9.4 A LARGE-SAMPLE TEST OF HYPOTHESIS FOR THE DIFFERENCE BETWEEN TWO POPULATION MEANS

In many situations, the statistical question to be answered involves a comparison of two population means. For example, the U.S. Postal Service is interested in reducing its massive 350 million gallons/year gasoline bill by replacing gasoline-powered trucks with electric-powered trucks. To determine whether significant savings in operating costs are achieved by changing to electric-powered trucks, a pilot study should be undertaken using, say, 100 conventional gasoline-powered mail trucks and 100 electric-powered mail trucks operated under similar conditions.

The statistic that summarizes the sample information regarding the difference in population means $(\mu_1 - \mu_2)$ is the difference in sample means $(\bar{x}_1 - \bar{x}_2)$. Therefore, in testing whether the difference in sample means indicates that the true difference in population means differs from a specified value, $(\mu_1 - \mu_2) = D_0$, you can use the standard error of $(\bar{x}_1 - \bar{x}_2)$,

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{estimated by} \quad SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

in the form of a z-statistic to measure how many standard deviations the difference $(\bar{x}_1 - \bar{x}_2)$ lies from the hypothesized difference D_0. The formal testing procedure is described next.

LARGE-SAMPLE STATISTICAL TEST FOR $(\mu_1 - \mu_2)$

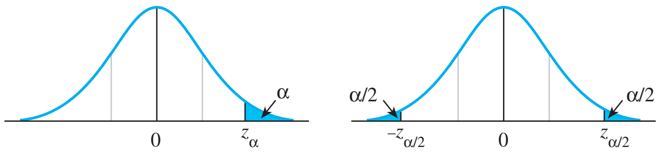
1. Null hypothesis: $H_0 : (\mu_1 - \mu_2) = D_0$, where D_0 is some specified difference that you wish to test. For many tests, you will hypothesize that there is no difference between μ_1 and μ_2; that is, $D_0 = 0$.
2. Alternative hypothesis:

One-Tailed Test	**Two-Tailed Test**
$H_a : (\mu_1 - \mu_2) > D_0$ [or $H_a : (\mu_1 - \mu_2) < D_0$]	$H_a : (\mu_1 - \mu_2) \neq D_0$

3. Test statistic: $z \approx \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{SE} = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$
4. Rejection region: Reject H_0 when

One-Tailed Test	**Two-Tailed Test**
$z > z_\alpha$ [or $z < -z_\alpha$ when the alternative hypothesis is $H_a : (\mu_1 - \mu_2) < D_0$]	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

or when p-value $< \alpha$



Assumptions: The samples are randomly and independently selected from the two populations and $n_1 \geq 30$ and $n_2 \geq 30$.

EXAMPLE 9.9

To determine whether car ownership affects a student's academic achievement, two random samples of 100 male students were each drawn from the student body. The grade point average for the $n_1 = 100$ non-owners of cars had an average and variance equal to $\bar{x}_1 = 2.70$ and $s_1^2 = .36$, while $\bar{x}_2 = 2.54$ and $s_2^2 = .40$ for the $n_2 = 100$ car owners. Do the data present sufficient evidence to indicate a difference in the mean achievements between car owners and nonowners of cars? Test using $\alpha = .05$.

Solution To detect a difference, if it exists, between the mean academic achievements for non-owners of cars μ_1 and car owners μ_2, you will test the null hypothesis that there is no difference between the means against the alternative hypothesis that $(\mu_1 - \mu_2) \neq 0$; that is,

$$H_0 : (\mu_1 - \mu_2) = D_0 = 0 \quad \text{versus} \quad H_a : (\mu_1 - \mu_2) \neq 0$$

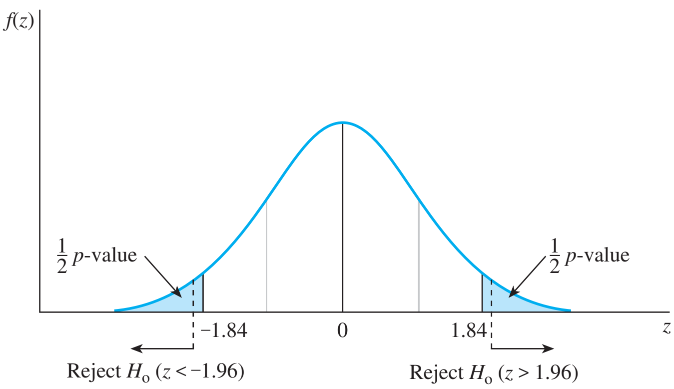
Substituting into the formula for the test statistic, you get

$$z \approx \frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{2.70 - 2.54}{\sqrt{\dfrac{.36}{100} + \dfrac{.40}{100}}} = 1.84$$

MY TIP
|test statistic| > |critical value| ⇔ reject H_0

- **The critical value approach:** Using a two-tailed test with significance level $\alpha = .05$, you place $\alpha/2 = .025$ in each tail of the z distribution and reject H_0 if $z > 1.96$ or $z < -1.96$. Since $z = 1.84$ does not exceed 1.96 and is not less than -1.96, H_0 cannot be rejected (see Figure 9.11). That is, there is insufficient evidence to declare a difference in the average academic achievements for the two groups. Remember that you should not be willing to *accept* H_0—declare the two means to be the same—until β is evaluated for some meaningful values of $(\mu_1 - \mu_2)$.

FIGURE 9.11
Rejection region and *p*-value for Example 9.9



- **The *p*-value approach:** Calculate the p-value, the probability that z is greater than $z = 1.84$ plus the probability that z is less than $z = -1.84$, as shown in Figure 9.11:

$$p\text{-value} = P(z > 1.84) + P(z < -1.84) = (1 - .9671) + .0329 = .0658$$

The p-value lies between .10 and .05, so you can reject H_0 at the .10 level but not at the .05 level of significance. Since the p-value of .0658 exceeds the specified significance level $\alpha = .05$, H_0 cannot be rejected. Again, you should not be willing to *accept* H_0 until β is evaluated for some meaningful values of $(\mu_1 - \mu_2)$.

