CHAPTER 21

More About Tests and Intervals

WHO	Florida motorcycle riders aged 20 and younger involved in motorcycle accidents
WHAT	% wearing helmets
WHEN	2001–2003
WHERE	Florida
WHY	Assessment of injury rates commissioned by the National Highway Traffic Safety Administration (NHTSA)

In 2000 Florida changed its motorcycle helmet law. No longer are riders 21 and older required to wear helmets. Under the new law, those under 21 still must wear helmets, but a report by the Preusser Group (www.preussergroup. com) suggests that helmet use may have declined in this group, too.

It isn't practical to survey young motorcycle riders. (For example, how can you construct a sampling frame? If you contacted licensed riders, would they admit to riding illegally without a helmet?) The researchers adopted a different strategy. Police reports of motorcycle accidents record whether the rider wore a helmet and give the rider's age. Before the change in the helmet law, 60% of youths involved in a motorcycle accident had been wearing their helmets. The Preusser study looked at accident reports during 2001–2003, the three years following the law change, considering these riders to be a representative sample of the larger population. They observed 781 young riders who were involved in accidents. Of these, 396 (or 50.7%) were wearing helmets. Is this evidence of a decline in helmet-wearing, or just the natural fluctuation of such statistics?

Zero In on the Null

Null hypotheses have special requirements. In order to perform a statistical test of the hypothesis, the null must be a statement about the value of a parameter for a model. We use this value to compute the probability that the observed sample statistic—or something even farther from the null value—might occur.

How do we choose the null hypothesis? The appropriate null arises directly from the context of the problem. It is dictated, not by the data, but by the situation. One good way to identify both the null and alternative hypotheses is to think about the *Why* of the situation. Typical null hypotheses might be that the proportion of patients recovering after receiving a new drug is the same as we would expect of patients receiving a placebo or that the mean strength attained by athletes training with new equipment is the same as with the old equipment. The alternative hypotheses would be that the new drug cures a higher proportion of patients or that the new equipment results in a greater mean strength.

To write a null hypothesis, you can't just choose any parameter value you like. The null must relate to the question at hand. Even though the null usually means no difference or no change, you can't automatically interpret "null" to mean zero. A claim that "nobody" wears a motorcycle helmet would be absurd. The null hypothesis for the Florida study could be that the true rate of helmet use remained the same among young riders after the law changed. You need to find the value for the parameter in the null hypothesis from the context of the problem.

There is a temptation to state your *claim* as the null hypothesis. As we have seen, however, you cannot prove a null hypothesis true any more than you can prove a defendant innocent. So, it makes more sense to use what you want to show as the *alternative*. This way, if you reject the null, you are left with what you want to show.

FOR EXAMPLE Writing hypotheses

The diabetes drug Avandia® was approved to treat Type 2 diabetes in 1999. But in 2007 an article in the *New England Journal of Medicine* (*NEJM*)[1] raised concerns that the drug might carry an increased risk of heart attack. This study combined results from a number of other separate studies to obtain an overall sample of 4485 diabetes patients taking Avandia. People with Type 2 diabetes are known to have about a 20.2% chance of suffering a heart attack within a seven-year period. According to the article's author, Dr. Steven E. Nissen,[2] the risk found in the *NEJM* study was equivalent to a 28.9% chance of heart attack over seven years. The FDA is the government agency responsible for relabeling Avandia to warn of the risk if it is judged to be unsafe. Although the statistical methods they use are more sophisticated, we can get an idea of their reasoning with the tools we have learned.

Question: What null hypothesis and alternative hypothesis about seven-year heart attack risk would you test? Explain.

$$H_0: p = 0.202$$
$$H_A: p > 0.202$$

The parameter of interest is the proportion of diabetes patients suffering a heart attack in seven years. The FDA is concerned only with whether Avandia *increases* the seven-year risk of heart attacks above the baseline value of 20.2%, so a one-sided upper-tail test is appropriate.

One-sided or two? In the 1930s, a series of experiments was performed at Duke University in an attempt to see whether humans were capable of extrasensory perception, or ESP. Psychologist Karl Zener designed a set of cards with 5 symbols, later made infamous in the movie *Ghostbusters*:

In the experiment, the "sender" selects one of the 5 cards at random from a deck and then concentrates on it. The "receiver" tries to determine which card it is. If we let p be the proportion of correct responses, what's the null hypothesis? The null hypothesis is that ESP makes no difference. Without ESP, the receiver would just be guessing, and since there are 5 possible responses, there would be a 20% chance of guessing each card correctly. So, H_0 is $p = 0.20$. What's the alternative? It seems that it should be $p > 0.20$, a one-sided alternative. But some ESP researchers have expressed the claim that if the proportion guessed were much *lower* than expected, that would show an "interference" and should be considered evidence for ESP as well. So they argue for a two-sided alternative.

[1] Steven E. Nissen, M.D., and Kathy Wolski, M.P.H., "Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes," *NEJM* 2007; 356.
[2] Interview reported in the *New York Times* [May 26, 2007].

STEP-BY-STEP EXAMPLE Another One-Proportion *z*-Test

Let's try to answer the question raised at the start of the chapter.

Question: Has helmet use in Florida declined among riders under the age of 21 subsequent to the change in the helmet laws?

THINK

Plan State the problem and discuss the variables and the W's.

I want to know whether the rate of helmet wearing among Florida's motorcycle riders under the age of 21 remained at 60% after the law changed to allow older riders to go without helmets. I have data from accident records showing 396 of 781 young riders were wearing helmets.

Hypotheses The null hypothesis is established by the rate set before the change in the law. The study was concerned with safety, so they'll want to know of any decline in helmet use, making this a lower-tail test.

$$H_O: p = 0.60$$
$$H_A: p < 0.60$$

SHOW

Model Check the conditions.

The Risky Behavior Surveillance survey is in fact a complex, multistage sample, but it is randomized and great effort is taken to make it representative. It is safe to treat it as though it were a random sample.

✓ **Independence Assumption:** The data are for riders involved in accidents during a three-year period. Individuals are independent of one another.

✗ **Randomization Condition:** No randomization was applied, but we are considering these riders involved in accidents to be a representative sample of all riders. We should take care in generalizing our conclusions.

