12

Testing Hypotheses About Proportions

LEARNING OBJECTIVES

In this chapter we show you how to test whether the proportion of a population with a certain characteristic is equal to, less than, or greater than a given value. After reading and studying this chapter, you should be able to:

1. Specify business issues in terms of hypothesis tests
2. Perform a hypothesis test about a proportion
3. See the relationship between hypothesis tests and confidence intervals
4. Estimate how powerful a hypothesis test is
5. Perform a hypothesis test comparing two proportions

Dow Jones

More than a hundred years ago, Charles Dow changed the way people look at the stock market. Surprisingly, he wasn't an investment wizard or a venture capitalist. He was a journalist who wanted to make investing understandable to ordinary people. Although he died at the relatively young age of 51 in 1902, his impact on how we track the stock market has been both long-lasting and far-reaching.

In the late 1800s, when Charles Dow reported on Wall Street, investors preferred bonds, not stocks. Bonds were reliable, backed by the real machinery and other hard assets the company owned. What's more, bonds were predictable; the bond owner knew when the bond would mature, and so knew when and how much the bond would pay. Stocks simply represented "shares" of ownership, which were risky and erratic. In May 1896, Dow and Edward Jones, whom he had known since their days as reporters for the *Providence Evening Press*, launched the now-famous Dow Jones Industrial Average (DJIA) to help the public understand stock market trends. The original DJIA averaged 11 stock prices. Of those original industrial stocks, only General Electric is still in the DJIA.

Since then, the DJIA has become synonymous with overall market performance and is often referred to simply as the Dow. Today the Dow is a weighted average of 30 industrial stocks, with weights used to account for splits and other adjustments. The "Industrial" part of the name is largely historical. Today's DJIA includes the service industry and financial companies and is much broader than just heavy industry.

Dow Jones also publishes international indices, including Canadian indices focusing on growth stocks and value stocks in large, medium, and small companies. Some of these, such as the Canada TMI (Total Market Index), provide a broad perspective on the Canadian equity market as a whole. Others, such as the Canada SDI (Select Dividend Index), focus on stocks that have dividends above the Canadian average. A rigorous screening process is applied before a stock is included in the index, including not just dividends paid out per share, but also sustainability of the dividends over the long term; the stock must also have a high trading volume. Dow Jones therefore provides investors with a range of indices that can be used as benchmarks for tracking the performance of the full range of Canadian companies.

Roadmap for Statistical Inference

Number of Variables	Objective	Large Sample or Normal Population		Small Sample and Non-normal Population or Non-numeric Data	
		Chapter	Parametric Method	Chapter	Nonparametric Method
1	Calculate confidence interval for a proportion	11			
1	Compare a proportion with a given value	12	*z*-test		
1	Calculate a confidence interval for a mean and compare it with a given value	13	*t*-test	17.2	Wilcoxon Signed-Rank Test
2	Compare two proportions	12.8	*z*-test		
2	Compare two means for independent samples	14.1–14.5	*t*-test	17.4, 17.5	Wilcoxon Rank-Sum (Mann-Whitney) Test Tukey's Quick Test
2	Compare two means for paired samples	14.6, 14.7	Paired *t*-test	17.2	Wilcoxon Signed-Rank Test
≥3	Compare multiple means	15	ANOVA: ANalysis Of VAriance	17.3 17.6	Friedman Test Kruskal-Wallis Test
≥3	Compare multiple counts (proportions)	16	χ^2 test		
2	Investigate the relationship between two variables	18	Correlation Regression	17.7, 17.8	Kendall's tau Spearman's rho
≥3	Investigate the relationship between multiple variables	20	Multiple Regression		

WHO	Days on which the stock market was open ("trading days")
WHAT	Closing price of the Dow Jones Industrial Average
UNITS	Points
WHEN	August 1982 to December 1986
WHY	To test a theory of stock market behaviour

How does the stock market move? Figure 12.1 shows the DJIA closing prices for the bull market that ran from mid-1982 to the end of 1986.

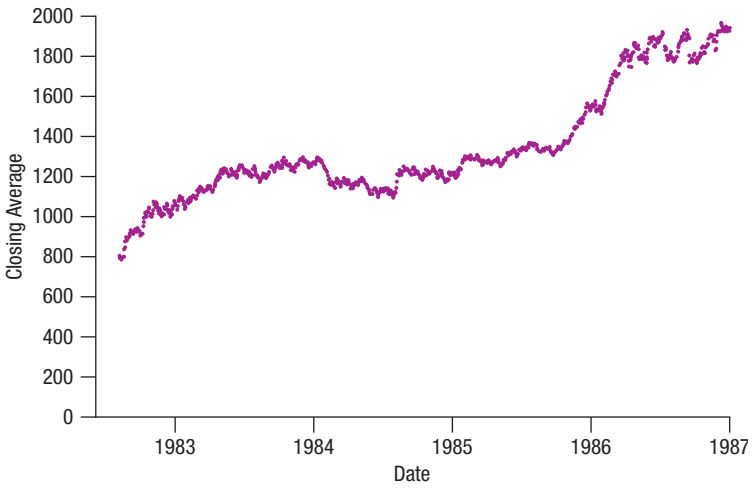


Figure 12.1 Daily closing prices of the Dow Jones Industrials from mid-1982 to the end of 1986.

The DJIA clearly increased during this famous bull market, more than doubling in value in less than five years. One common theory of market behaviour says that on a given day, the market is just as likely to move up as down. Another way of phrasing this is that the daily behaviour of the stock market is random. Can that be true during such periods of obvious increase? Let's investigate if the Dow is just as likely to move higher or lower on any given day. First we remove days on which the market was unchanged. Out of the 1112 trading days remaining, the average increased on 573 days, a sample proportion of 0.5153 or 51.53%. That's more "up" days than "down" days, but is it far enough from 50% to cast doubt on the assumption of an equally likely up or down movement?

LO ❶

12.1 Hypotheses

Hypothesis *n.*;
pl. {Hypotheses}.
A supposition; a proposition or principle which is supposed or taken for granted, in order to draw a conclusion or inference for proof of the point in question; something not proved, but assumed for the purpose of argument.
—*Webster's Unabridged Dictionary*, 1913

How can we state and test a hypothesis about daily changes in the DJIA? Hypotheses are working models that we adopt temporarily. To test whether the daily fluctuations are equally likely to be up as down, we assume that they are, and that any apparent difference from 50% is just random fluctuation. So, our starting hypothesis, called the null hypothesis, is that the proportion of days on which the DJIA increases is 50%. The **null hypothesis**, which we denote H_0, specifies a population model parameter and proposes a value for that parameter. We usually express a null hypothesis about a proportion in the form H_0: $p = p_0$. This is a concise way to specify the two things we need most: the identity of the parameter we hope to learn about (the true proportion) and a specific hypothesized value for that parameter (in this case, 50%). We need a hypothesized value so that we can compare our observed statistic to it. Which value to use for the hypothesis is not a statistical question. It may be obvious from the context of the data, but sometimes it takes a bit of thinking to translate the question we hope to answer into a hypothesis about a parameter. For our hypothesis about whether the DJIA moves up or down with equal likelihood, it's pretty clear that we need to test

$$H_0: p = 0.5.$$

Notation Alert!

Capital H is the standard letter for hypotheses. H_0 labels the null hypothesis, and H_A labels the alternative.

The **alternative hypothesis**, which we denote H_A, contains the values of the parameter that we consider plausible if we reject the null hypothesis. In our example, our null hypothesis is that the proportion, p, of "up" days is 0.5. What's the alternative? During a bull market, you might expect more up days than down, but we'll assume that we're interested in a deviation in either direction from the null hypothesis, so our alternative is

$$H_A: p \neq 0.5.$$

What would convince you that the proportion of up days was not 50%? If on 95% of the days the DJIA closed up, most people would be convinced that up and down days were not equally likely. But if the sample proportion of up days were only slightly higher than 50%, you'd be sceptical. After all, observations do vary, so we wouldn't be surprised to see some difference. How different from 50% must the proportion be before we *are* convinced that it has changed? Whenever we ask about the size of a statistical difference, we naturally think of the standard deviation. So let's start by finding the standard deviation of the sample proportion of days on which the DJIA increased.

We've seen 51.53% up days out of 1112 trading days. The sample size of 1112 is certainly big enough to satisfy the Success/Failure Condition. (We expect $0.50 \times 1112 = 556$ daily increases.) We suspect that the daily price changes are random and independent. And we know what hypothesis we're testing. To test a hypothesis we (temporarily) *assume* it's true so that we can see whether that description of the world is plausible. If we assume that the Dow increases or decreases with equal likelihood, we'll need to centre our Normal sampling model at a mean of 0.5. Then we can find the standard deviation of the sampling model as

$$\mathrm{SD}(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.5)(1 - 0.5)}{1112}} = 0.015.$$

- **Why is this a standard deviation and not a standard error?** This is a standard deviation because we haven't estimated anything. Once we assume that the null hypothesis is true, it gives us a value for the model parameter, p. With proportions, if we know p then we also automatically know its standard deviation. Because we find the standard deviation from the model parameter, this is a standard deviation and not a standard error. When we found a confidence interval for p, we could not assume that we knew its value, so we estimated the standard deviation from the sample value, $\hat{p}$.

To remind us that the parameter value comes from the null hypothesis, it's sometimes written as p_0 and the standard deviation as $\mathrm{SD}(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}$.

Now we know both parameters of the Normal sampling distribution model for our null hypothesis. For the mean, μ, we use $p = 0.50$, and for σ we use the standard deviation of the sample proportions $SD(\hat{p}) = 0.015$. We want to know how likely it would be to see the observed value $\hat{p}$ as far away from 50% as the value of 51.53% that we've actually observed. Looking first at a picture (Figure 12.2), we can see that 51.53% doesn't look very surprising. The more exact answer (from a calculator, a computer program, or the Normal table) is that the probability is about 0.308. This is the probability of observing more than 51.53% up days (or more than 51.53% down days) if the null model were true. In other words, if the chance of an up day for the Dow is 50%, we'd expect to see stretches of 1112 trading days with as many as 51.53% up days about 15.4% of the time and with as many as 51.53% down days about 15.4% of the time. That's not terribly unusual, so there's really no convincing evidence that the market did not act randomly.

It may surprise you that even during a bull market, the direction of daily movements is random. But the probability that any given day will end up or down appears to be about 0.5 regardless of the longer-term trends. It may be that when the stock market has a long run up (or possibly down, although we haven't checked that),

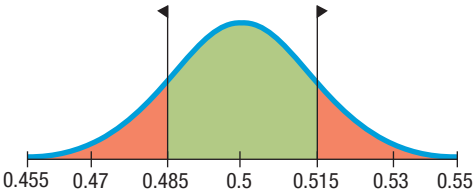


Figure 12.2 How likely is a proportion of more than 51.5% or less than 48.5% when the true mean is 50%? This is what it looks like. Each red area is 0.154 of the total area under the curve.

it does so not by having more days of increasing or decreasing value, but by the actual amounts of the increases or decreases being unequal.

In our example about the DJIA, we were equally interested in proportions that deviate from 50% in *either* direction. So we wrote our alternative hypothesis as H_A: $p \neq 0.5$. Such an alternative hypothesis is known as a **two-sided alternative**, because we are equally interested in deviations on either side of the null hypothesis value (see Figure 12.3). For two-sided alternatives, the P-value is the probability of deviating in *either* direction from the null hypothesis value.

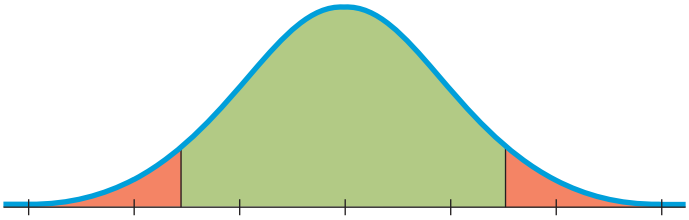


Figure 12.3 The P-value for a two-sided alternative adds the probabilities in both tails of the sampling distribution model outside the value that corresponds to the test statistic.

They make things admirably plain,
But one hard question will remain:
If one hypothesis you lose,
Another in its place you choose . . .

—James Russell Lowell,
credidimus jovem regnare

Suppose we want to test whether the proportion of customers returning merchandise has decreased under our new quality monitoring program. We know the quality has improved, so we can be pretty sure things haven't become worse. But have the customers noticed? We would only be interested in a sample proportion *smaller* than the null hypothesis value. We'd write our alternative hypothesis as H_A: $p < p_0$. An alternative hypothesis that focuses on deviations from the null hypothesis value in only one direction is called a **one-sided alternative** (see Figure 12.4).

The Three Alternative Hypotheses

Two-sided:
H_0: $p = p_0$
H_A: $p \neq p_0$
One-sided:
H_0: $p = p_0$
H_A: $p < p_0$
One-sided:
H_0: $p = p_0$
H_A: $p > p_0$

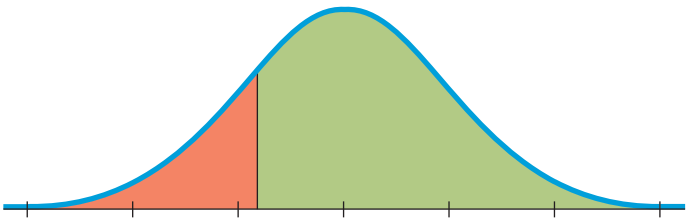


Figure 12.4 The P-value for a one-sided alternative considers only the probability of values beyond the test statistic value in the specified direction.

For a hypothesis test with a one-sided alternative, the P-value is the probability of deviating *only in the direction of the alternative* away from the null hypothesis value.

For Example Framing hypotheses

Summit Projects is a full-service interactive agency that offers companies a variety of website services. One of Summit's clients is SmartWool, which produces and sells wool apparel, including the famous SmartWool socks. Summit recently redesigned SmartWool's apparel website, and analysts at SmartWool wonder whether traffic has changed since the new website went live. In particular, an analyst might want to know if the proportion of visits resulting in a sale has changed since the new site went online.

Question: If the old site's proportion was 20%, frame appropriate null and alternative hypotheses for the proportion.

Answer: For the proportion, let p = proportion of visits that result in a sale.

$$H_0: p = 0.2 \text{ vs. } H_A \neq 0.2$$

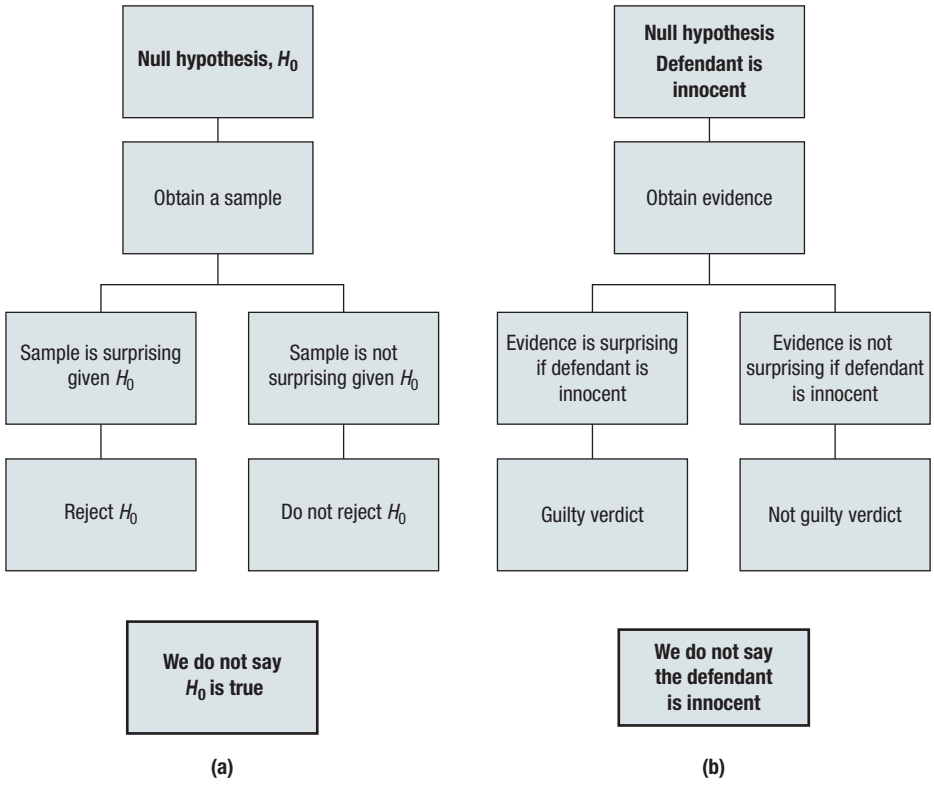
LO ❶ 12.2 A Trial as a Hypothesis Test

We started by assuming that the probability of an up day was 50%. Then we looked at the data and concluded that we couldn't say otherwise because the proportion we actually observed wasn't far enough from 50%. Does this reasoning of hypothesis tests seem backward? That could be because we usually prefer to think about getting things right rather than getting them wrong. But you've seen this reasoning before in a different context. This is the logic of jury trials.

