

Chapter 1

Significance: How strong is the evidence?

Chapter Overview

This chapter is about statistical significance, the first of the four pillars of inference: strength, size, breadth, and cause. **Statistical significance** indicates the **strength** of the evidence. For example, how strong is the evidence that the form of the question about organ donation affects the chance that a person agrees to be a donor? Do we have only a suggestion of an impact or are we overwhelmingly convinced?

The goal of the chapter is to explain how statisticians measure strength of evidence. In the organ donation study, the researchers wanted to determine whether the wording (opt in, versus opt out, versus forced choice) really does make a difference. The data in our study pointed in that direction, but are we convinced? How strong is that evidence? Are there other explanations for the results found in the study? One way to approach our investigation is similar to a criminal trial where innocence is assumed and then evidence is presented to help convince a jury that the assumption of innocence and the actual evidence are at odds. Likewise, we will assume that the wording does *not* make a difference and then see whether the evidence (our data) is dramatic enough to convince us otherwise. But we know the outcomes for any one study are random in the sense that there is “chance variability.” How, then, do we eliminate “they just got lucky” as an explanation? We have to compare our results to what we would expect to see, “by chance,” if the wording did not have an effect. If the actual data and this “by chance” explanation are at odds, then which do you believe? Short of a fluke outcome, the actual data trumps the chance explanation and we then we have strong evidence against the “wording does not have an effect” hypothesis and in favor of our research hypothesis that there is an effect.

Understanding this logic is the hardest challenge of the chapter.

Section 1.1: Introduction to Chance Models

Introduction

A key step in the Statistical Investigation Method is drawing conclusions beyond the observed data. Statisticians often call this “statistical inference.” There are four main types of conclusions (inferences) that statisticians can draw from data: significance, estimation, generalization, and causation. In the remainder of this chapter we will focus on **statistical significance**.

If you think back to the Organ Donor study from the Preliminaries, there were three groups: those in the neutral group were asked to make a yes/no choice about becoming a donor; those in the “opt in” group were told their default was not to be a donor, but they could choose to become a donor; and those in the “opt out” group were told their default was they were a donor, but they could choose not to be one if they wished. Let’s further examine two of those groups. We saw that 23 of 55 or 41.8% in the “opt in” group elected to become donors, compared with 44/56 or 78.6% in the neutral group. A key question here is whether we believe that the 78.6% is far enough away from the 41.8% to be considered **statistically significant**, meaning unlikely to have occurred by random chance alone. True, 78.6% looks very different from 41.8%, but it is

at least possible that the wording of the solicitation to donate actually makes no difference, and that the difference we observed happened by random chance.

To answer the question “Is our result unlikely to happen by random chance?,” our general strategy will be to consider what we expect the results to look like if any differences we are seeing are solely due to random chance. Exploring the random chance results is critical to our ability to draw meaningful conclusions from data. In this section we will provide a framework for assessing random chance explanations.

Example 1.1: Can Dolphins Communicate?

A famous study from the 1960s explored whether two dolphins (Doris and Buzz) could communicate abstract ideas. Researchers believed dolphins could communicate simple feelings like “Watch out!” or “I’m happy,” but Dr. Jarvis Bastian wanted to explore whether they could also communicate in a more abstract way, much like humans do. To investigate this, Dr. Bastian spent many years training Doris and Buzz and exploring the limits of their communicative ability.

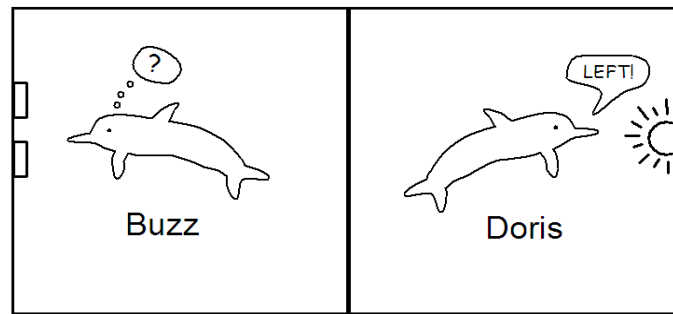
During a training period lasting many months, Dr. Bastian placed buttons underwater on each end of a large pool—two buttons for Doris and two buttons for Buzz. He then used an old automobile headlight as his signal. When he turned on the headlight and let it shine steadily, he intended for this signal to mean “push the button on the right.” When he let the headlight blink on and off, this was meant as a signal to “push the button on the left.” Every time the dolphins pushed the correct button, Dr. Bastian gave the dolphins a reward of some fish. Over time Doris and Buzz caught on and could earn their fish reward every time.

Then Dr. Bastian made things a bit harder. Now, Buzz had to push his button before Doris. If they didn’t push the buttons in the correct order—no fish. After a bit more training, the dolphins caught on again and could earn their fish reward every time. The dolphins were now ready to participate in the real study to examine whether they could communicate *with each other*.

Dr. Bastian placed a large canvas curtain in the middle of the pool. (See Figure 1.1.) Doris was on one side of the curtain and could see the headlight, whereas Buzz was on the other side of the curtain and could not see the headlight. Dr. Bastian turned on the headlight and let it shine steadily. He then watched to see what Doris would do. After looking at the light, Doris swam near the curtain and began to whistle loudly. Shortly after that, Buzz whistled back and then pressed the button on the right—he got it correct and so both dolphins got a fish. But this single attempt was not enough to convince Dr. Bastian that Doris had communicated with Buzz through her whistling. Dr. Bastian repeated the process several times, sometimes having the light blink (so Doris needed to let Buzz know to push the left button) and other times having it glow steadily (so Doris needed to let Buzz know to push the right button). He kept track of how often Buzz pushed the correct button.

In this scenario, even if Buzz and Doris can communicate, we don’t necessarily expect Buzz to push the correct button every time. We allow for some “randomness” in the process; maybe on one trial Doris was a bit more underwater when she whistled and the signal wasn’t as clear for Buzz. Or maybe Buzz and Doris aren’t communicating at all and Buzz guesses which button to push every time and just happens to guess correctly once in a while. Our goal is to get an idea of how likely Buzz is to push the correct button in the long run.

Figure 1.1: Depending whether or not the light was blinking or shown steadily, Doris had to communicate to Buzz as to which button to push.



Let's see how Dr. Bastian was applying the Six-Step Statistical Investigation Method.

Step 1: Ask a research question. Can dolphins communicate in an abstract manner?