✓ **10% Condition:** These 781 riders are a small sample of a larger population of all young motorcycle riders.

✓ **Success/Failure Condition:** We'd expect $np = 781(0.6) = 468.6$ helmeted riders and $nq = 781(0.4) = 312.4$ non-helmeted. Both are at least 10.

Specify the sampling distribution model and name the test.

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

SHOW

Mechanics Find the standard deviation of the sampling model using the hypothesized proportion.

There were 396 helmet wearers among the 781 accident victims.

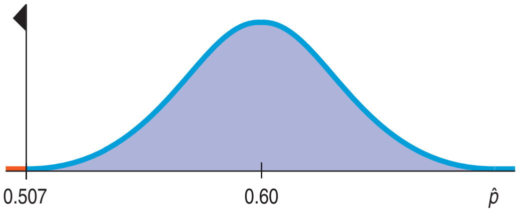
$$\hat{p} = \frac{396}{781} = 0.507$$

$$SD(\hat{p}) = \sqrt{\frac{p_0 q_0}{n}} = \sqrt{\frac{(0.60)(0.40)}{781}} = 0.0175$$

Find the *z*-score for the observed proportion.

$$z = \frac{\hat{p} - p_0}{SD(\hat{p})} = \frac{0.507 - 0.60}{0.0175} = -5.31$$

Make a picture. Sketch a Normal model centered at the hypothesized helmet rate of 60%. This is a lower-tail test, so shade the region to the left of the observed rate.



Given this *z*-score, the P-value is obviously very low.

The observed helmet rate is 5.31 standard deviations below the former rate. The corresponding P-value is less than 0.001.

TELL

Conclusion Link the P-value to your decision about the null hypothesis, and then state your conclusion in context.

The very small P-value says that if the true rate of helmet-wearing among riders under 21 were still 60%, the probability of observing a rate no higher than 50.7% in a sample like this is less than 1 chance in 1000, so I reject the null hypothesis. There is strong evidence that there has been a decline in helmet use among riders under 21.

How to Think About P-values

A P-value actually is a conditional probability. It tells us the probability of getting results at least as unusual as the observed statistic, *given* that the null hypothesis is true. We can write P-value $= P(\text{observed statistic value [or even more extreme]} \mid H_0)$.

Writing the P-value this way helps to make clear that the P-value is *not* the probability that the null hypothesis is true. It is a probability about the data. Let's say that again:

The P-value is not the probability that the null hypothesis is true.

The P-value is not even the conditional probability that the null hypothesis is true given the data. We would write that probability as $P(H_0 \mid \text{observed statistic value})$. This is a conditional probability but in reverse. It would be nice to know this, but it's impossible to calculate without making additional assumptions. As we saw in Chapter 15, reversing the order in a conditional probability is difficult, and the results can be counterintuitive.

We can find the P-value, $P(\text{observed statistic value} \mid H_0)$, because H_0 gives the parameter values that we need to find the required probability. But there's no direct way to find $P(H_0 \mid \text{observed statistic value})$.[3] As tempting as it may be to say that a P-value of 0.03 means there's a 3% chance that the null hypothesis is true, that just isn't right. All we can say is that, given the null hypothesis, there's a 3% chance of observing the statistic value that we have actually observed (or one more unlike the null value).

Which Conditional?

Suppose that as a political science major you are offered the chance to be a White House intern. There would be a very high probability that next summer you'd be in Washington, D.C. That is, $P(\text{Washington} \mid \text{Intern})$ would be high. But if we find a student in Washington, D.C., is it likely that he's a White House intern? Almost surely not; $P(\text{Intern} \mid \text{Washington})$ is low. You can't switch around conditional probabilities. The P-value is $P(\text{data} \mid H_0)$. We might wish we could report $P(H_0 \mid \text{data})$, but these two quantities are NOT the same.

[3] The approach to statistical inference known as Bayesian Statistics addresses the question in just this way, but it requires more advanced mathematics and more assumptions. See p. 358 for more about the founding father of this approach.

"The wise man proportions his belief to the evidence."

—David Hume,
"Enquiry Concerning Human
Understanding," 1748

How guilty is the suspect? We might like to know $P(\mathrm{H}_0 \mid \mathrm{data})$, but when you think about it, we can't talk about the probability that the null hypothesis is true. The null is not a random event, so either it is true or it isn't. The data, however, are random in the sense that if we were to repeat a randomized experiment or draw another random sample, we'd get different data and expect to find a different statistic value. So we can talk about the probability of the data given the null hypothesis, and that's the P-value.

But it does make sense that the smaller the P-value, the more confident we can be in declaring that we doubt the null hypothesis. Think again about the jury trial. Our null hypothesis is that the defendant is innocent. Then the evidence starts rolling in. A car the same color as his was parked in front of the bank. Well, there are lots of cars that color. The probability of that happening (given his innocence) is pretty high, so we're not persuaded that he's guilty. The bank's security camera showed the robber was male and about the dependant's height and weight. Hmmm. Could that be a coincidence? If he's innocent, then it's a little less likely that the car and description would *both* match, so our P-value goes down. We're starting to question his innocence a little. Witnesses said the robber wore a blue jacket just like the one the police found in a garbage can behind the defendant's house. Well, if he's innocent, then that doesn't seem very likely, does it? If he's really innocent, the probability that all of these could have happened is getting pretty low. Now our P-value may be small enough to be called "beyond a reasonable doubt" and lead to a conviction. Each new piece of evidence strains our skepticism a bit more. The more compelling the evidence—the more *unlikely* it would be were he innocent—the more convinced we become that he's guilty.

"You're so guilty now."

—Rearview Mirror

But even though it may make *us* more confident in declaring him guilty, additional evidence does not make *him* any guiltier. Either he robbed the bank or he didn't. Additional evidence (like the teller picking him out of a police lineup) just makes us more confident that we did the right thing when we convicted him. The lower the P-value, the more comfortable we feel about our decision to reject the null hypothesis, but the null hypothesis doesn't get any more false.

FOR EXAMPLE Thinking about the P-value

Recap: A *New England Journal of Medicine* paper reported that the seven-year risk of heart attack in diabetes patients taking the drug Avandia was increased from the baseline of 20.2% to an estimated risk of 28.9% and said the P-value was 0.03.

