what would happen just by chance if we randomly reassign the 21 subjects in the calcium and blood pressure experiment to the two groups many times, assuming the drug received *doesn't affect* each individual's change in systolic blood pressure. We used Fathom software to redo the random assignment 1000 times. The approximate *randomization distribution* of $\bar{x}_1 - \bar{x}_2$ is shown in Figure 10.11. It has an approximately Normal shape with mean 0 (no difference) and standard deviation 3.42.

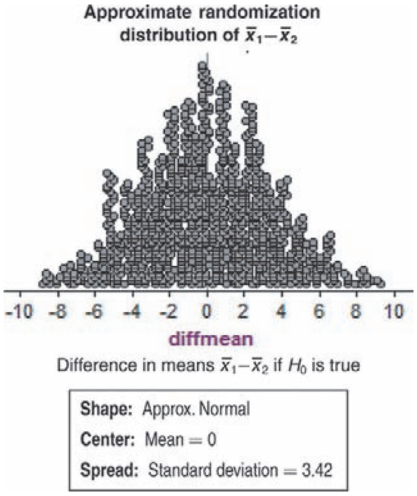


FIGURE 10.11 Fathom simulation showing the approximate randomization distribution of $\bar{x}_1 - \bar{x}_2$ from 1000 random reassignments of subjects to treatment groups in the calcium and blood pressure experiment.

In the actual experiment, the difference in the mean change in blood pressure in the calcium and placebo groups was $5.000 - (-0.273) = 5.273$. How likely is it that a difference this large or larger would happen just by chance when H_0 is true? Figure 10.11 provides a rough answer: 61 of the 1000 random reassignments yielded a difference in means greater than or equal to 5.273. That is, our estimate of the *P*-value is 0.061. This is quite close to the 0.0644 *P*-value that we obtained in the Technology Corner.

If Figure 10.11 displayed the results of *all* possible random reassignments of subjects to treatment groups, it would be the actual randomization distribution of $\bar{x}_1 - \bar{x}_2$. The *P*-value obtained from this distribution would be *exactly correct*. Using the two-sample *t* test to calculate the *P*-value gives only approximately correct results.

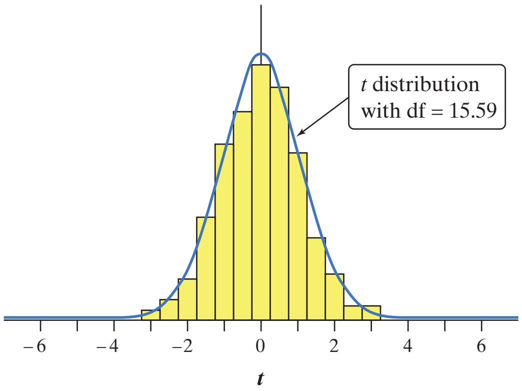


FIGURE 10.12 The distribution of the two-sample *t* test statistic for the 1000 random reassignments in Figure 10.11.

Figure 10.12 shows the value of the two-sample *t* test statistic for each of the 1000 re-randomizations, calculated using our familiar formula

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The density curve for the *t* distribution with df = 15.59 is shown in blue. We can see that the test statistic follows the *t* distribution quite closely in this case.

Whenever the conditions are met, the randomization distribution of $\bar{x}_1 - \bar{x}_2$ looks much like its sampling distribution. We are therefore safe using two-sample *t* procedures for comparing two means in a randomized experiment.

CHECK YOUR UNDERSTANDING

How quickly do synthetic fabrics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay. Lower strength means the fabric has decayed.

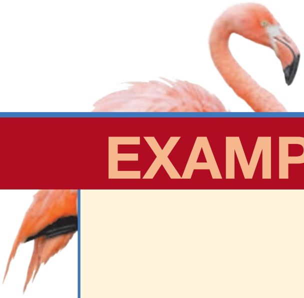
For one part of the study, the researcher buried 10 strips of polyester fabric in well-drained soil in the summer. The strips were randomly assigned to two groups: 5 of them were buried for 2 weeks and the other 5 were buried for 16 weeks. Here are the breaking strengths in pounds:[25]

Group 1 (2 weeks):	118	126	126	120	129
Group 2 (16 weeks):	124	98	110	140	110

Do the data give convincing evidence that polyester decays more in 16 weeks than in 2 weeks?

Using Two-Sample *t* Procedures Wisely

In Chapter 9, we used paired *t* procedures to compare the mean change in depression scores for a group of caffeine-dependent individuals when taking caffeine and a placebo. The inference involved paired data because the same 11 subjects received both treatments. In this chapter, we used two-sample *t* procedures to compare the mean change in blood pressure for a group of healthy black men when taking calcium and a placebo. This time, the inference involved two distinct groups of subjects. *The proper method of analysis depends on the design of the study.*



EXAMPLE Comparing Tires and Comparing Workers

Independent samples versus paired data

PROBLEM: In each of the following settings, decide whether you should use paired *t* procedures or two-sample *t* procedures to perform inference.[26] Explain your choice.

(a) To test the wear characteristics of two tire brands, A and B, one Brand A tire is mounted on one side of each car in the rear, while a Brand B tire is mounted on the other side. Which side gets which brand is determined by flipping a coin.

(b) Can listening to music while working increase productivity? Twenty factory workers agree to take part in a study to investigate this question. Researchers randomly assign 10 workers to do a repetitive task while listening to music and the other 10 workers to do the task in silence.

SOLUTION:

(a) Paired *t* procedures. This is a matched pairs experiment, with the two treatments (Brand A and Brand B) being randomly assigned to the rear pair of wheels on each car.

(b) Two-sample *t* procedures. The data are being produced using two distinct groups of workers in a randomized experiment.

For Practice *Try Exercise* **53**

The same logic applies when data are produced by random sampling. If independent random samples are taken from each of two populations, we should use two-sample *t* procedures to perform inference about $\mu_1 - \mu_2$ if conditions are met. If one random sample is taken, and two data values are recorded for each individual, we should use paired *t* procedures to perform inference about the population mean difference μ_D if conditions are met.

The Pooled Two-Sample *t* Procedures (Don't Use Them!) Most software offers a choice of two-sample *t* statistics. One is often labeled "unequal" variances; the other, "equal" variances. The "unequal" variance procedure uses our two-sample *t* statistic. *This test is valid whether or not the population variances are equal.*

The other choice is a special version of the two-sample t statistic that assumes that the two populations have the same variance. This procedure combines (the statistical term is *pools*) the two sample variances to estimate the common population variance. The resulting statistic is called the *pooled two-sample t statistic*.

The pooled t statistic has exactly the t distribution with $n_1 + n_2 - 2$ degrees of freedom *if* the two population variances really are equal and the population distributions are exactly Normal. This method offers more degrees of freedom than Option 1 (technology), which leads to narrower confidence intervals and smaller P-values. The pooled t procedures were in common use before software made it easy to use Option 1 for our two-sample t statistic.

In the real world, distributions are not exactly Normal, and population variances are not exactly equal. In practice, the Option 1 two-sample t procedures are almost always more accurate than the pooled procedures. Our advice: *Never use the pooled t procedures if you have software that will carry out Option 1.*

Remember, we always use the pooled sample proportion $\hat{p}_C$ when performing a significance test for comparing two proportions. But we don't recommend pooling when comparing two means.

caution

case closed

Fast-Food Frenzy!

Let's return to the chapter-opening Case Study (page 609) about drive-thru service at fast-food restaurants. Here, once again, are some results from the 2012 QSR study.

- For restaurants with order-confirmation boards, 1169 of 1327 visits (88.1%) resulted in accurate orders. For restaurants with no order-confirmation board, 655 of 726 visits (90.2%) resulted in accurate orders.
- McDonald's average service time for 362 drive-thru visits was 188.83 seconds with a standard deviation of 17.38 seconds. Burger King's service time for 318 drive-thru visits had a mean of 201.33 seconds and a standard deviation of 18.85 seconds.

You are now ready to use what you have learned about comparing population parameters to perform inference about accuracy and average service time in the drive-thru lane.

1. Is there a significant difference in order accuracy between restaurants with and without order-confirmation boards? Carry out an appropriate test at the $\alpha = 0.05$ level to help answer this question.

A 95% confidence interval for the difference in the population proportions of accurate orders at restaurants with and without order-confirmation boards is (−0.049, 0.00649).

2. Interpret the meaning of "95% confident" in the context of this study.

3. Explain how the confidence interval is consistent with your conclusion from Question 1.

Now turn your attention to the speed-of-service data.

4. Construct and interpret a 99% confidence interval for the difference in the mean service times at McDonald's and Burger King drive-thrus.

Section 10.2 Summary

- Choose independent SRSs of size n_1 from Population 1 and size n_2 from Population 2. The sampling distribution of $\bar{x}_1 - \bar{x}_2$ has the following properties:
    - **Shape:** Normal if both population distributions are Normal; approximately Normal otherwise if both samples are large enough ($n_1 \geq 30$ and $n_2 \geq 30$) by the central limit theorem.
    - **Center:** Its mean is $\mu_1 - \mu_2$.
    - **Spread:** As long as each sample is no more than 10% of its population, its standard deviation is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
- Confidence intervals and tests for the difference between the means of two populations or the mean responses to two treatments μ_1 and μ_2 are based on the difference $\bar{x}_1 - \bar{x}_2$ between the sample means.
- Because we almost never know the population standard deviations in practice, we use the **two-sample *t* statistic**

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

  This statistic has approximately a *t* distribution. There are two options for using a *t* distribution to approximate the distribution of the two-sample *t* statistic:
    - **Option 1 (Technology)** Use the *t* distribution with degrees of freedom calculated from the data by a somewhat messy formula. The degrees of freedom probably won't be a whole number.
    - **Option 2 (Conservative)** Use the *t* distribution with degrees of freedom equal to the *smaller* of $n_1 - 1$ and $n_2 - 1$. This method gives wider confidence intervals and larger *P*-values than Option 1.
- Before estimating or testing a claim about $\mu_1 - \mu_2$, check that these conditions are met:
    - **Random:** The data are produced by independent random samples of size n_1 from Population 1 and of size n_2 from Population 2 or by two groups of size n_1 and n_2 in a randomized experiment.

- ○ **10%:** When sampling without replacement, check that the two populations are at least 10 times as large as the corresponding samples.
- **Normal/Large Sample:** Both population distributions (or the true distributions of responses to the two treatments) are Normal or both sample sizes are large ($n_1 \geq 30$ and $n_2 \geq 30$). If either population (treatment) distribution has unknown shape and the corresponding sample size is less than 30, use a graph of the sample data to assess the Normality of the population (treatment) distribution. Do not use two-sample t procedures if the graph shows strong skewness or outliers.

- An approximate C% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t^* is the critical value with C% of its area between $-t^*$ and t^* for the t distribution with degrees of freedom from either Option 1 (technology) or Option 2 (the smaller of $n_1 - 1$ and $n_2 - 1$). This is called a **two-sample *t* interval for $\boldsymbol{\mu_1 - \mu_2}$**.

- To test H_0: $\mu_1 - \mu_2$ = hypothesized value, use a **two-sample *t* test for $\boldsymbol{\mu_1 - \mu_2}$**. The test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

P-values are calculated using the t distribution with degrees of freedom from either Option 1 (technology) or Option 2 (the smaller of $n_1 - 1$ and $n_2 - 1$).

- Inference about the difference $\mu_1 - \mu_2$ in the effectiveness of two treatments in a completely randomized experiment is based on the **randomization distribution** of $\bar{x}_1 - \bar{x}_2$. When the conditions are met, our usual inference procedures based on the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will be approximately correct.
- Don't use two-sample t procedures to compare means for paired data.
- STEP 4 Be sure to follow the four-step process whenever you construct a confidence interval or perform a significance test for comparing two means.