Step 2: Design a study and collect data. Notice Dr. Bastian took some time to train the dolphins in order to get them to a point where he could test a specific research conjecture. The research conjecture is that Buzz pushes the correct button more often than he would if he and Doris could not communicate. If Buzz and Doris could not communicate, Buzz would just be guessing which button to push. The *observational units* are Buzz's attempts and the *variable* for each attempt is whether or not on each attempt, Buzz pushes the correct button (a *categorical variable*).

Step 3: Explore the data. In one phase of the study, Dr. Bastian had Buzz attempt to push the correct button a total of 16 different times. In this **sample** of 16 attempts, Buzz pushed the correct button 15 out of 16 times. To summarize these results, we report the **statistic**, a numerical summary of the sample. For this example, we could report either 15, the number of correct pushes, or $15/16 = 0.9375$, as the statistic.

Definitions: The set of observational units on which we collect data is called the **sample**. The number of observational units in the sample is the **sample size**. A **statistic** is a number summarizing the results in the sample.

The **sample size** in this example is 16. Note that the word "sample" is used as both a noun (the set of observational units being studied) and as an adjective, for example to mean "computed from the observed data," as, for example, "sample statistic."

Step 4: Draw inferences beyond the data. These 16 observations are a mere snapshot of Buzz's overall selection process. We will consider this a *random process*. We are interested in Buzz's actual long-run *probability* of pushing the correct button based on Doris' whistles. This unknown long-run probability is called a **parameter**.

Definition: For a random process, a **parameter** is a long-run numerical property of the process.

Note that we are assuming this parameter is not changing over time, at least for the process used by Buzz in this phase of the study. Because we can't observe Buzz pushing the button forever, we need to draw conclusions (possibly incorrect, but hopefully not) about the value of the parameter based only on these 16 attempts. Buzz certainly pushed the correct button most of the time, so we might consider either:

- Buzz is doing something other than just guessing (his probability of a correct button push is larger than 0.50).
- Buzz is just guessing (his probability of a correct button push is 0.50) and he got lucky in these 16 attempts.

These are the two possible explanations to be evaluated. Because we can't collect more data, we have to base our conclusions only on the data we have. It's certainly *possible* that Buzz was just guessing and got lucky! But does this seem like a reasonable explanation to you? How would you argue against someone who thought this was the case?

Think about it: Based on these data, do you think Buzz somehow knew which button to push? Is 15 out of 16 correct pushes convincing to you? Or do you think that Buzz could have just been guessing? How might you justify your answer?

So how are we going to decide between these two possible explanations? One approach is to choose a **model** for the random process (repeated attempts to push the correct button) and then see whether our model is consistent with the observed data. If it is, then we will conclude that we have a reasonable model and we will use that model to answer our questions.

The Chance Model

Scientists use models to help understand complicated real world phenomena. Statisticians often employ **chance models** to generate data from random processes to help them investigate such processes. You did this with the Monty Hall Exploration (P.3) to investigate properties of the two strategies, switching and staying with your original choice of door. In that exploration it was clear how the underlying chance process worked, even though the probabilities themselves were not obvious. But here we don't know for sure what the underlying real world process is. We are trying to decide whether the process could be Buzz simply guessing or whether the process is something else, such as Buzz and Doris being able to communicate.

Let us first investigate the "Buzz was simply guessing" process. Because Buzz is choosing between two options, the simplest chance model to consider is a coin flip. We can flip a coin to represent or *simulate* Buzz's choice *assuming he is just guessing* which button to push. To generate this artificial data, we can let "heads" represent the outcome that Buzz pushes the correct button and let "tails" be the outcome that Buzz pushes the incorrect button. This gives Buzz a 50% chance of pushing the correct button. This can be used to represent the "Buzz was just guessing" or the "random chance alone" explanation. The correspondence between the real study and the physical simulation is shown in Table 1.1.

Table 1.1: Parallels between real study and physical simulation

Coin flip	=	guess by Buzz
Heads	=	correct guess
Tails	=	wrong guess
Chance of Heads	= 1/2 =	probability of correct button when Buzz is just guessing
One repetition	=	one set of 16 simulated attempts by Buzz

Now that we see how flipping a coin can simulate Buzz guessing, let's flip some coins to simulate Buzz's performance. Imagine that we get heads on the first flip. What does this mean? This would correspond to Buzz pushing the correct button! But, why did he push the correct

button? In this chance model, the only reason he pushed the correct button is because he happened to guess correctly—remember the coin is simulating what happens when Buzz is just guessing which button to push.

What if we keep flipping the coin? Each time we flip the coin we are simulating another attempt where Buzz guesses which button to push. Remember that heads represents Buzz guessing correctly and tails represents Buzz guessing incorrectly. How many times do we flip the coin? Sixteen, to match Buzz's 16 attempts in the actual study. After 16 tosses, we obtained the sequence of flips shown in Figure 1.2.

Figure 1.2: A sequence of 16 coin flips

Heads, Tails, Tails, Heads, Tails, Heads, Heads, Heads, Heads, Tails, Heads, Heads, Heads, Heads, Tails

Here we got 11 heads and 5 tails (11 out of 16 or 0.6875 is the simulated statistic). This gives us an idea of what could have happened in the study if Buzz had been randomly guessing which button to push each time.

Will we get this same result every time we flip a coin 16 times? Let's flip our coin another 16 times and see what happens. When we did this we got 7 heads and 9 tails as shown in the sequence of coin flips (7 out of 16 or 0.4375 is the simulated statistic) in Figure 1.3.

Figure 1.3: Another sequence of 16 coin flips

Heads, Tails, Heads, Heads, Tails, Tails, Tails, Heads, Tails, Tails, Heads, Tails, Heads, Heads, Tails, Tails

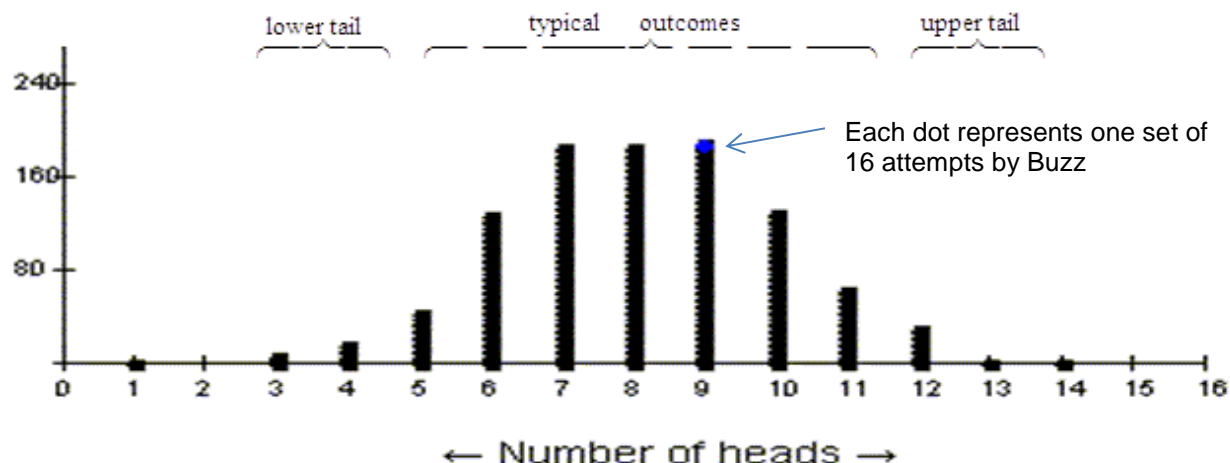
So can we learn anything from these coin tosses when the results vary between the sets of 16 tosses?