Hypothesis Testing and Confidence Intervals

Whether you use the critical value or the p-value approach for testing hypotheses about $(\mu_1 - \mu_2)$, you will always reach the same conclusion because the calculated value of the test statistic and the critical value are related *exactly* in the same way that the p-value and the significance level α are related. You might remember that the confidence intervals constructed in Chapter 8 could also be used to answer questions about the difference between two population means. In fact, for a two-tailed test, the $(1 - \alpha)100\%$ confidence interval for the parameter of interest can be used to test its value, just as you did informally in Chapter 8. The value of α indicated by the confidence coefficient in the confidence interval is equivalent to the significance level α in the statistical test. For a one-tailed test, the equivalent confidence interval approach would use the one-sided confidence bounds in Section 8.8 with confidence coefficient α. In addition, by using the confidence interval approach, you gain a range of possible values for the parameter of interest, regardless of the outcome of the test of hypothesis.

- If the confidence interval you construct *contains* the value of the parameter specified by H_0, then that value is one of the likely or possible values of the parameter and H_0 should not be rejected.
- If the hypothesized value *lies outside* of the confidence limits, the null hypothesis is rejected at the α level of significance.

EXAMPLE 9.10

Construct a 95% confidence interval for the difference in average academic achievements between car owners and non-owners. Using the confidence interval, can you conclude that there is a difference in the population means for the two groups of students?

Solution For the large-sample statistics discussed in Chapter 8, the 95% confidence interval is given as

$$\text{Point estimator} \pm 1.96 \times (\text{Standard error of the estimator})$$

For the difference in two population means, the confidence interval is approximated as

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(2.70 - 2.54) \pm 1.96\sqrt{\frac{.36}{100} + \frac{.40}{100}}$$

$$.16 \pm .17$$

or $-.01 < (\mu_1 - \mu_2) < .33$. This interval gives you a range of possible values for the difference in the population means. Since the hypothesized difference, $(\mu_1 - \mu_2) = 0$, is contained in the confidence interval, you should not reject H_0. Look at the signs of the possible values in the confidence interval. You cannot tell from the interval whether the difference in the means is negative (−), positive (+), or zero (0)—the latter of the three would indicate that the two means are the same. Hence, you can really reach no conclusion in terms of the question posed. There is not enough evidence to indicate that there is a difference in the average achievements for car owners versus non-owners. The conclusion is the same one reached in Example 9.9.

9.4 EXERCISES

BASIC TECHNIQUES

9.18 Independent random samples of 80 measurements were drawn from two quantitative populations, 1 and 2. Here is a summary of the sample data:

	Sample 1	Sample 2
Sample Size	80	80
Sample Mean	11.6	9.7
Sample Variance	27.9	38.4

a. If your research objective is to show that μ_1 is larger than μ_2, state the alternative and the null hypotheses that you would choose for a statistical test.

b. Is the test in part a one- or two-tailed?

c. Calculate the test statistic that you would use for the test in part a. Based on your knowledge of the standard normal distribution, is this a likely or unlikely observation, assuming that H_0 is true and the two population means are the same?

d. *p-value approach:* Find the p-value for the test. Test for a significant difference in the population means at the 1% significance level.

e. *Critical value approach:* Find the rejection region when $\alpha = .01$. Do the data provide sufficient evidence to indicate a difference in the population means?

9.19 Independent random samples of 36 and 45 observations are drawn from two quantitative populations, 1 and 2, respectively. The sample data summary is shown here:

	Sample 1	Sample 2
Sample Size	36	45
Sample Mean	1.24	1.31
Sample Variance	.0560	.0540

Do the data present sufficient evidence to indicate that the mean for population 1 is smaller than the mean for population 2? Use one of the two methods of testing presented in this section, and explain your conclusions.

9.20 Suppose you wish to detect a difference between μ_1 and μ_2 (either $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$) and, instead of running a two-tailed test using $\alpha = .05$, you use the following test procedure. You wait until you have collected the sample data and have calculated $\bar{x}_1$ and $\bar{x}_2$. If $\bar{x}_1$ is larger than $\bar{x}_2$, you choose the alternative hypothesis $H_a : \mu_1 > \mu_2$ and run a one-tailed test placing $\alpha_1 = .05$ in the upper tail of the z distribution. If, on the other hand, $\bar{x}_2$ is larger than $\bar{x}_1$, you reverse the procedure and run a one-tailed test, placing $\alpha_2 = .05$ in the lower tail of the z distribution. If you use this procedure and if μ_1 actually equals μ_2, what is the probability α that you will conclude that μ_1 is not equal to μ_2 (i.e., what is the probability α that you will incorrectly reject H_0 when H_0 is true)? This exercise demonstrates why statistical tests should be formulated *prior* to observing the data.

APPLICATIONS

9.21 Cure for the Common Cold? An experiment was planned to compare the mean time (in days) required to recover from a common cold for persons given a daily dose of 4 milligrams (mg) of vitamin C versus those who were not given a vitamin supplement. Suppose that 35 adults were randomly selected for each treatment category and that the mean recovery times and standard deviations for the two groups were as follows:

	No Vitamin Supplement	4 mg Vitamin C
Sample Size	35	35
Sample Mean	6.9	5.8
Sample Standard Deviation	2.9	1.2

a. Suppose your research objective is to show that the use of vitamin C reduces the mean time required to recover from a common cold and its complications. Give the null and alternative hypotheses for the test. Is this a one- or a two-tailed test?

b. Conduct the statistical test of the null hypothesis in part a and state your conclusion. Test using $\alpha = .05$.

9.22 Healthy Eating Americans are becoming more conscious about the importance of good nutrition, and some researchers believe we may be altering our diets to include less red meat and more fruits and vegetables. To test the theory that the consumption of red meat has decreased over the last 10 years, a researcher decides to select hospital nutrition records for 400 subjects surveyed 10 years ago and to compare their average amount of beef consumed per year to amounts consumed by an equal number of subjects interviewed this year. The data are given in the table.

