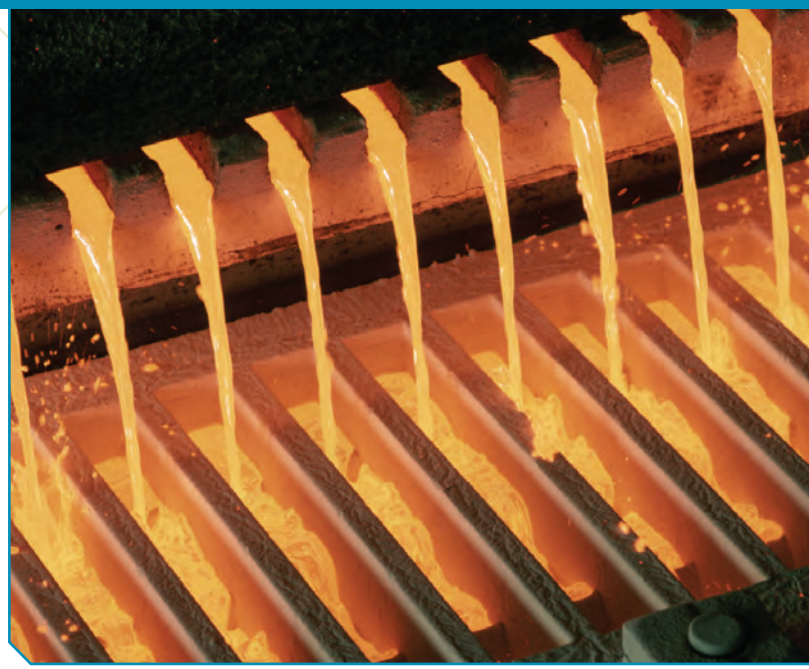
CHAPTER 20

Testing Hypotheses About Proportions



A S ***Activity:*** **Testing a Claim.** Can we really draw a reasonable conclusion from a random sample? Run this simulation before you read the chapter, and you'll gain a solid sense of what we're doing here.

"Half the money I spend on advertising is wasted; the trouble is I don't know which half."

—John Wanamaker (attributed)

Ingots are huge pieces of metal, often weighing more than 20,000 pounds, made in a giant mold. 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 can 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 is completely under control. In one plant, only about 80% of the ingots have been free of cracks. In an attempt to reduce the cracking proportion, the plant engineers and chemists recently tried out some changes in the casting process. Since then, 400 ingots have been cast and only 17% of them have cracked. Should management declare victory? Has the cracking rate really decreased, or was 17% just due to luck?

We can treat the 400 ingots cast with the new method as a random sample. We know that each random sample will have a somewhat different proportion of cracked ingots. Is the 17% we observe merely a result of natural sampling variability, or is this lower cracking rate strong enough evidence to assure management that the true cracking rate now is really below 20%?

People want answers to questions like these all the time. Has the president's approval rating changed since last month? Has teenage smoking decreased in the past five years? Is the global temperature increasing? Did the Super Bowl ad we bought actually increase sales? To answer such questions, we test *hypotheses* about models.

Hypotheses

How can we state and test a hypothesis about ingot cracking? Hypotheses are working models that we adopt temporarily. To test whether the changes made by the engineers have *improved* the cracking rate, we assume that they have in fact

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*

made no difference and that any apparent improvement is just random fluctuation (sampling error). So, our starting hypothesis, called the **null hypothesis,** is that the proportion of cracks is still 20%.

The null hypothesis, which we denote H_0, specifies a population model parameter of interest and proposes a value for that parameter. We usually write down the null hypothesis in the form H_0: *parameter* = *hypothesized value*. This is a concise way to specify the two things we need most: the identity of the parameter we hope to learn about and a specific hypothesized value for that parameter. (We need a hypothesized value so we can compare our observed statistic value to it.)

Which value to use is often obvious from the *Who* and *What* 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 the ingots we can write H_0: $p = 0.20$.

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 the ingots example, our null hypothesis is that $p = 0.20$. What's the alternative? Management is interested in *reducing* the cracking rate, so their alternative is H_A: $p < 0.20$.

NOTATION ALERT:

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

What would convince you that the cracking rate had actually gone down? If you observed a cracking rate *much lower* than 20% in your sample, you'd likely be convinced. If only 3 out of the next 400 ingots crack (for a rate of 0.75%), most folks would conclude that the changes helped. But if the sample cracking rate is only slightly lower than 20%, you should be skeptical. After all, observed proportions do vary, so we wouldn't be surprised to see some difference. How much smaller must the cracking rate be before we *are* convinced that it has changed? Whenever we ask about the size of a statistical difference, we naturally think of using the standard deviation as a ruler. So let's start by finding the standard deviation of the sample cracking rate.

Since the company changed the process, 400 new ingots have been cast. The sample size of 400 is big enough to satisfy the **Success/Failure Condition.** (We expect $0.20 \times 400 = 80$ ingots to crack.) We have no reason to think the ingots are not independent, so the Normal sampling distribution model should work well. The standard deviation of the sampling model is

To remind us that the parameter value comes from the null hypothesis, it is sometimes written as p_0 and the standard deviation as

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

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.20)(0.80)}{400}} = 0.02$$

Why is this a standard deviation and not a standard error? Because we haven't estimated anything. When 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. And 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 $\hat{p}$.