Using and evaluating the coin flip chance model

Because coin flipping is a random process, we know that we won't obtain the same number of heads with every set of 16 flips. But are some numbers of heads more likely than others? If we continue our repetitions of 16 tosses, we can start to see how the outcomes for the number of heads are distributed. Does the *distribution* of the number of heads that result in 16 flips have a predictable long-run pattern? In particular, how much *variability* is there in our simulated statistics between repetitions (sets of 16 flips) just by random chance?

In order to investigate these questions, we need to continue to flip our coin to get many, many sets of 16 flips (or many repetitions of the 16 choices where we are modeling Buzz simply guessing each time). We did this, and Figure 1.4 shows what we found when we graphed the number of heads from each set of 16 coin flips. Here, the process of flipping a coin 16 times was repeated 1,000 times—this number was chosen for convenience, but also appears to be large enough to give us a fairly accurate sense of the long-run behavior for the number of heads in 16 tosses.

Figure 1.4: A dotplot showing 1000 repetitions of flipping a coin 16 times and counting the number of heads



Let's think carefully about what the graph in Figure 1.4 shows. For this graph, each dot represents the number of heads in one set of 16 coin tosses. We see that the resulting number of heads follows a clear pattern: 7, 8, and 9 heads happened quite a lot, 10 was pretty common also (though less so than 8), 6 happened some of the time, 1 happened once. But we never got 15 heads in any set of 16 tosses! We might consider any outcome between about 5 and 11 heads to be typical, but getting fewer than 5 heads or more than 11 heads happened rarely enough we can consider it a bit unusual. We refer to these unusual results as being out in the 'tails' of the distribution.

Think about it: How does the analysis above help us address the strength of evidence for our research conjecture that Buzz was doing something other than just guessing?

What does this have to do with the dolphin communication study? We said that we would flip a coin to simulate what could happen if Buzz was really just guessing each time he pushed the button in 16 attempts. We saw that getting results like 15 heads out of 16 never happened in our 1000 repetitions. This shows us that 15 is a very unusual outcome—far out in the tail of the distribution of the simulated statistics—if Buzz is guessing. In short, even though we expect some variability in the results for different sets of 16 tosses, the pattern shown in this distribution indicates that an outcome of 15 heads is outside the typical chance variability we would expect to see when Buzz is simply guessing.

In the actual study, Buzz really did push the correct button 15 times out of 16, an outcome that we just determined would rarely occur if Buzz was just guessing. So, our coin flip chance model tells us that we have very strong evidence that Buzz was not just tossing a coin to make his choices. This means we have strong evidence that Buzz wasn't just guessing. Therefore, we don't believe the "by chance alone" explanation is a good one for Buzz. The results mean we have strong enough evidence to be **statistically significant**, not a result that happened by chance alone, and to convince us that something other than "random chance" was at play.

Definition: A result is **statistically significant** if it is unlikely to occur just by random chance. If our observed result appears to be consistent with the chance model, we say that the chance model is **plausible** or believable.

What next? A glimpse into Steps 5 and 6

The steps we went through above have helped us evaluate how strong the evidence is that Buzz is not guessing (Step 4 of the Statistical Investigation Method). In this case, the evidence provided by this sample is fairly strong that Buzz isn't guessing. Still, there are some important questions you should be asking right now, such as: If Buzz isn't guessing what is he doing?

Step 5: Formulate conclusions. We should also ask ourselves: if Buzz wasn't guessing, does this prove that Buzz and Doris can communicate? And if so, what does this say about other dolphins? As we'll find in later chapters, the answers to these questions hinge mainly on how the study was designed and how we view the 16 attempts that we observed (e.g., we assume Buzz couldn't see the light himself, the light signal displayed each time was chosen randomly so Buzz couldn't figure out a pattern to help him decide which button to push; Buzz's 16 attempts are a good representation of what Buzz would do given many more attempts under identical conditions; but we might still wonder whether Buzz's behavior is representative of dolphin behavior in general or are there key differences among individual dolphins).

Step 6: Look back and ahead. After completing Steps 1-5 of the Statistical Investigation Method, we need to revisit the big picture of the initial research question. First, we reflect on the limitations of the analysis, and think about future studies. In short, we are now stepping back and thinking about the initial research question more than the specific research conjecture being tested in the study. In some ways, this is the most important step of the whole study because it is where we think about the true implications of the scientific study we've conducted. For this study, we would reflect on Dr. Bastian's methods, summarize the results for Buzz, and reflect on ways to improve the study to enhance the conclusions we can draw.

The 3S Strategy

Let us summarize the overall approach to assessing *statistical significance* that we have been taking in this section. We observed a sample statistic (e.g., the number of "successes" or the proportion of "successes" in the sample). Then we simulated 'could-have-been' outcomes for that statistic under a specific chance model (just guessing). Then we used the information we gained about the random variation in the 'by-chance' values of the statistics to help us judge whether the observed value of the statistic is an unusual or a typical outcome. If it is unusual—we say the observed statistic is *statistically significant*—it provides strong evidence that the chance-alone explanation is wrong. If it is typical, we consider the chance model plausible.

You may have noticed that we only simulated results for one specific model. When we saw that the sample statistic observed in the study was not consistent with these simulated results, we rejected the chance-alone explanation. Often, research analyses stop here. Instead of trying to simulate results from other models (in particular we may not really have an initial idea what a more appropriate model might be), we are content to say there is something other than random chance at play here. This might lead the researchers to reformulate their conjectures and collect more data in order to investigate different models.

We will call the process of simulating could-have-been statistics under a specific chance model the **3S Strategy**. After forming our research conjecture and collecting the sample data, we will use the 3S strategy to weigh the evidence against the chance model. This 3S Strategy will

serve as the foundation for addressing the question of statistical significance in Step 4 of the Statistical Investigation Method.