	Ten Years Ago	This Year
Sample Mean	73	63
Sample Standard Deviation	25	28

a. Do the data present sufficient evidence to indicate that per-capita beef consumption has decreased in the last 10 years? Test at the 1% level of significance.

b. Find a 99% lower confidence bound for the difference in the average per-capita beef consumptions for the two groups. (This calculation was done as part of Exercise 8.76.) Does your confidence bound confirm your conclusions in part a? Explain. What additional information does the confidence bound give you?

9.23 Lead Levels in Drinking Water Analyses of drinking water samples for 100 homes in each of two different sections of a city gave the following means and standard deviations of lead levels (in parts per million):

	Section 1	Section 2
Sample Size	100	100
Mean	34.1	36.0
Standard Deviation	5.9	6.0

a. Calculate the test statistic and its p-value (observed significance level) to test for a difference in the two population means. Use the p-value to evaluate the statistical significance of the results at the 5% level.

b. Use a 95% confidence interval to estimate the difference in the mean lead levels for the two sections of the city.

c. Suppose that the city environmental engineers will be concerned only if they detect a difference of more than 5 parts per million in the two sections of the city. Based on your confidence interval in part b, is the statistical significance in part a of *practical significance* to the city engineers? Explain.

9.24 Starting Salaries, again In an attempt to compare the starting salaries for college graduates who majored in chemical engineering and computer science (see Exercise 8.45), random samples of 50 recent college graduates in each major were selected and the following information obtained.

Major	Mean	SD
Chemical Engineering	$53,659	2225
Computer Science	51,042	2375

a. Do the data provide sufficient evidence to indicate a difference in average starting salaries for college graduates who majored in chemical engineering and computer science? Test using $\alpha = .05$.

b. Compare your conclusions in part a with the results of part b in Exercise 8.45. Are they the same? Explain.

9.25 Hotel Costs In Exercise 8.18, we explored the average cost of lodging at three different hotel chains.[6] We randomly select 50 billing statements from the computer databases of the Marriott, Radisson, and Wyndham hotel chains, and record the nightly room rates. A portion of the sample data is shown in the table.

	Marriott	Radisson
Sample Average	$170	$145
Sample Standard Deviation	17.5	10

a. Before looking at the data, would you have any preconceived idea about the direction of the difference between the average room rates for these two hotels? If not, what null and alternative hypotheses should you test?

b. Use the *critical value* approach to determine if there is a significant difference in the average room rates for the Marriott and the Radisson hotel chains. Use $\alpha = .01$.

c. Find the p-value for this test. Does this p-value confirm the results of part b?

9.26 Hotel Costs II Refer to Exercise 9.25. The table below shows the sample data collected to compare the average room rates at the Wyndham and Radisson hotel chains.[6]

	Wyndham	Radisson
Sample Average	\$150	\$145
Sample Standard Deviation	16.5	10

a. Do the data provide sufficient evidence to indicate a difference in the average room rates for the Wyndham and the Radisson hotel chains? Use $\alpha = .05$.

b. Construct a 95% confidence interval for the difference in the average room rates for the two chains. Does this interval confirm your conclusions in part a?

9.27 MMT in Gasoline The addition of MMT, a compound containing manganese (Mn), to gasoline as an octane enhancer has caused concern about human exposure to Mn because high intakes have been linked to serious health effects. In a study of ambient air concentrations of fine Mn, Wallace and Slonecker (*Journal of the Air and Waste Management Association*) presented the accompanying summary information about the amounts of fine Mn (in nanograms per cubic meter) in mostly rural national park sites and in mostly urban California sites.[7]

	National Parks	California
Mean	.94	2.8
Standard Deviation	1.2	2.8
Number of Sites	36	26

a. Is there sufficient evidence to indicate that the mean concentrations differ for these two types of sites at the $\alpha = .05$ level of significance? Use the large-sample z-test. What is the p-value of this test?

b. Construct a 95% confidence interval for $(\mu_1 - \mu_2)$. Does this interval confirm your conclusions in part a?

9.28 Noise and Stress In Exercise 8.48, you compared the effect of stress in the form of noise on the ability to perform a simple task. Seventy subjects were divided into two groups; the first group of 30 subjects acted as a control, while the second group of 40 was the experimental group. Although each subject performed the task in the same control room, each of the experimental group subjects had to perform the task while loud rock music was played. The time to finish the task was recorded for each subject and the following summary was obtained:

	Control	Experimental
n	30	40
$\bar{x}$	15 minutes	23 minutes
s	4 minutes	10 minutes

a. Is there sufficient evidence to indicate that the average time to complete the task was longer for the experimental "rock music" group? Test at the 1% level of significance.

b. Construct a 99% one-sided upper bound for the difference (control − experimental) in average times for the two groups. Does this interval confirm your conclusions in part a?

9.29 What's Normal II Of the 130 people in Exercise 9.16, 65 were female and 65 were male.[3] The means and standard deviations of their temperatures are shown below.

	Men	Women
Sample Mean	98.11	98.39
Standard Deviation	0.70	0.74

a. Use the p-value approach to test for a significant difference in the average temperatures for males versus females.

b. Are the results significant at the 5% level? At the 1% level?

9.5 A LARGE-SAMPLE TEST OF HYPOTHESIS FOR A BINOMIAL PROPORTION

When a random sample of n identical trials is drawn from a binomial population, the sample proportion $\hat{p}$ has an approximately normal distribution when n is large, with mean p and standard error

$$\text{SE} = \sqrt{\frac{pq}{n}}$$

When you test a hypothesis about p, the proportion in the population possessing a certain attribute, the test follows the same general form as the large-sample tests in Sections 9.3 and 9.4. To test a hypothesis of the form

$$H_0 : p = p_0$$

versus a one- or two-tailed alternative

$$H_a : p > p_0 \quad \text{or} \quad H_a : p < p_0 \quad \text{or} \quad H_a : p \neq p_0$$

the test statistic is constructed using $\hat{p}$, the best estimator of the true population proportion p. The sample proportion $\hat{p}$ is standardized, using the hypothesized mean and standard error, to form a test statistic z, which has a standard normal distribution if H_0 is true. This large-sample test is summarized next.