Let's suppose a defendant has been accused of robbery. In British common law and those systems derived from it (including Canadian and U.S. law), the null hypothesis is that the defendant is innocent. Instructions to juries are quite explicit about this.

The evidence takes the form of facts that seem to contradict the presumption of innocence. For us, this means collecting data. In the trial, the prosecutor presents evidence. ("If the defendant were innocent, wouldn't it be remarkable that the police found him at the scene of the crime with a bag full of money in his hand, a mask on his face, and a getaway car parked outside?") The next step is to judge the evidence. Evaluating the evidence is the responsibility of the jury in a trial, but it falls on your shoulders in hypothesis testing. The jury considers the evidence in light of the *presumption* of innocence and judges whether the evidence against the defendant would be plausible *if the defendant were in fact innocent*.

Like the jury, we ask, "Could these data plausibly have happened by chance if the null hypothesis were true?" (See Figure 12.5.) If they're very unlikely to



I Figure 12.5 (a) Hypothesis testing. (b) Court case.

have occurred, then the evidence raises a reasonable doubt about the null hypothesis. Ultimately, *you* must make a decision. The standard of "beyond a reasonable doubt" is purposely ambiguous, because it leaves the jury to decide the degree to which the evidence contradicts the hypothesis of innocence. Juries don't explicitly use probability to help them decide whether to reject that hypothesis. But when you ask the same question of your null hypothesis, you have the advantage of being able to quantify exactly how surprising the evidence would be if the null hypothesis were true.

How unlikely is unlikely? Some people set rigid standards. Levels like 1 time out of 20 (0.05) or 1 time out of 100 (0.01) are common. But if *you* have to make the decision, you must judge for yourself in each situation whether the probability of observing your data is small enough to constitute "reasonable doubt."

LO ❶

12.3 P-Values

The fundamental step in our reasoning is the question "Are the data surprising, given the null hypothesis?" And the key calculation is to determine exactly how likely the data we observed would be if the null hypothesis were the true model of the world. So we need a *probability*. Specifically, we want to find the probability of seeing data like these (or something even less likely) *given* that we accept the null hypothesis. This probability is the value on which we base our decision, so statisticians give this probability a special name, the **P-value**.

Beyond a Reasonable Doubt

We ask whether the data were unlikely beyond a reasonable doubt. The probability that the observed statistic value (or an even more extreme value) could occur if the null model were true is the P-value.

A low enough P-value says that the data we've observed would be very unlikely if our null hypothesis were true. We started with a model, and now that same model tells us that the data we have are unlikely to have happened. That's surprising. In this case, the model and data are at odds with each other, so we have to make a choice. Either the null hypothesis is correct and we've just seen something remarkable, or the null hypothesis is wrong (and, in fact, we were wrong to use it as the basis for computing our P-value). If you believe in data more than in assumptions, then, given that choice, when you see a low P-value you should reject the null hypothesis.

When the P-value is *high* (or just not low *enough*), what do we conclude? In that case, we haven't seen anything unlikely or surprising at all. The data are consistent with the model from the null hypothesis, and we have no reason to reject the null hypothesis. Events that have a high probability of happening happen all the time. So when the P-value is high, does that mean we've proved the null hypothesis is true? No! We realize that many other similar hypotheses could also account for the data we've seen. The most we can say is that it doesn't appear to be false. Formally, we say that we "fail to reject" the null hypothesis. That may seem to be a pretty weak conclusion, but it's all we can say when the P-value isn't low enough. All that means is that the data are consistent with the model we started with.

What to Do with an "Innocent" Defendant

Let's see what that last statement means in a jury trial. If the evidence isn't strong enough to reject the defendant's presumption of innocence, what verdict does the jury return? They don't say that the defendant is innocent. They say "not guilty." All they're saying is that they haven't seen sufficient evidence to reject innocence and convict the defendant. The defendant may, in fact, be innocent, but the jury has no way to be sure.

Expressed statistically, the jury's null hypothesis is "innocent defendant." If the evidence is too unlikely (the P-value is low), then, given the assumption of innocence, the jury rejects the null hypothesis and finds the defendant guilty. But—and this is an important distinction—if there's *insufficient evidence* to convict the defendant (if the P-value is *not* low), the jury does not conclude that the null hypothesis is

Don't We Want to Reject the Null?

Often the people who collect the data or perform the experiment hope to reject the null. They hope the new drug is better than the placebo; they hope the new ad campaign is better than the old one; or they hope their candidate is ahead of the opponent. But when we practise Statistics, we can't allow that hope to affect our decision. The essential attitude for a hypothesis tester is scepticism. Until we become convinced otherwise, we cling to the null's assertion that there's nothing unusual, nothing unexpected, no effect, no difference, etc. As in a jury trial, the burden of proof rests with the alternative hypothesis—innocent until proven guilty. When you test a hypothesis, you must act as judge and jury; you're not the prosecutor.

Conclusion

If the P-value is "low," reject H_0 and conclude H_A.

If the P-value isn't "low enough," then fail to reject H_0 and the test is inconclusive.

true and declare that the defendant is innocent. Juries can only *fail to reject* the null hypothesis and declare the defendant "not guilty."

In the same way, if the data aren't particularly unlikely under the assumption that the null hypothesis is true, then the most we can do is to "fail to reject" our null hypothesis. We never declare the null hypothesis to be true. In fact, we simply don't know whether it's true or not. (After all, more evidence may come along later.)

Imagine a test of whether a company's new website design encourages a higher percentage of visitors to make a purchase (as compared with the site it's used for years). The null hypothesis is that the new site is no more effective at stimulating purchases than the old one. The test sends visitors randomly to one version of the website or the other. Of course, some will make a purchase, and others won't. If we compare the two websites on only 10 customers each, the results are likely *not to be clear*, and we'll be unable to reject the hypothesis. Does this mean the new design is a complete bust? Not necessarily. It simply means that we don't have enough evidence to reject our null hypothesis. That's why we don't start by assuming that the new design is *more* effective. If we were to do that, then we could test just a few customers, find that the results aren't clear, and claim that since we've been unable to reject our original assumption, the redesign must be effective. The board of directors is unlikely to be impressed by that argument.

For Example Conclusions from P-values

Question: The SmartWool analyst (see page 391) collects a representative sample of visits since the new website has gone online and finds that the P-value for the test of proportion is 0.0015. What conclusions can she draw?

Answer: The proportion of visits that resulted in a sale since the new website went online is very unlikely to still be 0.20. There is strong evidence to suggest that the proportion has changed. She should reject the null hypotheses.

Just Checking

1 A pharmaceutical firm wants to know whether aspirin helps to thin blood. The null hypothesis says that it doesn't. The firm's researchers test 12 patients, observe the proportion with thinner blood, and get a P-value of 0.32. They proclaim that aspirin doesn't work. What would you say?

2 An allergy drug has been tested and found to give relief to 75% of the patients in a large clinical trial. Now the scientists want to see whether a new, "improved" version works even better. What would the null hypothesis be?

3 The new allergy drug in Question 2 above is tested, and the P-value is 0.0001. What would you conclude about the drug?

LO ❷

12.4 The Reasoning of Hypothesis Testing

Hypothesis tests follow a carefully structured path. To avoid getting lost as we navigate down it, we divide that path into four distinct sections: hypotheses, model, mechanics, and conclusion.

Hypotheses

The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

—Sir Ronald Fisher, *The Design of Experiments*, 1931

First, state the null hypothesis. That's usually the sceptical claim that nothing's different. The null hypothesis assumes the default (often the status quo) is true (the defendant is innocent, the new method is no better than the old, customer preferences haven't changed since last year, the DJIA goes up as often as it goes down, etc.).

In statistical hypothesis testing, hypotheses are almost always about model parameters. To assess how unlikely our data may be, we need a null model. The null hypothesis specifies a particular parameter value to use in our model. In the usual notation, we write H_0: *parameter* = *hypothesized value*. The alternative hypothesis, H_A, contains the values of the parameter we consider plausible when we reject the null.

Model

When the Conditions Fail . . .

You might proceed with caution, explicitly stating your concerns. Or you may need to do the analysis with and without an outlier, or on different subgroups, or after re-expressing the response variable. Or you may not be able to proceed at all.

To plan a statistical hypothesis test, specify the *model* for the sampling distribution of the statistic you'll use to test the null hypothesis and the parameter of interest. For instance, the parameter might be the proportion of days on which the DJIA went up. For proportions, we use the Normal model for the sampling distribution. Of course, all models require assumptions, so you'll need to state them and check any corresponding conditions. For a test of a proportion, the assumptions and conditions are the same as for a one-proportion z-interval.

Your model step should end with a statement such as: *Because the conditions are satisfied, we can model the sampling distribution of the proportion with a Normal model.* Watch out, though. Your model step could end with: *Because the conditions are not satisfied, we can't proceed with the test.* (If that's the case, stop and reconsider.)

Each test we discuss in this book has a name that you should include in your report. We'll see many tests in the following chapters. Some will be about more than one sample, some will involve statistics other than proportions, and some will use models other than the Normal (and so will not use z-scores). The test about proportions is called a **one-proportion z-test**.[1]

One-Proportion z-Test

The conditions for the one-proportion z-test are the same as for the one-proportion z-interval. We test the hypothesis H_0: $p = p_0$ using the statistic

$$z = \frac{(\hat{p} - p_0)}{SD(\hat{p})}.$$

We use the hypothesized proportion to find the standard deviation: $SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}$. When the conditions are met and the null hypothesis is true, this statistic follows the standard Normal model, so we can use that model to obtain a P-value.

Conditional Probability

Did you notice that a P-value results from what we referred to as a "conditional probability" in Chapter 8? A P-value is a "conditional probability" because it's based on—or is conditional on—another event being true: It's the probability that the observed results could have happened *if the null hypothesis is true*.

Mechanics

Under "Mechanics" we perform the actual calculation of our test statistic from the data. Different tests we encounter will have different formulas and different test statistics. Usually, the mechanics are handled by a statistics program or calculator. The ultimate goal of the calculation is to obtain a P-value—the probability that the observed statistic value (or an even more extreme value) could occur if the null model were correct. If the P-value is small enough, we'll reject the null hypothesis.

[1] It's also called the "one-sample test for a proportion."

Assumptions and Conditions

Hypothesis testing requires the same four assumptions and conditions that we use in calculating confidence intervals:

- **Independence Assumption:** The individuals in the sample behave independently of each other.
- **Randomization Condition:** The individuals in each sample were selected at random.
- **10% Condition:** The sample is less than 10% of the population.
- **Success/Failure Condition:**

  $np_0 > 10$;

  $nq_0 > 10$.

Conclusion and Decisions

The primary conclusion in a formal hypothesis test is only a statement about the null hypothesis. It simply states whether we reject or fail to reject that hypothesis. As always, the conclusion should be stated in context, but your conclusion about the null hypothesis should never be the end of the process. You can't make a decision based solely on a P-value. Business decisions have consequences, with actions to take or policies to change. The conclusions of a hypothesis test can help *inform* your decision, but they shouldn't be the only basis for it.

Business decisions should always take into consideration three things: the statistical significance of the test, the *cost* of the proposed action, and the *effect size* of the statistic observed. For example, a cell phone provider finds that 30% of its customers switch providers (or *churn*) when their two-year subscription contract expires. The provider tries a small experiment and offers a random sample of customers a free $350 top-of-the-line phone if they renew their contracts for another two years. Not surprisingly, the provider finds that the new switching rate is lower by a statistically significant amount. Should it offer these free phones to all its customers? Obviously, the answer depends on more than the P-value of the hypothesis test. Even if the P-value is statistically significant, the correct business decision also depends on the cost of the free phones and by how much the churn rate is lowered (the effect size). It's rare that a hypothesis test alone is enough to make a sound business decision.

For Example — The reasoning of hypothesis tests

Question: The analyst at SmartWool (see page 394) selects 200 recent weblogs at random and finds that 58 of them have resulted in a sale. The null hypothesis is that $p = 0.20$. Would this be a surprising proportion of sales if the true proportion of sales were 20%?

Answer: To judge whether 58 is a surprising number of sales given the null hypothesis, we use the Normal model based on the null hypothesis. That is, we use 0.20 as the mean and $\sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.2)(0.8)}{200}} = 0.02828$ as the standard deviation.

58 sales is a sample proportion of $\hat{p} = \frac{58}{200} = 0.29$ or 29%.

The z-value for 0.29 is then $z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{0.29 - 0.20}{0.02828} = 3.182$.

In other words, given that the null hypothesis is true, our sample proportion is 3.182 standard deviations higher than the mean. That seems like a surprisingly large value, since the probability of being farther than three standard deviations from the mean is (from the 68-95-99.7 Rule) only 0.3%.

Guided Example — Home Field Advantage

Major league sports are big business. And the fans are more likely to come out to root for the team if the home team has a good chance of winning. Anyone who follows or plays sports has heard of the "home field advantage." It is said that teams are more likely to win when they play at home. That *would* be good for encouraging the fans to come to the games. But is it true?

In the 2006 Major League Baseball (MLB) season, there were 2429 regular season games. (One rained-out game was never made up.) It turns out that the home team won 1327 of the 2429 games, or 54.63% of the time. If there were no home field advantage, the home teams would win about half of all games played. Could this deviation from 50% be explained just from natural sampling variability, or does this evidence suggest that there really is a home field advantage, at least in professional baseball?

To test the hypothesis, we'll ask whether the observed rate of home team victories, 54.63%, is so much greater than 50% that we can't explain it away as just chance variation.

Remember the four main steps in performing a hypothesis test—hypotheses, model, mechanics, and conclusion? Let's put them to work and see what this will tell us about the home team's chances of winning a baseball game.

PLAN

Setup State what we want to know. Define the variables and discuss their context.

We want to know whether the home team in professional baseball is more likely to win. The data are all 2429 games from the 2006 Major League Baseball season. The variable is whether or not the home team won. The parameter of interest is the proportion of home team wins. If there is an advantage, we'd expect that proportion to be greater than 0.50. The observed statistic value is $\hat{p} = 0.5463$.

Hypotheses The null hypothesis makes the claim of no home field advantage.

$$H_0: p = 0.50$$

We're interested only in a home field *advantage*, so the alternative hypothesis is one-sided.

$$H_A: p > 0.50$$

Model Think about the assumptions and check the appropriate conditions.

- ✓ **Independence Assumption.** Generally, the outcome of one game has no effect on the outcome of another game. But this may not always be strictly true. For example, if a key player is injured, the probability that the team will win in the next couple of games may decrease slightly, but independence is still roughly true.
- ✓ **Randomization Condition.** We have results for all 2429 games of the 2006 season. But we're not just interested in 2006. While these games were not randomly selected, they may be reasonably representative of all recent professional baseball games.
- ✓ **10% Condition.** This is not a random sample, but these 2429 games are fewer than 10% of all games played over the years.
- ✓ **Success/Failure Condition.** Both $np_0 = 2429(0.50) = 1214.5$ and $nq_0 = 2429(0.50) = 1214.5$ are at least 10.

Consider the time frame carefully.

Specify the sampling distribution model.

Because the conditions are satisfied, we'll use a Normal model for the sampling distribution of the proportion and do a one-proportion z-test.

Tell what test you plan to use.

(continued)

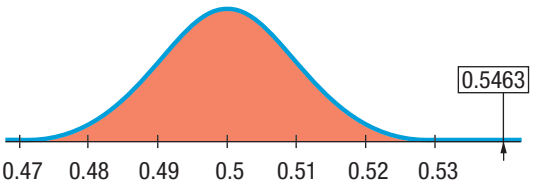
DO

Mechanics The null model gives us the mean, and (because we're working with proportions) the mean gives us the standard deviation.

The null model is a Normal distribution with a mean of 0.50. Since this is a hypothesis test, the standard deviation is calculated from p_0 (in the null hypothesis).

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.5)(1 - 0.5)}{2429}} = 0.01015$$

The observed proportion $\hat{p}$ is 0.5463.



From technology or numerical calculation and the Normal distribution table, we can find the P-value, which tells us the probability of observing a value that extreme (or more).

The probability of observing a $\hat{p}$ of 0.5463 or more in our Normal model can be found by computer, calculator, or table to be < 0.001.

$$z = \frac{0.5463 - 0.5}{0.01015} = 4.56$$

The corresponding P-value is < 0.001.

REPORT

Conclusion State your conclusion about the parameter—in context.

MEMO:

Re: Home Field Advantage

Our analysis of outcomes during the 2006 Major League Baseball season showed a statistically significant advantage to the home team ($P < 0.001$). We can conclude that playing at home gives a baseball team an advantage.

In the Guided Example on home field advantage, we never even considered home field disadvantage. Some statisticians build this into the null hypothesis and write

$$H_0: p \leq 0.50$$

$$H_A: p > 0.50,$$

which spells out the fact that there's a possibility of a home field disadvantage. The calculations are exactly the same; the only difference is the way the null hypothesis is written. In this book, we'll always use an exact value in our null hypotheses, since that corresponds to most practical situations. We usually have a number, p_0, and we're testing whether our proportion is different from that number. Table 12.1 summarizes the three types of hypothesis tests.