Question: How should the P-value be interpreted in this context?

The P-value $= P(\hat{p} \geq 28.9\% \mid p = 20.2\%)$. That is, it's the probability of seeing such a high heart attack rate among the people studied if, in fact, taking Avandia really didn't increase the risk at all.

What to Do with a High P-value

Therapeutic touch (TT), taught in many schools of nursing, is a therapy in which the practitioner moves her hands near, but does not touch, a patient in an attempt to manipulate a "human energy field." Therapeutic touch practitioners believe that by adjusting this field they can promote healing. However, no instrument has ever detected a human energy field, and no experiment has ever shown that TT practitioners can detect such a field.

In 1998, the *Journal of the American Medical Association* published a paper reporting work by a then nine-year-old girl.[4] She had performed a simple experiment in

[4] L. Rosa, E. Rosa, L. Sarner, and S. Barrett, "A Close Look at Therapeutic Touch," *JAMA* 279(13) [1 April 1998]: 1005–1010.

A S ***Video:*** **Is There Evidence for Therapeutic Touch?** This video shows the experiment and tells the story.

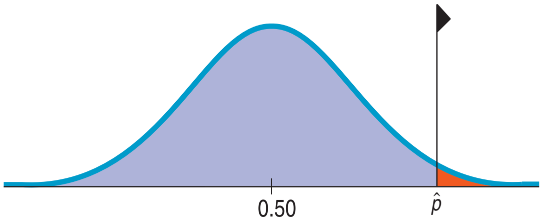
which she challenged 15 TT practitioners to detect whether her unseen hand was hovering over their left or right hand (selected by the flip of a coin).

The practitioners "warmed up" with a period during which they could see the experimenter's hand, and each said that they could detect the girl's human energy field. Then a screen was placed so that the practitioners could not see the girl's hand, and they attempted 10 trials each. Overall, of 150 trials, the TT practitioners were successful 70 times, for a success proportion of 46.7%. Is there evidence from this experiment that TT practitioners can successfully detect a "human energy field"?

A S ***Activity:*** **Testing Therapeutic Touch.** Perform the one-proportion *z*-test using *ActivStats* technology. The test in *ActivStats* is two-sided. Do you think this is the appropriate choice?

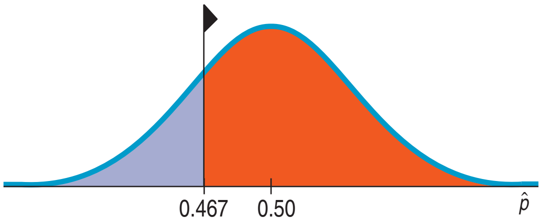
When we see a small P-value, we could continue to believe the null hypothesis and conclude that we just witnessed a rare event. But instead, we trust the data and use it as evidence to reject the null hypothesis.

In the therapeutic touch example, the null hypothesis is that the practitioners are guessing, so we expect them to be right about half the time by chance. That's why we say H_0: $p = 0.5$. They claim that they can detect a "human energy field" and that their success rate should be well above chance, so our alternative is that they would do *better* than guessing. That's a one-sided alternative hypothesis: H_A: $p > 0.5$. With a one-sided hypothesis, our P-value is the probability the practitioners could achieve the observed number of successes or *more* even if they were just guessing.



If the practitioners had been highly successful, that would have been unusually lucky for guessing, so we would have seen a correspondingly low P-value. Since we don't believe in rare events, we would then have concluded that they weren't guessing.

But that's not what happened. What we actually observed was that they did slightly *worse* than 50%, with a $\hat{p} = 0.467$ success rate.



As the figure shows, the probability of a success rate of 0.467 *or more* is even bigger than 0.5. In this case, it turns out to be 0.793. Obviously, we won't be rejecting the null hypothesis; for us to reject it, the P-value would have to be quite small. But a P-value of 0.788 seems so big it is almost awkward. With a success rate even lower than chance, we could have concluded right away that we have no evidence for rejecting H_0.

Big P-values just mean that what we've observed isn't surprising. That is, the results are in line with our assumption that the null hypothesis models the world, so we have no reason to reject it. A big P-value doesn't prove that the null hypothesis is true, but it certainly offers no evidence that it's *not* true. When we see a large P-value, all we can say is that we "don't reject the null hypothesis."

FOR EXAMPLE More about P-values

Recap: The question of whether the diabetes drug Avandia increased the risk of heart attack was raised by a study in the *New England Journal of Medicine*. This study estimated the seven-year risk of heart attack to be 28.9% and reported a P-value of 0.03 for a test of whether this risk was higher than the baseline seven-year risk of 20.2%. An earlier study (the ADOPT study) had estimated the seven-year risk to be 26.9% and reported a P-value of 0.27.

Question: Why did the researchers in the ADOPT study not express alarm about the increased risk they had seen?

A P-value of 0.27 means that a heart attack rate at least as high as the one they observed could be expected in 27% of similar experiments even if, in fact, there were no increased risk from taking Avandia. That's not remarkable enough to reject the null hypothesis. In other words, the ADOPT study wasn't convincing.

Alpha Levels

A S ***Activity:* Rejecting the Null Hypothesis.** See alpha levels at work in the animated hypothesis-testing tool.

NOTATION ALERT:

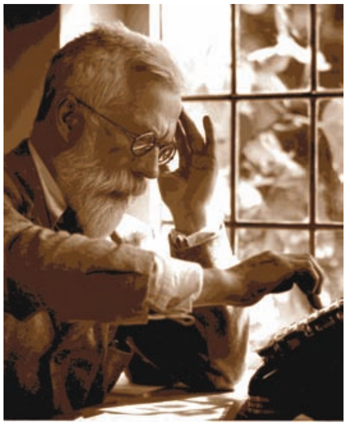
The first Greek letter, α, is used in Statistics for the threshold value of a hypothesis test. You'll hear it referred to as the alpha level. Common values are 0.10, 0.05, 0.01, and 0.001.

Sometimes we need to make a firm decision about whether or not 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 Statistics to enroll in.

When the P-value is small, it tells us that our data are rare, *given the null hypothesis*. As humans, we are 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"?