10.2 TECHNOLOGY CORNERS

TI-Nspire Instructions in Appendix B; HP Prime instructions on the book's Web site.

23. Two-sample *t* intervals on the calculator page 643
24. Two-sample *t* tests on the calculator page 647

Section 10.2 Exercises

STEP 4 *Remember: We are no longer reminding you to use the four-step process in exercises that require you to perform inference.*

31. **Cholesterol** The level of cholesterol in the blood for all men aged 20 to 34 follows a Normal distribution with mean 188 milligrams per deciliter (mg/dl) and standard deviation 41 mg/dl. For 14-year-old boys, blood cholesterol levels follow a Normal distribution with mean 170 mg/dl and standard deviation 30 mg/dl. Suppose we select independent SRSs of 25 men aged 20 to 34 and 36 boys aged 14 and calculate the sample mean cholesterol levels $\bar{x}_M$ and $\bar{x}_B$.

pg 638

(a) What is the shape of the sampling distribution of $\bar{x}_M - \bar{x}_B$? Why?

(b) Find the mean of the sampling distribution. Show your work.

(c) Find the standard deviation of the sampling distribution. Show your work.

32. **How tall?** The heights of young men follow a Normal distribution with mean 69.3 inches and standard deviation 2.8 inches. The heights of young women follow a Normal distribution with mean 64.5 inches and standard deviation 2.5 inches. Suppose we select independent SRSs of 16 young men and 9 young women and calculate the sample mean heights $\bar{x}_M$ and $\bar{x}_W$.

(a) What is the shape of the sampling distribution of $\bar{x}_M - \bar{x}_W$? Why?

(b) Find the mean of the sampling distribution. Show your work.

(c) Find the standard deviation of the sampling distribution. Show your work.

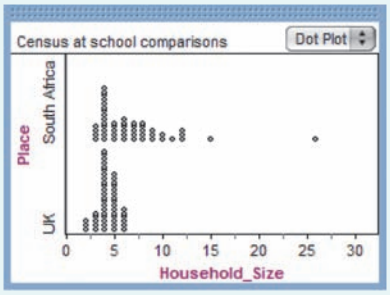
In Exercises 33 to 36, determine whether or not the conditions for using two-sample t procedures are met.

33. **Shoes** How many pairs of shoes do teenagers have? To find out, a group of AP® Statistics students conducted a survey. They selected a random sample of 20 female students and a separate random sample of 20 male students from their school. Then they recorded the number of pairs of shoes that each respondent reported having. The back-to-back stemplot displays the data.

Females		Males
	0	4
	0	555677778
333	1	0000124
95	1	
4332	2	2
66	2	
410	3	
8	3	58
	4	
9	4	
100	5	
7	5	

Key: 2|2 represents a male student with 22 pairs of shoes.

34. **Household size** How do the numbers of people living in households in the United Kingdom (U.K.) and South Africa compare? To help answer this question, we used CensusAtSchool's random data selector to choose independent samples of 50 students from each country. Here is a Fathom dotplot of the household sizes reported by the students in the survey.

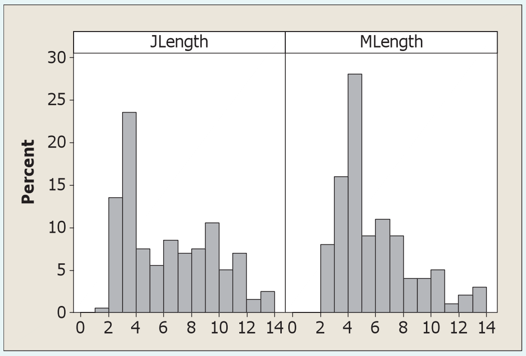


35. **Literacy rates** Do males have higher average literacy rates than females in Islamic countries? The table below shows the percent of men and women who were literate in the major Islamic nations at the time of this writing.[27] (We omitted countries with populations of less than 3 million.)

Country	Male (%)	Female (%)
Afghanistan	43	13
Algeria	80	60
Azerbaijan	99.9	99.7
Bangladesh	61	52
Egypt	80	64
Indonesia	94	86.8
Iran	84	70
Iraq	86	71
Jordan	96	89
Kazakhstan	100	99
Kyrgyzstan	99.3	98.1
Lebanon	93	82
Libya	96	83
Malaysia	92	85
Morocco	69	44
Pakistan	68.6	30.3
Saudi Arabia	90	81

Country	Male (%)	Female (%)
Syria	86	74
Tajikistan	100	100
Tunisia	83	65
Turkey	98	90
Turkmenistan	99.3	98.3
Uzbekistan	100	99
Yemen	81	47

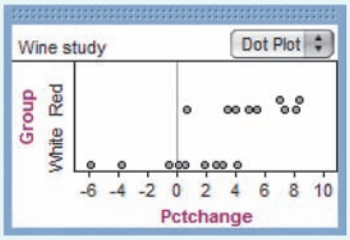
36. **Long words** Mary was interested in comparing the mean word length in articles from a medical journal and an airline's in-flight magazine. She counted the number of letters in the first 400 words of an article in the medical journal and in the first 100 words of an article in the airline magazine. Mary then used Minitab statistical software to produce the histograms shown. Note that J is for journal and M is for magazine.



37. pg 641 **Is red wine better than white wine?** Observational studies suggest that moderate use of alcohol by adults reduces heart attacks and that red wine may have special benefits. One reason may be that red wine contains polyphenols, substances that do good things to cholesterol in the blood and so may reduce the risk of heart attacks. In an experiment, healthy men were assigned at random to drink half a bottle of either red or white wine each day for two weeks. The level of polyphenols in their blood was measured before and after the two-week period. Here are the percent changes in level for the subjects in both groups:[28]

Red wine:	3.5	8.1	7.4	4.0	0.7	4.9	8.4	7.0	5.5
White wine:	3.1	0.5	−3.8	4.1	−0.6	2.7	1.9	−5.9	0.1

(a) A Fathom dotplot of the data is shown below. Write a few sentences comparing the distributions.



(b) Construct and interpret a 90% confidence interval for the difference in mean percent change in polyphenol levels for the red wine and white wine treatments.

(c) Does the interval in part (b) suggest that red wine is more effective than white wine? Explain.

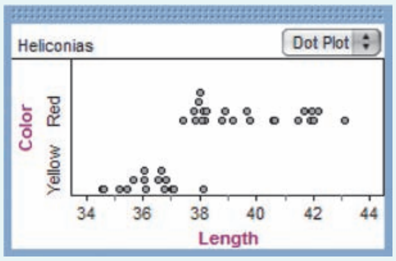
38. **Tropical flowers** Different varieties of the tropical flower *Heliconia* are fertilized by different species of hummingbirds.

Researchers believe that over time, the lengths of the flowers and the forms of the hummingbirds' beaks have evolved to match each other. Here are data on the lengths in millimeters for random samples of two color varieties of the same species of flower on the island of Dominica:[29]

H. *caribaea* red							
41.90	42.01	41.93	43.09	41.17	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	

H. *caribaea* yellow							
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	

(a) A Fathom dotplot of the data is shown below. Write a few sentences comparing the distributions.



(b) Construct and interpret a 95% confidence interval for the difference in the mean lengths of these two varieties of flowers.

(c) Does the interval support the researchers' belief that the two flower varieties have different average lengths? Explain.

39. **Paying for college** College financial aid offices expect students to use summer earnings to help pay for college. But how large are these earnings? One large university studied this question by asking a random sample of 1296 students who had summer jobs how much they earned. The financial aid office separated the responses into two groups based on gender. Here are the data in summary form:[30]

Group	n	$\bar{x}$	s_x
Males	675	\$1884.52	\$1368.37
Females	621	\$1360.39	\$1037.46

(a) How can you tell from the summary statistics that the distribution of earnings in each group is strongly skewed to the right? The use of two-sample *t* procedures is still justified. Why?

(b) Construct and interpret a 90% confidence interval for the difference between the mean summer earnings of male and female students at this university.

(c) Interpret the 90% confidence level in the context of this study.

40. **Happy customers** As the Hispanic population in the United States has grown, businesses have tried to understand what Hispanics like. One study interviewed a random sample of customers leaving a bank. Customers were classified as Hispanic if they preferred to be interviewed in Spanish or as Anglo if they preferred English. Each customer rated the importance of several aspects of bank service on a 10-point scale.[31] Here are summary results for the importance of "reliability" (the accuracy of account records and so on):

Group	n	$\bar{x}$	s_x
Anglo	92	6.37	0.60
Hispanic	86	5.91	0.93

(a) The distribution of reliability ratings in each group is not Normal. The use of two-sample *t* procedures is still justified. Why?

(b) Construct and interpret a 95% confidence interval for the difference between the mean ratings of the importance of reliability for Anglo and Hispanic bank customers.

(c) Interpret the 95% confidence level in the context of this study.

41. pg 645 **Baby birds** Do birds learn to time their breeding? Blue titmice eat caterpillars. The birds would like lots of caterpillars around when they have young to feed, but they must breed much earlier. Do the birds learn from one year's experience when to time their breeding next year? Researchers randomly assigned 7 pairs of birds to have the natural caterpillar supply supplemented while feeding their young and another 6 pairs to serve as a control group relying on natural food supply. The next year, they measured how many days after the caterpillar peak the birds produced their nestlings.[32] The investigators expected the control group to adjust their breeding date the next year, whereas the well-fed supplemented group had no reason to change. Here are the data (days after caterpillar peak):

Control:	4.6	2.3	7.7	6.0	4.6	−1.2	
Supplemented:	15.5	11.3	5.4	16.5	11.3	11.4	7.7

(a) Do the data provide convincing evidence to confirm the researchers' belief?

(b) Interpret the *P*-value from part (a) in the context of this study.

42. **DDT in rats** Poisoning by the pesticide DDT causes convulsions in humans and other mammals. Researchers seek to understand how the convulsions are caused. In a randomized comparative experiment, they compared 6 white rats poisoned with DDT with a control group of 6 unpoisoned rats. Electrical measurements of nerve activity are the main clue to the nature of DDT poisoning. When a nerve is stimulated, its electrical response shows a sharp spike followed by a much smaller second spike. The researchers measured the height of the second spike as a percent of the first spike when a nerve in the rat's leg was stimulated.[33] For the poisoned rats, the results were

12.207 16.869 25.050 22.429 8.456 20.589

The control group data were

11.074 9.686 12.064 9.351 8.182 6.642

(a) Do these data provide convincing evidence that DDT affects the mean relative height of the second spike's electrical response?

(b) Interpret the *P*-value from part (a) in the context of this study.

43. **Who talks more—men or women?** Researchers equipped random samples of 56 male and 56 female students from a large university with a small device that secretly records sound for a random 30 seconds during each 12.5-minute period over two days. Then they counted the number of words spoken by each subject during each recording period and, from this, estimated how many words per day each subject speaks. The female

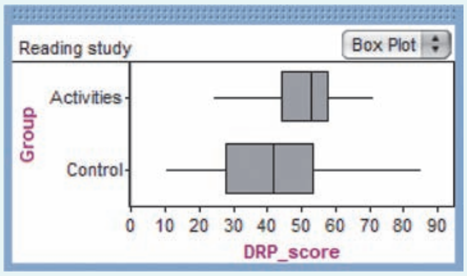
estimates had a mean of 16,177 words per day with a standard deviation of 7520 words per day. For the male estimates, the mean was 16,569 and the standard deviation was 9108. Do these data provide convincing evidence of a difference in the average number of words spoken in a day by male and female students at this university?