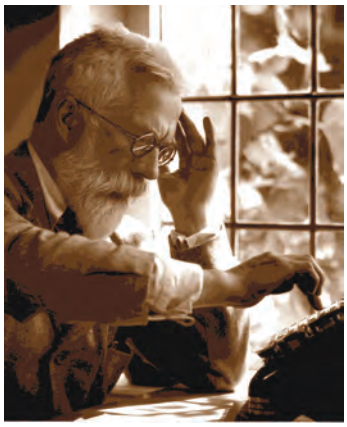
Notice that the null hypothesis always has an "equals" sign. The alternative hypothesis involves "less than," "greater than," or "not equal to."

	Two-Sided	One-Sided	One-Sided
How we write it in this book	$H_0: p = p_0$ $H_A: p \neq p_0$	$H_0: p = p_0$ $H_A: p > p_0$	$H_0: p = p_0$ $H_A: p < p_0$
How some people write it to spell out the details	No change, i.e., $H_0: p = p_0$ $H_A: p \neq p_0$	$H_0: p \leq p_0$ $H_0: p > p_0$	$H_0: p \geq p_0$ $H_0: p < p_0$
Practical example	Is the proportion of "up" days on the stock market different from the proportion of "down" days?	Is there a home field advantage?	Are customers returning fewer items this year than the 3% they returned last year?
p_0 in the example	0.5	0.5	0.03

Table 12.1 Three types of hypothesis test.

LO ❶

12.5 Alpha Levels and Significance



Sir Ronald Fisher (1890–1962) was one of the founders of modern Statistics.

Sometimes we need to make a firm decision about whether to reject the null hypothesis. A jury must *decide* whether the evidence reaches the level of "beyond a reasonable doubt." A business must *select* a Web design. You need to decide which section of a Statistics course to enrol in.

When the P-value is small, it tells us that our data are rare *given the null hypothesis*. As humans, we're suspicious of rare events. If the data are "rare enough," we just don't think that could have happened due to chance. Since the data *did* happen, something must be wrong. All we can do now is reject the null hypothesis.

But how rare is "rare"? How low does the P-value have to be?

We can define "rare event" arbitrarily by setting a threshold for our P-value. If our P-value falls below that point, we'll reject the null hypothesis. We call such results *statistically significant*. The threshold is called an **alpha level**. Not surprisingly, it's labelled with the Greek letter α. Common α-levels are 0.10, 0.05, and 0.01. You have the option—almost the *obligation*—to consider your alpha level carefully and choose an appropriate one for the situation. If you're assessing the safety of air bags, you'll want a low alpha level; even 0.01 might not be low enough. If you're just wondering whether folks prefer their pizza with or without pepperoni, you might be happy with $\alpha = 0.10$. It can be hard to justify your choice of α, though, so often we arbitrarily choose 0.05.

- **Where did the value 0.05 come from?** In 1931, in a famous book called *The Design of Experiments*, Sir Ronald Fisher discussed the amount of evidence needed to reject a null hypothesis. He said that it was situation dependent, but remarked, somewhat casually, that for many scientific applications, 1 out of 20 might be a reasonable value, especially in a first experiment—one that will be followed by confirmation. Since then, some people—indeed some entire disciplines—have acted as if the number 0.05 were sacrosanct.

Notation Alert!

The first Greek letter α is used in Statistics for the threshold value of a hypothesis test. You'll hear it referred to as the alpha level. Common values are 0.10, 0.05, 0.01, and 0.001.

The alpha level is also called the **significance level**. When we reject the null hypothesis, we say that the test is "significant at that level." For example, we might say that we reject the null hypothesis that the DJIA goes up on 50% of days "at the 5% level of significance." You must select the alpha level *before* you look at the data. Otherwise, you can be accused of finagling the conclusions by tuning the alpha level to the results after you've seen the data.

What can you say if the P-value does not fall below α? When you haven't found sufficient evidence to reject the null according to the standard you've established, you should say, "The data have failed to provide sufficient evidence to reject the null hypothesis." Don't say, "We accept the null hypothesis." You certainly haven't proven or established the null hypothesis; it was assumed to begin with. You *could* say that you have *retained* the null hypothesis, but it's better to say that you've failed to reject it.

It Could Happen to You!

Of course, if the null hypothesis *is* true, no matter what alpha level you choose, you still have a probability α of rejecting the null hypothesis by mistake. When we do reject the null hypothesis, no one ever thinks that *this* is one of those rare times. As statistician Stu Hunter notes, "The statistician says 'rare events do happen—but not to me!'"

Conclusion

If the P-value $< \alpha$, then reject H_0.
If the P-value $\geq \alpha$, then fail to reject H_0.

Look again at the home field advantage example. The P-value was < 0.001. This is so much smaller than any reasonable alpha level that we can reject H_0. We concluded: "We reject the null hypothesis. There is sufficient evidence to conclude that there is a home field advantage over and above what we expect with random variation."

The automatic nature of the reject/fail-to-reject decision when we use an alpha level may make you uncomfortable. If your P-value falls just slightly above your alpha level, you're not allowed to reject the null. Yet a P-value just barely below the alpha level leads to rejection. If this bothers you, you're in good company. Many statisticians think it better to report the P-value than to choose an alpha level and carry the decision through to a final reject/fail-to-reject verdict. So when you declare your decision, it's always a good idea to report the P-value as an indication of the strength of the evidence.

- **It's in the stars.** Some disciplines carry the idea further and code P-values by their size. In this scheme, a P-value between 0.05 and 0.01 gets highlighted by a single asterisk (*). A P-value between 0.01 and 0.001 gets two asterisks (**), and a P-value less than 0.001 gets three (***). This can be a convenient summary of the weight of evidence against the null hypothesis, but it isn't wise to take the distinctions too seriously and make black-and-white decisions near the boundaries. The boundaries are a matter of tradition, not science; there is nothing special about 0.05. A P-value of 0.051 should be looked at seriously and not casually thrown away just because it's larger than 0.05, and one that's 0.009 is not very different from one that's 0.011.

The importance of P-values is also clear in the common situation in which the person performing the statistical analysis isn't the decision maker. In many organizations, statistical results are reported to management, which then makes the decision on whether to accept the alternative hypothesis. Pharmaceutical companies developing drugs spend millions of dollars testing whether a new drug is more effective than existing drugs, and their reports are filled with P-values. But the decision on whether a new drug is better and whether to manufacture it is made by management taking into account all those P-values plus numerous other factors. Suppose management wants to be 95% sure the new drug is better. A statistical report shouldn't simply do a hypothesis test with a $\sigma = 0.05$ and state that the hypothesis test shows the new drug is better. It should also give the P-value. A P-value of 0.01 leads to the same hypothesis test result as a P-value of 0.045, but it gives the decision maker more confidence in the results.

Sometimes it's best to report that the conclusion is not yet clear and to suggest that more data be gathered. (In a trial, a jury may "hang" and be unable to return a verdict.) In such cases, it's an especially good idea to report the P-value, since it's the best summary we have of what the data say or fail to say about the null hypothesis.

What do we mean when we say that a test is statistically significant? All we mean is that the test statistic had a P-value lower than our alpha level. Don't be lulled into thinking that "statistical significance" necessarily carries with it any practical importance or impact.

For large samples, even small, unimportant ("insignificant") deviations from the null hypothesis can be statistically significant. On the other hand, if the sample isn't large enough, even large, financially or scientifically important differences may not be statistically significant.

It's good practice to report the magnitude of the difference between the observed statistic value and the null hypothesis value (in the data units) along with the P-value on which you've based your decision about statistical significance.

Practical vs. Statistical Significance

A large insurance company mined its data and found a statistically significant (P = 0.04) difference between the mean value of policies sold in 2010 and those sold in 2011. The difference in the mean values was $0.98. Even though it was statistically significant, management did not see this as an important difference when a typical policy sold for more than $1000. On the other hand, a marketable improvement of 10% in the relief rate for a new pain medicine may not be statistically significant unless a large number of people are tested. The effect, which is economically significant, might not be statistically significant.

For Example Setting the α level

Question: The manager of the analyst at SmartWool (see pages 394 and 396) wants her to use an α level of 0.05 for all her hypothesis tests. Would her conclusion have changed if she used an α level of 0.05?

Answer: Using $\alpha = 0.05$, we reject the null hypothesis when the P-value is less than 0.05 and fail to reject when the P-value is greater than or equal to 0.05. For the test of proportion, $p = 0.00146$, which is much less than 0.05 and so we reject; in other words, our conclusion is unchanged.

LO ❷

12.6 Critical Values

When building a confidence interval, we found a **critical value**, z^*, to correspond to our selected confidence level. Critical values can also be used as a shortcut for hypothesis tests. Any z-score larger in magnitude (i.e., more extreme) than a particular critical value has to be less likely, so it will have a P-value smaller than the corresponding alpha.

If we were willing to settle for a flat reject/fail-to-reject decision, comparing an observed z-score with the critical value for a specified alpha level would give a shortcut path to that decision. For the home field advantage example, if we choose $\alpha = 0.05$, then in order to reject H_0, our z-score has to be larger than the one-sided critical value of 1.645. The observed proportion was actually 4.56 standard deviations above 0.5, so we clearly reject the null hypothesis. This is perfectly correct and does give us a yes/no decision, but it gives us less information about the hypothesis because we don't have the P-value to think about. With technology, P-values are easy to find. And since they give more information about the strength of the evidence, you should report them.

Table 12.2 gives the traditional z^* critical values from the Normal model, as illustrated in Figures 12.6 and 12.7.[2]

If you need to make a decision on the fly with no technology, remember "2." That's our old friend from the 68-95-99.7 Rule. It's roughly the critical value for testing a hypothesis against a two-sided alternative at $\alpha = 0.05$. The exact critical value is 1.96, but 2 is close enough for most decisions.

[2]In a sense, these are the flip side of the 68-95-99.7 Rule. There we chose simple statistical distances from the mean and recalled the areas of the tails. Here we select convenient tail areas (0.05, 0.01, and 0.001, either on one side or adding the two together) and record the corresponding statistical distances.

α	One-Sided	Two-Sided
0.10	1.28	1.645
0.05	1.645	1.96
0.01	2.33	2.576
0.001	3.09	3.29

Table 12.2 Critical values, z^*, for different types of hypothesis test.

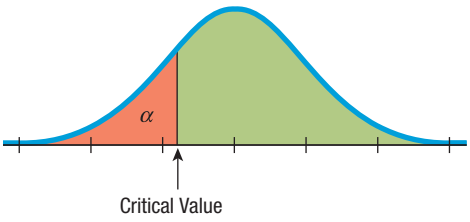


Figure 12.6 When the alternative is one-sided, the critical value puts all of α on one side.

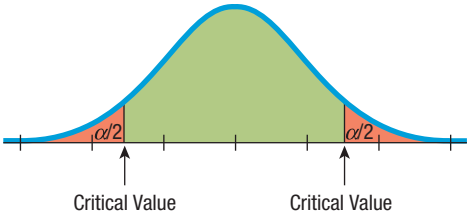


Figure 12.7 When the alternative is two-sided, the critical value splits α equally into two tails.

For Example Testing using critical values

Question: Find the critical z value for the SmartWool hypothesis (see pages 394 and 396) using $\alpha = 0.05$ and show that the same decision would have been made using critical values.

Answer: For the two-sided test of proportions, the critical z values at $\alpha = 0.05$ are ± 1.96. Because the z value was 3.182, much larger than 1.96, we reject the null hypothesis.

LO ❸

Notation Alert:

We've attached symbols to many of the p's. Let's keep them straight.

p is a population parameter—the true proportion in the population.

p_0 is a hypothesized value of p.

$\hat{p}$ is an observed proportion.

12.7 Confidence Intervals and Hypothesis Tests

Confidence intervals and hypothesis tests are built from similar calculations. They have the same assumptions and conditions. As we've just seen, you can approximate a hypothesis test by examining the confidence interval. Just ask whether the null hypothesis value is consistent with a confidence interval for the parameter at the corresponding confidence level. Because confidence intervals are naturally two-sided, they correspond to two-sided tests. For example, a 95% confidence interval corresponds to a two-sided hypothesis test at $\alpha = 5\%$. In general, a confidence interval with a confidence level of $C\%$ corresponds to a two-sided hypothesis test with an α level of $100 - C\%$.

The relationship between confidence intervals and one-sided hypothesis tests gives us a choice between one- and two-sided confidence intervals.

One-Sided Confidence Intervals

For a one-sided test with $\alpha = 5\%$, you could construct a one-sided confidence interval, leaving 5% in one tail and extending to infinity the other side. A one-sided confidence interval leaves one side unbounded. For example, in the home field scenario, we wondered whether the home field gave the home team an *advantage*, so our test was naturally one-sided. A 95% one-sided confidence interval would be constructed from one side of the associated two-sided confidence interval:

$$0.5463 - 1.645 \times 0.0101 = 0.530$$

Difference Between Hypothesis Tests and Confidence Intervals

Watch out for a subtle difference between the calculations for hypothesis tests and confidence intervals. Although they're very similar, they're not identical. An easy way to remember this difference is to focus on what information is available. For a confidence interval, all we have available is the proportion from our sample, whereas for a hypothesis test we *also* have the hypothesized value for the population.

For a confidence interval, we estimate the standard deviation of $\hat{p}$ from $\hat{p}$ itself, making it a *standard error*,

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

For the corresponding hypothesis test, we use the model's *standard deviation* for $\hat{p}$ based on the null hypothesis value p_0,

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}}.$$

When $\hat{p}$ and p_0 are close, these calculations give similar results. When they differ, you're likely to reject H_0 (because the observed proportion is far from your hypothesized value). In that case, you're better off building your confidence interval with a standard error estimated from the data rather than relying on the model you just rejected.

In order to leave 5% on one side, we used the z^* value 1.645, which leaves 5% in one tail. Writing the one-sided interval as 0.530, ∞ allows us to say with 95% confidence that we know the home team will win, on average, at least 53.0% of the time. To test the hypothesis H_0: $p = 0.50$, we note that the value 0.50 is not in this interval. The lower bound of 0.53 is clearly above 0.50, showing the connection between hypothesis and confidence intervals, as shown in Figure 12.8 (a).

Two-Sided Confidence Intervals

For convenience, and to provide more information, we sometimes report a two-sided confidence interval even though we're interested in a one-sided test. For the home field example, we could report a 90% confidence interval:

$$0.5463 \pm 1.645 \times 0.0101 = (0.530, 0.563)$$

Notice that we *matched* the left-end point by leaving α in *both* sides, which made the corresponding confidence level 90%. We can still see the correspondence that since the (two-sided) confidence interval for $\hat{p}$ doesn't contain 0.50, we reject the null hypothesis, but it also tells us that the home team winning percentage is unlikely to be greater than 56.3%, an added benefit to understanding. You can see the relationship between the two confidence intervals in Figure 12.8.

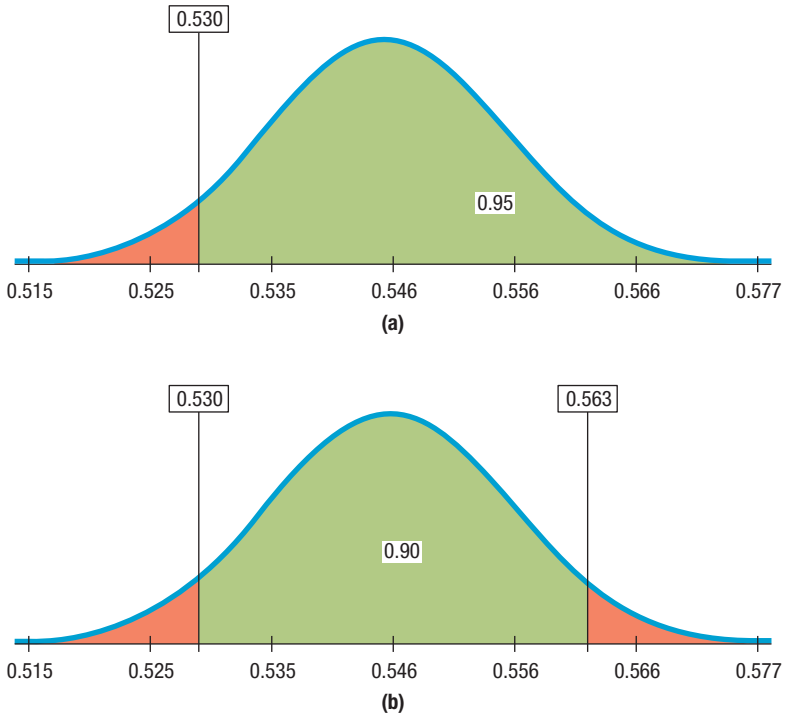


Figure 12.8 (a) The one-sided 95% confidence interval (top) leaves 5% on one side (in this case the left), but leaves the other side unbounded. (b) The 90% confidence interval is symmetric and matches the one-sided interval on the side of interest. Both intervals indicate that a one-sided test of p = 0.50 would be rejected at $\alpha = 0.05$.