Now we know both parameters of the Normal sampling distribution model: $p = 0.20$ and $SD(\hat{p}) = 0.02$, so we can find out how likely it would be to see the observed value of $\hat{p} = 17\%$. Since we are using a Normal model, we find the z-score:

$$z = \frac{0.17 - 0.20}{0.02} = -1.5$$

Then we ask, "How likely is it to observe a value at least 1.5 standard deviations below the mean of a Normal model?" The answer (from a calculator, computer program, or the Normal table) is about 0.067. This is the probability of observing a cracking rate of 17% or less in a sample of 400 if the null hypothesis is true.

Management now must decide whether an event that would happen 6.7% of the time by chance is strong enough evidence to conclude that the true cracking proportion has decreased.

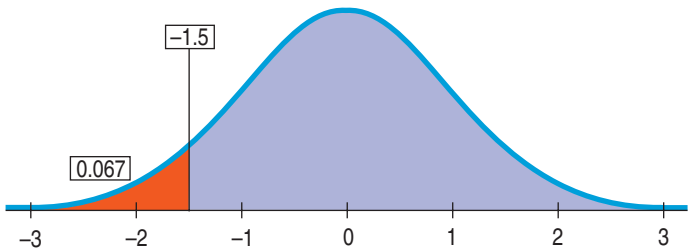


FIGURE 20.1
How likely is a z-score of −1.5 (or lower)? This is what it looks like. The red area is 0.067 of the total area under the curve.

A Trial as a Hypothesis Test

Does the reasoning of hypothesis tests seem backward? That could be because we usually prefer to think about getting things right rather than getting them wrong. You have 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 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, you ask, "Could these data plausibly have happened by chance if the null hypothesis were true?" If they are very unlikely to 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 wonderfully 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 were the null hypothesis true.

How unlikely is unlikely? Some people set rigid standards, like 1 time out of 20 (0.05) or 1 time out of 100 (0.01). 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."

A S ***Activity:* The Reasoning of Hypothesis Testing.** Our reasoning is based on a rule of logic that dates back to ancient scholars. Here's a modern discussion of it.

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 were the null hypothesis a 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 the null hypothesis is true. Statisticians are so thrilled with their ability to measure precisely

Beyond a Reasonable Doubt

We ask whether the data were unlikely beyond a reasonable doubt. We've just calculated that probability. The probability that the observed statistic value (or an even more extreme value) could occur if the null model were true—in this case, 0.067—is the P-value.

how surprised they are that they give this probability a special name. It's called a **P-value.**[1]

When the P-value is high, we haven't seen anything unlikely or surprising at all. Events that have a high probability of happening happen often. The data are thus consistent with the model from the null hypothesis, and we have no reason to reject the null hypothesis. But we realize that many other similar hypotheses could also account for the data we've seen, so *we haven't proven that the null hypothesis is true*. The most we can say is that it doesn't appear to be false. Formally, we "fail to reject" the null hypothesis. That's a pretty weak conclusion, but it's all we're entitled to.

When the P-value *is* low enough, it says that it's very unlikely we'd observe data like these if our null hypothesis were true. We started with a model. Now that model tells us that the data we have are unlikely to have happened. 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 we were wrong to use it as the basis for computing our P-value. Perhaps another model is correct, and the data really aren't that remarkable after all. If you believe in data more than in assumptions, then, given that choice, you should reject the null hypothesis.

NOTATION ALERT:

We have many P's to keep straight. We use an uppercase P for probabilities, as in $P(\mathbf{A})$, and for the special probability we care about in hypothesis testing, the P-value.

We use lowercase p to denote our model's underlying proportion parameter and $\hat{p}$ to denote our observed proportion statistic.

What to Do with an "Innocent" Defendant

"If the People fail to satisfy their burden of proof, you must find the defendant not guilty."

—NY state jury instructions

If the evidence is not strong enough to reject the defendant's presumption of innocence, what verdict does the jury return? They say "not guilty." Notice that they do not say that the defendant is innocent. All they say is that they have not seen sufficient evidence to convict, to reject innocence. The defendant may, in fact, be innocent, but the jury has no way to be sure.

Said statistically, the jury's null hypothesis is H_0: innocent defendant. If the evidence is too unlikely given this assumption, the jury rejects the null hypothesis and finds the defendant guilty. But—and this is an important distinction—if there is *insufficient evidence* to convict the defendant, the jury does not decide that H_0 is true and declare the defendant innocent. Juries can only *fail to reject* the null hypothesis and declare the defendant "not guilty."

In the same way, if the data are not 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 (or "accept" the null), because we simply do not know whether it's true or not. (After all, more evidence may come along later.)

Don't "Accept" the Null Hypothesis

Every child knows that he (or she) is at the "center of the universe," so it's natural to suppose that the sun revolves around the earth. The fact that the sun appears to rise in the east every morning and set in the west every evening is *consistent* with this hypothesis and *seems* to lend support to it, but it certainly doesn't prove it, as we all eventually come to understand.

In the trial, the burden of proof is on the prosecution. In a hypothesis test, the burden of proof is on the unusual claim. The null hypothesis is the ordinary state of affairs, so it's the alternative to the null hypothesis that we consider unusual and for which we must marshal evidence.