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 labeled 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. Note, however: You must select the alpha level *before* you look at the data. Otherwise you can be accused of cheating by tuning your alpha level to suit the data.



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

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. Since then, some people—indeed some entire disciplines—have treated the number 0.05 as sacrosanct.

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 "at the 5% level of significance."

What can you say if the P-value does not fall below α?

When you have not found sufficient evidence to reject the null according to the standard you have established, you should say that "The data have failed to provide sufficient evidence to reject the null hypothesis." Don't say that you "accept the null hypothesis." You certainly haven't proven or established it; it was merely assumed to begin with. Say that you've failed to reject it.

Think again about the therapeutic touch example. The P-value was 0.788. This is so much larger than any reasonable alpha level that we can't reject H_0. For this test, we'd conclude, "We fail to reject the null hypothesis. There is insufficient evidence to conclude that the practitioners are performing better than they would if they were just guessing."

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 base a decision on an arbitrary alpha level.

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. This is the rare event we want to protect ourselves against. 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!'*"

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 P-value between 0.01 and 0.001 gets **, and a P-value less than 0.001 gets ***. This can be a convenient summary of the weight of evidence against the null hypothesis if it's not taken too literally. But we warn you against taking the distinctions too seriously and against making a black-and-white decision 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 very 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.

When you decide to declare a verdict, it's always a good idea to report the P-value as an indication of the strength of the evidence. 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 these cases, the P-value is the best summary we have of what the data say or fail to say about the null hypothesis.

Significant vs. Important

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 carries with it any sense of 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 is not large enough, even large financially or scientifically "significant" 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 we base 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 2001 and 2002. The difference in the mean values was \$9.83. 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, even a clinically important improvement of 10% in cure rate with a new treatment is not likely to be statistically significant in a study of fewer than 225 patients. A small clinical trial would probably not be conclusive.

Confidence Intervals and Hypothesis Tests

For the motorcycle helmet example, a 95% confidence interval would give $0.507 \pm 1.96 \times 0.0179 = (0.472, 0.542)$, or 47.2% to 54.2%. If the previous rate of helmet compliance had been, say, 50%, we would not have been able to reject the null hypothesis because 50% is in the interval, so it's a plausible value. Indeed, *any* hypothesized value for the true proportion of helmet wearers in this interval is consistent with the data. Any value outside the confidence interval would make a null hypothesis that we would reject, but we'd feel more strongly about values far outside the interval.

Confidence intervals and hypothesis tests are built from the same calculations.[5] They have the same assumptions and conditions. As we have just seen, you can

[5] As we saw in Chapter 20, this is not *exactly* true for proportions. For a confidence interval, we estimate the standard deviation of $\hat{p}$ from $\hat{p}$ itself. Because we estimate it from the data, we have a *standard error*. For the corresponding hypothesis test, we use the model's standard deviation for $\hat{p}$, based on the null hypothesis value p_0. 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.

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 is a little more complicated. For a one-sided test with $\alpha = 5\%$, the corresponding confidence interval has a confidence level of 90%—that's 5% in each tail. In general, a confidence interval with a confidence level of $C\%$ corresponds to a one-sided hypothesis test with an α level of $\frac{1}{2}(100 - C)\%$.

FOR EXAMPLE Making a decision based on a confidence interval

Recap: The baseline seven-year risk of heart attacks for diabetics is 20.2%. In 2007 a *NEJM* study reported a 95% confidence interval equivalent to 20.8% to 40.0% for the risk among patients taking the diabetes drug Avandia.

Question: What did this confidence interval suggest to the FDA about the safety of the drug?

The FDA could be 95% confident that the interval from 20.8% to 40.0% included the true risk of heart attack for diabetes patients taking Avandia. Because the lower limit of this interval was higher than the baseline risk of 20.2%, there was evidence of an increased risk.

JUST CHECKING

1. An experiment to test the fairness of a roulette wheel gives a z-score of 0.62. What would you conclude?
2. In the last chapter we encountered a bank that wondered if it could get more customers to make payments on delinquent balances by sending them a DVD urging them to set up a payment plan. Well, the bank just got back the results on their test of this strategy. A 90% confidence interval for the success rate is (0.29, 0.45). Their old send-a-letter method had worked 30% of the time. Can you reject the null hypothesis that the proportion is still 30% at $\alpha = 0.05$? Explain.
3. Given the confidence interval the bank found in their trial of DVDs, what would you recommend that they do? Should they scrap the DVD strategy?

STEP-BY-STEP EXAMPLE Wear that Seatbelt!

Teens are at the greatest risk of being killed or injured in traffic crashes. According to the National Highway Traffic Safety Administration, 65% of young people killed were not wearing a safety belt. In 2001, a total of 3322 teens were killed in motor vehicle crashes, an average of 9 teenagers a day. Because many of these deaths could easily be prevented by the use of safety belts, several states have begun "Click It or Ticket" campaigns in which increased enforcement and publicity have resulted in significantly higher seatbelt use. Overall use in Massachusetts quickly increased from 51% in 2002 to 64.8% in 2006, with a goal of surpassing the national average of 82%. Recently, a local newspaper reported that a roadblock resulted in 23 tickets to drivers who were unbelted out of 134 stopped for inspection.

Question: Does this provide evidence that the goal of over 82% compliance was met?

Let's use a confidence interval to test this hypothesis.

THINK

Plan State the problem and discuss the variables and the W's.

The data come from a local newspaper report that tells the number of tickets issued and number of drivers stopped at a recent roadblock. I want to know whether the rate of compliance with the seatbelt law is greater than 82%.

Hypotheses The null hypothesis is that the compliance rate is only 82%. The alternative is that it is now 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.

$$H_0: p = 0.82$$
$$H_A: p > 0.82$$

Model Think about the assumptions and check the conditions.

We are finding a confidence interval, so we work from the data rather than the null model.

- ✓ **Independence Assumption:** Drivers are not likely to influence one another when it comes to wearing a seatbelt.
- ✓ **Randomization Condition:** This wasn't a random sample, but I assume these drivers are representative of the driving public.
- ✓ **10% Condition:** The police stopped fewer than 10% of all drivers.
- ✓ **Success/Failure Condition:** There were 111 successes and 23 failures, both at least 10. The sample is large enough.