Extraordinary claims require extraordinary proof.

—Carl Sagan

There's another good reason for finding a confidence interval along with a hypothesis test. Although the test can tell us whether the observed statistic differs from the hypothesized value, it doesn't say by how much. Often, business decisions depend not only on whether there's a statistically significant difference, but also on whether the difference is meaningful. For the home field advantage, the corresponding confidence interval shows that over a full season, home field advantage adds an average of about two to six extra victories for a team. That could make a meaningful difference in both the team's standing and the size of the crowd.

Just Checking

4 A bank is testing a new method for getting delinquent customers to pay their past-due credit card bills. The standard way was to send a letter (costing about $0.60 each) asking the customer to pay. That worked 30% of the time. The bank wants to test a new method that involves sending a DVD to customers encouraging them to contact the bank and set up a payment plan. Developing and sending the DVD costs about $10 per customer. What is the parameter of interest? What are the null and alternative hypotheses?

5 The bank sets up an experiment to test the effectiveness of the DVD. The DVD is mailed to several randomly selected delinquent customers, and employees keep track of how many customers then contact the bank to arrange payments. The bank just got back the results on its test of the DVD strategy: A 90% confidence interval for the success rate is (0.29, 0.45). Its old send-a-letter method had worked 30% of the time. Can you reject the null hypothesis and conclude that the method increases the proportion at $\alpha = 0.05$? Explain.

6 Given the confidence interval the bank found in the trial of the DVD mailing, what would you recommend be done? Should the bank scrap the DVD strategy?

Guided Example Credit Card Promotion

A credit card company plans to offer a special incentive program to customers who charge at least $500 next month. The marketing department has pulled a sample of 500 customers from the same month last year and noted that the mean amount charged was $478.19 and the median amount was $216.48. The finance department says that the only relevant quantity is the proportion of customers who spend more than $500. If that proportion is more than 25%, the program will make money.

Among the 500 customers, 148, or 29.6% of them, charged $500 or more. Can we use a confidence interval to test whether the goal of 25% for all customers was met?

PLAN

Setup State the problem and discuss the variables and the context.

We want to know whether more than 25% of customers will spend $500 or more in the next month and qualify for the special program. We will use the data from the same month a year ago to estimate the proportion and see whether it was at least 25%.

Hypotheses The null hypothesis is that the proportion qualifying is 25%. The alternative is that it's higher. It's clearly a one-sided test, so if we use a confidence interval, we'll have to be careful about what level we use.

The statistic is $\hat{p} = 0.296$, the proportion of customers who charged $500 or more.

$H_0: p = 0.25$

$H_A: p > 0.25$

Model Check the conditions.

State your method. Here we're using a confidence interval to test a hypothesis.

- ✓ **Independence Assumption.** Customers aren't likely to influence one another when it comes to spending on their credit cards.
- ✓ **Randomization Condition.** This is a random sample from the company's database.
- ✓ **10% Condition.** The sample is less than 10% of all customers.
- ✓ **Success/Failure Condition.**

  $np_0 = 500 \times 0.25 = 125$

  $nq_0 = 500 \times 0.75 = 375$

  Since both are > 10, our sample size is large enough.

Under these conditions, the sampling model is Normal. We'll create a one-proportion z-interval.

DO

Mechanics Write down the given information and determine the sample proportion.

To use a confidence interval, we need a confidence level that corresponds to the alpha level of the test. If we use $\alpha = 0.05$, we should construct a 90% confidence interval, because this is a one-sided test. That will leave 5% on *each* side of the observed proportion. Determine the standard error of the sample proportion and the margin of error. The critical value is $z^* = 1.645$.

The confidence interval is estimate $\pm$ margin of error.

$n = 500$, so

$$\hat{p} = \frac{148}{500} = 0.296$$

Since we're calculating a confidence interval, the standard error is obtained from $\hat{p}$. Contrast the hypothesis test in the previous Guided Example.

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.296)(0.704)}{500}} = 0.0204$$

$$ME = z^* \times SE(\hat{p})$$
$$= 1.645(0.0204) = 0.034$$

The 90% confidence interval is 0.296 ± 0.034, or $(0.262, 0.330)$.

REPORT

Conclusion Link the confidence interval to your decision about the null hypothesis, then state your conclusion in context.

MEMO:

Re: Credit Card Promotion

Our study of a sample of customer records indicates that between 26.2% and 33.0% of customers charge $500 or more. We are 90% confident that this interval includes the true value. Because the minimum suitable value of 25% is below this interval, we conclude that it's not a plausible value, and so we reject the null hypothesis that only 25% of customers charge more than $500 a month. The goal appears to have been met, assuming that the month we studied is typical.

For Example Confidence intervals and hypothesis tests

Question: Construct appropriate confidence intervals for testing the earlier two hypotheses (see page 401) and show how we could have reached the same conclusions from these intervals.

Answer: The test of proportion was two-sided, so we construct a 95% confidence interval for the true proportion:

$\hat{p} \pm 1.96SE(\hat{p}) = 0.29 \pm 1.96 \times \sqrt{\frac{(0.29)(0.71)}{200}} = (0.227, 0.353)$. Since 0.20 is not a plausible value, we reject the null hypothesis.

The test of means is one-sided, so we construct a one-sided 95% confidence interval, using the t critical value of 1.672:

$$(\bar{y} - t^*SE(\bar{y}), \infty) = (26.05 - 1.672 \times \frac{10.2}{\sqrt{58}}, \infty) = (23.81, \infty)$$

We can see that the hypothesized value of \$24.85 is in this interval, so we fail to reject the null hypothesis.

LO ❺

12.8 Comparing Two Proportions

A survey of 1003 Canadian adults by Angus Reid Strategies showed that 61% want same-sex marriage to continue to remain legal in Canada. We could use the methods covered so far in this chapter to check out a hypothesis as to whether the proportion of the whole population of Canada with that view is greater than, say, 50%.

Angus Reid then went a step further and also surveyed people in the United States and Britain in order to compare views on this issue among the three countries. In the U.S. it surveyed 1002 adults and found that 36% supported gay marriage. In Britain it surveyed 1980 adults and found that the level of support was 41%.

Is there a difference between Britain as a whole and the U.S. as a whole in the level of support for gay marriage? After all, 41% in one survey compared with 36% in another survey is not a big difference. Could it be due to sampling error, or is there a real difference between the British and American populations? Let's phrase this question in terms of a hypothesis test.

H_0: There is no difference between the percentage support for gay marriage in the U.S. and Britain.

H_A: There is a difference between the percentage support for gay marriage in the U.S. and Britain.

This is a different type of hypothesis test from the one we dealt with earlier about whether the percentage of up days for the DJIA was equal to 50% or whether there's a home team advantage. In those cases we were comparing sample results with a fixed number of 50%. In the case of gay marriage, there is no fixed number. Instead, we're comparing one sample with another. At first sight this may seem tough. We want to know whether the percentage support in the U.S. is different from what it is in Britain, but we don't know what the percentage support is in Britain. In fact, we can resolve this problem pretty fast by thinking instead about the *difference* in the percentage support between the two countries. Now we're comparing the percentage support in the U.S. *minus* the percentage support in Britain with a fixed number: zero.

If p_1 and p_2 are the population proportions supporting gay marriage in the U.S. and Britain, respectively, our original null hypothesis was:

$$H_0: p_1 = p_2$$

Two-Proportion *z*-Test

Testing whether the difference between two proportions is equal to a given number, K.

In order to test
$H_0: p_1 - p_2 = K$
$H_A: p_1 - p_2 \neq K$
we calculate the test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2 - k}{SD(\hat{p}_1 - \hat{p}_2)}$$

where

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

We then obtain the corresponding P-value from the table for the Normal distribution.

Now we have rephrased it as:

$$H_0: p_1 - p_2 = 0$$

Our estimate of p_1 is $\hat{p}_1$ (in our case, 0.36) and our estimate of p_2 is $\hat{p}_2$ (in our case, 0.41). So, using the approach described in Section 12.4, we calculate:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SD(\hat{p}_1 - \hat{p}_2)}$$

The standard deviation of the difference between $\hat{p}_1$ and $\hat{p}_2$ is obtained from the fact that these are independent random variables and that we can therefore add their variances:

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{SD(\hat{p}_1)^2 + SD(\hat{p}_2)^2}$$

$$= \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

This is known as the "Two-Proportion *z*-Test," and it can be used to test whether the difference between two proportions is any number, K:

$$H_0: p_1 - p_2 = K$$

In our case, K = 0, meaning that we're testing whether the two proportions are equal. This is a special case. Since the null hypothesis is $p_1 = p_2$, we don't really have two estimates $\hat{p}_1$ and $\hat{p}_2$ of different proportions. They're two estimates of the same proportion, $\hat{p}_1 = \hat{p}_2$. We can "pool" these two estimates into a single estimate. Suppose x_1 people out of n_1 support gay marriage in Britain (giving $p_1 = x_1/n_1$) and x_2 people out of n_2 support gay marriage in the U.S. (giving $p_2 = x_2/n_2$), and our null hypothesis says the support in the two countries is the same. Then we should use a "pooled" estimate of the support in both countries together:

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

Two-Proportion *z*-Test for equal proportions

Testing whether two proportions are equal.

In order to test
$H_0: p_1 - p_2 = 0$
$H_A: p_1 - p_2 \neq 0$
we calculate the test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SD(\hat{p}_1 - \hat{p}_2)}$$

where

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\bar{p}\,\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

and

$$p = \frac{x_1 + x_2}{n_1 + n_2} \text{ and } \bar{q} = 1 - \bar{p}.$$

We then obtain the corresponding P-value from the table for the Normal distribution.

Our standard deviation is now

$$SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}} = \sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

where $\bar{q} = 1 - \bar{p}$.

We now have two *z*-tests for two proportions, as summarized in the boxes in the margin. One of them tests whether the difference between two proportions is any number, K, and the other is specific to testing whether the two proportions are the same—i.e., K = 0.

These tests require the same four assumptions and conditions that we used in the case of the One-Proportion *z*-Test:

- **Independence Assumption:** The two samples are independent of each other.
- **Randomization Condition:** The people in each sample were selected at random.
- **10% Condition:** The sample is less than 10% of the population of the two countries.
- **Success/Failure Condition:**

  $n_1p_1 > 10$; $n_1q_1 > 10$;

  $n_2p_2 > 10$; $n_2q_2 > 10$.

We can be confident that the first two conditions are satisfied, since Angus Reid Strategies is a professional survey company. A quick calculation shows that the other two conditions are also satisfied.

Returning to our question about whether there's a difference between the support for gay marriage in the U.S. and Britain, we have:

$$H_0: p_1 = p_2$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{0.36 \times 1002 + 0.41 \times 1980}{1002 + 1980} = 0.3932$$

$$z = \frac{0.36 - 0.41}{\sqrt{0.3932 \times 0.6068 \times \left(\frac{1}{1002} + \frac{1}{1980}\right)}} = -2.64$$

The corresponding P-value is 0.0083, which is greater than 0.05. Clearly there *is* a difference between the levels of support for gay marriage in Britain and the U.S. at the 95% significance level.

For Example The effect of sample size when comparing two proportions

Survey companies like Angus Reid Strategies often survey about 1000 people in order to get a narrow standard deviation on their results and hence significant results. To see the effect of using a much smaller sample, let's suppose that only 30 people had been surveyed.

Question: If the survey of people's opinions about gay marriage had been done on only 30 people in Canada and 30 people in the U.S. and resulted in 61% and 36% in favour, respectively, would this indicate a significant difference between Canadians and Americans on this issue?

Answer: H_0: There is no difference between the percentage support for gay marriage in Canada and the U.S.—i.e., $p_1 = p_2$.

H_A: There is a difference between the percentage support for gay marriage in Canada and the U.S.—i.e., $p_1 - p_2 \neq 0$.

Checking the conditions, the Independence and Randomization Conditions are assumed true if this is a professionally designed survey. Certainly these small samples are less than 10% of the population of these countries. The Success/Failure Condition is only just satisfied, indicating that these samples are really only just large enough for us to use a test based on the Normal distribution:
$n_1p_1 = 30 \times 0.61 = 18.3 > 10$;
$n_1q_1 = 30 \times 0.39 = 11.7 > 10$;
$n_2p_2 = 30 \times 0.36 = 10.8 > 10$;
$n_2q_2 = 30 \times 0.64 = 19.2 > 10$.

First we calculate the pooled proportion:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{0.36 \times 30 + 0.61 \times 30}{30 + 30} = 0.485$$

Our test statistic is

$$z = \frac{0.61 - 0.36}{\sqrt{0.485 \times 0.515 \times \left(\frac{1}{30} + \frac{1}{30}\right)}}.$$
$$= 1.94$$

The corresponding P-value is 0.053, indicating that the difference is *not* significant at the 95% level.
This example shows that a difference that looks large may not be significant if the sample sizes are small.

LO ❹

12.9 Two Types of Error

Nobody's perfect. Even with lots of evidence, we can still make the wrong decision. In fact, when we perform a hypothesis test, we can make mistakes in *two* ways:

I. The null hypothesis is true, but we mistakenly reject it.

II. The null hypothesis is false, but we fail to reject it.

These are known as **Type I errors** and **Type II errors,** respectively. One way to keep the names straight is to remember that we start by assuming the null hypothesis is true, so a Type I error is the first kind of error we could make.

In medical disease testing, the null hypothesis is usually the assumption that a person is healthy. The alternative is that he or she has the disease we're testing for. So a Type I error is a *false positive*—a healthy person is diagnosed with the disease. A Type II error, in which an ill person is diagnosed as disease free, is a *false negative*. These errors have other names, depending on the particular discipline and context.

Which type of error is more serious depends on the situation. In a jury trial, a Type I error occurs if the jury convicts an innocent person. A Type II error occurs if the jury fails to convict a guilty person. Which seems more serious? In medical diagnosis, a false negative could mean that a sick patient goes untreated. A false positive might mean that a healthy person must undergo treatment.

In business planning, a false positive result could mean that money will be invested in a project that turns out not to be profitable. A false negative result might mean that money won't be invested in a project that would have been profitable. Which error is worse, the lost investment or the lost opportunity? The answer always depends on the situation, the cost, and your point of view.

Figure 12.9 gives an illustration of the situations:

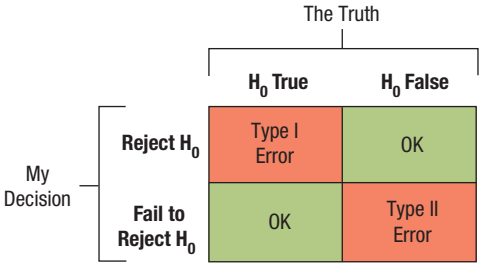


Figure 12.9 The two types of errors occur on the diagonal, where the truth and decision don't match. Remember that we start by assuming H_0 to be true, so an error made (rejecting it) when H_0 is true is called a Type I error. A Type II error is made when H_0 is false (and we fail to reject it).

How often will a Type I error occur? It happens when the null hypothesis is true but we've had the bad luck to draw an unusual sample. To reject H_0, the P-value must fall below α. When H_0 is true, that happens *exactly* with probability α. So when you choose level α, you're setting the probability of a Type I error to α.

What if H_0 is not true? Then we can't possibly make a Type I error. You can't get a false positive from a sick person. A Type I error can happen only when H_0 is true.

When H_0 is false and we reject it, we've done the right thing. A test's ability to detect a false hypothesis is called the **power** of the test. In a jury trial, power is a measure of the ability of the criminal justice system to convict people who are guilty. We'll have a lot more to say about power soon.

When H_0 is false but we fail to reject it, we've made a Type II error. We assign the letter β to the probability of this mistake. What's the value of β? That's harder to assess than α because we don't know what the value of the parameter really is. When H_0 is true, it specifies a single parameter value. But when H_0 is false, we don't have a specific one; we have many possible values. We can compute the probability β for any parameter value in H_A, but the choice of which one to pick is not always clear.

One way to focus our attention is by thinking about the *effect size*. That is, ask, "How big a difference would matter?" Suppose a charity wants to test whether placing personalized address labels in an envelope along with a request for a donation increases the response rate above the baseline of 5%. If the minimum response that would pay for the address labels is 6%, the charity would calculate β for the alternative $p = 0.06$.

Of course, we could reduce β for *all* alternative parameter values by increasing α. By making it easier to reject the null, we'd be more likely to reject it whether it's true or not. The only way to reduce *both* types of error is to collect more evidence

Notation Alert!

In Statistics, α is the probability of a Type I error and β is the probability of a Type II error.

The null hypothesis specifies a single value for the parameter. So it's easy to calculate the probability of a Type I error. But the alternative gives a whole range of possible values, and we may want to find a β for several of them.

We've seen ways to find a sample size by specifying the margin of error. Choosing the sample size to achieve a specified β (for a particular alternative value) is sometimes more appropriate, but the calculation is more complex and lies beyond the scope of this book.

or, in statistical terms, to collect more data. Otherwise, we just wind up trading off one kind of error against the other. Whenever you design a survey or experiment, it's a good idea to calculate β (for a reasonable α level). Use a parameter value in the alternative that corresponds to an effect size you want to be able to detect. Too often, studies fail because their sample sizes are too small to detect the change they're looking for.