3S Strategy for Measuring Strength of Evidence

- 1. Statistic:** Compute the statistic from the observed sample data.
- 2. Simulate:** Identify a “by chance alone” explanation for the data. Repeatedly simulate values of the statistic that could have happened when the chance model is true.
- 3. Strength of evidence:** Consider whether the value of the observed statistic from the research study is unlikely to occur if the chance model is true. If we decide the observed statistic is unlikely to occur by chance alone, then we can conclude that the observed data provide strong evidence against the plausibility of the chance model. If not, then we consider the chance model to be a plausible (believable) explanation for the observed data; in other words what we observed could plausibly have happened just by random chance.

Let’s illustrate how we implemented the 3S Strategy for the Doris and Buzz example.

1. Statistic: Our observed statistic was 15, the number of times Buzz pushed the correct button in 16 attempts.

2. Simulate: If Buzz was actually guessing, the parameter (the probability he would push the correct button) would equal 0.5. In other words, he would push the correct button 50% of the time in the long run. We used a coin flip to model what could have happened in 16 attempts when Buzz is just guessing. We flip the coin 16 times and count how many of the 16 flips are heads, meaning how many times Buzz pressed the correct button (“success”). We then repeat this process many more times, each time keeping track of the number of the 16 attempts that Buzz pushed the correct button. We end up with a distribution of could-have-been statistics representing typical values for the number of correct pushes when Buzz is just guessing.

3. Strength of evidence: Because 15 successes of 16 rarely happens by chance alone, we conclude that we have strong evidence that, in the long-run, Buzz is not just guessing.

Notice that we have used the result of 15 out of 16 correct attempts to *infer* that Buzz’s actual long-run probability of pushing the correct button was not simply 0.5.

Another Doris and Buzz study

One goal of statistical significance is to rule out random chance as a *plausible* (believable) explanation for what we have observed. We still need to worry about how well the study was conducted. For example, are we absolutely sure Buzz couldn’t see the headlight around the curtain? Are we sure there was no pattern to which headlight setting was displayed that he might have detected? And of course we haven’t completely ruled out random chance, he may have had an incredibly lucky day. But the chance of him being that lucky is so small, that we conclude that other explanations are more plausible or credible.

One option that Dr. Bastian pursued was to re-do the study except now he replaced the curtain with a wooden barrier between the two sides of the tank in order to ensure a more complete

separation between the dolphins to see whether that would diminish the effectiveness of their communication.

Step 1: Ask a research question. The research question remains the same: Can dolphins communicate in a deep abstract manner?

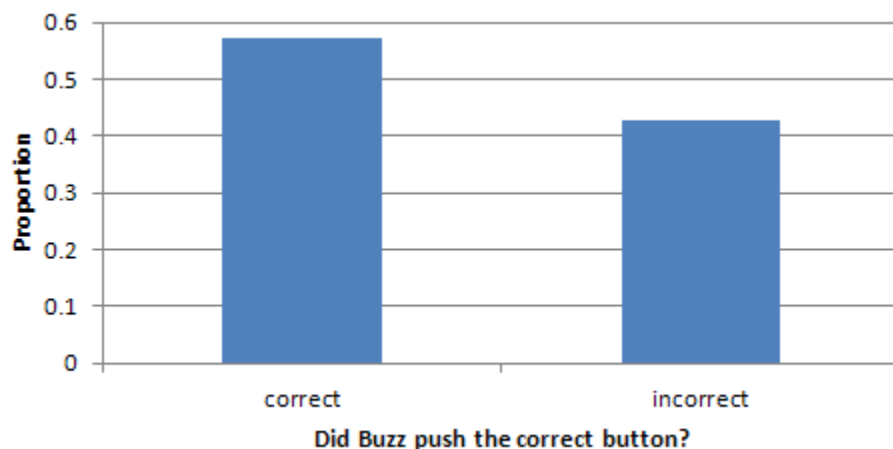
Step 2: Design a study and collect data. The study design is similar with some adjustments to the barrier between Doris and Buzz. The canvas curtain is replaced by a plywood board. The research conjecture, observational units, and variable remain the same.

In this case, Buzz pushed the correct button only 16 out of 28 times. The variable is the same (whether or not Buzz pushed the correct button), but the number of observational units (sample size) has changed to 28 (the number of attempts).

Think about it: Based on the results for this phase of the study, do you think that Doris could tell Buzz which button to push, even under these conditions? Or is it believable that Buzz could have just been guessing?

Step 3: Explore the data. So our observed statistic is 16 out of 28 correct attempts, which is $16/28 \times 100\% = 57.1\%$ of Buzz's attempts. This is again more than half the time, but not much larger than 50%. A simple **bar graph** of these results is shown in Figure 1.5.

Figure 1.5: Bar graph for Buzz's 28 attempts



Step 4: Draw inferences. Is it plausible (believable) that Buzz was simply

guessing in this set of attempts? How do we measure how much evidence these results provide against the chance model? Let's use the same chance model as we used earlier to see what could have happened if Buzz was just guessing. We will apply the 3S Strategy to this new study.

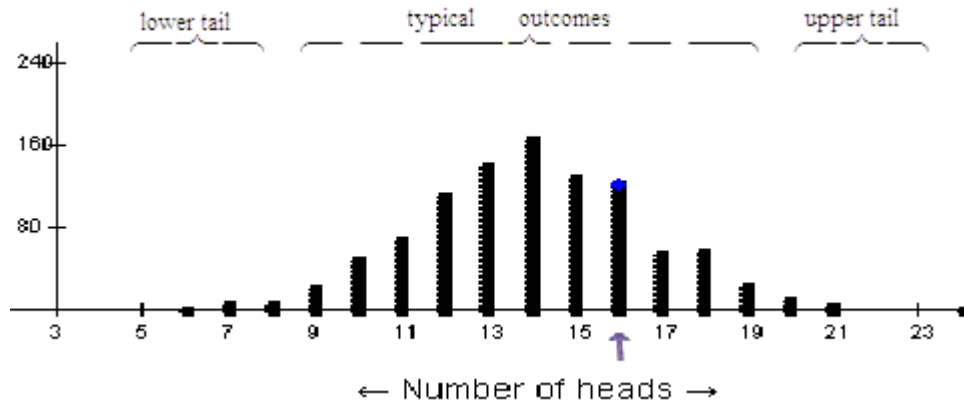
1. Statistic: The new observed sample statistic is 16 out of 28, or about 0.571.

Think about it: Consider again our simulation of the chance model assuming Buzz is guessing. What do we need to change for this new phase of the study?