State your method.

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

SHOW

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

$n = 134$, so

$$\hat{p} = \frac{111}{134} = 0.828 \text{ and}$$

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.

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{(0.828)(0.172)}{134}} = 0.033$$

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 $z^* = 1.645$.

$$ME = z^* \times SE(\hat{p})$$
$$= 1.645(0.033) = 0.054$$

The confidence interval is

estimate ± margin of error.

The 90% confidence interval is

0.828 ± 0.054 or

$(0.774, 0.882)$.

TELL

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

I am 90% confident that between 77.4% and 88.2% of all drivers wear their seatbelts. Because the hypothesized rate of 82% is within this interval, I do not reject the null hypothesis. There is insufficient evidence to conclude that the campaign was truly effective and now more than 82% of all drivers are wearing seatbelts.

The upper limit of the confidence interval shows it's possible that the campaign is quite successful, but the small sample size makes the interval too wide to be very specific.

*A 95% Confidence Interval for Small Samples

When the **Success/Failure Condition** fails, all is not lost. A simple adjustment to the calculation lets us make a 95% confidence interval anyway.

All we do is add four *phony* observations—two to the successes, two to the failures. So instead of the proportion $\hat{p} = \frac{y}{n}$, we use the adjusted proportion $\tilde{p} = \frac{y+2}{n+4}$ and, for convenience, we write $\tilde{n} = n + 4$. We modify the interval by using these adjusted values for both the center of the interval *and* the margin of error. Now the adjusted interval is

$$\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}.$$

This adjusted form gives better performance overall[6] and works much better for proportions near 0 or 1. It has the additional advantage that we no longer need to check the **Success/Failure Condition** that $n\hat{p}$ and $n\hat{q}$ are greater than 10.

FOR EXAMPLE An Agresti-Coull "plus-four" interval

Surgeons examined their results to compare two methods for a surgical procedure used to alleviate pain on the outside of the wrist. A new method was compared with the traditional "freehand" method for the procedure. Of 45 operations using the "freehand" method, three were unsuccessful, for a failure rate of 6.7%. With only 3 failures, the data don't satisfy the **Success/Failure Condition,** so we can't use a standard confidence interval.

Question: What's the confidence interval using the "plus-four" method?

[6] By "better performance," we mean that a 95% confidence interval has more nearly a 95% chance of covering the true population proportion. Simulation studies have shown that our original, simpler confidence interval in fact is less likely than 95% to cover the true population proportion when the sample size is small or the proportion very close to 0 or 1. The original idea for this method can be attributed to E. B. Wilson. The simpler approach discussed here was proposed by Agresti and Coull (A. Agresti and B. A. Coull, "Approximate Is Better Than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician,* 52[1998]: 119–129).

There were 42 successes and 3 failures. Adding 2 “pseudo-successes” and 2 “pseudo-failures,” we find

$$\tilde{p} = \frac{3 + 2}{45 + 4} = 0.102$$

A 95% confidence interval is then

$$0.102 \pm 1.96\sqrt{\frac{0.102(1 - 0.102)}{49}} = 0.102 \pm 0.085 \text{ or } (0.017, 0.187).$$

Notice that although the observed failure rate of 0.067 is contained in the interval, it is not at the center of the interval—something we haven’t seen with any of the other confidence intervals we’ve considered.

Making Errors

***Activity:* Type I and Type II Errors.** View an animated exploration of Type I and Type II errors—a good backup for the reading in this section.

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 two types of errors are known as **Type I** and **Type II errors.** 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.

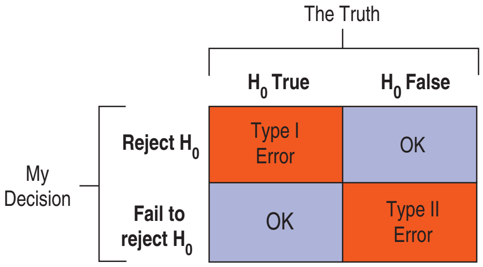
Some false-positive results mean no more than an unnecessary chest X-ray. But for a drug test or a disease like AIDS, a false-positive result that is not kept confidential could have serious consequences.

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 infected 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 the 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 the person must undergo further tests. In a Statistics final exam (with H_0: the student has learned only 60% of the material), a Type I error would be passing a student who in fact learned less than 60% of the material, while a Type II error would be failing a student who knew enough to pass. Which of these errors seems more serious? It depends on the situation, the cost, and your point of view.

Here’s an illustration of the situations:

***Activity:* Hypothesis Tests Are Random.** Simulate hypothesis tests and watch Type I errors occur. When you conduct real hypothesis tests you’ll never know, but simulation can tell you when you’ve made an error.



		The Truth	
		H_0 True	H_0 False
My Decision	Reject H_0	Type I Error	OK
	Fail to reject H_0	OK	Type II Error

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 have done the right thing. A test's ability to detect a false null hypothesis is called the **power** of the test. In a jury trial, power is the ability of the criminal justice system to convict people who are guilty—a good thing! We'll have a lot more to say about power soon.

When H_0 is false but we fail to reject it, we have 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 which one should we choose?

One way to focus our attention is by thinking about the *effect size.* That is, we ask *"How big a difference would matter?"* Suppose a charity wants to test whether placing personalized address labels in the 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%, they would calculate β for the alternative $p = 0.06$.

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. So we'd reduce β, the chance that we fail to reject a false null—but we'd make more Type I errors. This tension between Type I and Type II errors is inevitable. In the political arena, think of the ongoing debate between those who favor provisions to reduce Type I errors in the courts (supporting Miranda rights, requiring warrants for wiretaps, providing legal representation for those who can't afford it) and those who advocate changes to reduce Type II errors (admitting into evidence confessions made when no lawyer is present, eavesdropping on conferences with lawyers, restricting paths of appeal, etc.).

The only way to reduce *both* types of error is to collect more evidence or, in statistical terms, to collect more data. Too often, studies fail because their sample sizes are too small to detect the change they are looking for.

NOTATION ALERT:

In Statistics, α is almost always saved for the alpha level. But β has already been used for the parameters of a linear model. Fortunately, it's usually clear whether we're talking about a Type II error probability or the slope or intercept of a regression model.