Table 12.3 gives a summary of Type I and Type II errors:

Name	Also Known As:	Probability	Statistical Terminology	Business Example
Type I error	False positive	α	Reject a true null hypothesis	Invest in a project that is not successful
Type II error	False negative	β	Fail to reject a false null hypothesis	Fail to invest in a project that would have been successful

I Table 12.3 Type I and II errors.

For Example Type I and Type II errors

Question: Suppose that a year later, a full accounting of all the SmartWool transactions (see page 408) finds that 26.5% of visits resulted in sales. Have any errors been made?

Answer: We rejected the null hypothesis that $p = 0.20$ and in fact $p = 0.265$, so we did not make a Type I error (the only error we could have made when rejecting the null hypothesis).

LO ❹

*12.10 Power

Remember, we can never prove a null hypothesis true. We can only fail to reject it. But when we fail to reject a null hypothesis, it's natural to wonder whether we looked hard enough. Might the null hypothesis actually be false and our test too weak to tell?

When the null hypothesis actually *is* false, we hope our test is strong enough to reject it. We'd like to know how likely we are to succeed. The power of the test gives us a way to think about that. The power of a test is the probability that it correctly rejects a false null hypothesis. When the power is high, we can be confident that we've looked hard enough. We know that β is the probability that a test *fails* to reject a false null hypothesis, so the power of the test is the complement, $1 - \beta$. We might have just written $1 - \beta$, but power is such an important concept that it gets its own name.

Let's take the case of a pharmaceutical company that has invested millions of dollars in developing a new drug. The company wouldn't just test this drug on a few patients; it might not work on those patients, even though it's a good drug in general. So drug companies typically conduct a large trial involving thousands of patients in order to be pretty sure of spotting an effective drug when they have one. By using more patients, they're increasing the *power* of their test, so as to reduce the risk of failing to market an effective drug (Type II error).

The Canadian natural gas company Encana holds approximately 1 million acres of mineral rights in the Cutbank Ridge area of northeast British Columbia and northwest Alberta. The production of this area increased from 3,000,000 to 9,600,000 cubic metres per day between 2005 and 2009, due in part to successful

exploration. When Encana explores for sites that are going to be productive for natural gas, it wants to be pretty sure of finding the gas if it's there. Encana doesn't want to commit a Type II error and fail to find gas available in the land for which it owns the mineral rights. The statistical design behind its exploration technique therefore aims for *high-power* tests so that the chance of a false negative is low.

Whenever a study fails to reject its null hypothesis, the test's power comes into question. Was our sample size big enough to detect an effect, had there been one? Might we have missed an effect large enough to be interesting just because we failed to gather sufficient data or because there was too much variability in the data we could gather?

When we calculate power, we imagine that the null hypothesis is false. The value of the power depends on how far the truth lies from the null hypothesis value. We call the distance between the null hypothesis value, p_0, and the truth, p, the **effect size**. The power depends directly on the effect size. It's easier to see larger effects, so the further p_0 is from p, the greater the power.

How can we decide what power we need? Choice of power is more a financial or scientific decision than a statistical one, because to calculate the power, we need to specify the "true" parameter value we're interested in. In other words, power is calculated for a particular effect size, and it changes depending on the size of the effect we want to detect.

Just Checking

7 Remember our bank that's sending out DVDs to try to get customers to make payments on delinquent loans? It's looking for evidence that the costlier DVD strategy produces a higher success rate than the letters it's been sending. Explain what a Type I error is in this context and what the consequences would be to the bank.

8 What's a Type II error in the bank experiment context and what would the consequences be?

9 If the DVD strategy *really* works well—actually getting 60% of the people to pay off their balances—would the power of the test be higher or lower compared with a 32% payoff rate? Explain briefly.

Notation Alert!

Now we have four different types of proportion, p.

p is a population parameter—the true proportion in the population.

p_0 is a hypothesized value of p.

$\hat{p}$ is an observed proportion.

p^* is a critical value of a proportion corresponding to a specified α (see Figure 12.5).

Graph It!

It makes intuitive sense that the larger the effect size, the easier it should be to see it. Obtaining a larger sample size decreases the probability of a Type II error, so it increases the power. It also makes sense that the more we're willing to accept a Type I error, the less likely we'll be to make a Type II error.

Figure 12.10 may help you visualize the relationships among these concepts. Suppose we're testing $H_0: p = p_0$ against the alternative $H_A: p > p_0$. We'll reject the null if the observed proportion, $\hat{p}$, is big enough. By *big enough*, we mean $\hat{p} > p^*$ for some critical value p^* (shown as the red region in the right tail of the upper curve). The upper model shows a picture of the sampling distribution model for the proportion when the null hypothesis is true. If the null were true, then this would be a picture of that truth. We'd make a Type I error whenever the sample gave us $\hat{p} > p^*$ because we would reject the (true) null hypothesis. Unusual samples like that would happen only with probability α.

In reality, though, the null hypothesis is rarely *exactly* true. The lower probability model supposes that H_0 is not true. In particular, it supposes that the true value is p, not p_0. It shows a distribution of possible observed $\hat{p}$ values around this true value. Because of sampling variability, sometimes $\hat{p} < p^*$ and we fail to reject the (false) null hypothesis. Then we'd make a Type II error. The area under the curve to the left of p^* in the bottom model represents how often this happens. The

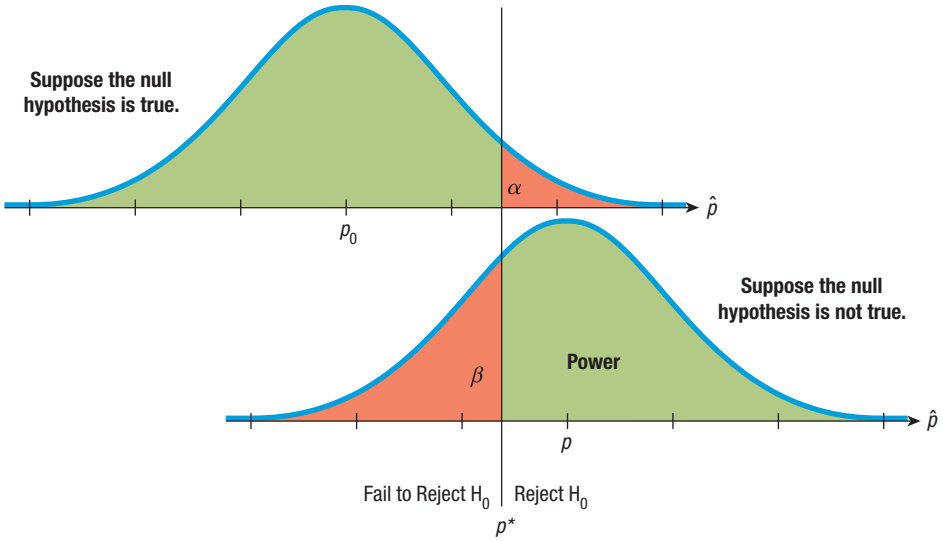


Figure 12.10 The power of a test is the probability that it rejects a false null hypothesis. The upper figure shows the null hypothesis model. We'd reject the null in a one-sided test if we observed a value in the red region to the right of the critical value, p^*. The lower figure shows the true model. If the true value of p is greater than p_0, then we're more likely to observe a value that exceeds the critical value and to make the correct decision to reject the null hypothesis. The power of the test is the green region on the right of the lower figure. Of course, even drawing samples whose observed proportions are distributed around p, we'll sometimes get a value in the red region on the left and make a Type II error of failing to reject the null.

probability is β. In this picture, β is less than half, so most of the time we *do* make the right decision. The *power* of the test—the probability that we make the right decision—is shown as the region to the right of p^*. It's $1 - \beta$.

We calculate p^* based on the upper model because p^* depends only on the null model and the alpha level. No matter what the true proportion, p^* doesn't change. After all, we don't *know* the truth, so we can't use it to determine the critical value. But we always reject H_0 when $\hat{p} > p^*$.

How often we reject H_0 when it's *false* depends on the effect size. We can see from the picture that if the true proportion were further from the hypothesized value, the bottom curve would shift to the right, making the power greater.

We can see several important relationships from this figure:

- Power = $1 - \beta$.
- Moving the critical value, p^*, to the right reduces α, the probability of a Type I error, but increases β, the probability of a Type II error. It correspondingly reduces the power.
- The larger the true effect size—the real difference between the hypothesized value, p_0, and the true population value, p—the smaller the chance of making a Type II error and the greater the power of the test.

If the two proportions are very far apart, the two models will barely overlap, and we wouldn't be likely to make any Type II errors at all—but then, we're unlikely to really need a formal hypothesis testing procedure to see such an obvious difference.

Reducing Both Type I and Type II Errors

Figure 12.10 seems to show that if we reduce Type I error, we must automatically increase Type II error. But there is a way to reduce both. Can you think of it?

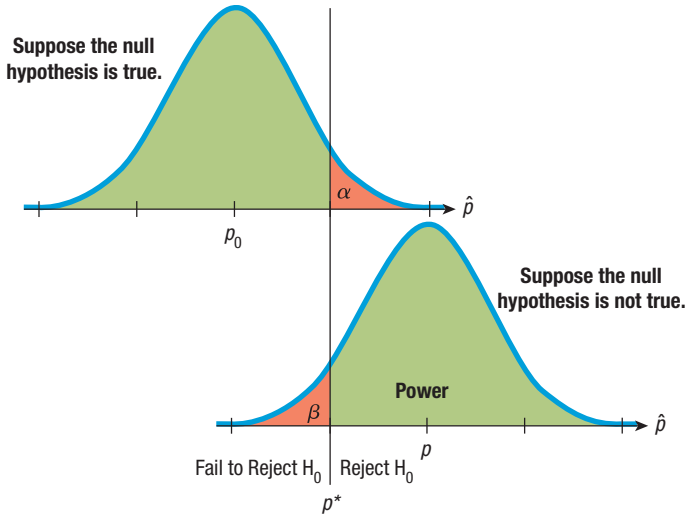


Figure 12.11 Making the standard deviations smaller increases the power without changing the alpha level or the corresponding *z*-critical value. The means are just as far apart as in Figure 12.10, but the error rates are reduced.

If we can make both curves narrower, as shown in Figure 12.11, then the probabilities of both Type I errors and Type II errors will decrease, and the power of the test will increase.

How can we do that? The only way is to reduce the standard deviations by increasing the sample size. (Remember, these are pictures of sampling distribution models, not of data.) Increasing the sample size works regardless of the true population parameters. But recall the curse of diminishing returns. The standard deviation of the sampling distribution model decreases only as the *square root* of the sample size, so to halve the standard deviations, we must *quadruple* the sample size.

What Can Go Wrong?

- **Don't base your null hypotheses on what you see in the data.** You're not allowed to look at the data first and then adjust your null hypothesis so that it will be rejected. If your sample value turns out to be $\hat{p} = 51.8\%$ with a standard deviation of 1%, don't form a null hypothesis as H_0: $p = 49.8\%$, knowing that this will enable you to reject it. Your null hypothesis describes the "nothing interesting" or "nothing has changed" scenario and should not be based on the data you collect.
- **Don't base your alternative hypothesis on the data, either.** You should always think about the situation you're investigating and base your alternative hypothesis on that. Are you interested only in knowing whether something has *increased*? Then write a one-tail (upper tail) alternative. Or would you be equally interested in a change in either direction? Then you want a two-tailed alternative. You should decide whether to do a one- or two-tailed test based on what results would be of interest to you, not on what you might see in the data.
- **Don't make your null hypothesis what you want to show to be true.** Remember, the null hypothesis is the status quo, the nothing-is-strange-here position that a sceptic would take. You wonder whether the data cast doubt on that. You can reject the null hypothesis, but you can never "accept" or "prove" the null.

- **Don't forget to check the conditions.** The reasoning of inference depends on randomization. No amount of care in calculating a test result can save you from a biased sample. The probabilities you compute depend on the Independence Assumption. And your sample must be large enough to justify your use of a Normal model.
- **Don't believe too strongly in arbitrary alpha levels.** There's not really much difference between a P-value of 0.051 and a P-value of 0.049, but sometimes it's regarded as the difference between night (having to retain H_0) and day (being able to shout to the world that your results are "statistically significant"). It may just be better to report the P-value and a confidence interval and let the world (perhaps your manager or client) decide along with you.
- **Don't confuse practical and statistical significance.** A large sample size can make it easy to discern even a trivial change from the null hypothesis value. On the other hand, you could miss an important difference if your test lacks sufficient power.
- **Don't forget that despite all your care, you might make a wrong decision.** No one can ever reduce the probability of a Type I error (α) or a Type II error (β) to zero (but increasing the sample size helps).

Ethics In Action

Many retailers have recognized the importance of staying connected to their in-store customers via the Internet. Retailers not only use the Internet to inform their customers about specials and promotions, but also to send them e-coupons redeemable for discounts. Shellie Cooper, longtime owner of a small organic food store in New Brunswick, specializes in locally produced organic foods and products. Over the years Shellie's customer base has been quite stable, consisting mainly of health-conscious individuals who tend not to be very price-sensitive, opting to pay higher prices for better-quality local, organic products. However, faced with increasing competition from grocery chains offering more organic choices, Shellie is now thinking of offering coupons. She needs to decide between the newspaper and the Internet. She recently read that the percentage of consumers who use printable Internet coupons is on the rise, but, at 15%, is much less than the 40% who clip and redeem newspaper coupons. Nonetheless, she's interested in learning more about the Internet and sets up a meeting with Jack Kasor, a Web consultant. She discovers that for an initial investment and continuing monthly fee, Jack would design Shellie's website, host it on his server, and broadcast e-coupons to her customers at regular intervals. While she was concerned about the difference in redemption rates for e-coupons and newspaper coupons, Jack assured her that e-coupon redemptions are continuing to rise and that she should expect between 15% and 40% of her customers to redeem them. Shellie agreed to give it a try. After the first six months, Jack informed Shellie that the proportion of her customers who redeemed e-coupons was significantly greater than 15%. He determined this by selecting several broadcasts at random and finding the number redeemed (483) out of the total number sent (3000). Shellie thought that this was positive and made up her mind to continue the use of e-coupons.

ETHICAL ISSUE *Statistical vs. practical significance. While it's true that the percentage of Shellie's customers redeeming e-coupons is significantly greater than 15% statistically, in fact the percentage is just over 16%. This difference amounts to about 33 customers, or more than 15%, which may not be of practical significance to Shellie (related to Item A, ASA Ethical Guidelines; see Appendix C, the American Statistical Association's Ethical Guidelines for Statistical Practice, also available online at www.amstat.org/about/ethicalguidelines.cfm). Mentioning a range of 15% to 40% may mislead Shellie into expecting a value somewhere in the middle.*

ETHICAL SOLUTION *Jack should report the difference between the observed value and the hypothesized value to Shellie, especially since there are costs associated with continuing e-coupons. Perhaps he should recommend that she reconsider using the newspaper.*

What Have We Learned?

Learning Objectives

❶ We've learned to use what we see in a random sample to test a particular hypothesis about the world. This is our second step in statistical inference, complementing our use of confidence intervals. We've learned that testing a hypothesis involves proposing a model and then seeing whether the data we observe are consistent with that model or so unusual that we must reject it. We do this by finding a P-value—the probability that data like ours could have occurred if the model is correct.

❷ If the data are out of line with the null hypothesis model, the P-value will be small, and we'll reject the null hypothesis. If the data are consistent with the null hypothesis model, the P-value will be large, and we won't reject the null hypothesis.

We've learned that:

- We start with a *null hypothesis* specifying the parameter of a model we'll test using our data.
- Our *alternative hypothesis* can be one- or two-sided, depending on what we want to learn.
- We must check the appropriate *assumptions* and *conditions* before proceeding with our test.
- The *significance level* of the test establishes the level of proof we'll require. That determines the critical value of z that will lead us to reject the null hypothesis.

❸ *Hypothesis tests* and *confidence intervals* are really two ways of looking at the same question. The hypothesis test gives us the answer to a decision about a parameter; the confidence interval tells us the plausible values of that parameter.

❹ If the null hypothesis is really true and we reject it, that's a *Type I error*; the alpha level of the test is the probability that this happens.

If the null hypothesis is really false but we fail to reject it, that's a *Type II error*.

❺ If we have independent samples from two different populations, we construct a hypothesis test to compare the two populations with each other.

*Optional Sections

- The *power* of the test is the probability that we reject the null hypothesis when it's false. The larger the size of the effect we're testing for, the greater the power of the test in detecting it.
- Tests with a greater likelihood of Type I error have more power and less chance of a Type II error. We can increase power while reducing the chances of both kinds of error by increasing the sample size.

Terms

Alpha level The threshold P-value that determines when we reject a null hypothesis. Using an alpha level of α, if we observe a statistic whose P-value based on the null hypothesis is less than α, we reject that null hypothesis.