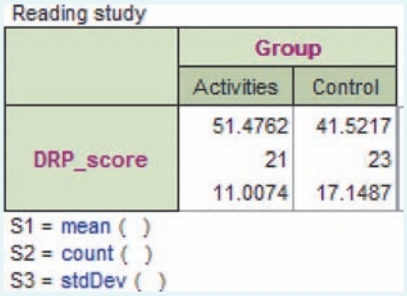
44. **Competitive rowers** What aspects of rowing technique distinguish between novice and skilled competitive rowers? Researchers compared two randomly selected groups of female competitive rowers: a group of skilled rowers and a group of novices. The researchers measured many mechanical aspects of rowing style as the subjects rowed on a Stanford Rowing Ergometer. One important variable is the angular velocity of the knee, which describes the rate at which the knee joint opens as the legs push the body back on the sliding seat. The data show no outliers or strong skewness. Here is the SAS computer output:[34]

```
                  TTEST PROCEDURE
Variable: KNEE
GROUP        N        Mean        Std Dev    Std Error
SKILLED      10       4.182       0.479      0.151
NOVICE       8        3.010       0.959      0.339
```

The researchers believed that the knee velocity would be higher for skilled rowers. Do the data provide convincing evidence to support this belief?

45. **Teaching reading** An educator believes that new reading activities in the classroom will help elementary school pupils improve their reading ability. She recruits 44 third-grade students and randomly assigns them into two groups. One group of 21 students does these new activities for an 8-week period. A control group of 23 third-graders follows the same curriculum without the activities. At the end of the 8 weeks, all students are given the Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the treatment is designed to improve. Comparative boxplots and summary statistics for the data from Fathom are shown below.[35]




Reading study

	Group	
	Activities	Control
DRP_score	51.4762	41.5217
	21	23
	11.0074	17.1487

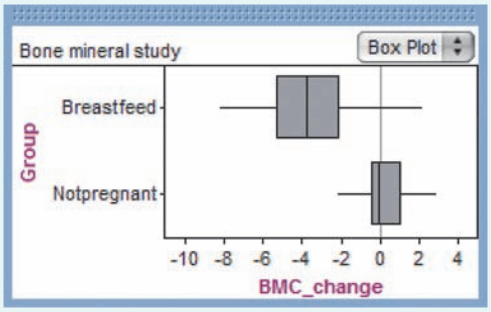
S1 = mean ( )
S2 = count ( )
S3 = stdDev ( )

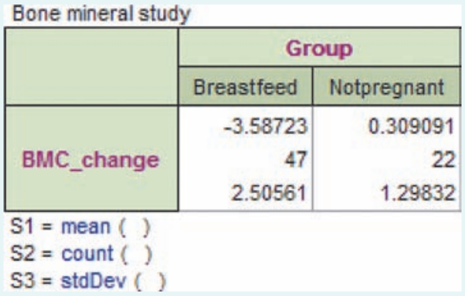
(a) Based on the graph and numerical summaries, write a few sentences comparing the DRP scores for the two groups.

(b) Is the mean DRP score significantly higher for the students who did the reading activities? Give appropriate evidence to justify your answer.

(c) Can we conclude that the new reading activities caused an increase in the mean DRP score? Explain.

46. **Does breast-feeding weaken bones?** Breast-feeding mothers secrete calcium into their milk. Some of the calcium may come from their bones, so mothers may lose bone mineral. Researchers compared a random sample of 47 breast-feeding women with a random sample of 22 women of similar age who were neither pregnant nor lactating. They measured the percent change in the bone mineral content (BMC) of the women's spines over three months. Comparative boxplots and summary statistics for the data from Fathom are shown below.[36]




Bone mineral study

	Group	
	Breastfeed	Notpregnant
BMC_change	-3.58723	0.309091
	47	22
	2.50561	1.29832

S1 = mean ( )
S2 = count ( )
S3 = stdDev ( )

(a) Based on the graph and numerical summaries, write a few sentences comparing the percent changes in BMC for the two groups.

(b) Is the mean change in BMC significantly lower for the mothers who are breast-feeding? Give appropriate evidence to justify your answer.

(c) Can we conclude that breast-feeding causes a mother's bones to weaken? Why or why not?

47. **Who talks more—men or women?** Refer to Exercise 43. Construct and interpret a 95% confidence interval for the difference in mean number of words spoken in a day. Explain how this interval provides more information than the significance test in Exercise 43.

48. **DDT in rats** Refer to Exercise 42. Construct and interpret a 95% confidence interval for the difference in mean relative height of the second spike's electrical response. Explain how this interval provides more information than the significance test in Exercise 42.

49. **A better drug?** In a pilot study, a company's new cholesterol-reducing drug outperforms the currently available drug. If the data provide convincing evidence that the mean cholesterol reduction with the new drug is more than 10 milligrams per deciliter of blood (mg/dl) greater than with the current drug, the company will begin the expensive process of mass-producing the new drug. For the 14 subjects who were assigned at random to the current drug, the mean cholesterol reduction was 54.1 mg/dl with a standard deviation of 11.93 mg/dl. For the 15 subjects who were randomly assigned to the new drug, the mean cholesterol reduction was 68.7 mg/dl with a standard deviation of 13.3 mg/dl. Graphs of the data reveal no outliers or strong skewness.

(a) Carry out an appropriate significance test. What conclusion would you draw? (Note that the null hypothesis is *not* H_0: $\mu_1 - \mu_2 = 0$.)

(b) Based on your conclusion in part (a), could you have made a Type I error or a Type II error? Justify your answer.

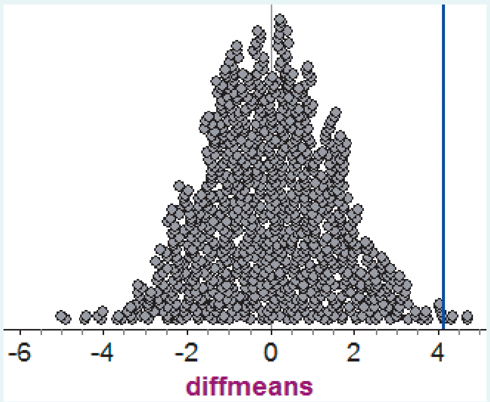
50. **Down the toilet** A company that makes hotel toilets claims that its new pressure-assisted toilet reduces the average amount of water used by more than 0.5 gallon per flush when compared to its current model. To test this claim, the company randomly selects 30 toilets of each type and measures the amount of water that is used when each toilet is flushed once. For the current-model toilets, the mean amount of water used is 1.64 gal with a standard deviation of 0.29 gal. For the new toilets, the mean amount of water used is 1.09 gal with a standard deviation of 0.18 gal.

(a) Carry out an appropriate significance test. What conclusion would you draw? (Note that the null hypothesis is *not* H_0: $\mu_1 - \mu_2 = 0$.)

(b) Based on your conclusion in part (a), could you have made a Type I error or a Type II error? Justify your answer.

51. **Rewards and creativity** Dr. Teresa Amabile conducted a study involving 47 college students who were randomly assigned to two treatment groups. The 23 students in one group were given a list of statements about external reasons (E) for writing, such as public recognition, making money, or pleasing their parents. The 24 students in the other group were given a list of statements about internal reasons (I) for writing, such as expressing yourself and enjoying playing with words. Both groups were then instructed to write a poem about laughter. Each student's poem was rated separately by 12 different poets using a creativity scale.[37] The 12 poets' ratings of each student's poem were averaged to obtain an overall creativity score.

We used Fathom software to randomly reassign the 47 subjects to the two groups 1000 times, assuming the treatment received doesn't affect each individual's average creativity rating. The dotplot shows the approximate randomization distribution of $\bar{x}_I - \bar{x}_E$.



(a) Why did researchers randomly assign the subjects to the two treatment groups?

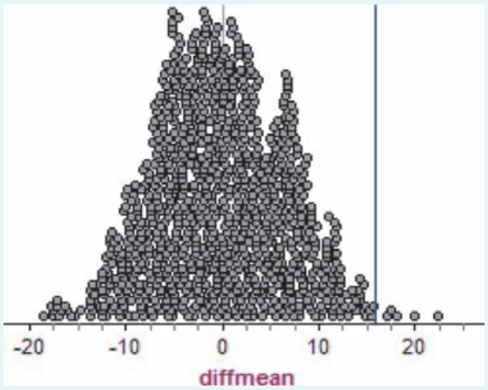
(b) In the actual experiment, $\bar{x}_I - \bar{x}_E = 4.15$. This value is marked with a blue line in the figure. What conclusion would you draw? Justify your answer with appropriate evidence.

(c) Based on your conclusion in part (b), could you have made a Type I error or a Type II error? Justify your answer.

52. **Sleep deprivation** Does sleep deprivation linger for more than a day? Researchers designed a study using 21 volunteer subjects between the ages of 18 and 25. All 21 participants took a computer-based visual discrimination test at the start of the study. Then the subjects were randomly assigned into two groups. The 11 subjects in one group, D, were deprived of sleep for an entire night in a laboratory setting. The

10 subjects in the other group, A, were allowed unrestricted sleep for the night. Both groups were allowed as much sleep as they wanted for the next two nights. On Day 4, all the subjects took the same visual discrimination test on the computer. Researchers recorded the improvement in time (measured in milliseconds) from Day 1 to Day 4 on the test for each subject.[38]

We used Fathom software to randomly reassign the 21 subjects to the two groups 1000 times, assuming the treatment received doesn't affect each individual's time improvement on the test. The dotplot shows the approximate randomization distribution of $\bar{x}_A - \bar{x}_D$.



(a) Explain why the researchers didn't let the subjects choose whether to be in the sleep deprivation group or the unrestricted sleep group.

(b) In the actual experiment, $\bar{x}_A - \bar{x}_D = 15.92$. This value is marked with a blue line in the figure. What conclusion would you draw? Justify your answer with appropriate evidence.

(c) Based on your conclusion in part (b), could you have made a Type I error or a Type II error? Justify your answer.

53. **Paired or unpaired?** In each of the following settings, decide whether you should use paired t procedures or two-sample t procedures to perform inference. Explain your choice.[39]

pg 650

(a) To test the wear characteristics of two tire brands, A and B, each brand of tire is randomly assigned to 50 cars of the same make and model.

(b) To test the effect of background music on productivity, factory workers are observed. For one month, each subject works without music. For another month, the subject works while listening to music on an MP3 player. The month in which each subject listens to music is determined by a coin toss.

(c) A study was designed to compare the effectiveness of two weight-reducing diets. Fifty obese women who volunteered to participate were randomly assigned into two equal-sized groups. One group used Diet A and the other used Diet B. The weight of each woman was measured before the assigned diet and again after 10 weeks on the diet.

54. **Paired or unpaired?** In each of the following settings, decide whether you should use paired t procedures or two-sample t procedures to perform inference. Explain your choice.[40]

(a) To compare the average weight gain of pigs fed two different rations, nine pairs of pigs were used. The pigs in each pair were littermates. A coin toss was used to decide which pig in each pair got Ration A and which got Ration B.

(b) Separate random samples of male and female college professors are taken. We wish to compare the average salaries of male and female teachers.

(c) To test the effects of a new fertilizer, 100 plots are treated with the new fertilizer, and 100 plots are treated with another fertilizer. A computer's random number generator is used to determine which plots get which fertilizer.

Exercises 55 and 56 refer to the following setting. Coaching companies claim that their courses can raise the SAT scores of high school students. Of course, students who retake the SAT without paying for coaching generally raise their scores. A random sample of students who took the SAT twice found 427 who were coached and 2733 who were uncoached.[41] Starting with their Verbal scores on the first and second tries, we have these summary statistics:

	Try 1			Try 2		Gain	
	n	$\bar{x}$	s_x	$\bar{x}$	s_x	$\bar{x}$	s_x
Coached	427	500	92	529	97	29	59
Uncoached	2733	506	101	527	101	21	52

55. **Coaching and SAT scores** Let's first ask if students who are coached increased their scores significantly.