Imagine a clinical trial testing the effectiveness of a new headache remedy. In Chapter 13 we saw the value of comparing such treatments to a placebo. The null hypothesis, then, is that the new treatment is no more effective than the placebo. This is important, because some patients will improve even when administered the placebo treatment. If we use only six people to test the drug, the results are likely *not to be clear* and we'll be unable to reject the hypothesis. Does this mean the drug doesn't work? Of course not. It simply means that we don't have enough

[1] You'd think if they were so excited, they'd give it a better name, but "P-value" is about as excited as statisticians get.

evidence to reject our assumption. That's why we don't start by assuming that the drug *is more effective*. If we were to do that, then we could test just a few people, find that the results aren't clear, and claim that since we've been unable to reject our original assumption the drug must be effective. The FDA is unlikely to be impressed by that argument.

JUST CHECKING

1. A research team wants to know if aspirin helps to thin blood. The null hypothesis says that it doesn't. They 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 if the new, improved version works even better. What would the null hypothesis be?
3. The new drug is tested and the P-value is 0.0001. What would you conclude about the new drug?

The Reasoning of Hypothesis Testing

"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*

Some folks pronounce the hypothesis labels "Ho!" and "Ha!" (but it makes them seem overexcitable). We prefer to pronounce H_0 "H naught" (as in "all is for naught").

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

1. HYPOTHESES

First we state the null hypothesis. That's usually the skeptical claim that nothing's different. Are we considering a (New! Improved!) possibly better method? The null hypothesis says, "Oh yeah? Convince me!" To convert a skeptic, we must pile up enough evidence against the null hypothesis that we can reasonably reject it.

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 shorthand, 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.

FOR EXAMPLE Writing hypotheses

A large city's Department of Motor Vehicles claimed that 80% of candidates pass driving tests, but a newspaper reporter's survey of 90 randomly selected local teens who had taken the test found only 61 who passed.

Question: Does this finding suggest that the passing rate for teenagers is lower than the DMV reported? Write appropriate hypotheses.

I'll assume that the passing rate for teenagers is the same as the DMV's overall rate of 80%, unless there's strong evidence that it's lower.

$$H_0: p = 0.80$$
$$H_A: p < 0.80$$

2. Model

To plan a statistical hypothesis test, specify the *model* you will use to test the null hypothesis and the parameter of interest. Of course, all models require assumptions, so you will need to state them and check any corresponding conditions.

Your Model step should end with a statement such as

Because the conditions are satisfied, I 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, I can't proceed with the test. (If that's the case, stop and reconsider.)

Each test in the book has a name that you should include in your report. We'll see many tests in the chapters that follow. 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.**[2]

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.

A S ***Activity:*** **Was the Observed Outcome Unlikely?** Complete the test you started in the first activity for this chapter. The narration explains the steps of the hypothesis test.

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 $z = \dfrac{(\hat{p} - p_0)}{SD(\hat{p})}$. We use the hypothesized proportion to find the standard deviation, $SD(\hat{p}) = \sqrt{\dfrac{p_0 q_0}{n}}$.

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.

FOR EXAMPLE Checking the conditions

Recap: A large city's DMV claimed that 80% of candidates pass driving tests. A reporter has results from a survey of 90 randomly selected local teens who had taken the test.

Question: Are the conditions for inference satisfied?

- ✔ The 90 teens surveyed were a random sample of local teenage driving candidates.
- ✔ 90 is fewer than 10% of the teenagers who take driving tests in a large city.
- ✔ We expect $np_0 = 90(0.80) = 72$ successes and $nq_0 = 90(0.20) = 18$ failures. Both are at least 10.

The conditions are satisfied, so it's okay to use a Normal model and perform a one-proportion z-test.

Conditional Probability
Did you notice that a P-value is a conditional probability? It's the probability that the observed results could have happened *if the null hypothesis is true.*

3. Mechanics

Under "Mechanics," we place 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, but it's good to have the formulas recorded for reference and to know what's

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

being computed. 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) occur if the null model is correct. If the P-value is small enough, we'll reject the null hypothesis.

FOR EXAMPLE Finding a P-value

Recap: A large city's DMV claimed that 80% of candidates pass driving tests, but a survey of 90 randomly selected local teens who had taken the test found only 61 who passed.

Question: What's the P-value for the one-proportion *z*-test?

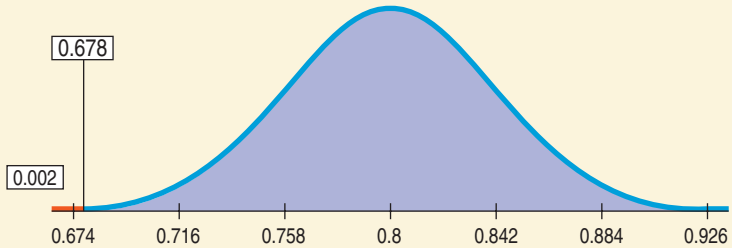
I have $n = 90$, $x = 61$, and a hypothesized $p = 0.80$.

$$\hat{p} = \frac{61}{90} \approx 0.678$$

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.8)(0.2)}{90}} \approx 0.042$$

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{0.678 - 0.800}{0.042} \approx -2.90$$

$$P\text{-value} = P(z < -2.90) = 0.002$$



4. Conclusion

The conclusion in a hypothesis test is always a statement about the null hypothesis. The conclusion must state either that we reject or that we fail to reject the null hypothesis. And, as always, the conclusion should be stated in context.