LARGE-SAMPLE STATISTICAL TEST FOR p

1. Null hypothesis: $H_0 : p = p_0$
2. Alternative hypothesis:

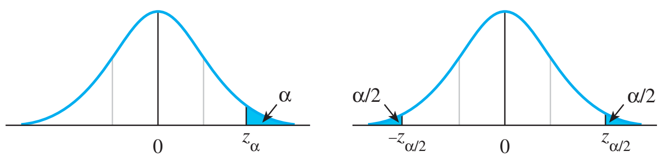
One-Tailed Test	Two-Tailed Test
$H_a : p > p_0$ (or, $H_a : p < p_0$)	$H_a : p \neq p_0$

3. Test statistic: $z = \dfrac{\hat{p} - p_0}{\text{SE}} = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$ with $\hat{p} = \dfrac{x}{n}$

   where x is the number of successes in n binomial trials.†
4. Rejection region: Reject H_0 when

One-Tailed Test	Two-Tailed Test
$z > z_\alpha$ (or $z < -z_\alpha$ when the alternative hypothesis is $H_a : p < p_0$)	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

   or when p-value $< \alpha$



Assumption: The sampling satisfies the assumptions of a binomial experiment (see Section 5.2), and n is large enough so that the sampling distribution of $\hat{p}$ can be approximated by a normal distribution ($np_0 > 5$ and $nq_0 > 5$).

†An equivalent test statistic can be found by multiplying the numerator and denominator by z by n to obtain

$$z = \frac{x - np_0}{\sqrt{np_0 q_0}}$$

EXAMPLE 9.11

Regardless of age, about 20% of American adults participate in fitness activities at least twice a week. However, these fitness activities change as the people get older, and occasionally participants become nonparticipants as they age. In a local survey of $n = 100$ adults over 40 years old, a total of 15 people indicated that they participated in a fitness activity at least twice a week. Do these data indicate that the participation rate for adults over 40 years of age is significantly less than the 20% figure? Calculate the p-value and use it to draw the appropriate conclusions.

Solution Assuming that the sampling procedure satisfies the requirements of a binomial experiment, you can answer the question posed using a one-tailed test of hypothesis:

$$H_0: p = .2 \quad \text{versus} \quad H_a: p < .2$$

Begin by assuming that H_0 is true—that is, the true value of p is $p_0 = .2$. Then $\hat{p} = x/n$ will have an approximate normal distribution with mean p_0 and standard error $\sqrt{p_0 q_0/n}$. (NOTE: This is different from the estimation procedure in which the unknown standard error is estimated by $\sqrt{\hat{p}\hat{q}/n}$.) The observed value of $\hat{p}$ is $15/100 = .15$ and the test statistic is

MY TIP

p-value $\le \alpha \Leftrightarrow$ reject H_0

p-value $> \alpha \Leftrightarrow$ do not reject H_0

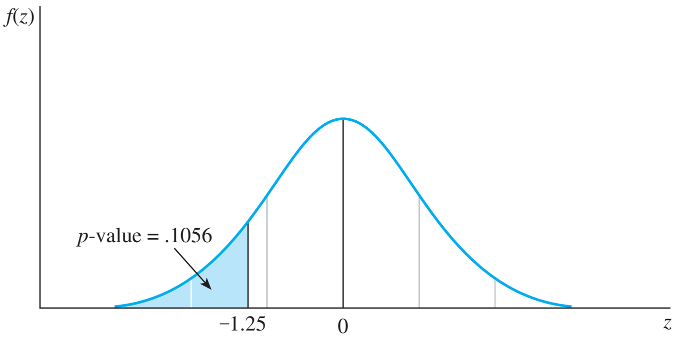
$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}} = \frac{.15 - .20}{\sqrt{\dfrac{(.20)(.80)}{100}}} = -1.25$$

The p-value associated with this test is found as the area under the standard normal curve to the left of $z = -1.25$ as shown in Figure 9.12. Therefore,

$$p\text{-value} = P(z < -1.25) = .1056$$

FIGURE 9.12

p-value for Example 9.11



If you use the guidelines for evaluating p-values, then .1056 is greater than .10, and you would not reject H_0. There is insufficient evidence to conclude that the percentage of adults over age 40 who participate in fitness activities twice a week is less than 20%.

Statistical Significance and Practical Importance

It is important to understand the difference between results that are "significant" and results that are practically "important." In statistical language, the word *significant* does not necessarily mean "important," but only that the results could not have occurred by chance. For example, suppose that in Example 9.11, the researcher had used

$n = 400$ adults in her experiment and had observed the same sample proportion. The test statistic is now

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{.15 - .20}{\sqrt{\frac{(.20)(.80)}{400}}} = -2.50$$

with

$$p\text{-value} = P(z < -2.50) = .0062$$

Now the results are *highly significant:* H_0 is rejected, and there is sufficient evidence to indicate that the percentage of adults over age 40 who participate in physical fitness activities is less than 20%. However, is this drop in activity really *important*? Suppose that physicians would be concerned only about a drop in physical activity of more than 10%. If there had been a drop of more than 10% in physical activity, this would imply that the true value of p was less than .10. What is the largest possible value of p? Using a 95% upper one-sided confidence bound, you have

$$\hat{p} + 1.645\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$.15 + 1.645\sqrt{\frac{(.15)(.85)}{400}}$$

$$.15 + .029$$

or $p < .179$. The physical activity for adults aged 40 and older has dropped from 20%, but you cannot say that it has dropped below 10%. So, the results, although *statistically significant,* are not *practically important.*

In this book, you will learn how to determine whether results are statistically significant. When you use these procedures in a practical situation, however, you must also make sure the results are practically important.

9.5 EXERCISES

BASIC TECHNIQUES

9.30 A random sample of $n = 1000$ observations from a binomial population produced $x = 279$.

a. If your research hypothesis is that p is less than .3, what should you choose for your alternative hypothesis? Your null hypothesis?

b. What is the critical value that determines the rejection region for your test with $\alpha = .05$?

c. Do the data provide sufficient evidence to indicate that p is less than .3? Use a 5% significance level.