(a) You could use the information on the Coached line to carry out either a two-sample t test comparing Try 1 with Try 2 for coached students or a paired t test using Gain. Which is the correct test? Why?

(b) Carry out the proper test. What do you conclude?

56. **Coaching and SAT scores** What we really want to know is whether coached students improve more than uncoached students, and whether any advantage is large enough to be worth paying for. Use the information above to answer these questions:

(a) How much more do coached students gain on the average? Construct and interpret a 99% confidence interval.

(b) Does the interval in part (a) give convincing evidence that coached students gain more, on average, than uncoached students? Explain.

(c) Based on your work, what is your opinion: do you think coaching courses are worth paying for?

Multiple choice: Select the best answer for Exercises 57 to 60.

57. There are two common methods for measuring the concentration of a pollutant in fish tissue. Do the two methods differ, on average? You apply both methods to each fish in a random sample of 18 carp and use

(a) the paired t test for $\mu_{d.}$

(b) the one-sample z test for p.

(c) the two-sample t test for $\mu_1 - \mu_2$.

(d) the two-sample z test for $p_1 - p_2$.

(e) none of these.

Exercises 58 to 60 refer to the following setting. A study of road rage asked random samples of 596 men and 523 women about their behavior while driving. Based on their answers, each person was assigned a road rage score on a scale of 0 to 20. The participants were chosen by random digit dialing of phone numbers. The researchers performed a test of the following hypotheses: H_0: $\mu_M = \mu_F$ versus H_a: $\mu_M \neq \mu_F$.

58. Which of the following describes a Type II error in the context of this study?

(a) Finding convincing evidence that the true means are different for males and females, when in reality the true means are the same

(b) Finding convincing evidence that the true means are different for males and females, when in reality the true means are different

(c) Not finding convincing evidence that the true means are different for males and females, when in reality the true means are the same

(d) Not finding convincing evidence that the true means are different for males and females, when in reality the true means are different

(e) Not finding convincing evidence that the true means are different for males and females, when in reality there is convincing evidence that the true means are different

59. The P-value for the stated hypotheses is 0.002. Interpret this value in the context of this study.

(a) Assuming that the true mean road rage score is the same for males and females, there is a 0.002 probability of getting a difference in sample means.

(b) Assuming that the true mean road rage score is the same for males and females, there is a 0.002 probability of getting an observed difference at least as extreme as the observed difference.

(c) Assuming that the true mean road rage score is different for males and females, there is a 0.002 probability of getting an observed difference at least as extreme as the observed difference.

(d) Assuming that the true mean road rage score is the same for males and females, there is a 0.002 probability that the null hypothesis is true.

(e) Assuming that the true mean road rage score is the same for males and females, there is a 0.002 probability that the alternative hypothesis is true.

60. Based on the P-value in Exercise 59, which of the following must be true?

(a) A 90% confidence interval for $\mu_M - \mu_F$ will contain 0.

(b) A 95% confidence interval for $\mu_M - \mu_F$ will contain 0.

(c) A 99% confidence interval for $\mu_M - \mu_F$ will contain 0.

(d) A 99.9% confidence interval for $\mu_M - \mu_F$ will contain 0.

(e) It is impossible to determine whether any of these statements is true based only on the P-value.

In each part of Exercises 61 and 62, state which inference procedure from Chapter 8, 9, or 10 you would use. Be specific. For example, you might say, "Two-sample z test for the difference between two proportions." You do not need to carry out any procedures.

61. **Which inference method?**

(a) Drowning in bathtubs is a major cause of death in children less than 5 years old. A random sample of parents was asked many questions related to bathtub safety. Overall, 85% of the sample said they used baby bathtubs for infants. Estimate the percent of all parents of young children who use baby bathtubs.

(b) How seriously do people view speeding in comparison with other annoying behaviors? A large random sample of adults was asked to rate a number of behaviors on a scale of 1 (no problem at all) to 5 (very severe problem). Do speeding drivers get a higher average rating than noisy neighbors?

(c) You have data from interviews with a random sample of students who failed to graduate from a particular college in 7 years and also from a random sample of students who entered at the same time and did graduate. You will use these data to compare the percents of students from rural backgrounds among dropouts and graduates.

(d) Do experienced computer game players earn higher scores when they play with someone present to cheer them on or when they play alone? Fifty teenagers with experience playing a particular computer game have volunteered for a study. We randomly assign 25 of them to play the game alone and the other 25 to play the game with a supporter present. Each player's score is recorded.

62. **Which inference method?**

(a) How do young adults look back on adolescent romance? Investigators interviewed 40 couples in their midtwenties. The female and male partners were interviewed separately. Each was asked about his or her current relationship and also about a romantic relationship that lasted at least two months when they were aged 15 or 16. One response variable was a measure on a numerical scale of how much the attractiveness of the adolescent partner mattered. You want to find out how much men and women differ on this measure.

(b) Are more than 75% of Toyota owners generally satisfied with their vehicles? Let's design a study to find out. We'll select a random sample of 400 Toyota owners. Then we'll ask each individual in the sample: "Would you say that you are generally satisfied with your Toyota vehicle?"

(c) Are male college students more likely to binge drink than female college students? The Harvard School of Public Health surveys random samples of male and female undergraduates at four-year colleges and universities about whether they have engaged in binge drinking.

(d) A bank wants to know which of two incentive plans will most increase the use of its credit cards and by how much. It offers each incentive to a group of current credit card customers, determined at random, and compares the amount charged during the following six months.

63. **Quality control** (2.2, 5.3, 6.3) Many manufacturing companies use statistical techniques to ensure that the products they make meet standards. One common way to do this is to take a random sample of products at regular intervals throughout the production shift. Assuming that the process is working properly, the mean measurement $\bar{x}$ from a random sample varies according to a Normal distribution with mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$. For each question that follows, assume that the process is working properly.

(a) What's the probability that at least one of the next two sample means will fall more than $2\sigma_{\bar{x}}$ from the target mean $\mu_{\bar{x}}$? Show your work.

(b) What's the probability that the first sample mean that is greater than $\mu_{\bar{x}} + 2\sigma_{\bar{x}}$ is the one from the fourth sample taken?

Plant managers are trying to develop a criterion for determining when the process is not working properly. One idea they have is to look at the 5 most recent sample means. If at least 4 of the 5 fall outside the interval $(\mu_{\bar{x}} - \sigma_{\bar{x}}, \mu_{\bar{x}} + \sigma_{\bar{x}})$, they will conclude that the process isn't working.

(c) Find the probability that at least 4 of the 5 most recent sample means fall outside the interval, assuming the process is working properly. Is this a reasonable criterion? Explain.

64. **Information online** (8.2, 10.1) A random digit dialing sample of 2092 adults found that 1318 used the Internet.[42] Of the users, 1041 said that they expect businesses to have Web sites that give product information; 294 of the 774 nonusers said this.

(a) Construct and interpret a 95% confidence interval for the proportion of all adults who use the Internet.

(b) Construct and interpret a 95% confidence interval to compare the proportions of users and nonusers who expect businesses to have Web sites that give product information.

65. **Coaching and SAT scores: Critique** (4.1, 4.3) The data in Exercises 55 and 56 came from a random sample of students who took the SAT twice. The response rate was 63%, which is fairly good for nongovernment surveys.

(a) Explain how nonresponse could lead to bias in this study.

(b) We can't be sure that coaching actually *caused* the coached students to gain more than the uncoached students. Explain briefly but clearly why this is so.

FRAPPY! Free Response AP® Problem, Yay!

The following problem is modeled after actual AP® Statistics exam free response questions. Your task is to generate a complete, concise response in 15 minutes.

Directions: Show all your work. Indicate clearly the methods you use, because you will be scored on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.

Will using name-brand microwave popcorn result in a greater percentage of popped kernels than using store-brand microwave popcorn? To find out, Briana and Maggie randomly selected 10 bags of name-brand microwave popcorn and 10 bags of store-brand microwave popcorn. The chosen bags were arranged in a random order. Then each bag was popped for 3.5 minutes, and the percentage of popped kernels was calculated. The results are displayed in the following table.

Name-brand	Store-brand	Name-brand	Store-brand
95	91	90	78
88	89	97	84
84	82	93	86
94	82	91	86
81	77	86	90

Do the data provide convincing evidence that using name-brand microwave popcorn will result in a greater mean percentage of popped kernels?

After you finish, you can view two example solutions on the book's Web site (www.whfreeman.com/tps5e). Determine whether you think each solution is "complete," "substantial," "developing," or "minimal." If the solution is not complete, what improvements would you suggest to the student who wrote it? Finally, your teacher will provide you with a scoring rubric. Score your response and note what, if anything, you would do differently to improve your own score.

Chapter Review

Section 10.1: Comparing Two Proportions

In this section, you learned how to construct confidence intervals and perform significance tests for a difference between two proportions. Inference for a difference in proportions is based on the sampling distribution of $\hat{p}_1 - \hat{p}_2$. When the conditions are met, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately Normal with a mean of $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$ and a standard deviation of $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$.

The conditions for inference about a difference in proportions are the same for confidence intervals and significance tests. The Random condition says that the data must be from two independent random samples or two groups in a randomized experiment. The 10% condition says that each sample size should be less than 10% of the corresponding population size when sampling without replacement. The Large Counts condition says that the number of successes and number of failures from each sample/group should be at least 10. That is, $n_1\hat{p}_1, n_1(1-\hat{p}_1), n_2\hat{p}_2, n_2(1-\hat{p}_2)$ are all ≥ 10.

A confidence interval for a difference between two proportions provides an interval of plausible values for the true difference in proportions. The formula is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

The logic of confidence intervals, including how to interpret the confidence interval and the confidence level, is the same as it was in Chapter 8, when you first learned about confidence intervals.

Likewise, a significance test for a difference between two proportions uses the same logic as the significance tests you learned about in Chapter 9. We start by assuming the null hypothesis is true and asking how likely it would be to get results at least as unusual as the results observed in a study by chance alone. If it is plausible that a difference in proportions could be the result of sampling variability or the chance variation due to random assignment, we do not have convincing evidence that the alternative hypothesis is true. However, if the difference is too big to attribute to chance, there is convincing evidence to believe that the alternative hypothesis is true. For a test of $H_0: p_1 - p_2 = 0$, the test statistic is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\dfrac{\hat{p}_C(1-\hat{p}_C)}{n_1} + \dfrac{\hat{p}_C(1-\hat{p}_C)}{n_2}}}$$

where $\hat{p}_C$ is the combined (overall) proportion of successes: $\hat{p}_C = \dfrac{X_1 + X_2}{n_1 + n_2}$.

Finally, you learned that the inference techniques used for analyzing a difference in proportions from two independent random samples work very well for analyzing a difference in proportions from two groups in a completely randomized experiment.

Section 10.2: Comparing Two Means

In this section, you learned how to construct confidence intervals and perform significance tests for a difference in two means. Inference for a difference in means is based on the sampling distribution of $\bar{x}_1 - \bar{x}_2$. When the conditions are met, the sampling distribution of $\bar{x}_1 - \bar{x}_2$ is approximately Normal with a mean of $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ and a standard deviation of $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$.

The conditions for inference about a difference in means are the same for confidence intervals and significance tests. The Random condition says that the data must be from two independent random samples or two groups in a randomized experiment. The 10% condition says that each sample size should be less than 10% of the corresponding population size when sampling without replacement. The Normal/Large Sample condition says that the two populations are Normal or that the two sample/group sizes are large ($n_1 \geq 30$, $n_2 \geq 30$). If the sample/group sizes are small and the population shapes are unknown, graph both sets of data to make sure there is no strong skewness or outliers.