FOR EXAMPLE Stating the conclusion

Recap: A large city's DMV claimed that 80% of candidates pass driving tests. Data from a reporter's survey of randomly selected local teens who had taken the test produced a P-value of 0.002.

Question: What can the reporter conclude? And how might the reporter explain what the P-value means for the newspaper story?

Because the P-value of 0.002 is very low, I reject the null hypothesis. These survey data provide strong evidence that the passing rate for teenagers taking the driving test is lower than 80%.

If the passing rate for teenage driving candidates were actually 80%, we'd expect to see success rates this low in only about 1 in 500 samples (0.2%). This seems quite unlikely, casting doubt that the DMV's stated success rate applies to teens.

". . . 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*

Your conclusion about the null hypothesis should never be the end of a testing procedure. Often there are actions to take or policies to change. In our ingot example, management must decide whether to continue the changes proposed by the engineers. The decision always includes the practical consideration of whether the new method is worth the cost. Suppose management decides to reject the null hypothesis of 20% cracking in favor of the alternative that the percentage has been reduced. They must still evaluate how much the cracking rate has been reduced and how much it cost to accomplish the reduction. The *size of the effect* is always a concern when we test hypotheses. A good way to look at the effect size is to examine a confidence interval.

How much does it cost? Formal tests of a null hypothesis base the decision of whether to reject the null hypothesis solely on the size of the P-value. But in real life, we want to evaluate the costs of our decisions as well. How much would you be willing to pay for a faster computer? Shouldn't your decision depend on how much faster? And on how much more it costs? Costs are not just monetary either. Would you use the same standard of proof for testing the safety of an airplane as for the speed of your new computer?

Alternative Alternatives

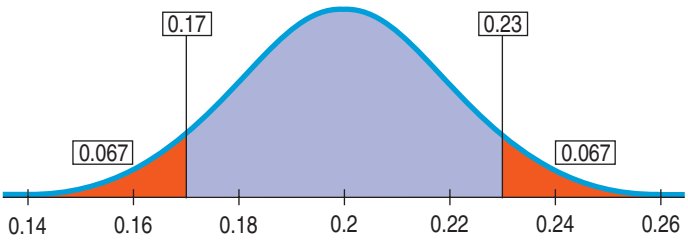
Tests on the ingot data can be viewed in two different ways. We know the old cracking rate is 20%, so the null hypothesis is

$$H_0: p = 0.20$$

A S ***Activity:* the Alternative Hypotheses.** This interactive tool provides easy ways to visualize how one- and two-tailed alternative hypotheses work.

But we have a choice of alternative hypotheses. A metallurgist working for the company might be interested in *any* change in the cracking rate due to the new process. Even if the rate got worse, she might learn something useful from it. She's interested in possible changes on both sides of the null hypothesis. So she would write her alternative hypothesis as

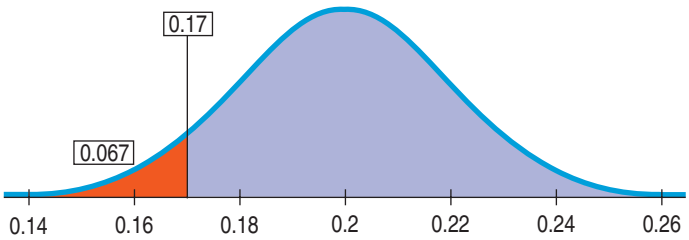
$$H_A: p \neq 0.20$$



An alternative hypothesis such as this is known as a **two-sided alternative**,[3] because we are equally interested in deviations on either side of the null hypothesis value. For two-sided alternatives, the P-value is the probability of deviating in *either* direction from the null hypothesis value.

But management is really interested only in *lowering* the cracking rate below 20%. The scientific value of knowing how to *increase* the cracking rate may not appeal to them. The only alternative of interest to them is that the cracking rate *decreases.* They would write their alternative hypothesis as

$$H_A: p < 0.20$$



An alternative hypothesis that focuses on deviations from the null hypothesis value in only one direction is called a **one-sided alternative.**

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 the same data, the one-sided P-value is half the two-sided P-value. So, a one-sided test will reject the null hypothesis more often. If you aren't sure which to use, a two-sided test is always more conservative. Be sure you can justify the choice of a one-sided test from the *Why* of the situation.

[3] It is also called a **two-tailed alternative,** because the probabilities we care about are found in both tails of the sampling distribution.

STEP-BY-STEP EXAMPLE Testing a Hypothesis

Anyone who plays or watches sports has heard of the "home field advantage." Teams tend to win more often when they play at home. Or do they?

If there were no home field advantage, the home teams would win about half of all games played. In the 2007 Major League Baseball season, there were 2431 regular-season games. (Tied at the end of the regular season, the Colorado Rockies and San Diego Padres played an extra game to determine who won the Wild Card playoff spot.) It turns out that the home team won 1319 of the 2431 games, or 54.26% of the time.