Alternative hypothesis The hypothesis that proposes what we should conclude if we find the null hypothesis to be unlikely.

Critical value The value in the sampling distribution model of the statistic whose P-value is equal to the alpha level. Any statistic value further from the null hypothesis value than the critical value will have a smaller P-value than α and will lead to rejecting the null hypothesis. The critical value is often denoted with an asterisk, as z^*, for example.

Effect size The difference between the null hypothesis value and the true value of a model parameter.

Null hypothesis The claim being assessed in a hypothesis test. Usually, the null hypothesis is a statement of "no change from the traditional value," "no effect," "no difference," or "no relationship." For a claim to be a testable null hypothesis, it must specify a value for some population parameter that can form the basis for assuming a sampling distribution for a test statistic.

One-proportion *z*-test A test of the null hypothesis that the proportion of a single sample equals a specified value (H_0: $p = p_0$) by comparing the statistic $z = \dfrac{\hat{p} - p_0}{SD(\hat{p})}$ to a standard Normal model.

One-sided alternative An alternative hypothesis is one-sided (e.g., H_A: $p > p_0$ or H_A: $p < p_0$) when we're interested in deviations in *only one* direction away from the hypothesized parameter value.

P-value The probability of observing a value for a test statistic at least as far from the hypothesized value as the statistic value actually observed if the null hypothesis is true. A small P-value indicates that the observation obtained is improbable given the null hypothesis and thus provides evidence against the null hypothesis.

Power The probability that a hypothesis test will correctly reject a false null hypothesis. To find the power of a test, we must specify a particular alternative parameter value as the "true" value. For any specific value in the alternative, the power is $1 - \beta$.

Significance level Another term for the alpha level, used most often in a phrase such as "at the 5% significance level."

Two-sided alternative An alternative hypothesis is two-sided (H_A: $p \neq p_0$) when we're interested in deviations in *either* direction away from the hypothesized parameter value.

Type I error The error of rejecting a null hypothesis when in fact it is true (also called a "false positive"). The probability of a Type I error is α.

Type II error The error of failing to reject a null hypothesis when in fact it is false (also called a "false negative"). The probability of a Type II error is commonly denoted β and depends on the effect size.

Skills

Plan

- Be able to state the null and alternative hypotheses for a one-proportion *z*-test.
- Know how to think about the assumptions and their associated conditions. Examine your data for violations of those conditions.
- Be able to identify and use the alternative hypothesis when testing hypotheses. Understand how to choose between a one-sided and two-sided alternative hypothesis and be able to explain your choice.

Do

- Know how to perform a one-proportion *z*-test.

Report

- Be able to interpret the results of a one-proportion *z*-test.
- Be able to interpret the meaning of a P-value in nontechnical language, making clear that the probability claim is about computed values under the assumption that the null model is true and not about the population parameter of interest.

Technology Help: Testing Hypothesis About Proportions

Hypothesis tests for proportions are so easy and natural that many statistics packages don't offer special commands for them. Most statistics programs want to know the "success" and "failure" status for each case. Usually these are given as 1 or 0, but they might be category names like "yes" and "no." Often we just know the proportion of successes, $\hat{p}$, and the total count, *n*. Computer packages don't usually deal naturally with summary data like this, but see below for a couple of important exceptions (Minitab and JMP).

In some programs you can reconstruct the original values. But even when you've reconstructed (or can reconstruct) the raw data values, often you won't get *exactly* the same test statistic from a computer package as you would from working by hand. The reason is that when the packages treat the proportion as a mean, they make some approximations. The result is very close, but not exactly the same. If you use a computer package, you may notice slight discrepancies between your answers and the answers in the back of the book, but they're not important.

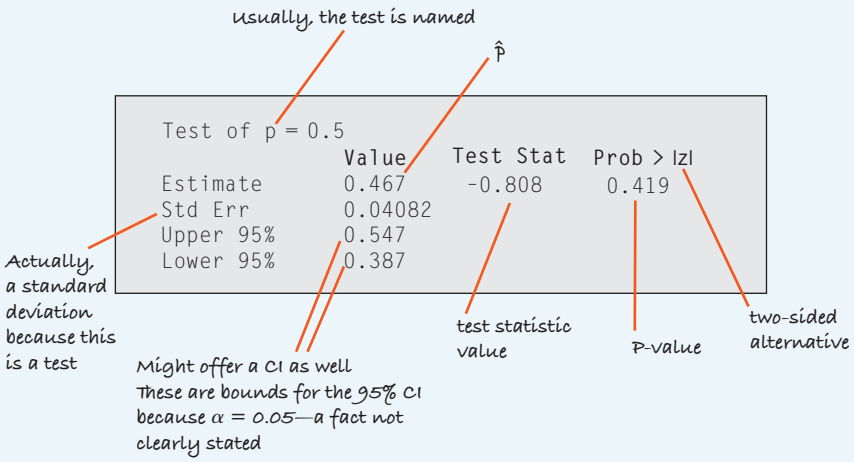
Reports about hypothesis tests generated by technologies don't follow a standard form. Most will name the test and provide the test statistic value, its standard deviation, and the P-value. But these elements may not be labelled clearly. For example, the expression "Prob > |z|" means the probability (the "Prob") of observing a test statistic whose magnitude (the absolute value tells us this) is larger than that of the one (the "z") found in the data (which, because it's written as "z," we know follows a Normal model). That is a fancy (and not very clear) way of saying P-value. In some packages, you can specify that the test be one-sided. Others might report three P-values, covering the ground for both one-sided tests and two-sided tests.

Sometimes a confidence interval and hypothesis test are automatically given together. The confidence interval ought to be for the corresponding confidence level: $1 - \alpha$.

Often, the standard deviation of the statistic is called the "standard error," and usually that's appropriate because we've had to estimate its value from the data. That's not the case for proportions, however: We get the standard deviation for a proportion from the null hypothesis value. Nevertheless, you may see the standard deviation called a "standard error," even for tests with proportions.

It's common for statistics packages and calculators to report more digits of "precision" than could possibly have been found from the data. You can safely ignore them. Round up values such as the standard deviation to one digit more than the number of digits reported in your data.

Here are the kind of results you might see in typical computer output.



EXCEL

Inference methods for proportions are not part of the standard Excel tool set.

Comments

For summarized data, type the calculation into any cell and evaluate it.

MINITAB

Choose **Basic Statistics** from the **Stat** menu.

- Choose **1 Proportion** from the Basic Statistics submenu.
- If the data are category names in a variable, assign the variable from the variable list box to the **Samples in columns** box.
- If you have summarized data, click the **Summarized Data** button and fill in the number of trials and the number of successes.
- Click the **Options** button and specify the remaining details.
- If you have a large sample, check **Use test and interval based on Normal distribution**.
- Click the **OK** button.

Comments

When working from a variable that names categories, MINITAB treats the last category as the "success" category. You can specify how the categories should be ordered.

(continued)

SPSS

SPSS does not find hypothesis tests for proportions.

JMP

For a **categorical** variable that holds category labels, the **Distribution** platform includes tests and intervals of proportions. For summarized data, put the category names in one variable and the frequencies in an adjacent variable. Designate the frequency column to have the **role** of **frequency**. Then use the **Distribution** platform.

Comments

JMP uses slightly different methods for proportion inferences than those discussed in this text. Your answers are likely to be slightly different.

MINI case studies

Common-Law Couples in Quebec

According to the 2006 Canadian census, 29% of all families in Quebec were common-law couples in 2006 (**Source:** Based on Statistics Canada. [2007, Winter]. Census Snapshot of Canada: Families. *Canadian Social Trends*, Catalogue No. 11-008). In order to test-market products to common-law couples, you need to select a city with a large percentage of this type of family. Suppose you survey 100 randomly selected families in Montreal and find that 35% of them are common-law couples. After completing your survey, you read in a newspaper about another survey done in Montreal by a reputable survey company that used a sample size of 400 families and found that 33% of them are common-law couples. Using hypothesis tests *and* confidence intervals, estimate the proportion of common-law couples in Montreal and whether it's higher than the provincial average. Comment on the different results from the two surveys.

Metal Production

Ingots are huge pieces of metal, often weighing in excess of 9000 kilograms, made in a giant mould. They must be cast in one large piece for use in fabricating large structural parts for cars and planes. If they crack while being made, the crack may propagate into the zone required for the part, compromising its integrity. Airplane manufacturers insist that metal for their planes be defect-free, so the ingot must be made over if any cracking is detected.

Even though the metal from the cracked ingot is recycled, the scrap cost runs into the tens of thousands of dollars. Metal manufacturers would like to avoid cracking if at all possible. But the casting process is complicated, and not everything can be controlled completely. In one plant, only about 75% of the ingots have been free of cracks. So, in an attempt to reduce the cracking proportion, the plant engineers and chemists made changes to the casting process. The data from 5000 ingots produced since the changes can be found in the file **ch12_MCSP_Ingots**. The variable *Crack* indicates whether a crack was found (1) or not (0). Select a random sample of 100 ingots and test the claim that the cracking rate

has decreased from 25%. Find a confidence interval for the cracking rate as well. Now select a random sample of 1000 ingots, test the claim, and find the confidence interval again. Compare the two tests and intervals and prepare a short report about your findings, including the differences (if any) you see in the two samples.

Loyalty Program

An airline marketing manager sent out 10,000 mail pieces to a random sample of customers to test a new Web-based loyalty program. The customers either received nothing (No Offer), a free companion airline ticket (Free Flight), or free flight insurance on their next flight (Free Insurance). The person in charge of selecting the 10,000 customers has assured the marketing manager that the sample is representative of the various marketing segments in the customer base. However, the manager is worried that the offer wasn't sent out to enough customers in the *Travel* segment, which represents 25% of the entire customer base (variable *Spending.Segment*). In addition, he's worried that fewer than one-third of customers in that segment actually received no offer. Using the data found in the file **ch12_MCSP_Loyalty_Program,** write a short report to the manager testing the appropriate hypotheses and summarizing your findings. Include in your report a 95% confidence interval for the proportion of customers who responded to the offer by signing up for the loyalty program. (The variable *Response* indicates a 1 for responders and 0 for nonresponders.)

MyStatLab Students! Save time, improve your Grades with MyStatLab. You can practise many of this chapter's exercises as often as you want, and most feature step-by-step guided solutions to help you find the right answer. You'll find a personalized study plan available to you too!

Exercises

SECTION 12.1

1. For each of the following situations, define the parameter and write the null and alternative hypotheses in terms of parameter values. Example: We want to know if the proportion of up days in the stock market is 50%.

Answer: Let p = the proportion of up days. $H_0: p = 0.5$ vs, $H_A: p \neq 0.5$.

a) A casino wants to know if its slot machine really delivers the 1 in 100 win rate that it claims.
b) A pharmaceutical company wonders if its new drug has a cure rate different from the 30% reported by the placebo.
c) A bank wants to know if the percentage of customers using its website has changed from the 40% who used it before the bank's system crashed last week. **LO ❶**

2. As in Exercise 1, for each of the following situations, define the parameter and write the null and alternative hypotheses in terms of parameter values.

a) Seat-belt compliance was 65% last year. We want to know if it's changed this year.
b) Last year, a survey found that 45% of the employees were willing to pay for on-site day care. The company wants to know if that has changed.
c) Regular card customers have a default rate of 6.7%. A credit card company wants to know if that rate is different for its Gold card customers. **LO ❶**

SECTION 12.2

3. For each of the following, write out the null and alternative hypothesis, making sure to state whether it's one-sided or two-sided.

a) A company reports that last year 40% of its reports in accounting were on time. From a random sample this year, it wants to know if that proportion has changed.
b) Last year, 42% of the employees enrolled in at least one wellness class at the company's site. Using a survey, the company wants to know if a greater percentage is planning to take a wellness class this year.
c) A political candidate wants to know from recent polls if she's going to garner a majority of votes in next week's election. **LO ❶**

4. For each of the following, write out the alternative hypothesis, being sure to indicate whether it is one-sided or two-sided.

a) *Consumer Reports* discovered that 20% of a certain computer model had warranty problems over the first three months. From a random sample, the manufacturer wants to know if a new model has improved that rate.
b) The last time a philanthropic agency requested donations, 4.75% of people responded. From a recent pilot mailing, it wonders if that rate has increased.
c) A student wants to know if other students on her campus prefer Coke or Pepsi. **LO ❶**

SECTION 12.3

5. Which of the following are true? If false, explain briefly.

a) A very high P-value is strong evidence that the null hypothesis is false.
b) A very low P-value proves that the null hypothesis is false.
c) A high P-value shows that the null hypothesis is true.
d) A P-value below 0.05 is always considered sufficient evidence to reject a null hypothesis. **LO ❶**

6. Which of the following are true? If false, explain briefly.

a) A very low P-value provides evidence against the null hypothesis.
b) A high P-value is strong evidence in favour of the null hypothesis.
c) A P-value above 0.10 shows that the null hypothesis is true.
d) If the null hypothesis is true, you can't get a P-value below 0.01. **LO ❶**

SECTION 12.4

7. A consulting firm had predicted that 35% of the employees at a large firm would take advantage of a new company credit union, but management is sceptical. They doubt the rate is that high. A survey of 300 employees shows that 138 of them are currently taking advantage of the credit union. From the sample proportion

a) Find the standard deviation of the sample proportion based on the null hypothesis.
b) Find the z-statistic.
c) Does the z-statistic seem like a particularly large or small value? **LO ❷**

8. A survey of 100 CEOs finds that 60 think the economy will improve next year. Is there evidence that the rate is higher among all CEOs than the 55% reported by the public at large?

a) Find the standard deviation of the sample proportion based on the null hypothesis.
b) Find the z-statistic.
c) Does the z-statistic seem like a particularly large or small value? **LO ❷**

SECTION 12.5

9. Which of the following statements are true? If false, explain briefly.

a) Using an alpha level of 0.05, a P-value of 0.04 results in rejecting the null hypothesis.
b) The alpha level depends on the sample size.
c) With an alpha level of 0.01, a P-value of 0.10 results in rejecting the null hypothesis.
d) Using an alpha level of 0.05, a P-value of 0.06 means the null hypothesis is true. **LO ❶**

10. Which of the following statements are true? If false, explain briefly.

a) It's better to use an alpha level of 0.05 than an alpha level of 0.01.
b) If we use an alpha level of 0.01, then a P-value of 0.001 is statistically significant.
c) If we use an alpha level of 0.01, then we reject the null hypothesis if the P-value is 0.001.
d) If the P-value is 0.01, we reject the null hypothesis for any alpha level greater than 0.01. **LO ❶**

SECTION 12.6

11. For each of the following situations, find the critical value (s) for z.

a) $H_0: p - 0.5$ vs. $H_A: p \neq 0.5$ at $\alpha = 0.05$.
b) $H_0: p = 0.4$ vs. $H_A: p > 0.4$ at $\alpha = 0.05$.
c) $H_0: p = 0.5$ vs. $H_A: p > 0.5$ at $\alpha = 0.01$; $n = 345$.
LO ❷

12. For each of the following situations, find the critical value for z.

a) $H_0: p = 0.5$ vs. $H_A: p > 0.5$ at $\alpha = 0.05$.
b) $H_0: p = 0.6$ vs. $H_A: p \neq 0.6$ at $\alpha = 0.01$.
c) $H_0: p = 0.5$ vs. $H_A: p < 0.5$ at $\alpha = 0.01$; $n = 500$.
d) $H_0: p = 0.2$ vs. $H_A: p < 0.2$ at $\alpha = 0.01$. **LO ❷**

SECTION 12.7

13. Suppose you're testing the hypotheses H_0: $p = 0.20$ vs. $p \neq 0.20$. A sample size of 250 results in a sample proportion of 0.25.

a) Construct a 95% confidence interval for P.
b) Based on the confidence interval, at $\alpha = 0.05$ can you reject H_0? Explain.
c) What's the difference between the standard error and standard deviation of the sample proportion?
d) Which is used in computing the confidence interval?
LO ❸

14. Suppose you're testing the hypotheses $H_0: p = 0.40$ vs. $H_A: p > 0.40$. A sample size of 200 results in a sample proportion of 0.55.

a) Construct a 90% confidence interval for p.
b) Based on the confidence interval, at $\alpha = 0.05$ can you reject H_0? Explain.

c) What's the difference between the standard error and standard deviation of the sample proportion?
d) Which is used in computing the confidence interval? **LO ❸**

15. Suppose you're testing the hypotheses $H_0: \mu = 16$ vs. $H_A: \mu < 16$.A sample size of 25 results in a sample mean of 16.5 and a standard deviation of 2.0.

a) What is the standard error of the mean?
b) What is the critical value of t^* for a 90% confidence interval?
c) Construct a 90% confidence interval for μ.
d) Based on the confidence interval, at $\alpha = .05$ can you reject H_0? Explain. **LO ❸**

16. Suppose you're testing the hypotheses $H_0: \mu = 80$ vs. $H_A: \mu \neq 80$. A sample size of 61 results in a sample mean of 75 and a standard deviation of 1.5.