As in Chapters 8 and 9, inference techniques for means are based on the t distributions. There are two options for calculating the number of degrees of freedom to use. The first option is to use technology to calculate the degrees of freedom. The second option is to use the smaller of $n_1 - 1$ and $n_2 - 1$. The technology option is preferred because it produces a larger number of degrees of freedom, resulting in narrower confidence intervals and smaller *P*-values. If you are using technology, always choose the *un*pooled option.

A confidence interval for a difference between two means provides an interval of plausible values for the true difference in means. The formula is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}$$

Use a significance test to decide between two competing hypotheses about a difference in true means. The test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

where $\mu_1 - \mu_2$ is the difference specified by the null hypothesis.

When constructing confidence intervals or performing significance tests for a difference in means, make sure that the data are not paired. If the data are paired, use the paired t procedures from Chapter 9.

What Did You Learn?

Learning Objective	**Section**	**Related Example on Page(s)**	**Relevant Chapter Review Exercise(s)**
Describe the shape, center, and spread of the sampling distribution of $\hat{p}_1 - \hat{p}_2$.	10.1	615	R10.2
Determine whether the conditions are met for doing inference about $p_1 - p_2$.	10.1	617	R10.5, R10.6
Construct and interpret a confidence interval to compare two proportions.	10.1	617	R10.2
Perform a significance test to compare two proportions.	10.1	622, 625	R10.5
Describe the shape, center, and spread of the sampling distribution of $\bar{x}_1 - \bar{x}_2$.	10.2	638	R10.3
Determine whether the conditions are met for doing inference for $\mu_1 - \mu_2$.	10.2	641	R10.3, R10.4, R10.6
Construct and interpret a confidence interval to compare two means.	10.2	641	R10.4
Perform a significance test to compare two means.	10.2	645	R10.7
Determine when it is appropriate to use two-sample t procedures versus paired t procedures.	10.2	650	R10.1, R10.7

Chapter 10 Chapter Review Exercises

These exercises are designed to help you review the important ideas and methods of the chapter.

R10.1 **Which procedure?** For each of the following settings, say which inference procedure from Chapter 8, 9, or 10 you would use. Be specific. For example, you might say, "Two-sample *z* test for the difference between two proportions." You do not need to carry out any procedures.[43]

(a) Do people smoke less when cigarettes cost more? A random sample of 500 smokers was selected. The number of cigarettes each person smoked per day was recorded over a one-month period before a 30% cigarette tax was imposed and again for one month after the tax was imposed.

(b) How much greater is the percent of senior citizens who attend a play at least once per year than the percent of people in their twenties who do so? Random samples of 100 senior citizens and 100 people in their twenties were surveyed.

(c) You have data on rainwater collected at 16 locations in the Adirondack Mountains of New York State. One measurement is the acidity of the water, measured by pH on a scale of 0 to 14 (the pH of distilled water is 7.0). Estimate the average acidity of rainwater in the Adirondacks.

(d) Consumers Union wants to see which of two brands of calculator is easier to use. They recruit 100 volunteers and randomly assign them to two equal-sized groups. The people in one group use Calculator A and those in the other group use Calculator B. Researchers record the time required for each volunteer to carry out the same series of routine calculations (such as figuring discounts and sales tax, totaling a bill) on the assigned calculator.

R10.2 **Seat belt use** The proportion of drivers who use seat belts depends on things like age (young people are more likely to go unbelted) and gender (women are more likely to use belts). It also depends on local law. In New York City, police can stop a driver who is not belted. In Boston at the time of the study, police could cite a driver for not wearing a seat belt only if the driver had been stopped for some other violation. Here are data from observing random samples of female Hispanic drivers in these two cities:[44]

City	Drivers	Belted
New York	220	183
Boston	117	68

(a) Calculate the standard error of the sampling distribution of the difference in the proportions of female Hispanic drivers in the two cities who wear seat belts. What information does this value provide?

(b) Construct and interpret a 95% confidence interval for the difference in the proportions of female Hispanic drivers in the two cities who wear seat belts.

R10.3 **Expensive ads** Consumers who think a product's advertising is expensive often also think the product must be of high quality. Can other information undermine this effect? To find out, marketing researchers did an experiment. The subjects were 90 women from the clerical and administrative staff of a large organization. All subjects read an ad that described a fictional line of food products called "Five Chefs." The ad also described the major TV commercials that would soon be shown, an unusual expense for this type of product. The 45 women who were randomly assigned to the control group read nothing else. The 45 in the "undermine group" also read a news story headlined "No Link between Advertising Spending and New Product Quality." All the subjects then rated the quality of Five Chefs products on a 7-point scale. The study report said, "The mean quality ratings were significantly lower in the undermine treatment ($\bar{x}_A = 4.56$) than in the control treatment ($\bar{x}_C = 5.05$; $t = 2.64$, $P < 0.01$)."[45]

(a) The 90 women who participated in the study were not randomly selected from a population. Explain why the Random condition is still satisfied.

(b) The distribution of individual responses is not Normal, because there is only a 7-point scale. What is the shape of the sampling distribution of $\bar{x}_C - \bar{x}_A$? Explain.

(c) Interpret the *P*-value in context.

R10.4 **Men versus women** The National Assessment of Educational Progress (NAEP) Young Adult Literacy Assessment Survey interviewed a random sample of 1917 people 21 to 25 years old. The sample contained 840 men and 1077 women.[46] The mean and standard deviation of scores on the NAEP's test of quantitative skills were $\bar{x}_1 = 272.40$ and $s_1 = 59.2$ for the men in the sample. For the women, the results were $\bar{x}_2 = 274.73$ and $s_2 = 57.5$.

(a) Construct and interpret a 90% confidence interval for the difference in mean score for male and female young adults.

(b) Based only on the interval from part (a), is there convincing evidence of a difference in mean score for male and female young adults?

R10.5 **Treating AIDS** The drug AZT was the first drug that seemed effective in delaying the onset of AIDS. Evidence for AZT's effectiveness came from a large randomized comparative experiment. The subjects were 870 volunteers who were infected with HIV, the virus that causes AIDS, but did not yet have AIDS. The study assigned 435 of the subjects at random to take 500 milligrams of AZT each day and another 435 to take a placebo. At the end of the study, 38 of the placebo subjects and 17 of the AZT subjects had developed AIDS.

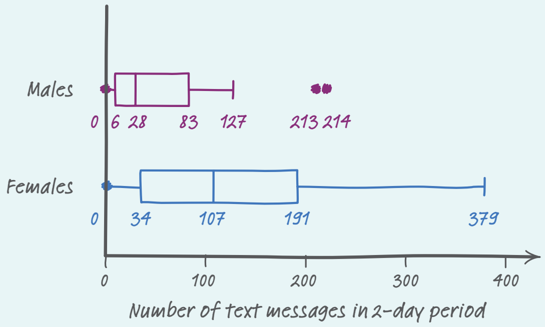
(a) Do the data provide convincing evidence at the $\alpha = 0.05$ level that taking AZT lowers the proportion of infected people who will develop AIDS in a given period of time?

(b) Describe a Type I error and a Type II error in this setting and give a consequence of each error. Based on your conclusion in part (a), which error could have been made in this study?

R10.6 **Conditions** Explain why it is not safe to use the methods of this chapter to perform inference in each of the following settings.

(a) Lyme disease is spread in the northeastern United States by infected ticks. The ticks are infected mainly by feeding on mice, so more mice result in more infected ticks. The mouse population in turn rises and falls with the abundance of acorns, their favored food. Experimenters studied two similar forest areas in a year when the acorn crop failed. They added hundreds of thousands of acorns to one area to imitate an abundant acorn crop, while leaving the other area untouched. The next spring, 54 of the 72 mice trapped in the first area were in breeding condition, versus 10 of the 17 mice trapped in the second area.[47]

(b) Who texts more—males or females? For their final project, a group of AP® Statistics students investigated their belief that females text more than males. They asked a random sample of 31 students—15 males and 16 females—from their school to record the number of text messages sent and received over a 2-day period. Boxplots of their data are shown below.

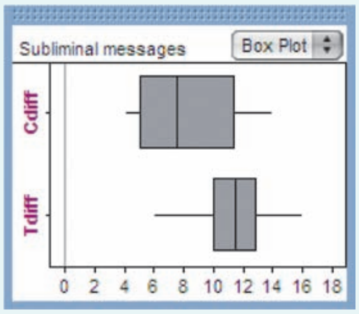


R10.7 **Each day I am getting better in math** A "subliminal" message is below our threshold of awareness but may nonetheless influence us. Can subliminal messages help students learn math? A group of 18 students who had failed the mathematics part of the City University of New York Skills Assessment Test agreed to participate in a study to find out. All received a daily subliminal message, flashed on a screen too rapidly to be consciously read. The treatment group of 10 students (assigned at random) was exposed to "Each day I am getting better in math." The control group of 8 students was exposed to a neutral message, "People are walking on the street." All 18 students participated in a summer program designed to raise their math skills, and all took the assessment test again at the end of the program. The table below gives data on the subjects' scores before and after the program.[48]

Treatment Group			Control Group		
Pretest	Posttest	Difference	Pretest	Posttest	Difference
18	24	6	18	29	11
18	25	7	24	29	5
21	33	12	20	24	4
18	29	11	18	26	8
18	33	15	24	38	14
20	36	16	22	27	5
23	34	11	15	22	7
23	36	13	19	31	12
21	34	13			
17	27	10			

(a) Explain why a two-sample t test is more appropriate than a paired t test for analyzing these data.

(b) The Fathom boxplots below display the differences in pretest and posttest scores for the students in the control (Cdiff) and treatment (Tdiff) groups. Write a few sentences comparing the performance of these two groups.



(c) Do the data provide convincing evidence that subliminal messages help students learn math?

(d) Can we generalize these results to the population of all students who failed the mathematics part of the City University of New York Skills Assessment Test? Why or why not?

Chapter 10 AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question.*

T10.1 A study of road rage asked separate random samples of 596 men and 523 women about their behavior while driving. Based on their answers, each respondent was assigned a road rage score on a scale of 0 to 20. Are the conditions for performing a two-sample t test satisfied?

(a) Maybe; we have independent random samples, but we need to look at the data to check Normality.

(b) No; road rage scores in a range between 0 and 20 can't be Normal.

(c) No; we don't know the population standard deviations.

(d) Yes; the large sample sizes guarantee that the corresponding population distributions will be Normal.

(e) Yes; we have two independent random samples and large sample sizes.

T10.2 Thirty-five people from a random sample of 125 workers from Company A admitted to using sick leave when they weren't really ill. Seventeen employees from a random sample of 68 workers from Company B admitted that they had used sick leave when they weren't ill. A 95% confidence interval for the difference in the proportions of workers at the two companies who would admit to using sick leave when they weren't ill is

(a) $0.03 \pm \sqrt{\dfrac{(0.28)(0.72)}{125} + \dfrac{(0.25)(0.75)}{68}}$

(b) $0.03 \pm 1.96\sqrt{\dfrac{(0.28)(0.72)}{125} + \dfrac{(0.25)(0.75)}{68}}$

(c) $0.03 \pm 1.645\sqrt{\dfrac{(0.28)(0.72)}{125} + \dfrac{(0.25)(0.75)}{68}}$

(d) $0.03 \pm 1.96\sqrt{\dfrac{(0.269)(0.731)}{125} + \dfrac{(0.269)(0.731)}{68}}$

(e) $0.03 \pm 1.645\sqrt{\dfrac{(0.269)(0.731)}{125} + \dfrac{(0.269)(0.731)}{68}}$

T10.3 The power takeoff driveline on tractors used in agriculture is a potentially serious hazard to operators of farm equipment. The driveline is covered by a shield in new tractors, but for a variety of reasons, the shield is often missing on older tractors. Two types of shields are the bolt-on and the flip-up. It was believed that the bolt-on shield was perceived as a nuisance by the operators and deliberately removed, but the flip-up shield is easily lifted for inspection and maintenance and may be left in place. In a study initiated by the U.S. National Safety Council, random samples of older tractors with both types of shields were taken to see what proportion of shields were removed. Of 183 tractors designed to have bolt-on shields, 35 had been removed. Of the 136 tractors with flip-up shields, 15 were removed. We wish to perform a test of $H_0: p_b = p_f$ versus $H_a: p_b > p_f$, where p_b and p_f are the proportions of all tractors with the bolt-on and flip-up shields removed, respectively. Which of the following is not a condition for performing the significance test?