Question: Could this deviation from 50% be explained just from natural sampling variability, or is it evidence to suggest that there really is a home field advantage, at least in professional baseball?

Plan State what we want to know.

Define the variables and discuss the W's.

I want to know whether the home team in professional baseball is more likely to win. The data are all 2431 games from the 2007 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's no advantage, I'd expect that proportion to be 0.50.

Hypotheses The null hypothesis makes the claim of no difference from the baseline. Here, that means no home field advantage.

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

$$H_0: p = 0.50$$
$$H_A: p > 0.50$$

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 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. The data come from one entire season, but I expect other seasons to be similar.
- ✓ **Randomization Condition:** I have results for all 2431 games of the 2007 season. But I'm not just interested in 2007, and those games, while not randomly selected, should be a reasonable representative sample of all Major League Baseball games in the recent past and near future.
- ✓ **10% Condition:** We are interested in home field advantage for Major League Baseball for all seasons. While not a random sample, these 2431 games are fewer than 10% of all games played over the years.
- ✓ **Success/Failure Condition:** Both $np_0 = 2431(0.50) = 1215.5$ and $nq_0 = 2431(0.50) = 1215.5$ are at least 10.

A S *Activity:* **Practice with Testing Hypotheses About Proportions.** Here's an interactive tool that makes it easy to see what's going on in a hypothesis test.

Specify the sampling distribution model.

State what test you plan to use.

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

SHOW

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

The null model is a Normal distribution with a mean of 0.50 and a standard deviation of

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.5)(1 - 0.5)}{2431}}$$
$$= 0.01014$$

Next, we find the *z*-score for the observed proportion, to find out how many standard deviations it is from the hypothesized proportion.

The observed proportion, $\hat{p}$, is 0.5426.

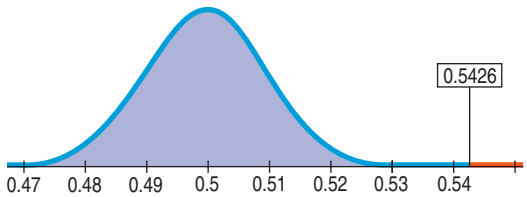
So the z-value is

$$z = \frac{0.5426 - 0.5}{0.01014} = 4.20$$

From the *z*-score, we can find the P-value, which tells us the probability of observing a value that extreme (or more).

The sample proportion lies 4.20 standard deviations above the mean.

The probability of observing a value 4.20 or more standard deviations above the mean of a Normal model can be found by computer, calculator, or table to be < 0.001.



The corresponding P-value is < 0.001.

TELL

Conclusion State your conclusion about the parameter—in context, of course!

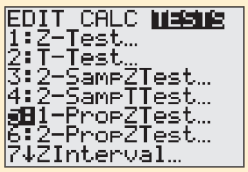
The P-value of < 0.001 says that if the true proportion of home team wins were 0.50, then an observed value of 0.5426 (or larger) would occur less than 1 time in 1000. With a P-value so small, I reject H_0. I have evidence that the true proportion of home team wins is not 50%. It appears there is a home field advantage.

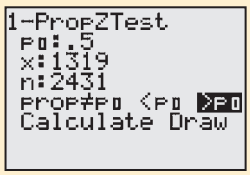
Ok, but how *big* is the home field advantage? Measuring the size of the effect involves a confidence interval. (Use your calculator.)

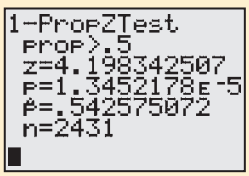
TI Tips

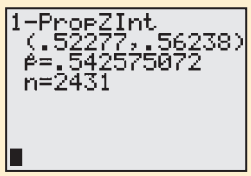
Testing a hypothesis

By now probably nothing surprises you about your calculator. Of course it can help you with the mechanics of a hypothesis test. But that's not much. It cannot write the correct hypotheses, check the appropriate conditions, interpret the results, or state a conclusion. You have to do the tough stuff!









Let's do the mechanics of the Step-By-Step example about home field advantage in baseball. We hypothesized that home teams would win 50% of all games, but during this 2431-game season they actually won 54.26% of the time.

- Go to the STAT TESTS menu. Scroll down the list and select 5:1-Prop ZTest.
- Specify the hypothesized proportion p0.
- Enter x, the observed number of wins: 1319.
- Specify the sample size.
- Since this is a one-tail upper tail test, indicate that you want to see if the observed proportion is significantly greater than what was hypothesized.
- Calculate the result.

Ok, the rest is up to you. The calculator reports a z-score of 4.20 and a P-value (in scientific notation) of 1.35×10^{-5}, or about 0.00001. Such a small P-value indicates that the high percentage of home team wins is highly unlikely to be sampling error. State your conclusion in the appropriate context.

And how big is the advantage for the home team? In the last chapter you learned to create a 95% confidence interval. Try it here.

Looks like we can be 95% confident that in major league baseball games the home team wins between 52.3% and 56.2% of the time. Over a full season, the low end of this interval, 52.3% of the 81 home games, is nearly 2 extra victories, on average. The upper end, 56.2%, is 5 extra wins.