2. Simulation: This time we need to do repetitions of 28 coin flips, not just 16. A distribution of the number of heads in 1000 repetitions of 28 coin flips is shown in Figure 1.6. This models

1000 repetitions of 28 attempts with Buzz randomly pushing one of the buttons (guessing) each time.

Figure 1.6: A graph showing 1000 repetitions of flipping a coin 28 times and counting the number of heads. This models the number of correct pushes in 28 attempts when Buzz is guessing each time.



3. Strength of evidence: Now we need to consider the new observed statistic (16 out of 28 or 0.571). We see from the graph that 16 out of 28 is a fairly typical outcome if Buzz is just randomly guessing. What does this tell us? It tells us that the results of this study are something that could easily have happened if Buzz was just randomly guessing. So what can we conclude? We can say his 16 successes are not convincing evidence against the “by chance alone” model.

The graph in Figure 1.6 shows what happens for the hypothetical Buzz who just guesses. An actual outcome far out in the tail of that distribution would be strong evidence against the “just guessing” hypothesis. But be careful: The opposite result – an actual outcome near the center – is *not* strong evidence in support of the guessing hypothesis. Yes, the result is *consistent* with that hypothesis, but it is also consistent with many other hypotheses as well.

Bottom line: In this second study we conclude that there is not enough evidence that the “by chance alone” model is wrong. That model is still a *plausible* explanation for the statistic we observed in the study (16 out of 28). Based on this set of attempts, we do not have convincing evidence against the possibility that Buzz is just guessing, but other explanations also remain plausible. For example, the results are consistent with very weak communication between the dolphins. All we know from this analysis is that one plausible explanation for the observed data is that Buzz was guessing.

In fact, Dr. Bastian soon discovered that in this set of attempts, the equipment malfunctioned and the food dispenser for Doris did not operate and so Doris was not receiving her fish rewards during the study. Because of this malfunction, it’s not so surprising that removing the incentive hindered the communication between the dolphins and we cannot refute that Buzz was just guessing for these attempts.

Dr. Bastian fixed the equipment and ran the study again. This time he found convincing evidence that Buzz was not guessing.

For a bit more discussion on processes and parameters, see FAQ 1.1.1.

FAQ 1.1.1: What is a random process?

Q: So a *parameter* is a numerical property of a process... but what is the process? All I have is Buzz's 16 attempts.

A: We can think of Buzz's 16 attempts as a sample from his long-run selection process. He has some underlying probability of pushing the correct button, but he's unlikely to be correct every time, even if the dolphins are communicating. There is *randomness* in his choices. His choices might be affected by Doris' squeaks, but also by how hungry he is, how tired he is, how warm the water is. We are trying to see how much Doris' communication is influencing him. Everything else gets lumped together into "random chance." Any one outcome from a random process is unknown to us in advance (like tossing a coin), but if we observe outcomes from the process for long enough, we should start to see some patterns amidst the randomness. So you can think of a random process as an unlimited source of potential observations for your sample.

Q: Can I use *any* old set of observations from a process as a sample?

A: Some samples are good and others are terrible.

Q: How do I tell the good apples from the rotten ones?

A: For a Yes/No process like Buzz's selections, we need to be willing to make certain assumptions along the way. For example, we assumed Buzz's probability of a correct guess was the same every time. In fact, under the null hypothesis, we assumed it was 50/50 every time. But even if the probability is not 0.5, we are assuming it is not changing. In other words, we assume that Buzz doesn't get better through learning and that he doesn't get tired or bored or quit trying. We are also assuming that his previous trials don't influence his future guesses – like a coin, he would have no memory of his past guesses.

Q: So how do I know if I have a good sample?

A: You need to believe that nothing about the process is changing and that each outcome does not depend on past outcomes. If you're a coin, your chance of heads doesn't change, and you have no memory of how you've landed in the past. If this is true about the process, then you will have a good sample! On the other hand, if Buzz's probability of success is different in the morning than in the afternoon, then only observing him in the morning will not give you a good representation of his overall probability of success.

Q: So then how does the parameter come into play?

A: Parameter is a hard concept, precisely because we never see it! For the dolphin example, we want to know whether Buzz is just guessing. We think of his guesses as a potentially never-ending process (like coin tossing), and the parameter is the *probability* that Buzz will be correct in his next attempt. As you saw in the Preliminaries, this means if we were to observe the random process forever, this probability is the long-run proportion of times that Buzz pushes the correct button.

Q: So I see why the parameter isn't just 15/16, that is his sample proportion and we may not have hit that long-run probability yet. So the parameter is just 0.5?

A: Not necessarily. The parameter equals 0.5 if he is just guessing. That's the chance model we

simulated. But his probability may actually be different from 0.5.

Q: So the parameter is his actual probability of pushing the correct button. It's a number, but we don't know its value.

A: Correct! But keep in mind that we will define lots of different kinds of parameters in future chapters. The "probability of success" parameter only applies in this type of Yes/No random process.

Exploration 1.1: Can Dogs Understand Human Cues?

Dogs have been domesticated for about 14,000 years. In that time, have they been able to develop an understanding of human gestures such as pointing or glancing? How about similar non-human cues? Researchers Udell, Giglio, and Wynne tested a small number of dogs in order to answer these questions.

In this exploration, we will first see whether dogs can understand human gestures as well as non-human gestures. To test this, the researchers positioned the dogs about 2.5 meters from the experimenter. On each side of the experimenter were two cups. The experimenter would perform some sort of gesture (pointing, bowing, looking) toward one of the cups or there would be some other non-human gesture (a mechanical arm pointing, a doll pointing, or a stuffed animal looking) toward one of the cups. The researchers would then see whether the dog would go to the cup that was indicated. There were six dogs tested. We will look at one of the dogs in two of his sets of trials. This dog, a 4 year-old mixed breed, was named Harley. Each trial involved one gesture and one pair of cups, with a total of ten trials in a set.

We will start out by looking at one set of trials where the experimenter bowed toward one of the cups to see whether Harley would go to that cup.

Step 1. State the research question.

1. Based on the description of the study, state the research question.

Step 2. Design a study and collect data. Harley was tested 10 times and 9 of those times he chose the correct cup.

2. What are the observational units?
3. Identify the variable in the study. What are the possible outcomes of this variable? Is this variable quantitative or categorical?

Step 3. Explore the data.

Definitions: The set of observational units on which we collect data is called the **sample**. The number of observational units in the sample is the **sample size**. A **statistic** is a number summarizing the results in the sample.

With categorical data, we typically report the number of “successes” or the proportion of successes as the statistic.