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 have 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.

FOR EXAMPLE Thinking about errors

Recap: A published study found the risk of heart attack to be increased in patients taking the diabetes drug Avandia. The issue of the *New England Journal of Medicine* (*NEJM*) in which that study appeared also included an editorial that said, in part, "A few events either way might have changed the findings for myocardial infarction[7] or for death from cardiovascular causes. In this setting, the possibility that the findings were due to chance cannot be excluded."

Question: What kind of error would the researchers have made if, in fact, their findings were due to chance? What could be the consequences of this error?

The null hypothesis said the risk didn't change, but the researchers rejected that model and claimed evidence of a higher risk. If these findings were just due to chance, they rejected a true null hypothesis—a Type I error.

If, in fact, Avandia carried no extra risk, then patients might be deprived of its benefits for no good reason.

[7] Doctorese for "heart attack."

Power

When we failed to reject the null hypothesis about TT practitioners, did we prove that they were just guessing? No, it could be that they actually *can* discern a human energy field but we just couldn't tell. For example, suppose they really have the ability to get 53% of the trials right but just happened to get only 47% in our experiment. Our confidence interval shows that with these data we wouldn't have rejected the null. And if we retained the null even though the true proportion was actually greater than 50%, we would have made a Type II error because we failed to detect their ability.

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 probability that it *does* reject: $1 - \beta$.

Whenever a study fails to reject its null hypothesis, the test's power comes into question. Was the 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? The therapeutic touch experiment failed to reject the null hypothesis that the TT practitioners were just guessing. Might the problem be that the experiment simply lacked adequate power to detect their ability?

FOR EXAMPLE Errors and power

Recap: The study of Avandia published in the *NEJM* combined results from 47 different trials—a method called *meta-analysis*. The drug's manufacturer, GlaxoSmithKline (GSK), issued a statement that pointed out, "Each study is designed differently and looks at unique questions: For example, individual studies vary in size and length, in the type of patients who participated, and in the outcomes they investigate." Nevertheless, by combining data from many studies, meta-analyses can achieve a much larger sample size.

Question: How could this larger sample size help?

If Avandia really did increase the seven-year heart attack rate, doctors needed to know. To overlook that would have been a Type II error (failing to detect a false null hypothesis), resulting in patients being put at greater risk. Increasing the sample size could increase the power of the analysis, making it more likely that researchers will detect the danger if there is one.

A S *Activity:* **The Power of a Test.** Power is a concept that's much easier to understand when you can visualize what's happening.

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 farther p_0 is from p, the greater the power. If the therapeutic touch practitioners were in fact able to detect human energy fields 90% of the time, it should be easy to see that they aren't guessing. With an effect size this large, we'd have a powerful test. If their true success rate were only 53%, however, we'd need a larger sample size to have a good chance of noticing that (and rejecting H_0).

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. For example, do you think that health insurance companies should pay for therapeutic touch if practitioners could detect a human energy field only 53% of the time—just slightly better than chance? That doesn't seem clinically useful.[8] How about 75% of the time? No therapy works all the time, and insurers might be quite willing to pay for such a success rate. Let's take 75% as a reasonably interesting effect size (keeping in mind that 50% is the level of guessing). With 150 trials, the TT experiment would have been able to detect such an ability with a power of 99.99%. So power was not an issue in this study. There is only a very small chance that the study would have failed to detect a practitioner's ability, had it existed. The sample size was clearly big enough.

JUST CHECKING

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

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

6. For the bank, which situation has higher power: a strategy that works really well, actually getting 60% of people to pay off their balances, or a strategy that barely increases the payoff rate to 32%? Explain briefly.

A Picture Worth $\dfrac{1}{P(z > 3.09)}$ Words

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 will be to make a Type II error.

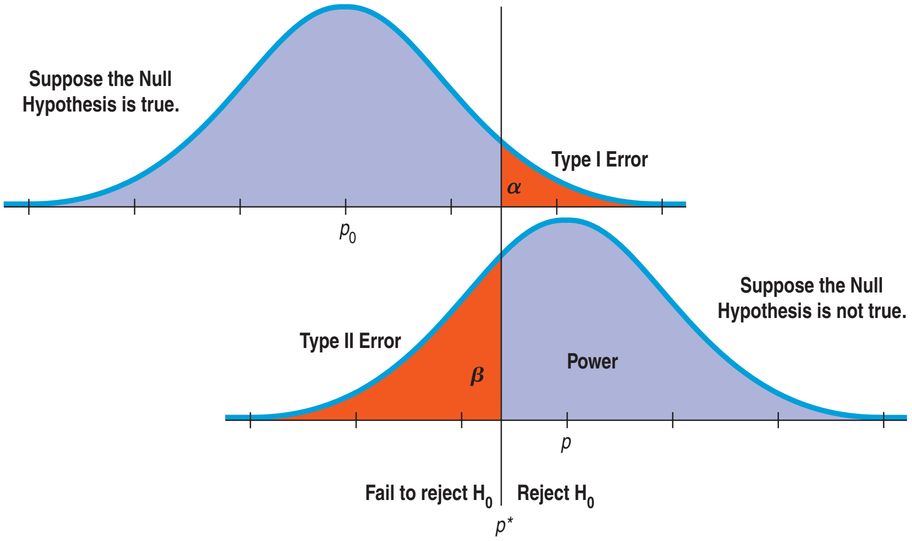


FIGURE 21.1

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 of $\hat{p}$ 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 make the correct decision to reject the null hypothesis. The power of the test is the purple 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.

[8] On the other hand, a scientist might be interested in anything clearly different from the 50% guessing rate because that might suggest an entirely new physics at work. In fact, it could lead to a Nobel prize.

NOTATION ALERT:

We've attached symbols to many of the p's. Let's keep them straight. p is a true proportion parameter. 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.