P-Values and Decisions: What to Tell About a Hypothesis Test

Hypothesis tests are particularly useful when we must make a decision. Is the defendant guilty or not? Should we choose print advertising or television? Questions like these cannot always be answered with the margins of error of confidence intervals. The absolute nature of the hypothesis test decision, however, makes some people (including the authors) uneasy. If possible, it's often a good idea to report a confidence interval for the parameter of interest as well.

How small should the P-value be in order for you to reject the null hypothesis? A jury needs enough evidence to show the defendant guilty "beyond a reasonable doubt." How does that translate to P-values? The answer is that it's highly context-dependent. When we're screening for a disease and want to be sure we treat all those who are sick, we may be willing to reject the null hypothesis of no disease with a P-value as large as 0.10. We would rather treat the occasional healthy person than fail to treat someone who was really sick. But a long-standing hypothesis, believed by many to be true, needs stronger evidence (and a correspondingly small P-value) to reject it.

See if you require the same P-value to reject each of the following null hypotheses:

- A renowned musicologist claims that she can distinguish between the works of Mozart and Haydn simply by hearing a randomly selected 20 seconds of music from any work by either composer. What's the null hypothesis? If she's just guessing, she'll get 50% of the pieces correct, on average. So our null hypothesis is that p is 50%. If she's for real, she'll get more than 50% correct. Now, we present her with 10 pieces of Mozart or Haydn chosen at random. She gets 9 out of 10 correct. It turns out that the P-value associated with

that result is 0.011. (In other words, if you tried to just guess, you'd get at least 9 out of 10 correct only about 1% of the time.) What would *you* conclude? Most people would probably reject the null hypothesis and be convinced that she has some ability to do as she claims. Why? Because the P-value is small and we don't have any particular reason to doubt the alternative.

"Extraordinary claims require extraordinary proof."
—Carl Sagan

- On the other hand, imagine a student who bets that he can make a flipped coin land the way he wants just by thinking hard. To test him, we flip a fair coin 10 times. Suppose he gets 9 out of 10 right. This also has a P-value of 0.011. Are you willing now to reject this null hypothesis? Are you convinced that he's not just lucky? What amount of evidence *would* convince you? We require more evidence if rejecting the null hypothesis would contradict long-standing beliefs or other scientific results. Of course, with sufficient evidence we would revise our opinions (and scientific theories). That's how science makes progress.

Another factor in choosing a P-value is the importance of the issue being tested. Consider the following two tests:

- A researcher claims that the proportion of college students who hold part-time jobs now is higher than the proportion known to hold such jobs a decade ago. You might be willing to believe the claim (and reject the null hypothesis of no change) with a P-value of 10%.
- An engineer claims that the proportion of rivets holding the wing on an airplane that are likely to fail is below the proportion at which the wing would fall off. What P-value would be small enough to get you to fly on that plane?

A S ***Activity:* Hypothesis Tests for Proportions.** You've probably noticed that the tools for confidence intervals and for hypothesis tests are similar. See how tests and intervals for proportions are related—and an important way in which they differ.

Your conclusion about any null hypothesis should be accompanied by the P-value of the test. Don't just declare the null hypothesis rejected or not rejected. Report the P-value to show the strength of the evidence against the hypothesis and the effect size. This will let each reader decide whether or not to reject the null hypothesis and whether or not to consider the result important if it is statistically significant.

To complete your analysis, follow your test with a confidence interval for the parameter of interest, to report the size of the effect.

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 $0.40) asking the customer to pay. That worked 30% of the time. They want 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 video costs about $10.00 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. They mail it out to several randomly selected delinquent customers and keep track of how many actually do contact the bank to arrange payments. The bank's statistician calculates a P-value of 0.003. What does this P-value suggest about the DVD?

6. The statistician tells the bank's management that the results are clear and that they should switch to the DVD method. Do you agree? What else might you want to know?

STEP-BY-STEP EXAMPLE — Tests and Intervals

Advances in medical care such as prenatal ultrasound examination now make it possible to determine a child's sex early in a pregnancy. There is a fear that in some cultures some parents may use this technology to select the sex of their children. A study from Punjab, India (E. E. Booth, M. Verma, and R. S. Beri, "Fetal Sex Determination in Infants in Punjab, India: Correlations and Implications," *BMJ* 309 [12 November 1994]: 1259–1261), reports that, in 1993, in one hospital, 56.9% of the 550 live births that year were boys. It's a medical fact that male babies are slightly more common than female babies. The study's authors report a baseline for this region of 51.7% male live births.

Question: Is there evidence that the proportion of male births has changed?

Plan State what we want to know.

Define the variables and discuss the W's.

I want to know whether the proportion of male births has changed from the established baseline of 51.7%. The data are the recorded sexes of the 550 live births from a hospital in Punjab, India, in 1993, collected for a study on fetal sex determination. The parameter of interest, p, is the proportion of male births:

$$H_0: p = 0.517$$
$$H_A: p \neq 0.517$$

Hypotheses The null hypothesis makes the claim of no difference from the baseline.

Before seeing the data, we were interested in any change in male births, so the alternative hypothesis is two-sided.

Model Think about the assumptions and check the appropriate conditions.

For testing proportions, the conditions are the same ones we had for making confidence intervals, except that we check the **Success/Failure Condition** with the *hypothesized* proportions rather than with the *observed* proportions.