4. What is the number of observational units (sample size) in this study?
5. Determine the observed statistic and produce a simple bar graph of the data (have one bar for the proportion of times Harley picked the correct cup, and another for the proportion of times he picked the wrong cup)
6. If the research conjecture is that Harley can understand what the experimenter means when they bow toward an object, is the statistic in the direction suggested by the research conjecture?
7. Could Harley have gotten 9 out of 10 correct even if he really didn’t understand the human gesture and so was randomly guessing between the two cups?
8. Do you think it is likely Harley would have gotten 9 out of 10 correct if he was just guessing randomly each time?

Step 4. Draw inferences beyond the data.

There are two possibilities for why Harley chose the correct cup nine out of ten times:

- He is merely picking a cup at random, and in these 10 trials happened to guess correctly in 9 of them. That is, he got more than half correct just by random chance alone.
- He is doing something other than merely guessing and perhaps understands what the experimenter means when they bow towards the cup.

The unknown long-run probability that Harley will chose the correct cup is called a ***parameter***.

Definition: For a random process, a ***parameter*** is a long-run numerical property of the process.

We don't know the value of the parameter, but the two possibilities listed above suggest two different possibilities.

9. What is the value of the parameter if Harley is picking a cup at random? Give a specific value.
10. What is the possible range of values (greater than or less than some value) for the parameter Harley is not just guessing and perhaps understands the experimenter?

We will now show you how statisticians use simulation to make a statement about the strength of evidence for these two possible statements about the parameter's value.

The Chance Model

Statisticians often use *chance models* to generate data from random processes to help them investigate the process. In particular, they can see whether the observed statistic is consistent with the values of the statistic simulated by the chance model. If we determine that Harley's results are not consistent with the results from the chance model, we will consider this to be evidence against the chance model and in favor of the research conjecture, that he understands the bowing gesture. In this case, we would say Harley's results are *statistically significant*, meaning unlikely to have occurred by chance alone.

We can't perform the actual study more times in order to assess the second possibility, but we can simulate the behavior of Harley's choices if we were to assume the first possibility (that he is simply guessing every time).

11. Explain how you could use a coin toss to represent Harley's choices if he is guessing between the two cups each time. How many times do you have to flip the coin to represent one set of Harley's attempts? What does heads represent?
12. If Harley was guessing randomly each time, on average, how many out of the 10 times would you expect him to choose the correct cup?

13. Simulate one repetition of Harley guessing randomly by flipping a coin 10 times (why 10?) and letting heads represent selecting the correct cup ("success") and tails represent selecting the incorrect cup ("failure"). Count the number of heads in your 10 flips. Combine your results with the rest of the class to create a *dotplot* of the distribution for the number of heads out of 10 flips of a coin.
- Where does 9 heads fall in the distribution? Would you consider it an unusual outcome or a fairly typical outcome for the number of heads in 10 flips?
 - Based on your answer to the previous question, do you think it is plausible (believable) that Harley was just guessing which cup to choose?

Using an applet to simulate flipping a coin many times

To really assess the typical values for the number of heads in 10 coin tosses (number of correct picks by Harley assuming he is guessing at random), we need to simulate many more outcomes of the chance model. Open the **One Proportion** applet from the textbook webpage.

Notice that the probability of heads has been set to be 0.50, representing the guessing between the two cups. Set the number of tosses to 10 and press the **Toss Coins** button. What was the resulting number of heads?

Notice that the number of heads in this set of 10 tosses is then displayed by a dot on the graph. Uncheck the **Animate** box and press the **Toss Coins** button 9 more times. This will demonstrate how the number of heads varies randomly across each set of 10 tosses. Nine more dots have been added to your dotplot. Is a pattern starting to emerge?

Now change the **Number of repetitions** from 1 to 990 and press **Toss Coins**. The applet will now show the results for the number of heads in 1000 different sets of 10 coin tosses. So each dot represents the number of times Harley chooses the correct cup out of ten attempts assuming he is just guessing.

Remember why we conducted this simulation: to assess whether Harley's result (9 correct in 10 attempts) would be unlikely to occur by chance alone if he were just guessing between the pair of cups for each attempt.

14. Locate the result of getting 9 heads in the dotplot created by the applet. Would you consider this an unlikely result, in the tail of the distribution of the number of heads?

15. Based on the results of 1000 simulated sets of 10 coin flips each, would you conclude that Harley would be very unlikely to have picked the correct cup in 9 times in 10 attempts if he was randomly guessing between the two cups each time? Explain how your answer relates to the applet's dotplot.

Definition: A result is **statistically significant** if it is unlikely to occur just by random chance. If our observed result appears to be consistent with the chance model, we say that the chance model is **plausible** or believable.

16. Do the results of this study appear to be statistically significant?
17. Do the results of this study suggest that Harley just guessing is a plausible explanation for Harley picking the correct cup 9 out of 10 times?

Summarizing your understanding

18. To make sure that you understand the coin flipping chance model fill in the following table indicating what parts of the real study correspond to the physical (coin-flipping) simulation.

Table 1.2: Parallels between real study and physical simulation

Coin flip	=	
Heads	=	
Tails	=	
	$\pi = 1/2$	
Chance of Heads	=	
One repetition	=	one set of ____ simulated attempts by Harley

The 3S Strategy

We will call the process of simulating could-have-been statistics under a specific chance model the **3S Strategy**. After forming our research conjecture and collecting the sample data, we will use the 3S strategy to weigh the evidence against the chance model. This 3S Strategy will

serve as the foundation for addressing the question of statistical significance in Step 4 of the Statistical Investigation Method.

3S Strategy for Measuring Strength of Evidence

1. Statistic: Compute the statistic from the observed sample data.

2. Simulate: Identify a “by chance alone” explanation for the data. Repeatedly simulate values of the statistic that could have happened when the chance model is true.

3. Strength of evidence: Consider whether the value of the observed statistic from the research study is unlikely to occur if the chance model is true. If we decide the observed statistic is unlikely to occur by chance alone, then we can conclude that the observed data provide strong evidence against the plausibility of the chance model. If not, then we consider the chance model to be a plausible (believable) explanation for the observed data; in other words what we observed could plausibly have happened just by random chance.

Let’s review how we have already applied the 3S strategy to this study.

19. **Statistic.** What is the statistic in this study?

20. **Simulate.** Fill in the blanks to describe the simulation.

We flipped a coin _____ times and kept track of how many times it came up heads. We then repeated this process _____ more times, each time keeping track of how many heads were obtained in each of the _____ flips.

21. **Strength of evidence.** Fill in the blanks to summarize how we are assessing the strength of evidence for this study.

Because we rarely obtained a value of _____ heads when flipping the coin _____ times, this means that it is _____ (believable/unlikely) that Harley is just guessing, since if Harley was just guessing he _____ (rarely/often) would get a value like _____ correct out of _____ attempts.