9.31 A random sample of $n = 1400$ observations from a binomial population produced $x = 529$.

a. If your research hypothesis is that p differs from .4, what hypotheses should you test?

b. Calculate the test statistic and its p-value. Use the p-value to evaluate the statistical significance of the results at the 1% level.

c. Do the data provide sufficient evidence to indicate that p is different from .4?

9.32 A random sample of 120 observations was selected from a binomial population, and 72 successes were observed. Do the data provide sufficient evidence to indicate that p is greater than .5? Use one of the two methods of testing presented in this section, and explain your conclusions.

APPLICATIONS

9.33 Childhood Obesity According to *PARADE* magazine's "What America Eats" survey involving

$n = 1015$ adults, almost half of parents say their children's weight is fine.[8] Only 9% of parents describe their children as overweight. However, the American Obesity Association says the number of overweight children and teens is at least 15%. Suppose that the number of parents in the sample is $n = 750$ and the number of parents who describe their children as overweight is $x = 68$.

a. How would you proceed to test the hypothesis that the proportion of parents who describe their children as overweight is less than the actual proportion reported by the American Obesity Association?

b. What conclusion are you able to draw from these data at the $\alpha = .05$ level of significance?

c. What is the p-value associated with this test?

9.34 Plant Genetics A peony plant with red petals was crossed with another plant having streaky petals. A geneticist states that 75% of the offspring resulting from this cross will have red flowers. To test this claim, 100 seeds from this cross were collected and germinated, and 58 plants had red petals.

a. What hypothesis should you use to test the geneticist's claim?

b. Calculate the test statistic and its p-value. Use the p-value to evaluate the statistical significance of the results at the 1% level.

9.35 Early Detection of Breast Cancer Of those women who are diagnosed to have early-stage breast cancer, one-third eventually die of the disease. Suppose a community public health department instituted a screening program to provide for the early detection of breast cancer and to increase the survival rate p of those diagnosed to have the disease. A random sample of 200 women was selected from among those who were periodically screened by the program and who were diagnosed to have the disease. Let x represent the number of those in the sample who survive the disease.

a. If you wish to detect whether the community screening program has been effective, state the null hypothesis that should be tested.

b. State the alternative hypothesis.

c. If 164 women in the sample of 200 survive the disease, can you conclude that the community screening program was effective? Test using $\alpha = .05$ and explain the practical conclusions from your test.

d. Find the p-value for the test and interpret it.

9.36 Sweet Potato Whitefly Suppose that 10% of the fields in a given agricultural area are infested with the sweet potato whitefly. One hundred fields in this area are randomly selected, and 25 are found to be infested with whitefly.

a. Assuming that the experiment satisfies the conditions of the binomial experiment, do the data indicate that the proportion of infested fields is greater than expected? Use the p-value approach, and test using a 5% significance level.

b. If the proportion of infested fields is found to be significantly greater than .10, why is this of practical significance to the agronomist? What practical conclusions might she draw from the results?

9.37 Brown or Blue? An article in the *Washington Post* stated that nearly 45% of the U.S. population is born with brown eyes, although they don't necessarily stay that way.[9] To test the newspaper's claim, a random sample of 80 people was selected, and 32 had brown eyes. Is there sufficient evidence to dispute the newspaper's claim regarding the proportion of brown-eyed people in the United States? Use $\alpha = .01$.

9.38 Colored Contacts Refer to Exercise 9.37. Contact lenses, worn by about 26 million Americans, come in many styles and colors. Most Americans wear soft lenses, with the most popular colors being the blue varieties (25%), followed by greens (24%), and then hazel or brown. A random sample of 80 tinted contact lens wearers was checked for the color of their lenses. Of these people, 22 wore blue lenses and only 15 wore green lenses.[9]

a. Do the sample data provide sufficient evidence to indicate that the proportion of tinted contact lens wearers who wear blue lenses is different from 25%? Use $\alpha = .05$.

b. Do the sample data provide sufficient evidence to indicate that the proportion of tinted contact lens wearers who wear green lenses is different from 24%? Use $\alpha = .05$.

c. Is there any reason to conduct a one-tailed test for either part a or b? Explain.

9.39 A Cure for Insomnia An experimenter has prepared a drug-dose level that he claims will induce sleep for at least 80% of people suffering from insomnia. After examining the dosage we feel that his claims regarding the effectiveness of his dosage are inflated. In an attempt to disprove his claim, we administer his prescribed dosage to 50 insomniacs and observe that

37 of them have had sleep induced by the drug dose. Is there enough evidence to refute his claim at the 5% level of significance?

9.40 Who Votes? About three-fourths of voting age Americans are registered to vote, but many do not bother to vote on Election Day. Only 64% voted in 1992, and 60% in 2000, but turnout in off-year elections is even lower. An article in *Time* stated that 35% of adult Americans are registered voters who always vote.[10] To test this claim, a random sample of $n = 300$ adult Americans was selected and $x = 123$ were registered regular voters who always voted. Does this sample provide sufficient evidence to indicate that the percentage of adults who say that they always vote is different from the percentage reported in *Time*? Test using $\alpha = .01$.

9.41 Man's Best Friend The Humane Society reports that there are approximately 65 million dogs owned in the United States and that approximately 40% of all U.S. households own at least one dog.[11] In a random sample of 300 households, 114 households said that they owned at least one dog. Does this data provide sufficient evidence to indicate that the proportion of households with at least one dog is different from that reported by the Humane Society? Test using $\alpha = .05$.