- ✓ **Independence Assumption:** There is no reason to think that the sex of one baby can affect the sex of other babies, so births can reasonably be assumed to be independent with regard to the sex of the child.
- ✓ **Randomization Condition:** The 550 live births are not a random sample, so I must be cautious about any general conclusions. I hope that this is a representative year, and I think that the births at this hospital may be typical of this area of India.
- ✓ **10% Condition:** I would like to be able to make statements about births at similar hospitals in India. These 550 births are fewer than 10% of all of those births.
- ✓ **Success/Failure Condition:** Both $np_0 = 550(0.517) = 284.35$ and $nq_0 = 550(0.483) = 265.65$ are greater than 10; I expect the births of at least 10 boys and at least 10 girls, so the sample is large enough.

Specify the sampling distribution model.

Tell what test you plan to use.

The conditions are satisfied, so I can use a Normal model and perform a **one-proportion z-test.**

SHOW

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

We find the *z*-score for the observed proportion to find out how many standard deviations it is from the hypothesized proportion.

Make a picture. Sketch a Normal model centered at $p_0 = 0.517$. Shade the region to the right of the observed proportion, and because this is a two-tail test, also shade the corresponding region in the other tail.

From the *z*-score, we can find the P-value, which tells us the probability of observing a value that extreme (or more). Use technology or a table (see p. 473.).

Because this is a two-tail test, the P-value is the probability of observing an outcome more than 2.44 standard deviations from the mean of a Normal model *in either direction.* We must therefore *double* the probability we find in the upper tail.

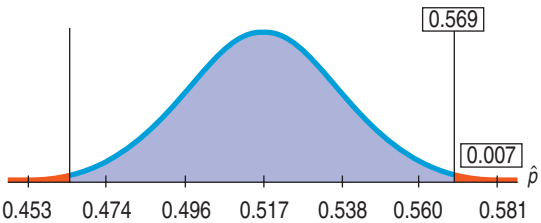
The null model is a Normal distribution with a mean of 0.517 and a standard deviation of

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.517)(1 - 0.517)}{550}} = 0.0213$$

The observed proportion, $\hat{p}$, is 0.569, so

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{0.569 - 0.517}{0.0213} = 2.44$$

The sample proportion lies 2.44 standard deviations above the mean.



$$P = 2P(z > 2.44) = 2(0.0073) = 0.0146$$

TELL

Conclusion State your conclusion in context.

This P-value is roughly 1 time in 70. That's clearly significant, but don't jump to other conclusions. We can't be sure how this deviation came about. For instance, we don't know whether this hospital is typical, or whether the time period studied was selected at random.

The P-value of 0.0146 says that if the true proportion of male babies were still at 51.7%, then an observed proportion as different as 56.9% male babies would occur at random only about 15 times in 1000. With a P-value this small, I reject H_0. This is strong evidence that the birth ratio of boys to girls is not equal to its natural level. It appears that the proportion of boys may have increased.

How big an increase are we talking about? Let's find a confidence interval for the proportion of male births.

THINK AGAIN

Model Check the conditions.

The conditions are identical to those for the hypothesis test, with one difference. Now we are not given a hypothesized proportion, p_0, so we must instead work with the observed proportion $\hat{p}$.

✓ **Success/Failure Condition:** Both $n\hat{p} = 550(0.569) = 313$ and $n\hat{q} = 237$ are at least 10.

Specify the sampling distribution model.

Tell what method you plan to use.

The conditions are satisfied, so I can model the sampling distribution of the proportion with a Normal model and find a ***one-proportion z-interval.***

SHOW MORE

Mechanics We can't find the sampling model standard deviation from the null model proportion. (In fact, we've just rejected it.) Instead, we find the standard error of $\hat{p}$ from the *observed* proportions. Other than that substitution, the calculation looks the same as for the hypothesis test.

With this large a sample size, the difference is negligible, but in smaller samples, it could make a bigger difference.

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.569)(1 - 0.569)}{550}}$$
$$= 0.0211$$

The sampling model is Normal, so for a 95% confidence interval, the critical value $z^* = 1.96$.

The margin of error is

$$ME = z^* \times SE(\hat{p}) = 1.96(0.0211) = 0.041$$

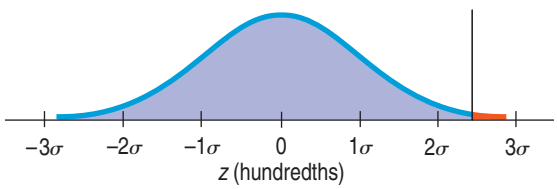
So the 95% confidence interval is

$$0.569 \pm 0.041 \text{ or } (0.528, 0.610).$$

TELL ALL

Conclusion Confidence intervals help us think about the size of the effect. Here we can see that the change from the baseline of 51.7% male births might be quite substantial.

We are 95% confident that the true proportion of male births is between 52.8% and 61.0%.



z	0.00	0.01	0.02	0.03	0.04	0.05
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960

Here's a portion of a Normal table that gives the probability we needed for the hypothesis test. At $z = 2.44$, the table gives the percentile as 0.9927. The upper-tail probability (shaded red) is, therefore, $1 - 0.9927 = 0.0073$; so, for our two-sided test, the P-value is $2(0.0073) = 0.0146$.