Step 5: Formulate conclusions.

22. Based on this analysis, are you convinced that Harley can understand human cues? Why or why not?

Another Study

One important step in a statistical investigation is to consider other models and whether the results can be confirmed in other settings.

23. In a different study, the researchers used a mechanical arm (roughly the size of a human arm) to point at one of the two cups. The researchers tested this to see whether dogs understood non-human gestures. In 10 trials, Harley chose the correct cup 6 times.

- a. Using the dotplot you obtained when you simulated 1000 sets of 10 coin flips assuming Harley was just guessing, locate the result of getting 6 heads. Would you consider this an unlikely result, in the tail of the distribution?
- b. Based on the results of 1000 simulated sets of 10 coin flips each, would you conclude that Harley would be very unlikely to have picked the correct cup in 6 times in 10 attempts, if he was randomly guessing between the two cups each time? Explain how your answer relates to the applet's dotplot.
- c. Is this study's result statistically significant?
- d. Do the results of this study suggest that Harley just guessing is a plausible explanation for Harley picking the correct cup 6 out of 10 times?
- e. Does this study prove that Harley cannot understand the mechanical arm?

Step 6: Look back and ahead.

24. Compare the analyses between the two studies. How does the unusualness of the observed statistic compare between the two studies? Does this make sense based on the value of the observed statistic in the two studies? Does this make sense based on how the two studies were designed? Explain. (*Hint*. Why might the results differ for human and mechanical arms? Why would this matter?)

25. A single study will not provide all of the information needed to fully understand a broad, complex research question. Thinking back to the original research question, what additional studies would you suggest conducting next?

Section 1.1 Summary

The set of observational units on which we collect data is called a **sample**. The number of observational units is the **sample size**. A number computed to summarize the variable measured on a sample is called a **statistic**.

For a chance process, a **parameter** is a long-run numerical property of that process, such as a probability (long-run proportion).

A simulation analysis based on a chance model can assess the strength of evidence provided by sample data against a particular claim about the chance model. The logic of assessing statistical significance employs what we call the **3S Strategy**:

- **Statistic:** Compute an observed statistic from the data.
- **Simulate:** Identify a model for the “by chance alone” explanation. Repeatedly simulate values of the statistic that could have occurred from that chance model.
- **Strength of evidence:** Examine how unusual the observed value of the statistic would be under repeated application of the chance model.
 - If the observed value of the sample statistic is *unlikely* to have occurred from the chance model, then the data provide *strong* evidence against the chance model as the explanation.
 - If the observed value of the sample statistic is *not* unlikely to have occurred from the chance model, then the chance model is a *plausible* explanation for the observed data.

The **chance model** considered in this section involved tossing a fair coin. This chance model allowed for assessing whether an observed number of “successes” in a study provided strong evidence that the two outcomes of a categorical variable were not equally likely. In the next section you will consider other chance models, but the reasoning process will remain the same.

Section 1.2: Measuring the Strength of Evidence

Introduction

In the previous section, we discussed testing whether or not a chance model of *equally choosing between two options* (e.g., *flipping a coin*) was plausible (believable) based on some observed sample data. Not all situations call for a coin flipping model. For example, if you were guessing at the answers to a true-false test, you would expect to have a 50% chance of getting a question correct. However, if you were guessing at the answers to a multiple choice test where each question had four possible answers, you should expect to have only a 25% chance of guessing a particular question correctly. Other events, as you will soon see, can be a bit more complicated than this. In this section, we will apply the Six-Step Statistical Investigation Method in the more general setting of an arbitrary probability of “success.”

We will also learn helpful terminology (null and alternative hypotheses, p-value) that are commonly used in statistical investigations to describe the process of drawing conclusions from data. These things will both help us formalize the procedure of a **test of significance** and give us some guidelines to help us determine when we have strong enough evidence that our chance model is not correct. We will also introduce some new symbols, for convenience, so try to keep the big picture in mind.

Example 1.2: Rock Paper Scissors

Have you ever played Rock-Paper-Scissors (or *Rochambeau*)? It's considered a “fair game” in that the two players are equally likely to win (like a coin toss). Both players simultaneously display one of three hand gestures (rock, paper or scissors), and the objective is to display a gesture that defeats that of your opponent. Official rules are available from the World RPS Society (www.worldrps.com), but the main gist is that rocks break scissors, scissors cut paper, and paper covers rock.

But is it really a fair game? Do some players exhibit patterns in their behavior that an opponent can exploit? An article published in *College Mathematics Journal* (Eyler, Shalla, Doumaux, and McDevitt, 2009) found that players, particularly novices, tend to not prefer scissors. Suppose you decide to investigate this tendency with a friend of yours who hasn't played the game before. You explain the rules of the game and play 12 rounds. Suppose your friend only shows scissors twice in those 12 plays.

Think about it:

- What are the observational units? What is the variable?
- What is the underlying process from which the researchers gathered their sample data?
- What is the parameter of interest?
- What is the observed statistic?

A similarity between this study and ones we've looked at previously (e.g., Buzz and Doris) is that we have repeated outcomes from the same random process. In this study, the individual plays of the game are the observational units, and the variable is whether or not the player chooses scissors. This categorical variable has just two outcomes (scissors or not scissors), and so is sometimes called a **binary variable**.

Definition: A **binary variable** is a categorical variable with only two outcomes. Often we convert categorical variables with more than two outcomes (e.g., rock, paper, or scissors) into

binary variables (e.g., scissors or not scissors). In this case we also define one outcome to be a “success” and one to be a “failure.”

The underlying process here is choosing a hand gesture in each repeated play of the game. The parameter of interest is the long-run probability that any player picks scissors. The observed statistic is the 2 scissor choices in the 12 plays, or $1/6$.

Think about it: What is the research conjecture about the long-run probability?

We are told that players seem to pick scissors less than the other two gestures. This would imply that the long-run probability for picking scissors is less than one-third. And in this study, that’s what you found. Your friend only chose scissors $1/6$ of the time.

But perhaps that was just an unlucky occurrence in this study? Maybe your friend would play scissors $1/3$ of the time if he played the game for a very long time, and, just by chance, you happened to observe less than one-third in the first 12 games? So, again, we have two possible explanations for why our statistic ($1/6$) is below what we think is the true parameter ($1/3$) if the random chance model is correct.

- Novice players, like your friend, pick equally among the 3 gestures and we just happened to observe fewer scissor choices in this study by chance alone.
- Novice players really do tend to pick scissors less than $1/3$ of the time.