(a) Both populations are Normally distributed.

(b) The data come from two independent samples.

(c) Both samples were chosen at random.

(d) The counts of successes and failures are large enough to use Normal calculations.

(e) Both populations are at least 10 times the corresponding sample sizes.

T10.4 A quiz question gives random samples of $n = 10$ observations from each of two Normally distributed populations. Tom uses a table of t distribution critical values and 9 degrees of freedom to calculate a 95% confidence interval for the difference in the two population means. Janelle uses her calculator's two-sample t interval with 16.87 degrees of freedom to compute the 95% confidence interval. Assume that both students calculate the intervals correctly. Which of the following is true?

(a) Tom's confidence interval is wider.

(b) Janelle's confidence interval is wider.

(c) Both confidence intervals are the same.

(d) There is insufficient information to determine which confidence interval is wider.

(e) Janelle made a mistake; degrees of freedom has to be a whole number.

Exercises T10.5 and T10.6 refer to the following setting. A researcher wished to compare the average amount of time spent in extracurricular activities by high school students in a suburban school district with that in a school district of a large city. The researcher obtained an SRS of 60 high school students in a large suburban school district and found the mean time spent in extracurricular activities per week to be 6 hours with a standard deviation of 3 hours. The researcher also obtained an independent SRS of 40 high school students in a large city school district and found the mean time spent in extracurricular activities per week to be 5 hours with a standard deviation of 2 hours. Suppose that the researcher decides to carry out a significance test of H_0: $\mu_{\text{suburban}} = \mu_{\text{city}}$ versus a two-sided alternative.

T10.5 The correct test statistic is

(a) $z = \dfrac{(6-5)-0}{\sqrt{\dfrac{3}{60}+\dfrac{2}{40}}}$

(b) $z = \dfrac{(6-5)-0}{\sqrt{\dfrac{3^2}{60}+\dfrac{2^2}{40}}}$

(c) $t = \dfrac{(6-5)-0}{\dfrac{3}{\sqrt{60}}+\dfrac{2}{\sqrt{40}}}$

(d) $t = \dfrac{(6-5)-0}{\sqrt{\dfrac{3}{60}+\dfrac{2}{40}}}$

(e) $t = \dfrac{(6-5)-0}{\sqrt{\dfrac{3^2}{60}+\dfrac{2^2}{40}}}$

T10.6 The *P*-value for the test is 0.048. A correct conclusion is to

(a) fail to reject H_0 at the $\alpha = 0.05$ level. There is convincing evidence of a difference in the average time spent on extracurricular activities by students in the suburban and city school districts.

(b) fail to reject H_0 at the $\alpha = 0.05$ level. There is not convincing evidence of a difference in the average time spent on extracurricular activities by students in the suburban and city school districts.

(c) fail to reject H_0 at the $\alpha = 0.05$ level. There is convincing evidence that the average time spent on extracurricular activities by students in the suburban and city school districts is the same.

(d) reject H_0 at the $\alpha = 0.05$ level. There is not convincing evidence of a difference in the average time spent on extracurricular activities by students in the suburban and city school districts.

(e) reject H_0 at the $\alpha = 0.05$ level. There is convincing evidence of a difference in the average time spent on extracurricular activities by students in the suburban and city school districts.

T10.7 At a baseball game, 42 of 65 randomly selected people own an iPod. At a rock concert occurring at the same time across town, 34 of 52 randomly selected people own an iPod. A researcher wants to test the claim that the proportion of iPod owners at the two venues is different. A 90% confidence interval for the difference in population proportions (game − concert) is (−0.154, 0.138). Which of the following gives the correct outcome of the researcher's test of the claim?

(a) Because the confidence interval includes 0, the researcher can conclude that the proportion of iPod owners at the two venues is the same.

(b) Because the center of the interval is −0.008, the researcher can conclude that a higher proportion of people at the rock concert own iPods than at the baseball game.

(c) Because the confidence interval includes 0, the researcher cannot conclude that the proportion of iPod owners at the two venues is different.

(d) Because the confidence interval includes more negative than positive values, the researcher can conclude that a higher proportion of people at the rock concert own iPods than at the baseball game.

(e) The researcher cannot draw a conclusion about a claim without performing a significance test.

T10.8 An SRS of size 100 is taken from Population A with proportion 0.8 of successes. An independent SRS of size 400 is taken from Population B with proportion 0.5 of successes. The sampling distribution for the difference (Population A – Population B) in sample proportions has what mean and standard deviation?

(a) mean = 0.3; standard deviation = 1.3

(b) mean = 0.3; standard deviation = 0.40

(c) mean = 0.3; standard deviation = 0.047

(d) mean = 0.3; standard deviation = 0.0022

(e) mean = 0.3; standard deviation = 0.0002

T10.9 How much more effective is exercise and drug treatment than drug treatment alone at reducing the rate of heart attacks among men aged 65 and older? To find out, researchers perform a completely randomized experiment involving 1000 healthy males in this age group. Half of the subjects are assigned to receive drug treatment only, while the other half are assigned to exercise regularly and to receive drug treatment. The most appropriate inference method for answering the original research question is

(a) one-sample z test for a proportion.

(b) two-sample z interval for $p_1 - p_2$.

(c) two-sample z test for $p_1 - p_2$.

(d) two-sample t interval for $\mu_1 - \mu_2$.

(e) two-sample t test for $\mu_1 - \mu_2$.

T10.10 Researchers are interested in evaluating the effect of a natural product on reducing blood pressure. This will be done by comparing the mean reduction in blood pressure of a treatment (natural product) group and a placebo group using a two-sample *t* test. The researchers would like to be able to detect whether the natural product reduces blood pressure by at least 7 points more, on average, than the placebo. If groups of size 50 are used in the experiment, a two-sample *t* test using $\alpha = 0.01$ will have a power of 80% to detect a 7-point difference in mean blood pressure reduction. If the researchers want to be able to detect a 5-point difference instead, then the power of the test

(a) would be less than 80%.

(b) would be greater than 80%.

(c) would still be 80%.

(d) could be either less than or greater than 80%, depending on whether the natural product is effective.

(e) would vary depending on the standard deviation of the data.

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

T10.11 Researchers wondered whether maintaining a patient's body temperature close to normal by heating the patient during surgery would affect wound infection rates. Patients were assigned at random to two groups: the normothermic group (patients' core temperatures were maintained at near normal, 36.5°C, with heating blankets) and the hypothermic group (patients' core temperatures were allowed to decrease to about 34.5°C). If keeping patients warm during surgery alters the chance of infection, patients in the two groups should have hospital stays of very different lengths. Here are summary statistics on hospital stay (in number of days) for the two groups:

Group	n	$\bar{x}$	s_x
Normothermic	104	12.1	4.4
Hypothermic	96	14.7	6.5

(a) Construct and interpret a 95% confidence interval for the difference in the true mean length of hospital stay for normothermic and hypothermic patients.

(b) Does your interval in part (a) suggest that keeping patients warm during surgery affects the average length of patients' hospital stays? Justify your answer.

(c) Interpret the meaning of "95% confidence" in the context of this study.

T10.12 A random sample of 100 of a certain popular car model last year found that 20 had a certain minor defect in the brakes. The car company made an adjustment in the production process to try to reduce the proportion of cars with the brake problem. A random sample of 350 of this year's model found that 50 had the minor brake defect.

(a) Was the company's adjustment successful? Carry out an appropriate test to support your answer.

(b) Describe a Type I error and a Type II error in this setting, and give a possible consequence of each.

T10.13 Pat wants to compare the cost of one- and two-bedroom apartments in the area of her college campus. She collects data for a random sample of 10 advertisements of each type. The table below shows the rents (in dollars per month) for the selected apartments.

1 bedroom:	500	650	600	505	450	550	515	495	650	395
2 bedroom:	595	500	580	650	675	675	750	500	495	670

Pat wonders if two-bedroom apartments rent for significantly more, on average, than one-bedroom apartments.

(a) State an appropriate pair of hypotheses for a significance test. Be sure to define any parameters you use.

(b) Name the appropriate test and show that the conditions for carrying out this test are met.

(c) The appropriate test from part (b) yields a *P*-value of 0.029. Interpret this *P*-value in context.

(d) What conclusion should Pat draw at the $\alpha = 0.05$ significance level? Explain.

Cumulative AP® Practice Test 3

Section I: Multiple Choice *Choose the best answer.*

AP3.1 Suppose the probability that a softball player gets a hit in any single at-bat is 0.300. Assuming that her chance of getting a hit on a particular time at bat is independent of her other times at bat, what is the probability that she will not get a hit until her fourth time at bat in a game?

(a) $\binom{4}{3}(0.3)^1(0.7)^3$ (d) $(0.3)^3(0.7)^1$

(b) $\binom{4}{3}(0.3)^3(0.7)^1$ (e) $(0.3)^1(0.7)^3$

(c) $\binom{4}{1}(0.3)^3(0.7)^1$

AP3.2 The probability that Color Me Dandy wins a horse race at Batavia Downs given good track conditions is 0.60. The probability of good track conditions on any given day is 0.85. What is the probability that Color Me Dandy wins or the track conditions are good?

(a) 0.94 (b) 0.51 (c) 0.49 (d) 0.06

(e) The answer cannot be determined from the given information.

AP3.3 *Sports Illustrated* planned to ask a random sample of Division I college athletes, "Do you believe performance-enhancing drugs are a problem in college sports?" How many athletes must be interviewed to estimate the proportion concerned about use of drugs within ±2% with 90% confidence?

(a) 17 (c) 1680 (e) 2401

(b) 21 (d) 1702

AP3.4 The distribution of grade point averages for a certain college is approximately Normal with a mean of 2.5 and a standard deviation of 0.6. Within which of the following intervals would we expect to find approximately 81.5% of all GPAs for students at this college?

(a) (0.7, 3.1) (c) (1.9, 3.7) (e) (0.7, 4.3)

(b) (1.3, 3.7) (d) (1.9, 4.3)

AP3.5 Which of the following will increase the power of a significance test?

(a) Increase the Type II error probability.

(b) Decrease the sample size.

(c) Reject the null hypothesis only if the *P*-value is smaller than the level of significance.

(d) Increase the significance level α.

(e) Select a value for the alternative hypothesis closer to the value of the null hypothesis.

AP3.6 You can find some interesting polls online. Anyone can become part of the sample just by clicking on a response. One such poll asked, "Do you prefer watching first-run movies at a movie theater, or waiting until they are available to watch at home or on a digital device?" In all, 8896 people responded, with only 12% (1118 people) saying they preferred theaters. You can conclude that

(a) American adults strongly prefer watching movies at home or on their digital devices.

(b) the high nonresponse rate prevents us from drawing a conclusion.

(c) the sample is too small to draw any conclusion.

(d) the poll uses voluntary response, so the results tell us little about all American adults.

(e) American adults strongly prefer seeing movies at a movie theater.