Figure 21.1 shows a good way to visualize the relationships among these concepts. Suppose we are 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). For example, we might be willing to believe the ability of therapeutic touch practitioners if they were successful in 65% of our trials. This is what the upper model shows. It's a picture of the sampling distribution model for the proportion if the null hypothesis were true. We'd make a Type I error whenever the sample gave us $\hat{p} > p^*$, because we would reject the (true) null hypothesis. And 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. (Perhaps the TT practitioner really can detect the human energy field 72% of the time.) 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. Suppose a TT practitioner with a true ability level of 72% is actually successful on fewer than 65% of our tests. 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 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, no matter whether the practitioners can detect a human energy field 90%, 53%, or 2% of the time, 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 correctly reject H_0 when it's *false* depends on the effect size. We can see from the picture that if the effect size were larger (the true proportion were farther above 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 $= 1 - \beta$.
- Reducing α to lower the chance of committing a Type I error will move the critical value, p^*, to the right (in this example). This will have the effect of increasing β, the probability of a Type II error, and correspondingly reducing the power.
- The larger 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 will not be likely to make any Type II errors at all—but then, we are unlikely to really need a formal hypothesis-testing procedure to see such an obvious difference. If the TT practitioners were successful almost all the time, we'd be able to see that with even a small experiment.

Fisher and $\alpha = 0.05$

Why did Sir Ronald Fisher suggest 0.05 as a criterion for testing hypotheses? It turns out that he had in mind small initial studies. Small studies have relatively little power. Fisher was concerned that they might make too many Type II errors—failing to discover an important effect—if too strict a criterion were used. Once a test failed to reject a null hypothesis, it was unlikely that researchers would return to that hypothesis to try again.

On the other hand, the increased risk of Type I errors arising from a generous criterion didn't concern him as much for exploratory studies because these are ordinarily followed by a replication or a larger study. The probability of a Type I error is α—in this case, 0.05. The probability that two independent studies would both make Type I errors is $0.05 \times 0.05 = 0.0025$, so Fisher was confident that Type I errors in initial studies were not a major concern.

The widespread use of the relatively generous 0.05 criterion even in large studies is most likely not what Fisher had in mind.

Reducing Both Type I and Type II Errors

A S ***Activity:* Power and Sample Size.** Investigate how the power of a test changes with the sample size. The interactive tool is really the only way you can see this easily.

Figure 21.1 seems to show that if we reduce Type I error, we automatically must increase Type II error. But there is a way to reduce both. Can you think of it?

If we can make both curves narrower, as shown in Figure 21.2, then both the probability of Type I errors and the probability of Type II errors will decrease, and the power of the test will increase.

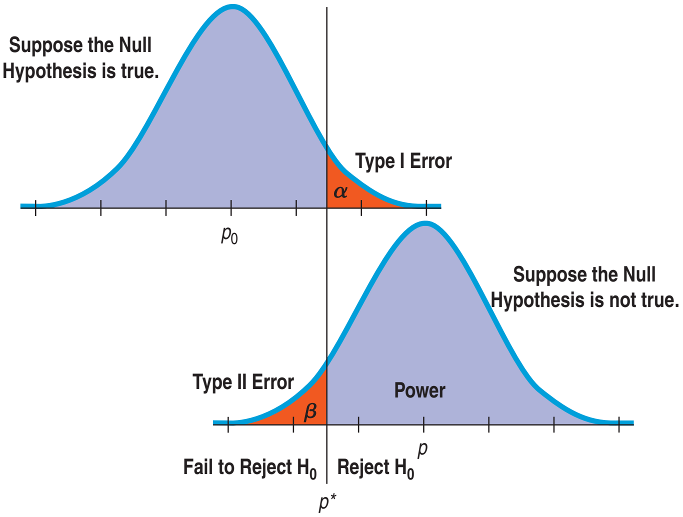


FIGURE 21.2

Making the standard deviations smaller increases the power without changing the corresponding critical value. The means are just as far apart as in Figure 21.1, but the error rates are reduced.

TI-*nspire*

Errors and power. Explore the relationships among Type I and Type II errors, sample size, effect size, and the power of a test.

How can we accomplish 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.

FOR EXAMPLE Sample size, errors, and power

Recap: The meta-analysis of the risks of heart attacks in patients taking the diabetes drug Avandia combined results from 47 smaller studies. As GlaxoSmith-Kline (GSK), the drug's manufacturer, pointed out in their rebuttal, "Data from the ADOPT clinical trial did show a small increase in reports of myocardial infarction among the *Avandia*-treated group . . . however, the number of events is too small to reach a reliable conclusion about the role any of the medicines may have played in this finding."

Question: Why would this smaller study have been less likely to detect the difference in risk? What are the appropriate statistical concepts for comparing the smaller studies?

Smaller studies are subject to greater sampling variability; that is, the sampling distributions they estimate have a larger standard deviation for the sample proportion. That gives small studies less power: They'd be less able to discern whether an apparently higher risk was merely the result of chance variation or evidence of real danger. The FDA doesn't want to restrict the use of a drug that's safe and effective (Type I error), nor do they want patients to continue taking a medication that puts them at risk (Type II error). Larger sample sizes can reduce the risk of both kinds of error. Greater power (the probability of rejecting a false null hypothesis) means a better chance of spotting a genuinely higher risk of heart attacks.

WHAT CAN GO WRONG?

- **Don't interpret the P-value as the probability that H_0 is true.** The P-value is about the data, not the hypothesis. It's the probability of observing data this unusual, *given* that H_0 is true, not the other way around.
- **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 refrain from rejecting 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 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, an important difference can be missed if your test lacks sufficient power.
- **Don't forget that in spite of all your care, you might make a wrong decision.** We can never reduce the probability of a Type I error (α) or of a Type II error (β) to zero (but increasing the sample size helps).

CONNECTIONS

All of the hypothesis tests we'll see boil down to the same question: "Is the difference between two quantities large?" We always measure "how large" by finding a ratio of this difference to the standard deviation of the sampling distribution of the statistic. Using the standard deviation as our ruler for inference is one of the core ideas of statistical thinking.

We've discussed the close relationship between hypothesis tests and confidence intervals. They are two sides of the same coin.

This chapter also has natural links to the discussion of probability, to the Normal model, and to the two previous chapters on inference.

WHAT HAVE WE LEARNED?

We've learned that there's a lot more to hypothesis testing than a simple yes/no decision.

- We've learned that the P-value can indicate evidence against the null hypothesis when it's small, but it does not tell us the probability that the null hypothesis is true.
- We've learned that the alpha 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.
- We've also learned more about the connection between hypothesis tests and confidence intervals; they're 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.