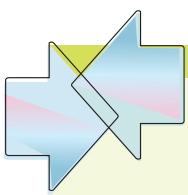
WHAT CAN GO WRONG?

Hypothesis tests are so widely used—and so widely misused—that we've devoted all of the next chapter to discussing the pitfalls involved, but there are a few issues that we can talk about already.

- **Don't base your null hypotheses on what you see in the data.** You are not allowed to look at the data first and then adjust your null hypothesis so that it will be rejected. When your sample value turns out to be $\hat{p} = 51.8\%$, with a standard deviation of 1%, don't form a null hypothesis like $H_0: p = 49.8\%$, knowing that you can reject it. You should always *Think* about the situation you are investigating and make your null hypothesis describe the "nothing interesting" or "nothing has changed" scenario. No peeking at the data!
- **Don't base your alternative hypothesis on the data, either.** Again, you need to *Think* about the situation. Are you interested only in knowing whether something has *increased*? Then write a one-sided (upper-tail) alternative. Or would you be equally interested in a change in either direction? Then you want a two-sided alternative. You should decide whether to do a one- or two-sided test based on what results would be of interest to you, not what you 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 a skeptic 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 recover from biased sampling. The probabilities we compute depend on the independence assumption. And our sample must be large enough to justify our use of a Normal model.
- **Don't accept the null hypothesis.** You may not have found enough evidence to reject it, but you surely have *not* proven it's true!
- **If you fail to reject the null hypothesis, don't think that a bigger sample would be more likely to lead to rejection.** If the results you looked at were "almost" significant, it's enticing to think that because you would have rejected the null had these same observations come from a larger sample, then a larger sample would surely lead to rejection. Don't be misled. Remember, each sample is different, and a larger sample won't necessarily duplicate your current observations. Indeed, the Central Limit Theorem tells us that statistics will vary *less* in larger samples. We should therefore expect such results to be less extreme. Maybe they'd be statistically significant but maybe (perhaps even probably) not. Even if you fail to reject the null hypothesis, it's a good idea to examine a confidence interval. If none of the plausible parameter values in the interval would matter to you (for example, because none would be *practically* significant), then even a larger study with a correspondingly smaller standard error is unlikely to be worthwhile.

Don't We Want to Reject the Null?

Often the folks who collect the data or perform the experiment hope to reject the null. (They hope the new drug is better than the placebo, or new ad campaign is better than the old one.) But when we practice Statistics, we can't allow that hope to affect our decision. The essential attitude for a hypothesis tester is skepticism. Until we become convinced otherwise, we cling to the null's assertion that there's nothing unusual, 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, but you are not the prosecutor.



CONNECTIONS

Hypothesis tests and confidence intervals share many of the same concepts. Both rely on sampling distribution models, and because the models are the same and require the same assumptions, both check the same conditions. They also calculate many of the same statistics. Like confidence intervals, hypothesis tests use the standard deviation of the sampling distribution as a ruler, as we first saw in Chapter 6.

For testing, we find ourselves looking once again at *z*-scores, and we compute the P-value by finding the distance of our test statistic from the center of the null model. P-values are conditional probabilities. They give the probability of observing the result we have seen (or one even more extreme) *given* that the null hypothesis is true.

The Standard Normal model is here again as our connection between *z*-score values and probabilities.

WHAT HAVE WE LEARNED?

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, then seeing whether the data we observe are consistent with that model or are 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.

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.
- If the data are out of line with the null hypothesis model, the P-value will be small and we will reject the null hypothesis.
- If the data are consistent with the null hypothesis model, the P-value will be large and we will not reject the null hypothesis.
- We must always state our conclusion in the context of the original question.

And we've learned that confidence intervals and hypothesis tests go hand in hand in helping us think about models. A hypothesis test makes a yes/no decision about the plausibility of a parameter value. The confidence interval shows us the range of plausible values for the parameter.

Terms

Null hypothesis — 460. The claim being assessed in a hypothesis test is called the null hypothesis. 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.

Alternative hypothesis — 460. The alternative hypothesis proposes what we should conclude if we find the null hypothesis to be unlikely.

Two-sided alternative (Two-tailed alternative) — 466. An alternative hypothesis is two-sided ($H_A: p \neq p_0$) when we are interested in deviations in *either* direction away from the hypothesized parameter value.

One-sided alternative (One-tailed alternative) — 466. An alternative hypothesis is one-sided (e.g., $H_A: p > p_0$ or $H_A: p < p_0$) when we are interested in deviations in *only one* direction away from the hypothesized parameter value.

P-value — 461. 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 either that the observation is improbable or that the probability calculation was based on incorrect assumptions. The assumed truth of the null hypothesis is the assumption under suspicion.

One-proportion *z*-test — 464. A test of the null hypothesis that the proportion of a single sample equals a specified value ($H_0: p = p_0$) by referring the statistic $z = \dfrac{\hat{p} - p_0}{SD(\hat{p})}$ to a Standard Normal model.

Skills

THINK

- Be able to state the null and alternative hypotheses for a one-proportion z-test.
- Know the conditions that must be true for a one-proportion z-test to be appropriate, and know how to 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.

SHOW

- Be able to perform a one-proportion z-test.

TELL

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