a) What is the standard error of the mean?
b) What is the critical value of t^* for a 95% confidence interval?
c) Construct a 95% confidence interval for μ.
d) Based on the confidence interval, at $\alpha = 0.05$ can you reject H_0? Explain. **LO ❸**

SECTION 12.8

17. In an Angus Reid survey of opinions regarding same-sex marriage, 13% of a sample of 1003 Canadian adults said that same-sex couples should not have any legal recognition. The corresponding percentage for a survey of 1980 Britons was 15%. Does a greater percentage of Britons hold this view than Canadians? **LO ❺**

18. An Angus Reid survey of 1003 Canadian adults and 1980 British adults found that 59% of the Canadian sample and 56% of the British sample thought that homosexuality is something people are born with as opposed to being a preference they have chosen. Does a smaller percentage of Britons hold this view than Canadians? **LO ❺**

SECTIONS 12.9 AND 12.10

19. For each of the following situations, state whether a Type I, a Type II, or neither error has been made. Explain briefly.

a) A bank wants to know if the enrollment on its website is above 30% based on a small sample of customers. It tests $H_0: p = 0.3$ vs. $H_A: p > 0.3$ and rejects the null hypothesis. Later it finds out that actually 28% of all customers enrolled.
b) A student tests 100 students to determine whether other students on her campus prefer Coke or Pepsi and finds no evidence that preference for Coke is not 0.5. Later, a marketing company tests all students on campus and finds no difference.
c) A human resource analyst wants to know if the applicants this year score, on average, higher on their placement exam than the 52.5 points the candidates averaged last year. She samples 50 recent tests and finds the average to be 54.1 points. She fails to reject the null hypothesis that the mean is 52.5 points. At the end of the year, they find that the candidates this year had a mean of 55.3 points.
d) A pharmaceutical company tests whether a drug lifts the headache relief rate from the 25% achieved by the placebo. It fails to reject the null hypothesis because the P-value is 0.465. Further testing shows that the drug actually relieves headaches in 38% of people. **LO ❹**

20. For each of the following situations, state whether a Type I, a Type II, or neither error has been made.

a) A test of $H_0: p = 0.8$ vs. $H_A: p < 0.08$ fails to reject the null hypothesis. Later it's discovered that $p = 0.9$.
b) A test of $H_0: p = 0.5$ vs. $H_A: p \neq 0.5$ rejects the null hypothesis. Later it's discovered that $p = 0.65$.
c) A test of $H_0: p = 0.7$ vs. $H_A: p < 0.7$ fails to reject the null hypothesis. Later it's discovered that $p = 0.6$. **LO ❹**

CHAPTER EXERCISES

21. Hypotheses. Write the null and alternative hypotheses to test each of the following situations.

a) An online clothing company is concerned about the timeliness of the delivery of its products. The VP of Operations and Marketing recently stated that she wanted the percentage of products delivered on time to be more than 90%, and she wants to know if the company has succeeded.
b) A realty company recently announced that the proportion of houses taking more than three months to sell is now greater than 50%.
c) A financial firm's accounting reports have an error rate below 2%. **LO ❶**

22. More hypotheses. Write the null and alternative hypotheses to test each of the following situations.

a) A business magazine article reports that, in 1990, 35% of CEOs had an MBA degree. Has the percentage changed?
b) Recently, 20% of cars of a certain model have needed costly transmission work after being driven between 50,000 and 100,000 miles. The car manufacturer hopes that the redesign of a transmission component has solved this problem.
c) A market researcher for a cola company decides to field-test a new-flavour soft drink, planning to market it only if he's sure that over 60% of the people like the flavour. **LO ❶**

23. Deliveries. The clothing company in Exercise 21a looks at a sample of delivery reports. The company tests the hypothesis that 90% of the deliveries are on time against the alternative that greater than 90% are on time, and finds a P-value of 0.22. Which of these conclusions is appropriate?

a) There's a 22% chance that 90% of the deliveries are on time.
b) There's a 78% chance than 90% of the deliveries are on time.

c) There's a 22% chance that the sample the company drew shows the correct percentage of on-time deliveries.
d) There's a 22% chance that natural sampling variation could produce a sample of on-time deliveries at least as extreme as the one obtained if, in fact, 90% of deliveries are on time. **LO ❶**

24. House sales. The realty company in Exercise 21b looks at a recent sample of houses that have sold. On testing the null hypothesis that 50% of the houses take more than three months to sell against the hypothesis that more than 50% of the houses take more than three months to sell, the company finds a P-value of 0.034. Which of these conclusions is appropriate?

a) There's a 3.4% chance that 50% of the houses take more than three months to sell.
b) If 50% of the houses take more than three months to sell, there's a 3.4% chance that a random sample would produce a sample proportion as high as the one obtained.
c) There's a 3.4% chance that the null hypothesis is correct.
d) There's a 96.6% chance that 50% of the houses take more than three months to sell. **LO ❶**

25. P-value. Have harsher penalties and ad campaigns increased seat belt use among drivers and passengers? Observations of commuter traffic have failed to find evidence of a significant change compared with three years ago. Explain what the study's P-value of 0.17 means in this context. **LO ❶**

26. Another P-value. A company developing scanners to search for hidden weapons at airports has concluded that a new device is significantly better than the current scanner. The company made this decision based on a P-value of 0.03. Explain the meaning of the P-value in this context. **LO ❶**

27. Ad campaign. An information technology analyst believes that her company is losing customers on its website who find the checkout and purchase system too complicated. She adds a one-click feature to the website to make it easier, but finds that only about 10% of the customers are using it. She decides to launch an ad awareness campaign to tell customers about the new feature in the hope of increasing the percentage. She doesn't see much of a difference, so she hires a consultant to help her. The consultant selects a random sample of recent purchases, tests the hypothesis that the ads produced no change against the alternative (that the percentage who use the one-click feature is now greater than 10%), and finds a P-value of 0.22. Which conclusion is appropriate? Explain.

a) There's a 22% chance that the ads worked.
b) There's a 78% chance that the ads worked.
c) There's a 22% chance that the null hypothesis is true.
d) There's a 22% chance that natural sampling variation could produce poll results at least as extreme as these if the use of the one-click feature has increased.
e) There's a 22% chance that natural sampling variation could produce poll results at least as extreme as these if there's really no change in website use. **LO ❶**

28. Mutual funds. A mutual fund manager claims that at least 70% of the stocks she selects will increase in price over the next year. We examined a sample of 200 of her selections over the past three years. Our P-value turns out to be 0.03. Test an appropriate hypothesis. Which conclusion is appropriate? Explain.

a) There's a 3% chance that the fund manager is correct.
b) There's a 97% chance that the fund manager is correct.
c) There's a 3% chance that a random sample could produce results at least as extreme as we observed if $p = 0.7$, so it's reasonable to conclude that the fund manager is correct.
d) There's a 3% chance that a random sample could produce results at least as extreme as we observed if $p = 0.7$, so it's reasonable to conclude that the fund manager is not correct.
e) There's a 3% chance that the null hypothesis is correct. **LO ❶**

29. Product effectiveness. A pharmaceutical company's old antacid formula provided relief for 70% of the people who used it. The company tests a new formula to see if it's better and gets a P-value of 0.27. Is it reasonable to conclude that the new formula and the old one are equally effective? Explain. **LO ❶**

30. Car sales. A German automobile company is counting on selling more cars to the younger market segment—drivers under the age of 20. The company's market researchers survey to investigate whether the proportion of today's high school seniors who own their own cars is higher than it was a decade ago. They find a P-value of 0.017. Is it reasonable to conclude that more high school seniors have cars? Explain. **LO ❶**

31. False claims? A candy company claims that in a large bag of holiday M&M'S, half the candies are red and half the candies are green. You pick candies at random from a bag and discover that of the first 20 you eat, 12 are red.

a) If it were true that half are red and half are green, what's the probability of finding that at least 12 out of 20 candies were red?
b) Do you think that half of the M&M'S candies in the bag are really red? Explain. **LO ❷**

32. Stocks. A young investor is concerned that investing in the stock market is actually gambling, since the chance of the stock market going up on any given day is 50%. She decides to track her favourite stock for 250 days and finds that on 140 days, the stock was "up."

a) Find a 95% confidence interval for the proportion of days the stock was "up." Don't forget to check the conditions first.
b) Does your confidence interval provide any evidence that the market is not random? Explain.
c) What is the significance level of this test? Explain. **LO ❸**

33. Economy. In 2008, a Gallup Poll asked 2336 U.S. adults aged 18 or over how they rated economic conditions. In a poll conducted from January 27 through February 1, 2008, 24% rated the economy as "Excellent/Good." A recent media outlet claimed that the percentage of Americans who felt the economy was in "Excellent/Good" shape was, in fact, 28%. Does the Gallup Poll support this claim?

a) Find a 95% confidence interval for the sample proportion of U.S. adults who rated the economy as "Excellent/Good." Check conditions.
b) Does your confidence interval provide evidence to support the claim?
c) What is the significance level of the test in part b)? Explain. **LO ❸**

34. Economy, part 2. The same Gallup Poll data from Exercise 33 also reported that 33% of those surveyed rated the economy as "Poor." The same media outlet claimed the true proportion to be 30%. Does the Gallup Poll support this claim?

a) Find a 95% confidence interval for the sample proportion of U.S. adults who rated the economy as "Poor." Check conditions.
b) Does your confidence interval provide evidence to support the claim?
c) What is the significance level of the test in part b)? Explain. **LO ❸**

35. Convenient alpha. An enthusiastic junior executive has run a test of his new marketing program. He reports that it resulted in a "significant" increase in sales. A footnote on his report explains that he used an alpha level of 7.2% for his test. Presumably, he performed a hypothesis test against the null hypothesis of no change in sales.

a) If instead he had used an alpha level of 5%, is it more or less likely that he would have rejected his null hypothesis? Explain.
b) If he chose the alpha level 7.2% so that he could claim statistical significance, explain why this is not an ethical use of statistics. **LO ❶**

36. Safety. The manufacturer of a new sleeping pill suspects that it may increase the risk of sleepwalking, which could be dangerous. A test of the drug fails to reject the null hypothesis of no increase in sleepwalking when tested at $\alpha = 0.01$.

a) If the test had been performed at $\alpha = 0.05$, would the test have been more or less likely to reject the null hypothesis of no increase in sleepwalking?
b) Which alpha level do you think the company should use? Why? **LO ❶**

37. Product testing. Since many people have trouble programming their digital video recorders (DVRs), an electronics company has developed what it hopes will be easier instructions. The goal is to have at least 96% of customers succeed at programming their DVRs. The company tests the new system on 200 people, 188 of whom were successful. Is this strong evidence that the new system fails to meet the company's goal? A student's test of this hypothesis is shown here. How many mistakes can you find?

$H_0: \hat{p} = 0.96$
$H_A: \hat{p} \neq 0.96$
SRS, $0.96(200) > 10$

$$\frac{188}{200} = 0.94; \text{SD}(\hat{p}) = \sqrt{\frac{(0.94)(0.06)}{200}} = 0.017$$

$$z = \frac{0.96 - 0.94}{0.017} = 1.18$$

$$P = P(z > 1.18) = 0.12$$

There is strong evidence that the new system doesn't work. **LO ❷**

38. Marketing. A newsletter reported that 90% of adults drink milk. A regional farmers' organization planning a new marketing campaign across its multicounty area polls a random sample of 750 adults living there. In this sample, 657 people said that they drink milk. Do these responses provide strong evidence that the 90% figure isn't accurate for this region? Correct the mistakes you find in a student's following attempt to test an appropriate hypothesis.

$H_0: \hat{p} = 0.9$
$H_A: \hat{p} < 0.9$
SRS, $750 > 10$

$$\frac{657}{750} = 0.876; \text{SD}(\hat{p}) = \sqrt{\frac{(0.88)(0.12)}{750}} = 0.012$$

$$z = \frac{0.876 - 0.94}{0.012} = -2$$

$$P = P(z > -2) = 0.977$$

There is more than a 97% chance that the stated percentage is correct for this region. **LO ❷**

39. Environment. In the 1980s, it was generally believed that congenital abnormalities affected about 5% of the nation's children. Some people believe that the increase in the number of chemicals in the environment has led to an increase in the incidence of abnormalities. A recent study examined 384 children and found that 46 of them showed signs of an abnormality. Is this strong evidence that the risk has increased? (We consider a P-value of around 5% to represent reasonable evidence.)

a) Write appropriate hypotheses.
b) Check the necessary assumptions.
c) Perform the mechanics of the test. What is the P-value?
d) Explain carefully what the P-value means in this context.
e) What's your conclusion?
f) Do environmental chemicals cause congenital abnormalities? **LO ❷**

40. Spike poll. In August 2004, *Time* magazine reported the results of a random U.S. telephone poll commissioned by the Spike network. Of the 1302 men who responded, only 39 said that their most important measure of success was their work.

a) Estimate the percentage of all American males who measure success primarily by their work. Use a 98% confidence interval. Don't forget to check the conditions first.
b) Some believe that few contemporary men judge their success primarily by their work. Suppose we wished to conduct a hypothesis test to see if the fraction has fallen below the 5% mark. What does your confidence interval indicate? Explain.
c) What is the significance level of this test? Explain. **LO ❸**

41. Education. In 2006, 34% of students had not been absent from school even once during the previous month. In a 2011 survey, responses from 8302 students showed that this figure had slipped to 33%. Officials would be concerned if student attendance were declining. Do these figures give evidence of a decrease in student attendance?

a) Write appropriate hypotheses.
b) Check the assumptions and conditions.
c) Perform the test and find the P-value.
d) State your conclusion.
e) Do you think this difference is meaningful? Explain. **LO ❷**

42. Customer satisfaction. A company hopes to improve customer satisfaction, setting as a goal no more than 5% negative comments. A random survey of 350 customers found only 10 with complaints.

a) Create a 95% confidence interval for the true level of dissatisfaction among customers.
b) Does this provide evidence that the company has reached its goal? Using your confidence interval, test an appropriate hypothesis, and state your conclusion. **LO ❸**

43. Maintenance costs. A limousine company is concerned with increasing costs of maintaining its fleet of 150 cars. After testing, the company found that the emissions systems of 7 out of the 22 cars it tested failed to meet pollution control guidelines. The company had forecasted costs assuming that a total of 30 cars would need updating to meet the latest guidelines. Is this strong evidence that more than 20% of the fleet might be out of compliance? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed. **LO ❷**

44. Damaged goods. An appliance manufacturer stockpiles washers and dryers in a large warehouse for shipment to retail stores. Sometimes in handling the appliances get damaged. Even though the damage may be minor, the company must sell those machines at drastically reduced prices. The company goal is to keep the proportion of damaged machines below 2%. One day an inspector randomly checks 60 washers and finds that 5 of them have scratches or dents. Is this strong evidence that the warehouse is failing to meet the company goal? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed. **LO ❷**

45. Defective products. An internal report from a manufacturing company indicated that about 3% of all products were defective. Data from one batch found only 7 defective products out of 469 products. Is this consistent with the report? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed. **LO ❷**

46. Jobs. The accounting department of a major Canadian university would like to advertise that more than 50% of its graduates obtained a job offer prior to graduation. A sample of 240 recent graduates indicated that 138 of these graduates had a job offer prior to graduation. Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed. **LO ❷**

47. *WebZine.* A magazine called *WebZine* is considering the launch of an online edition. The magazine plans to go ahead only if it's convinced that more than 25% of current readers would subscribe. The magazine contacts a simple random sample of 500 current subscribers, and 137 of those surveyed expressed interest. What should the magazine do? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed. **LO ❷**

48. Truth in advertising. A garden centre wants to store leftover packets of vegetable seeds for sale the following spring, but the centre is concerned that the seeds may not germinate at the same rate a year later. The manager finds a packet of last year's green bean seeds and plants them as a test. Although the packet claims a germination rate of 92%, only 171 of 200 test seeds sprout. Is this evidence that the seeds have lost viability during a year in storage? Test an appropriate hypothesis and state your conclusion. Be sure the appropriate assumptions and conditions are satisfied before you proceed. **LO ❷**

49. Women executives. A company is criticized because only 13 of 43 people in executive-level positions are women. The company explains that although this proportion is lower than it might wish, it's not surprising given that only 40% of its employees are women. What do you think? Test an appropriate hypothesis and state your conclusion. Be

sure the appropriate assumptions and conditions are satisfied before you proceed. **LO ❷**

50. Jury. Census data for a certain county show that 19% of the adult residents are Hispanic. Suppose 72 people are called for jury duty, and only 9 of them are Hispanic. Does this apparent underrepresentation of Hispanics call into question the fairness of the jury selection system? Explain. **LO ❷**

51. Nonprofit. A nonprofit company concerned with high school dropout rates has designed a tutoring program aimed at students between 16 and 18 years old. Nationally, the high school dropout rate for the year 2013 was 10.9%. One school district, which adopted the use of the nonprofit's tutoring program and has always had a dropout rate very close to the national average, reported in 2013 that 175 of its 1782 students dropped out. Is their experience evidence that the tutoring program has been effective? Explain. **LO ❷**