AP3.7 A certain candy has different wrappers for various holidays. During Holiday 1, the candy wrappers are 30% silver, 30% red, and 40% pink. During Holiday 2, the wrappers are 50% silver and 50% blue. Forty pieces of candy are randomly selected from the Holiday 1 distribution, and 40 pieces are randomly selected from the Holiday 2 distribution. What are the expected value and standard deviation of the total number of silver wrappers?

(a) 32, 18.4 (c) 32, 4.29 (e) 80, 4.29

(b) 32, 6.06 (d) 80, 18.4

AP3.8 A beef rancher randomly sampled 42 cattle from her large herd to obtain a 95% confidence interval to estimate the mean weight of the cows in the herd. The interval obtained was (1010, 1321). If the rancher had used a 98% confidence interval instead, the interval would have been

(a) wider and would have less precision than the original estimate.

(b) wider and would have more precision than the original estimate.

(c) wider and would have the same precision as the original estimate.

(d) narrower and would have less precision than the original estimate.

(e) narrower and would have more precision than the original estimate.

AP3.9 School A has 400 students and School B has 2700 students. A local newspaper wants to compare the distributions of SAT scores for the two schools. Which of the following would be the most useful for making this comparison?

(a) Back-to-back stemplots for A and B

(b) A scatterplot of A versus B

(c) Dotplots for A and B drawn on the same scale

(d) Two relative frequency histograms of A and B drawn on the same scale

(e) Two bar graphs for A and B drawn on the same scale

AP3.10 Let X represent the outcome when a fair six-sided die is rolled. For this random variable, $\mu_X = 3.5$ and $\sigma_X = 1.71$. If the die is rolled 100 times, what is the approximate probability that the total score is at least 375?

(a) 0.0000 (c) 0.0721 (e) 0.9279

(b) 0.0017 (d) 0.4420

AP3.11 An agricultural station is testing the yields for six different varieties of seed corn. The station has four large fields available, which are located in four distinctly different parts of the county. The agricultural researchers consider the climatic and soil conditions in the four parts of the county as being unequal but are reasonably confident that the conditions within each field are fairly similar throughout. The researchers divide each field into six sections and then randomly assign one variety of corn seed to each section in that field. This procedure is done for each field. At the end of the growing season, the corn will be harvested, and the yield, measured in tons per acre, will be compared. Which one of the following statements about the design is correct?

(a) This is an observational study because the researchers are watching the corn grow.

(b) This a randomized block design with fields as blocks and seed types as treatments.

(c) This is a randomized block design with seed types as blocks and fields as treatments.

(d) This is a completely randomized design because the six seed types were randomly assigned to the four fields.

(e) This is a completely randomized design with 24 treatments—6 seed types and 4 fields.

AP3.12 The correlation between the heights of fathers and the heights of their (grownup) sons is $r = 0.52$, both measured in inches. If fathers' heights were measured in feet instead, the correlation between heights of fathers and heights of sons would be

(a) much smaller than 0.52.

(b) slightly smaller than 0.52.

(c) unchanged; equal to 0.52.

(d) slightly larger than 0.52.

(e) much larger than 0.52.

AP3.13 A random sample of 200 New York State voters included 88 Republicans, while a random sample of 300 California voters produced 141 Republicans. Which of the following represents the 95% confidence interval that should be used to estimate the true difference in the proportions of Republicans in New York State and California?

(a) $(0.44 - 0.47) \pm 1.96\dfrac{(0.44)(0.56) + (0.47)(0.53)}{\sqrt{200 + 300}}$

(b) $(0.44 - 0.47) \pm 1.96\dfrac{(0.44)(0.56)}{\sqrt{200}} + \dfrac{(0.47)(0.53)}{\sqrt{300}}$

(c) $(0.44 - 0.47) \pm 1.96\sqrt{\dfrac{(0.44)(0.56)}{200} + \dfrac{(0.47)(0.53)}{300}}$

(d) $(0.44 - 0.47) \pm 1.96\sqrt{\dfrac{(0.44)(0.56) + (0.47)(0.53)}{200 + 300}}$

(e) $(0.44 - 0.47) \pm 1.96\sqrt{\dfrac{(0.45)(0.55)}{200} + \dfrac{(0.45)(0.55)}{300}}$

AP3.14 Which of the following is *not* a property of a binomial setting?

(a) Outcomes of different trials are independent.

(b) The chance process consists of a fixed number of trials, *n*.

(c) The probability of success is the same for each trial.

(d) Trials are repeated until a success occurs.

(e) Each trial can result in either a success or a failure.

AP3.15 Mrs. Woods and Mrs. Bryan are avid vegetable gardeners. They use different fertilizers, and each claims that hers is the best fertilizer to use when growing tomatoes. Both agree to do a study using the weight of their tomatoes as the response variable. They had each planted the same varieties of tomatoes on the same day and fertilized the plants on the same schedule throughout the growing season. At harvest time, they each randomly select 15 tomatoes from their respective gardens and weigh them. After performing a two-sample *t* test on the difference in mean weights of tomatoes, they get $t = 5.24$ and $P = 0.0008$. Can the gardener with the larger mean claim that her fertilizer caused her tomatoes to be heavier?

(a) Yes, because a different fertilizer was used on each garden.

(b) Yes, because random samples were taken from each garden.

(c) Yes, because the *P*-value is so small.

(d) No, because the soil conditions in the two gardens is a potential confounding variable.

(e) No, because there was no replication.

AP3.16 The Environmental Protection Agency is charged with monitoring industrial emissions that pollute the atmosphere and water. So long as emission levels stay within specified guidelines, the EPA does not take action against the polluter. If the polluter is in violation of the regulations, the offender can be fined, forced to clean up the problem, or possibly closed. Suppose that for a particular industry the acceptable emission level has been set at no more than 5 parts per million (5 ppm). The null and alternative hypotheses are $H_0: \mu = 5$ versus $H_a: \mu > 5$. Which of the following describes a Type II error?

(a) The EPA fails to find convincing evidence that emissions exceed acceptable limits when, in fact, they are within acceptable limits.

(b) The EPA finds convincing evidence that emissions exceed acceptable limits when, in fact, they are within acceptable limits.

(c) The EPA finds convincing evidence that emissions exceed acceptable limits when, in fact, they do exceed acceptable limits.

(d) The EPA takes more samples to ensure that they make the correct decision.

(e) The EPA fails to find convincing evidence that emissions exceed acceptable limits when, in fact, they do exceed acceptable limits.

AP3.17 Which of the following is *false*?

(a) A measure of center alone does not completely describe the characteristics of a set of data. Some measure of spread is also needed.

(b) If the original measurements are in inches, converting them to centimeters will not change the mean or standard deviation.

(c) One of the disadvantages of a histogram is that it doesn't show each data value.

(d) Between the range and the interquartile range, the *IQR* is a better measure of spread if there are outliers.

(e) If a distribution is skewed, the median and interquartile range should be reported rather than the mean and standard deviation.

AP3.18 A 96% confidence interval for the proportion of the labor force that is unemployed in a certain city is (0.07, 0.10). Which of the following statements about this interval is true?

(a) The probability is 0.96 that between 7% and 10% of the labor force is unemployed.

(b) About 96% of the intervals constructed by this method will contain the true proportion of unemployed in the city.

(c) In repeated samples of the same size, there is a 96% chance that the sample proportion will fall between 0.07 and 0.10.

(d) The true rate of unemployment lies within this interval 96% of the time.

(e) Between 7% and 10% of the labor force is unemployed 96% of the time.

AP3.19 A large toy company introduces a lot of new toys to its product line each year. The company wants to predict the demand as measured by *y*, first-year sales (in millions of dollars) using *x*, awareness of the product (as measured by the percent of customers who had heard of the product by the end of the second month after its introduction). A random sample of 65 new products was taken, and a correlation of 0.96 was computed. Which of the following is a correct interpretation of this value?

(a) Ninety-six percent of the time, the least-squares regression line accurately predicts first-year sales.

(b) About 92% of the time, the percent of people who have heard of the product by the end of the second month will correctly predict first-year sales.

(c) About 92% of first-year sales can be accounted for by the percent of people who have heard of the product by the end of the second month.

(d) For each increase of 1% in awareness of the new product, the predicted sales will go up by 0.96 million dollars.

(e) About 92% of the variation in first-year sales can be accounted for by the least-squares regression line with percent of people who have heard of the product by the end of the second month as the explanatory variable.

AP3.20 Final grades for a class are approximately Normally distributed with a mean of 76 and a standard deviation of 8. A professor says that the top 10% of the class will receive an A, the next 20% a B, the next 40% a C, the next 20% a D, and the bottom 10% an F. What is the approximate maximum grade a student could attain and still receive an F for the course?

(a) 70

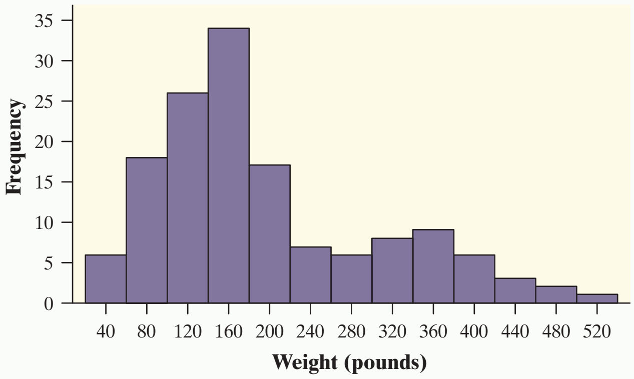
(b) 69.27

(c) 65.75

(d) 62.84

(e) 57

AP3.21 National Park rangers keep data on the bears that inhabit their park. Below is a histogram of the weights of 143 bears measured in a recent year.



Which statement below is correct?

(a) The median will lie in the interval (140, 180), and the mean will lie in the interval (180, 220).

(b) The median will lie in the interval (140, 180), and the mean will lie in the interval (260, 300).

(c) The median will lie in the interval (100, 140), and the mean will lie in the interval (180, 220).

(d) The mean will lie in the interval (140, 180), and the median will lie in the interval (260, 300).

(e) The mean will lie in the interval (100, 140), and the median will lie in the interval (180, 220).

AP3.22 A random sample of size n will be selected from a population, and the proportion of those in the sample who have a Facebook page will be calculated. How would the margin of error for a 95% confidence interval be affected if the sample size were increased from 50 to 200?

(a) It remains the same.

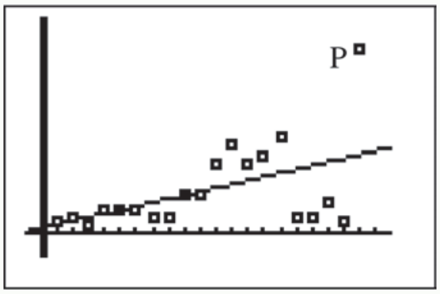
(b) It is multiplied by 2.

(c) It is multiplied by 4.

(d) It is divided by 2.

(e) It is divided by 4.

AP3.23 A scatterplot and a least-squares regression line are shown in the figure below. What effect does point P have on the slope of the regression line and the correlation?



(a) Point P increases the slope and increases the correlation.

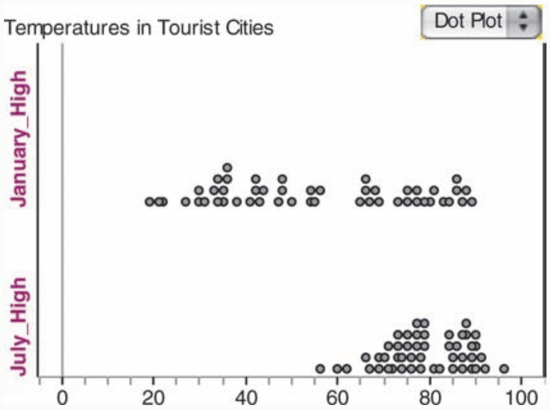
(b) Point P increases the slope and decreases the correlation.