9.6 A LARGE-SAMPLE TEST OF HYPOTHESIS FOR THE DIFFERENCE BETWEEN TWO BINOMIAL PROPORTIONS

When random and independent samples are selected from two *binomial* populations, the focus of the experiment may be the difference $(p_1 - p_2)$ in the proportions of individuals or items possessing a specified characteristic in the two populations. In this situation, you can use the difference in the sample proportions $(\hat{p}_1 - \hat{p}_2)$ along with its standard error,

$$\text{SE} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

in the form of a z-statistic to test for a significant difference in the two population proportions. The null hypothesis to be tested is usually of the form

$$H_0 : p_1 = p_2 \quad \text{or} \quad H_0 : (p_1 - p_2) = 0$$

MY TIP

Remember: Each trial results in one of two outcomes (S or F).

versus either a one- or two-tailed alternative hypothesis. The formal test of hypothesis is summarized in the next display. In estimating the standard error for the z-statistic, you should use the fact that when H_0 is true, the two population proportions are equal to some common value—say, p. To obtain the best estimate of this common value, the sample data are "pooled" and the estimate of p is

$$\hat{p} = \frac{\text{Total number of successes}}{\text{Total number of trials}} = \frac{x_1 + x_2}{n_1 + n_2}$$

Remember that, in order for the difference in the sample proportions to have an approximately normal distribution, the sample sizes must be large and the proportions should not be too close to 0 or 1.

LARGE-SAMPLE STATISTICAL TEST FOR $(p_1 - p_2)$

1. Null hypothesis: $H_0 : (p_1 - p_2) = 0$ or equivalently $H_0 : p_1 = p_2$
2. Alternative hypothesis:

One-Tailed Test	**Two-Tailed Test**
$H_a : (p_1 - p_2) > 0$ [or $H_a : (p_1 - p_2) < 0$]	$H_a : (p_1 - p_2) \neq 0$

3. Test statistic: $z = \dfrac{(\hat{p}_1 - \hat{p}_2) - 0}{\text{SE}} = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{p_1 q_1}{n_1} + \dfrac{p_2 q_2}{n_2}}} = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{pq}{n_1} + \dfrac{pq}{n_2}}}$

   where $\hat{p}_1 = x_1/n_1$ and $\hat{p}_2 = x_2/n_2$. Since the common value of $p_1 = p_2 = p$ (used in the standard error) is unknown, it is estimated by

   $$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

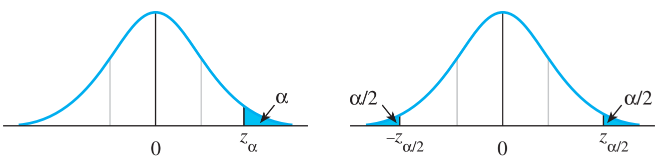
   and the test statistic is

   $$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\dfrac{\hat{p}\hat{q}}{n_1} + \dfrac{\hat{p}\hat{q}}{n_2}}} \quad \text{or} \quad z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

4. Rejection region: Reject H_0 when

One-Tailed Test	**Two-Tailed Test**
$z > z_\alpha$ [or $z < -z_\alpha$ when the alternative hypothesis is $H_a : (p_1 - p_2) < 0$]	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$

   or when p-value $< \alpha$



Assumptions: Samples are selected in a random and independent manner from two binomial populations, and n_1 and n_2 are large enough so that the sampling distribution of $(\hat{p}_1 - \hat{p}_2)$ can be approximated by a normal distribution. That is, $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$, and $n_2\hat{q}_2$ should all be greater than 5.

EXAMPLE 9.12

The records of a hospital show that 52 men in a sample of 1000 men versus 23 women in a sample of 1000 women were admitted because of heart disease. Do these data present sufficient evidence to indicate a higher rate of heart disease among men admitted to the hospital? Use $\alpha = .05$.

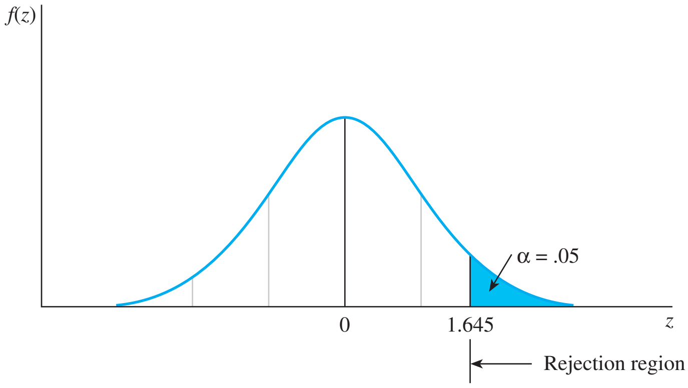
Solution Assume that the number of patients admitted for heart disease has an approximate binomial probability distribution for both men and women with parameters

p_1 and p_2, respectively. Then, since you wish to determine whether $p_1 > p_2$, you will test the null hypothesis $p_1 = p_2$—that is, $H_0 : (p_1 - p_2) = 0$—against the alternative hypothesis $H_a : p_1 > p_2$ or, equivalently, $H_a : (p_1 - p_2) > 0$. To conduct this test, use the z-test statistic and approximate the standard error using the pooled estimate of p. Since H_a implies a one-tailed test, you can reject H_0 only for large values of z. Thus, for $\alpha = .05$, you can reject H_0 if $z > 1.645$ (see Figure 9.13).

The pooled estimate of p required for the standard error is

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{52 + 23}{1000 + 1000} = .0375$$

FIGURE 9.13
Location of the rejection region in Example 9.12



and the test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{.052 - .023}{\sqrt{(.0375)(.9625)\left(\frac{1}{1000} + \frac{1}{1000}\right)}} = 3.41$$

Since the computed value of z falls in the rejection region, you can reject the hypothesis that $p_1 = p_2$. The data present sufficient evidence to indicate that the percentage of men entering the hospital because of heart disease is higher than that of women. (NOTE: This does not imply that the *incidence* of heart disease is higher in men. Perhaps fewer women enter the hospital when afflicted with the disease!)

How *much higher* is the proportion of men than women entering the hospital with heart disease? A 95% lower one-sided confidence bound will help you find the lowest likely value for the difference.

$$(\hat{p}_1 - \hat{p}_2) - 1.645\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$(.052 - .023) - 1.645\sqrt{\frac{.052(.948)}{1000} + \frac{.023(.977)}{1000}}$$

$$.029 - .014$$

or $(p_1 - p_2) > .015$. The proportion of men is roughly 1.5% higher than women. Is this of *practical importance*? This is a question for the researcher to answer.