52. Real estate. A national real estate magazine advertised that 15% of first-home buyers have a family income below $40,000. A national real estate firm believes this percentage is too low and samples 100 of its records. The firm finds that 25 of its first-home buyers did have a family income below $40,000. Does the sample suggest that the proportion of first-home buyers with an income less than $40,000 is more than 15%? Comment and write up your own conclusions based on an appropriate confidence interval as well as a hypothesis test. Include any assumptions you made about the data. **LO ❸**

53. Public relations. An airline's public relations department says that the airline rarely loses luggage. Furthermore, it claims that when it does, 90% of the time the bags are recovered and delivered within 24 hours. A consumer group surveys a large number of air travellers and finds that 103 of 122 people who lost luggage were reunited with their missing items within 24 hours. Does this cast doubt on the airline's claim? Explain. **LO ❷**

54. TV ads. A startup company is about to market a new computer printer. It decides to gamble by running commercials during the Super Bowl. The company hopes that name recognition will be worth the high cost of the ads. The goal of the company is that over 40% of the public recognize its brand name and associate it with computer equipment. The day after the game, a pollster contacts 420 randomly chosen adults and finds that 181 of them know that this company manufactures printers. Would you recommend that the company continue to advertise during the Super Bowl? Explain. **LO ❷**

55. Business ethics. A study reports that 30% of newly hired MBAs are confronted with unethical business practices during their first year of employment. One business school dean wondered if her MBA graduates had had similar experiences. She surveyed recent graduates to find that 27% of the 120 graduates from the previous year claim to have encountered unethical business practices in the workplace. Can she conclude that her graduates' experiences are different? **LO ❷**

56. Stocks, part 2. A young investor believes he can beat the market by picking stocks that will increase in value. Assume that, on average, 50% of the stocks selected by a portfolio manager will increase over 12 months. Of the 25 stocks that the young investor bought over the past 12 months, 14 have increased. Can he claim that he's better at predicting increases than the typical portfolio manager? **LO ❷**

57. Retirement. A survey of 1000 workers indicated that approximately 520 have invested in an individual retirement account. National data suggest that 44% of workers invest in individual retirement accounts.

a) Create a 95% confidence interval for the proportion of workers who have invested in individual retirement accounts based on the survey.
b) Does this provide evidence of a change in behaviour among workers? Using your confidence interval, test an appropriate hypothesis and state your conclusion. **LO ❸**

58. iPod reliability. MacInTouch reported that several versions of the iPod reported failure rates of 20% or more. From a customer survey, the colour iPod, first released in 2004, showed 64 failures out of 517. Is there any evidence that the failure rate for this model may be lower than the 20% rate of previous models?

a) State the hypotheses.
b) Find the z-score of the observed proportion.
c) Compare the z-score to the critical value for a 0.1% significance level using a one-sided alternative.
d) Explain your conclusion. **LO ❷**

59. Testing cars. A clean air standard requires that vehicle exhaust emissions not exceed specified limits for various pollutants. Suppose government regulators double-check a random sample of cars that a suspect repair shop has certified as okay. They will revoke the shop's licence if they find significant evidence that the shop is certifying vehicles that don't meet standards.

a) In this context, what is a Type I error?
b) In this context, what is a Type II error?
c) Which type of error would the shop's owner consider more serious?
d) Which type of error might environmentalists consider more serious? **LO ❹**

60. Quality control. Production managers on an assembly line must monitor the output to be sure that the level of defective products remains small. They periodically inspect a random sample of the items produced. If they find a significant increase in the proportion of items that must be

rejected, they'll halt the assembly process until the problem can be identified and repaired.

a) Write null and alternative hypotheses for this problem.
b) What are the Type I and Type II errors in this context?
c) Which type of error would the factory owner consider more serious?
d) Which type of error might customers consider more serious? **LO ❹**

61. Testing cars, again. As in Exercise 59, regulators are checking up on repair shops to see if they're certifying vehicles that don't meet pollution standards.

a) In this context, what is meant by the power of the test the regulators are conducting?
b) Will the power be greater if they test 20 or 40 cars? Why?
c) Will the power be greater if they use a 5% or a 10% level of significance? Why?
d) Will the power be greater if the repair shop's inspectors are only a little out of compliance or a lot? Why? **LO ❹**

62. Quality control, part 2. Consider again the task of the quality control inspectors in Exercise 60.

a) In this context, what is meant by the power of the test the inspectors conduct?
b) They're currently testing five items each hour. Someone has proposed that they test 10 items each hour instead. What are the advantages and disadvantages of such a change?
c) Their test currently uses a 5% level of significance. What are the advantages and disadvantages of changing to a significance level of 1%?
d) Suppose that as a day passes one of the machines on the assembly line produces more and more items that are defective. How will this affect the power of the test? **LO ❹**

63. Statistics software. A Statistics professor has observed that for several years about 13% of the students who initially enrol in his Introductory Statistics course withdraw before the end of the semester. A salesperson suggests that he try a statistics software package that gets students more involved with computers, predicting that it will cut the dropout rate. The software is expensive, and the salesperson offers to let the professor use it for a semester to see if the dropout rate goes down significantly. The professor will have to pay for the software only if he chooses to continue using it.

a) Is this a one-tailed or two-tailed test? Explain.
b) Write the null and alternative hypotheses.
c) In this context, explain what would happen if the professor makes a Type I error.
d) In this context, explain what would happen if the professor makes a Type II error.
e) What is meant by the power of this test? **LO ❹**

64. Radio ads. A company is willing to renew its advertising contract with a local radio station only if the station can prove that more than 20% of the residents of the city have heard the ad and recognize the company's product. The radio station conducts a random phone survey of 400 people.

a) What are the hypotheses?
b) The station plans to conduct this test using a 10% level of significance, but the company wants the significance level lowered to 5%. Why?
c) What is meant by the power of this test?
d) For which level of significance will the power of this test be higher? Why?
e) The station finally agrees to use $\alpha = 0.05$, but the company proposes that the station call 600 people instead of the 400 initially proposed. Will that make the risk of Type II error higher or lower? Explain. **LO ❹**

65. Statistics software, part 2. Initially, 203 students signed up for the Statistics course in Exercise 63. They used the software suggested by the salesperson, and only 11 dropped out of the course.

a) Should the professor spend the money for this software? Support your recommendation with an appropriate test.
b) Explain what your P-value means in this context. **LO ❷**

66. Radio ads, part 2. The company in Exercise 64 contacts 600 people selected at random, and 133 can remember the ad.

a) Should the company renew the contract? Support your recommendation with an appropriate test.
b) Explain carefully what your P-value means in this context. **LO ❷**

T 67. Customer spending The data set provided contains last month's credit card purchases of 500 customers randomly chosen from a segment of a major credit card issuer. The marketing department is considering a special offer for customers who spend more than $1000 per month on their card. Historically, the percentage has been 11%, and the finance department wonders if it has increased. Test the appropriate hypothesis and write up a few sentences with your conclusions. **LO ❷**

68. Fundraising. A philanthropic organization knows that its donors have an average age near 60, and so is considering taking out an ad in the *Canadian Association of Retired Persons* (CARP) magazine. The head of finance says that the CARP advertisement won't be worth the money unless more than two-thirds of the donors are 50 or older. Test the appropriate hypothesis and write up a few sentences with your conclusions. **LO ❷**

69. Drugs in Canada. The Liberal government that decriminalized the possession of small quantities of marijuana was replaced by a Conservative government in the 2006 federal election. The new government considered scrapping the legislation, and in 2007 Angus Reid conducted a survey of 1028 adult Canadians and found that only

38% would support that move. However, pollsters also found that 59% of Conservative voters supported scrapping the legislation. Suppose the Conservative government wanted to adopt policies supported by more than 50% of its own voters. Should it have scrapped the legislation in 2007? Assume 40% of the people surveyed were Conservative voters. (**Source:** Based on Angus Reid Strategies. [2007]. Illegal drugs: Drugs a national problem—Canadians would mix Grit and Tory policies. Retrieved from http://juror.ca/Angus_Reid_55%25.pdf.) **LO ❷**

70. Drugs in Canada, part 2. In 2002, the Liberal government introduced "harm reduction" programs for drug users, such as supervised injection sites and needle-exchange programs. However, a Conservative government was elected in 2006. The new government considered eliminating the harm reduction programs, and in 2007 Angus Reid conducted a survey of 1028 adult Canadians and found that only 37% would support that move. Fifty-four percent of Conservative voters supported eliminating the programs, however. Suppose the Conservative government wanted to adopt policies supported by more than 50% of its own voters. Should it have eliminated the harm reduction programs in 2007? Assume 40% of the people surveyed were Conservative voters. (**Source:** Angus Reid Strategies. [2007]. Illegal drugs: Drugs a national problem—Canadians would mix Grit and Tory policies. Retrieved from http://juror.ca/Angus_Reid_55%25.pdf.) **LO ❷**

71. The Canadian monarchy. In May 2010, in advance of the visit of Queen Elizabeth II for Canada Day celebrations, Angus Reid conducted a survey of 1005 adult Canadians to determine their views on the monarchy. The results were that 36% preferred Canada to have an elected head of state, whereas a similar survey conducted in November 2009 had resulted in a figure of 43% for the same question. (**Source:** Angus Reid Strategies. [2010]. The monarchy: Canadians divided over having a monarch or an elected head of state.)

a) How sure can we be that the percentage of adult Canadians preferring an elected head of state in May 2010 was in fact less than 43%?

b) In the Atlantic provinces, those preferring an elected head of state represented 25% of those surveyed in May 2010. How sure can we be that the percentage preferring an elected head of state in the Atlantic provinces was in fact more than 20%? Assume that 7% of the survey respondents were in the Atlantic provinces. **LO ❷**

72. Sex offenders in Canada. In May 2010, Public Safety Minister Vic Toews introduced a bill in Parliament proposing that sex offenders who have abused children should not be able to get a government pardon. Angus Reid conducted a survey of 1013 Canadian adults and found that 81% supported the bill. (**Source:** Angus Reid Strategies. [2010]. Crime: Canadians support proposed new regulations for pardons. Retrieved from http://www.angus-reid.com/polls/43065/canadians-support-proposed-new-regulations-for-pardons.)

a) How sure can we be that the percentage of Canadian adults supporting the bill was in fact more than 80%?

b) The survey found that 76% of Liberal Party voters supported the bill. How sure can we be that the percentage of Liberal voters supporting the bill is in fact less than 80%? Assume that 40% of the survey respondents were Liberal voters. **LO ❷**

73. Canadian Aboriginals. Aboriginals make up 3.9% of the Canadian population, and in 2012, 7 out of the 308 federal Members of Parliament were Aboriginal. Using the methods given in this chapter, compare these pieces of information to answer the question "Are Aboriginals underrepresented in Parliament?" State your assumptions clearly. **LO ❷**

74. Canadian Aboriginals again. During the entire history of Canada, 0.74% of the 4201 federal Members of Parliament have been Aboriginals. In 2012, 7 out of the 308 federal Members of Parliament were Aboriginal. Using the methods given in this chapter, compare these pieces of information to answer the question "Were Aboriginals better represented in Parliament in 2012 than previously?" State your assumptions clearly. **LO ❷**

75. Canadian Senate. In July 2011, Angus Reid surveyed 1000 adult Canadians about reform of the Canadian Senate.

a) 72% answered "Yes" to the question "Do you support allowing Canadians to directly elect their senators?" Does this indicate that the proportion of the adult Canadian population who would answer "Yes" to this question is over 70% at the 95% significance level?

b) 71% said that a referendum should be held to determine the future of the Canadian Senate. Does this indicate that the proportion of Canadians who favour a referendum is over 70% at the 90% significance level?

c) A statistician commented to the press: "Over 70% of Canadians want a referendum on Senate reform and want to elect their senators directly." Comment on the ethics of this statement in relation to the ASA Ethical Guidelines summarized in Appendix C. **LO ❷**

76. Molson. Molson has been in business in Canada for over two and a quarter centuries, a fact that speaks to the consistent quality of the beer. Suppose Molson's aim is that at least 98% of customers say the taste hasn't changed. One way to ensure consistency is to run taste tests with a random selection of customers. Ninety-nine percent of a random sample of 850 customers say the taste hasn't changed in the past 10 years. Is that sufficient for Molson's purposes? **LO ❷**

77. The BC carbon tax. British Columbia introduced a carbon tax of $10 per tonne of carbon in 2008, rising by $5 per tonne each year until 2012. In 2011, with the tax at $25 per tonne,

Environics surveyed 1025 British Columbians about the tax and found that 54% supported it. A politician claimed, "Over half of British Columbians support the carbon tax."

a) At what significance level is this statement statistically sound?
b) Comment on the ethics of this statement in relation to the ASA Ethical Guidelines summarized in Appendix C. How could the statement be improved? **LO ❷**

78. The Canadian penny. It costs 1.5 cents to produce the one-cent coin, otherwise known as the "penny." Angus Reid Strategies surveyed 1016 Canadian adults and found that 55% were in favour of scrapping the penny. A reporter commented, "Over half Canadian adults want to scrap the penny."

a) Comment on the statistical validity of this statement.
b) Comment on the ethics of this statement as it relates to the ASA Ethical Guidelines in Appendix C. **LO ❷**

79. The Canadian penny again. Angus Reid Strategies surveyed 1016 Canadian adults and found that support for scrapping the penny was high in British Columbia (62%) and Ontario (55%). Do these survey results lead to the conclusion that a greater percentage of British Columbian adults support scrapping the penny than Ontarian adults? (*Hint:* Assume that 13.3% of the people surveyed were in British Columbia and 38.8% were in Ontario.) **LO ❺**

80. The Canadian penny, third time. Angus Reid Strategies surveyed 1016 Canadian adults (assume a 50/50 split between men and women) and found that 65% of men and 45% of women support scrapping the penny.

a) At what level of significance can we conclude that more Canadian men than women support scrapping the penny?
b) If we wanted to check out the conclusion in a) at the 95% level without going to the expense of surveying 1000 people, how many people would we need to survey? **LO ❺**

81. Canadian Senate again. In July 2011, Angus Reid surveyed 1000 adult Canadians about reform of the Canadian Senate. Thirty-two percent of people in Ontario supported abolishing the Senate of Canada, whereas the percentage in Quebec was 43%. Do more people in Quebec support abolishing the Canadian Senate than in Ontario? (Assume that 38.4% of the people surveyed were in Ontario and 23.6% were in Quebec.) **LO ❺**

82. Single-parent families in Canada. A random sample of 1000 families in Alberta reported 14% single-parent families. Another random sample of 700 families in Nova Scotia reported 17% single-parent families. Is there a lower percentage of single-parent families in Alberta than in Nova Scotia at the 95% level of significance? **LO ❺**

83. Adults living with parents in Canada. A random survey of 1000 Canadians aged 20–24 in 2008 found that 61.4% lived with their parents. In 2013 a similar survey found 65.3% of Canadians aged 20–24 living with their parents. Is there a difference between the percentages of Canadians aged 20–24 living with their parents in 2008 and 2013 at the 95% level of significance? **LO ❺**

84. Mature middle class in India. The National Sample Survey identifies a segment of Indian households as the "Mature Middle Class" with an income of around 170,000 rupees per year. The rapid growth of the Indian economy has resulted in an increase in the proportion of households in this group, and the survey estimates that it increased from 27% in 2005 to 50% in 2011. What sample size is necessary in order to conclude that the proportion of the Indian population in this group increased between 2005 and 2011 at the 99% significance level? You can assume that the size of the sample is the same in 2005 and in 2011. **LO ❺**

Just Checking Answers

1 You can't conclude that the null hypothesis is true. You can conclude only that the experiment was unable to reject the null hypothesis. They were unable, on the basis of 12 patients, to show that aspirin was effective.

2 The null hypothesis is $H_0: p = 0.75$.

3 With a P-value of 0.0001, this is very strong evidence against the null hypothesis. We can reject H_0 and conclude that the improved version of the drug gives relief to a higher proportion of patients.

4 The parameter of interest is the proportion, p, of all delinquent customers who will pay their bills. $H_0: p = 0.30$ and $H_A: p > 0.30$.

5 At $\alpha = 0.05$, you can't reject the null hypothesis because 0.30 is contained in the 90% confidence interval—it's plausible that sending the DVDs is no more effective than sending letters.

6 The confidence interval is from 29% to 45%. The DVD strategy is more expensive and may not be worth it. We can't distinguish the success rate from 30% given the results of this experiment, but 45% would represent a large improvement. The bank should consider another trial, increasing the sample size to get a narrower confidence interval .

7 A Type I error would mean deciding that the DVD success rate is higher than 30% when it isn't. The bank would adopt a more expensive method for collecting payments that's no better than its original, less expensive strategy.

8 A Type II error would mean deciding that there's not enough evidence to say the DVD strategy works when in fact it does. The bank would fail to discover an effective method for increasing revenue from delinquent accounts.

9 Higher; the larger the effect size, the greater the power. It's easier to detect an improvement to a 60% success rate than to a 32% rate.