We've learned about the two kinds of errors we might make, and we've seen why in the end we're never sure we've made the right decision.

- 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 could happen.
- If the null hypothesis is really false but we fail to reject it, that's a Type II error.
- 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 to detect it.
- We've seen that 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

486. The threshold P-value that determines when we reject a null hypothesis. If we observe a statistic whose P-value based on the null hypothesis is less than α, we reject that null hypothesis.

Statistically significant

486. When the P-value falls below the alpha level, we say that the test is "statistically significant" at that alpha level.

Significance level 486. The alpha level is also called the significance level, most often in a phrase such as a conclusion that a particular test is "significant at the 5% significance level."

Type I error 491. 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 491. 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.

Power 492, 493. The probability that a hypothesis test will correctly reject a false null hypothesis is the power of the test. To find power, we must specify a particular alternative parameter value as the "true" value. For any specific value in the alternative, the power is $1 - \beta$.

Effect size 493. The difference between the null hypothesis value and true value of a model parameter is called the effect size.

Skills

THINK

- Understand that statistical significance does not measure the importance or magnitude of an effect. Recognize when others misinterpret statistical significance as proof of practical importance.
- Understand the close relationship between hypothesis tests and confidence intervals.
- 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 know how to defend the choice of a one-sided alternative.
- Understand how the critical value for a test is related to the specified alpha level.
- Understand that the power of a test gives the probability that it correctly rejects a false null hypothesis when a specified alternative is true.
- Understand that the power of a test depends in part on the sample size. Larger sample sizes lead to greater power (and thus fewer Type II errors).

SHOW

- Know how to complete a hypothesis test for a population proportion.

TELL

- Be able to interpret the meaning of a P-value in nontechnical language.
- Understand that the P-value of a test does not give the probability that the null hypothesis is correct.
- Know that we do not "accept" a null hypothesis if we cannot reject it but, rather, that we can only "fail to reject" the hypothesis for lack of evidence against it.

HYPOTHESIS TESTS ON THE COMPUTER

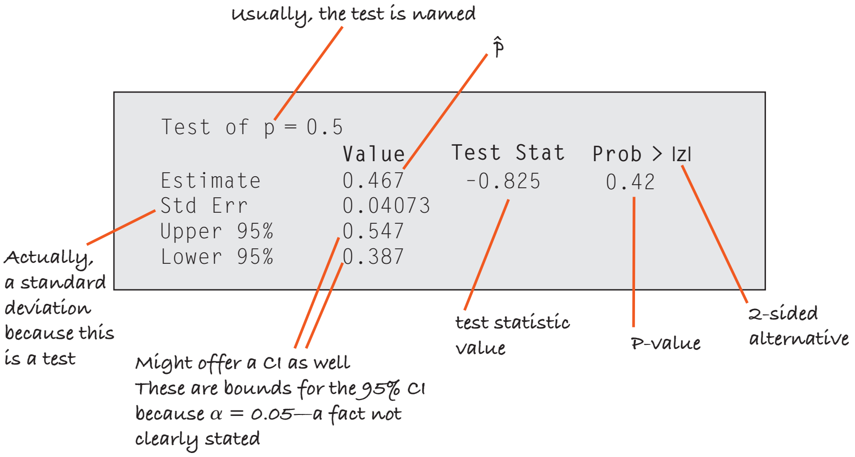
Reports about hypothesis tests generated by technologies don't follow a standard form. Most will name the test and provide the test statistic value, its standard deviation, and the P-value. But these elements may not be labeled clearly. For example, the expression "Prob $> |z|$" means the probability (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 is 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 the two-sided test.

Sometimes a confidence interval and hypothesis test are automatically given together. The CI ought to be for the corresponding confidence level: $1 - \alpha$ for 2-tailed tests, $1 - 2\alpha$ for 1-tailed tests.

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 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. This is not from any program or calculator we know of, but it shows some of the things you might see in typical computer output.



```
Test of p = 0.5
                 Value    Test Stat   Prob > |z|
Estimate         0.467    -0.825       0.42
Std Err          0.04073
Upper 95%        0.547
Lower 95%        0.387
```

Usually, the test is named
$\hat{p}$
Actually, a standard deviation because this is a test
Might offer a CI as well
These are bounds for the 95% CI because $\alpha = 0.05$—a fact not clearly stated
test statistic value
P-value
2-sided alternative

For information on hypothesis testing with particular statistics packages, see the table for Chapter 20 in Appendix B.

EXERCISES

1. One sided or two? In each of the following situations, is the alternative hypothesis one-sided or two-sided? What are the hypotheses?

a) A business student conducts a taste test to see whether students prefer Diet Coke or Diet Pepsi.

b) PepsiCo recently reformulated Diet Pepsi in an attempt to appeal to teenagers. They run a taste test to see if the new formula appeals to more teenagers than the standard formula.

c) A budget override in a small town requires a two-thirds majority to pass. A local newspaper conducts a poll to see if there's evidence it will pass.

d) One financial theory states that the stock market will go up or down with equal probability. A student collects data over several years to test the theory.

2. Which alternative? In each of the following situations, is the alternative hypothesis one-sided or two-sided? What are the hypotheses?

a) A college dining service conducts a survey to see if students prefer plastic or metal cutlery.

b) In recent years, 10% of college juniors have applied for study abroad. The dean's office conducts a survey to see if that's changed this year.

c) A pharmaceutical company conducts a clinical trial to see if more patients who take a new drug experience headache relief than the 22% who claimed relief after taking the placebo.

d) At a small computer peripherals company, only 60% of the hard drives produced passed all their performance tests the first time. Management recently invested a lot of resources into the production system and now conducts a test to see if it helped.

3. P-value. A medical researcher tested a new treatment for poison ivy against the traditional ointment. He concluded that the new treatment is more effective. Explain what the P-value of 0.047 means in this context.

4. Another P-value. Have harsher penalties and ad campaigns increased seat-belt use among drivers and passengers? Observations of commuter traffic 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.

5. Alpha. A researcher developing scanners to search for hidden weapons at airports has concluded that a new device is significantly better than the current scanner. He