In some situations, you may need to test for a difference D_0 (other than 0) between two binomial proportions. If this is the case, the test statistic is modified for testing

$H_0 : (p_1 - p_2) = D_0$, and a pooled estimate for a common p is no longer used in the standard error. The modified test statistic is

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}}$$

Although this test statistic is not used often, the procedure is no different from other large-sample tests you have already mastered!

9.6 EXERCISES

BASIC TECHNIQUES

9.42 Independent random samples of $n_1 = 140$ and $n_2 = 140$ observations were randomly selected from binomial populations 1 and 2, respectively. Sample 1 had 74 successes, and sample 2 had 81 successes.

a. Suppose you have no preconceived idea as to which parameter, p_1 or p_2, is the larger, but you want to detect only a difference between the two parameters if one exists. What should you choose as the alternative hypothesis for a statistical test? The null hypothesis?

b. Calculate the standard error of the difference in the two sample proportions, $(\hat{p}_1 - \hat{p}_2)$. Make sure to use the pooled estimate for the common value of p.

c. Calculate the test statistic that you would use for the test in part a. Based on your knowledge of the standard normal distribution, is this a likely or unlikely observation, assuming that H_0 is true and the two population proportions are the same?

d. *p-value approach:* Find the p-value for the test. Test for a significant difference in the population proportions at the 1% significance level.

e. *Critical value approach:* Find the rejection region when $\alpha = .01$. Do the data provide sufficient evidence to indicate a difference in the population proportions?

9.43 Refer to Exercise 9.42. Suppose, for practical reasons, you know that p_1 cannot be larger than p_2.

a. Given this knowledge, what should you choose as the alternative hypothesis for your statistical test? The null hypothesis?

b. Does your alternative hypothesis in part a imply a one- or two-tailed test? Explain.

c. Conduct the test and state your conclusions. Test using $\alpha = .05$.

9.44 Independent random samples of 280 and 350 observations were selected from binomial populations 1 and 2, respectively. Sample 1 had 132 successes, and sample 2 had 178 successes. Do the data present sufficient evidence to indicate that the proportion of successes in population 1 is smaller than the proportion in population 2? Use one of the two methods of testing presented in this section, and explain your conclusions.

APPLICATIONS

9.45 Treatment versus Control An experiment was conducted to test the effect of a new drug on a viral infection. The infection was induced in 100 mice, and the mice were randomly split into two groups of 50. The first group, the *control group,* received no treatment for the infection. The second group received the drug. After a 30-day period, the proportions of survivors, $\hat{p}_1$ and $\hat{p}_2$, in the two groups were found to be .36 and .60, respectively.

a. Is there sufficient evidence to indicate that the drug is effective in treating the viral infection? Use $\alpha = .05$.

b. Use a 95% confidence interval to estimate the actual difference in the cure rates for the treated versus the control groups.

9.46 Movie Marketing Marketing to targeted age groups has become a standard method of advertising, even in movie theater advertising. Advertisers use computer software to track the demographics of moviegoers and then decide on the type of products to advertise before a particular movie.[12] One statistic that might be of interest is how frequently adults with children under 18 attend movies as compared to those without children. Suppose that a theater database is used to randomly select 1000 adult ticket purchasers. These adults are then surveyed and asked whether they were frequent moviegoers—that is, do they attend movies 12 or more times a year? The results are shown in the table:

	With Children under 18	Without Children
Sample Size	440	560
Number Who Attend 12^+ Times per Year	123	145

a. Is there a significant difference in the population proportions of frequent moviegoers in these two demographic groups? Use $\alpha = .01$.

b. Why would a statistically significant difference in these population proportions be of *practical importance* to the advertiser?

9.47 M&M'S In Exercise 8.53, you investigated whether Mars, Inc., uses the same proportion of red M&M'S in its plain and peanut varieties. Random samples of plain and peanut M&M'S provide the following sample data for the experiment:

	Plain	Peanut
Sample Size	56	32
Number of Red M&M'S	12	8

Use a test of hypothesis to determine whether there is a significant difference in the proportions of red candies for the two types of M&M'S. Let $\alpha = .05$ and compare your results with those of Exercise 8.53.

9.48 Hormone Therapy and Alzheimer's Disease In the last few years, many research studies have shown that the purported benefits of hormone replacement therapy (HRT) do not exist, and in fact, that hormone replacement therapy actually increases the risk of several serious diseases. A four-year experiment involving 4532 women, reported in *The Press Enterprise,* was conducted at 39 medical centers. Half of the women took placebos and half took Prempro, a widely prescribed type of hormone replacement therapy. There were 40 cases of dementia in the hormone group and 21 in the placebo group.[13] Is there sufficient evidence to indicate that the risk of dementia is higher for patients using Prempro? Test at the 1% level of significance.

9.49 HRT, continued Refer to Exercise 9.48. Calculate a 99% lower one-sided confidence bound for the difference in the risk of dementia for women using hormone replacement therapy versus those who do not. Would this difference be of *practical importance* to a woman considering HRT? Explain.

9.50 Clopidogrel and Aspirin A large study was conducted to test the effectiveness of clopidogrel in combination with aspirin in warding off heart attacks and strokes.[14] The trial involved more than 15,500 people 45 years of age or older from 32 countries, including the United States, who had been diagnosed with cardiovascular disease or had multiple risk factors. The subjects were randomly assigned to one of two groups. After two years, there was no difference in the risk of heart attack, stroke, or dying from heart disease between those who took clopidogrel and low-dose aspirin daily and those who took low-dose aspirin plus a dummy pill. The two-drug combination actually increased the risk of dying (5.4% versus 3.8%) or dying specifically from cardiovascular disease (3.9% versus 2.2%).

a. The subjects were randomly assigned to one of the two groups. Explain how you could use the random number table to make these assignments.

b. No sample sizes were given in the article: however, let us assume that the sample sizes for each group were $n_1 = 7720$ and $n_2 = 7780$. Determine whether the risk of dying was significantly different for the two groups.

c. What do the results of the study mean in terms of *practical significance*?