(c) Point P decreases the slope and decreases the correlation.

(d) Point P decreases the slope and increases the correlation.

(e) No conclusion can be drawn because the other coordinates are unknown.

AP3.24 The following dotplots show the average high temperatures (in degrees Fahrenheit) for a sample of tourist cities from around the world. Both the January and July average high temperatures are shown. What is one statement that can be made with certainty from an analysis of the graphical display?



(a) Every city has a larger average high temperature in July than in January.

(b) The distribution of temperatures in July is skewed right, while the distribution of temperatures in January is skewed left.

(c) The median average high temperature for January is higher than the median average high temperature for July.

(d) There appear to be outliers in the average high temperatures for January and July.

(e) There is more variability in average high temperatures in January than in July.

AP3.25 Suppose the null and alternative hypotheses for a significance test are defined as

$$H_0 : \mu = 40$$
$$H_a : \mu < 40$$

Which of the following specific values for H_a will give the highest power?

(a) $\mu = 38$ (c) $\mu = 40$ (e) $\mu = 42$

(b) $\mu = 39$ (d) $\mu = 41$

AP3.26 A large university is considering the establishment of a schoolwide recycling program. To gauge interest in the program by means of a questionnaire, the university takes separate random samples of undergraduate students, graduate students, faculty, and staff. This is an example of what type of sampling design?

(a) Simple random sample

(b) Stratified random sample

(c) Convenience sample

(d) Cluster sample

(e) Randomized block design

AP3.27 Suppose the true proportion of people who use public transportation to get to work in the Washington, D.C., area is 0.45. In a simple random sample of 250 people who work in Washington, about how far do you expect the sample proportion to be from the true proportion?

(a) 0.4975 (c) 0.0315 (e) 0

(b) 0.2475 (d) 0.0009

Questions 28 and 29 refer to the following setting. According to sleep researchers, if you are between the ages of 12 and 18 years old, you need 9 hours of sleep to be fully functional. A simple random sample of 28 students was chosen from a large high school, and these students were asked how much sleep they got the previous night. The mean of the responses was 7.9 hours, with a standard deviation of 2.1 hours.

AP3.28 If we are interested in whether students at this high school are getting too little sleep, which of the following represents the appropriate null and alternative hypotheses?

(a) $H_0: \mu = 7.9$ and $H_a: \mu < 7.9$

(b) $H_0: \mu = 7.9$ and $H_a: \mu \neq 7.9$

(c) $H_0: \mu = 9$ and $H_a: \mu \neq 9$

(d) $H_0: \mu = 9$ and $H_a: \mu < 9$

(e) $H_0: \mu \leq 9$ and $H_a: \mu \geq 9$

AP3.29 Which of the following is the test statistic for the hypothesis test?

(a) $t = \dfrac{7.9 - 9}{\dfrac{2.1}{\sqrt{28}}}$ (b) $t = \dfrac{9 - 7.9}{\dfrac{2.1}{\sqrt{28}}}$ (c) $t = \dfrac{7.9 - 9}{\sqrt{\dfrac{2.1}{28}}}$

(d) $t = \dfrac{7.9 - 9}{\dfrac{2.1}{\sqrt{27}}}$ (e) $t = \dfrac{9 - 7.9}{\dfrac{2.1}{\sqrt{27}}}$

AP3.30 Shortly before the 2012 presidential election, a survey was taken by the school newspaper at a very large state university. Randomly selected students were asked, "Whom do you plan to vote for in the upcoming presidential election?" Here is a two-way table of the responses by political persuasion for 1850 students:

Candidate of	Political Persuasion			
choice	**Democrat**	**Republican**	**Independent**	**Total**
Obama	925	78	26	**1029**
Romney	78	598	19	**695**
Other	2	8	11	**21**
Undecided	32	28	45	**105**
Total	**1037**	**712**	**101**	**1850**

Which of the following statements about these data is true?

(a) The percent of Republicans among the respondents is 41%.

(b) The marginal distribution of the variable choice of candidate is given by Obama: 55.6%; Romney: 37.6%; Other: 1.1%; Undecided: 5.7%.

(c) About 11.2% of Democrats reported that they planned to vote for Romney.

(d) About 44.6% of those who are undecided are Independents.

(e) The conditional distribution of political persuasion among those for whom Romney is the candidate of choice is Democrat: 7.5%; Republican: 84.0%; Independent: 18.8%

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations.*

AP3.31 A researcher wants to determine whether or not a five-week crash diet is effective over a long period of time. A random sample of 15 dieters is selected. Each person's weight is recorded before starting the diet and one year after it is concluded. Based on the data shown at right (weight in pounds), can we conclude that the diet has a long-term effect, that is, that dieters manage to not regain the weight they lose? Include appropriate statistical evidence to justify your answer.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Before	158	185	176	172	164	234	258	200
After	163	182	188	150	161	220	235	191

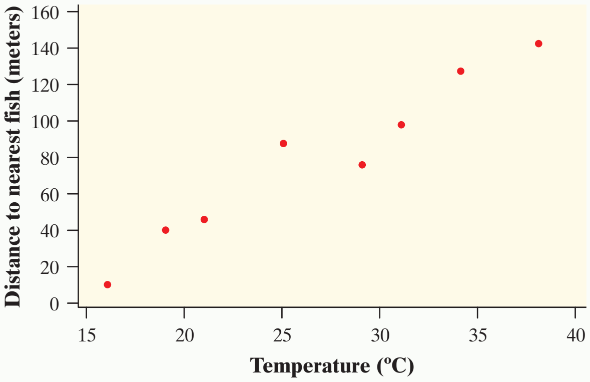
	9	**10**	**11**	**12**	**13**	**14**	**15**
Before	228	246	198	221	236	255	231
After	228	237	209	220	222	268	234

AP3.32 Starting in the 1970s, medical technology allowed babies with very low birth weight (VLBW, less than 1500 grams, or about 3.3 pounds) to survive without major handicaps. It was noticed that these children nonetheless had difficulties in school and as adults. A long study has followed 242 randomly selected VLBW babies to age 20 years, along with a control group of 233 randomly selected babies from the same population who had normal birth weight.[49]

(a) Is this an experiment or an observational study? Why?

(b) At age 20, 179 of the VLBW group and 193 of the control group had graduated from high school. Is the graduation rate among the VLBW group significantly lower than for the normal-birth-weight controls? Give appropriate statistical evidence to justify your answer.

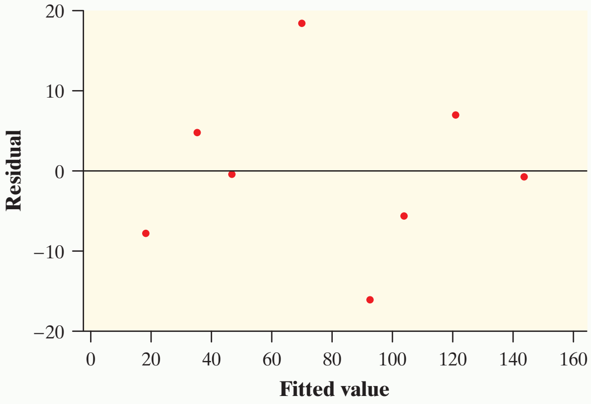
AP3.33 A nuclear power plant releases water into a nearby lake every afternoon at 4:51 P.M. Environmental researchers are concerned that fish are being driven away from the area around the plant. They believe that the temperature of the water discharged may be a factor. The scatterplot below shows the temperature of the water (°C) released by the plant and the measured distance (in meters) from the outflow pipe of the plant to the nearest fish found in the water on eight randomly chosen afternoons.



Computer output from a least-squares regression analysis on these data and a residual plot are shown below.

```
Predictor        Coef    SE Coef       T        P
Constant       -73.64     15.48    -4.76    0.003
Temperature    5.7188    0.5612   -10.19    0.000

S = 11.4175 R-Sq = 94.5% R-Sq(adj) = 93.6%
```



(a) Write the equation of the least-squares regression line. Define any variables you use.

(b) Interpret the slope of the regression line in context.

(c) Is a linear model appropriate for describing the relationship between temperature and distance to the nearest fish? Justify your answer.

(d) Compute the residual for the point (29, 78). Interpret this residual in context.

AP3.34 The Candy Shoppe assembles gift boxes that contain 8 chocolate truffles and 2 handmade caramel nougats. The truffles have a mean weight of 2 ounces with a standard deviation of 0.5 ounce, and the nougats have a mean weight of 4 ounces with a standard deviation of 1 ounce. The empty boxes weigh 3 ounces with a standard deviation of 0.2 ounce.

(a) Assuming that the weights of the truffles, nougats, and boxes are independent, what are the mean and standard deviation of the weight of a box of candy?

(b) Assuming that the weights of the truffles, nougats, and boxes are approximately Normally distributed, what is the probability that a randomly selected box of candy will weigh more than 30 ounces?

(c) If five gift boxes are randomly selected, what is the probability that at least one of them will weigh more than 30 ounces?

(d) If five gift boxes are randomly selected, what is the probability that the mean weight of the five boxes will be more than 30 ounces?

AP3.35 An investor is comparing two stocks, A and B. She wants to know if over the long run, there is a significant difference in the return on investment as measured by the percent increase or decrease in the price of the stock from its date of purchase. The investor takes a random sample of 50 annualized daily returns over the past five years for each stock. The data are summarized below.

Stock	Mean return	Standard deviation
A	11.8%	12.9%
B	7.1%	9.6%

(a) Is there a significant difference in the mean return on investment for the two stocks? Support your answer with appropriate statistical evidence. Use a 5% significance level.

(b) The investor believes that although the return on investment for Stock A usually exceeds that of Stock B, Stock A represents a riskier investment, where the risk is measured by the price volatility of the stock. The standard deviation is a statistical measure of the price volatility and indicates how much an investment's actual performance during a specified period varies from its average performance over a longer period. Do the price fluctuations in Stock A significantly exceed those of Stock B, as measured by their standard deviations? Identify an appropriate set of hypotheses that the investor is interested in testing.

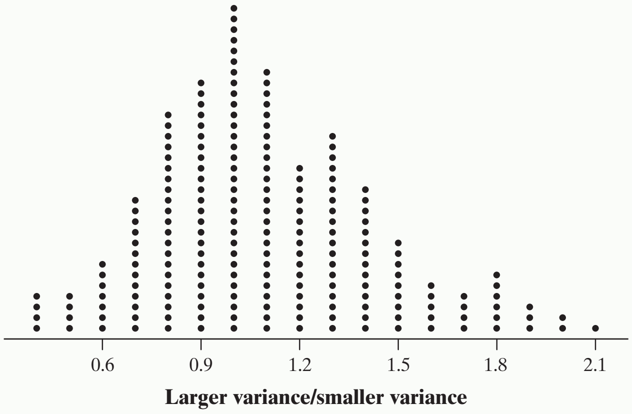
(c) To measure this, we will construct a test statistic defined as

$$F = \frac{\text{large sample variance}}{\text{smaller sample variance}}$$

What value(s) of the statistic would indicate that the price fluctuations in Stock A significantly exceed those of Stock B? Explain.

(d) Calculate the value of the F statistic using the information given in the table.

(e) Two hundred simulated values of this test statistic, F, were calculated assuming no difference in the standard deviations of the returns for the two stocks. The results of the simulation are displayed in the following dotplot.



Use these simulated values and the test statistic that you calculated in part (d) to determine whether the observed data provide convincing evidence that Stock A is a riskier investment than Stock B. Explain your reasoning.