9.51 Baby's Sleeping Position Does a baby's sleeping position affect the development of motor skills? In one study, published in the *Archives of Pediatric Adolescent Medicine,* 343 full-term infants were examined at their 4-month checkups for various developmental milestones, such as rolling over, grasping a rattle, reaching for an object, and so on.[15] The baby's predominant sleep position—either prone (on the stomach) or supine (on the back) or side—was determined by a telephone interview with the parent. The sample results for 320 of the 343 infants for whom information was received are shown here:

	Prone	Supine or Side
Number of Infants	121	199
Number That Roll Over	93	119

The researcher reported that infants who slept in the side or supine position were less likely to roll over at the 4-month checkup than infants who slept primarily in the prone position ($P < .001$). Use a large-sample test of hypothesis to confirm or refute the researcher's conclusion.

9.7 SOME COMMENTS ON TESTING HYPOTHESES

A statistical test of hypothesis is a fairly clear-cut procedure that enables an experimenter to either reject or accept the null hypothesis H_0, with measured risks α and β. The experimenter can control the risk of falsely rejecting H_0 by selecting an appropriate value of α. On the other hand, the value of β depends on the sample size and the values of the parameter under test that are of practical importance to the experimenter. When this information is not available, an experimenter may decide to select an affordable sample size, in the hope that the sample will contain sufficient information to reject the null hypothesis. The chance that this decision is in error is given by α, whose value has been set in advance. If the sample does not provide sufficient evidence to reject H_0, the experimenter may wish to state the results of the test as "The data do not support the rejection of H_0" rather than accepting H_0 without knowing the chance of error β.

Some experimenters prefer to use the observed p-value of the test to evaluate the strength of the sample information in deciding to reject H_0. These values can usually be generated by computer and are often used in reports of statistical results:

- If the p-value is greater than .05, the results are reported as NS—not significant at the 5% level.
- If the p-value lies between .05 and .01, the results are reported as $P < .05$—significant at the 5% level.
- If the p-value lies between .01 and .001, the results are reported as $P < .01$—"highly significant" or significant at the 1% level.
- If the p-value is less than .001, the results are reported as $P < .001$—"very highly significant" or significant at the .1% level.

Still other researchers prefer to construct a confidence interval for a parameter and perform a test informally. If the value of the parameter specified by H_0 is included within the upper and lower limits of the confidence interval, then "H_0 is not rejected." If the value of the parameter specified by H_0 is not contained within the interval, then "H_0 is rejected." These results will agree with a two-tailed test; one-sided confidence bounds are used for one-tailed alternatives.

Finally, consider the choice between a one- and two-tailed test. In general, experimenters wish to know whether a treatment causes what could be a beneficial increase in a parameter or what might be a harmful decrease in a parameter. Therefore, most tests are two-tailed unless a one-tailed test is strongly dictated by practical considerations. For example, assume you will sustain a large financial loss if the mean μ is greater than μ_0 but not if it is less. You will then want to detect values larger than μ_0 with a high probability and thereby use a right-tailed test. In the same vein, if pollution levels higher than μ_0 cause critical health risks, then you will certainly wish to detect levels higher than μ_0 with a right-tailed test of hypothesis. In any case, the choice of a one- or two-tailed test should be dictated by the practical consequences that result from a decision to reject or not reject H_0 in favor of the alternative.

CHAPTER REVIEW

Key Concepts and Formulas

I. Parts of a Statistical Test

1. **Null hypothesis:** a contradiction of the alternative hypothesis
2. **Alternative hypothesis:** the hypothesis the researcher wants to support
3. **Test statistic** and its ***p*-value:** sample evidence calculated from the sample data
4. **Rejection region—critical values** and **significance levels:** values that separate rejection and nonrejection of the null hypothesis
5. **Conclusion:** Reject or do not reject the null hypothesis, stating the practical significance of your conclusion

II. Errors and Statistical Significance

1. The **significance level** α is the probability of rejecting H_0 when it is in fact true.
2. The ***p*-value** is the probability of observing a test statistic as extreme as or more extreme than the one observed; also, the smallest value of α for which H_0 can be rejected.
3. When the ***p*-value** is less than the **significance level** α, the null hypothesis is rejected. This happens when the **test statistic** exceeds the **critical value.**
4. In a **Type II error,** β is the probability of accepting H_0 when it is in fact false. The **power of the test** is $(1 - \beta)$, the probability of rejecting H_0 when it is false.

III. Large-Sample Test Statistics Using the z Distribution

To test one of the four population parameters when the sample sizes are large, use the following test statistics:

Parameter	Test Statistic
μ	$z = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
p	$z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$
$\mu_1 - \mu_2$	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$
$p_1 - p_2$	$z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ or $z = \dfrac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\dfrac{\hat{p}_1\hat{q}_1}{n_1} + \dfrac{\hat{p}_2\hat{q}_2}{n_2}}}$

Supplementary Exercises

Starred (*) exercises are optional.

9.52 **a.** Define α and β for a statistical test of hypothesis.

b. For a fixed sample size n, if the value of α is decreased, what is the effect on β?

c. In order to decrease both α and β for a particular alternative value of μ, how must the sample size change?

9.53 What is the p-value for a test of hypothesis? How is it calculated for a large-sample test?

9.54 What conditions must be met so that the z test can be used to test a hypothesis concerning a population mean μ?

9.55 Define the power of a statistical test. As the alternative value of μ gets farther from μ_0, how is the power affected?

9.56 **Acidity in Rainfall** Refer to Exercise 8.31 and the collection of water samples to estimate the mean acidity (in pH) of rainfalls in the northeastern United States. As noted, the pH for pure rain falling through clean air is approximately 5.7. The sample of $n = 40$ rainfalls produced pH readings with $\bar{x} = 3.7$ and $s = .5$. Do the data provide sufficient evidence to indicate that the mean pH for rainfalls is more acidic ($H_a: \mu < 5.7$ pH) than pure rainwater? Test using $\alpha = .05$. Note that this inference is appropriate only for the area in which the rainwater specimens were collected.