We can rewrite these two possible explanations as two competing hypotheses:

Null hypothesis: Novice players pick equally between the 3 gestures in the long-run (picking scissors one-third of the time in the long run).

Alternative hypothesis: Novice players pick scissors less than one-third of the time in the long run.

Definition:

- The **null hypothesis** typically represents the “by random chance alone” explanation. The chance model (or “null model”) is chosen to reflect this hypothesis.
- The **alternative hypothesis** typically represents the “there is an effect” explanation that contradicts the null hypothesis. It represents what the researchers are hoping to gather evidence to support (the research conjecture).

Notice that the null and alternative hypotheses are statements about the parameter (probability of choosing scissors) and the underlying process (in the long-run), not just about what was observed in this study. In fact, we should state the hypotheses prior to conducting the study, before we ever gather any data! Our goal is to use the sample data to estimate the parameter and to infer whether this unknown value is less than one-third.

Use of Symbols

We can use mathematical symbols to represent quantities and simplify our writing. Throughout the book we will emphasize written explanations, but will also show you mathematical symbols which you are free to use as a short-hand once you are comfortable with the material. The distinction between parameter and statistic is so important that we always use different symbols to refer to them.

We will use the Greek letter for p which is π (pronounced "pie") to represent a parameter that is a probability. For example, the long-run probability that a novice player picks scissors can be represented by π . We will then use the symbol \hat{p} (pronounced "p-hat") to represent the proportion in the sample. In this example, $\hat{p} = 1/6 \approx 0.167$, the proportion of times that your friend chose scissors. In fact, one way to distinguish between the parameter and the statistic is verb tense! The statistic is the proportion of times that your friend did (past tense, observed) show scissors. The parameter is the long-run proportion he would throw scissors (future tense, unobserved) if he played the game forever.

We can also use symbols for the hypotheses. The null hypothesis is often written as H_0 and the alternative as H_a . Finally, the sample size, which is the same as the number of observational units, also has a symbol: the letter " n ."

We can use these symbols to rewrite the null and alternative hypotheses.

$$H_0: \pi = 1/3$$

$$H_a: \pi < 1/3$$

where π represents your friend's true probability of throwing scissors.

Notice that both of these hypothesis statements are claims about the parameter π . When testing a single probability as we are doing in this example, we compare the actual (but unknown) value of the parameter (π) to the same numerical value in both the null and alternative hypotheses. In this case that value is $1/3$ (often called the "hypothesized" probability). What differs between the two statements is the inequality symbol. The null hypothesis will always contain an equals sign and the alternative hypothesis will contain either a (strictly) greater than sign (as it would have in the Doris and Buzz example from the previous question), a (strictly) less than sign (as it does in this example), or a not equal to sign like we will see in Section 1.4. Which inequality symbol to use in the alternative hypothesis is determined by the research conjecture.

Applying the 3S Strategy

Now, let's apply the 3S Strategy to evaluate the evidence against the null hypothesis in favor of the alternative hypothesis, that your friend is less likely to choose scissors than the other options.

1. Statistic: Your friend showed scissors $1/6$ of the time in the first 12 plays.

2. Simulation: We will again focus on the chance alone explanation, to see whether we have strong evidence that your friend chooses scissors less than $1/3$ of the time in the long run. So we will use a chance model that assumes the probability of your friend choosing scissors is one-third, and then examine the types of results we get for the sample proportion of times he chooses scissors to see whether our observed statistic is consistent with that chance variability.

Think about it: Can we use a coin to represent the chance model specified by the null hypothesis like before? If not, can you suggest a different random device we could use? What

needs to be different about our simulation this time?

We cannot directly use a coin to simulate the chance model for this scenario. The key difference is now we want to sample from a process where the probability of “success” is not 0.5 (as with coin tossing) but $1/3$. This will model your friend choosing scissors one-third of the time. We could use dice (e.g., rolls of 1 or 2 represent scissors) or cards (like with Monty Hall), or the computer can do this for us with correspondences as outlined in Table 1.2.

Table 1.2: Parallels between actual study and simulation.

Observational unit	=	One round of the game
Success	=	Plays scissors
Failure	=	Doesn't play scissors
Chance of success	$\pi = 1/3 =$	Three gestures are equally likely; chance of scissors is one out of three
One repetition	=	A simulated set of 12 rounds of the game

For the first 12 plays of the game, your friend only showed scissors twice. If scissors were chosen as often as the other two options, we would expect four of the 12 plays to be scissors, so maybe this is not all that different, especially in the short run. Let's first evaluate how surprising it is for your friend to only play scissors twice, looking at results from a simulation that models your friend choosing scissors one-third of the time in the long-run.

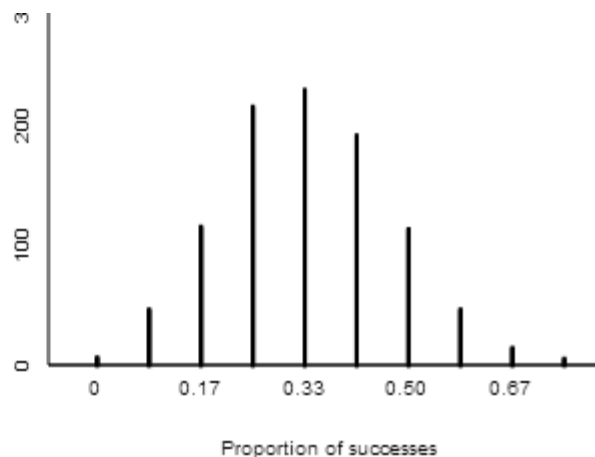
In the **One Proportion** applet you can change the probability of success to any number between 0 and 1, and the applet will generate successful attempts at that rate. One way to think about it is as a spinner that will spin and land in one area one-third of the time and the other area two-thirds of the time. So we now have a chance model that generates successes with a long-run probability equal to the null hypothesis probability of $1/3$.

Figure 1.7 shows us the distribution of the simulated sample proportion of scissor choices (successes) that could have happened in 12 rounds of the game assuming this chance model. We can use this distribution to assess the strength of evidence against the $1/3$ chance model. The distributions of simulated statistics we have been looking at represent what could have happened in the study assuming the null hypothesis was true. For that reason, we will now refer to this distribution as the **null distribution**.

Figure 1.7: The null distribution of simulated sample proportion of successes in 12 plays assuming scissors is chosen $1/3$ (≈ 0.3333) of the time in the long run

Probability of success (π): Sample size(n): Number of samples: ☐ Animate

Total = 1000



Notice that we are using the proportion of successes as the statistic instead of the number of successes, but these are equivalent. This distribution is centered around $1/3$ and has a very slight skew to the right. But what does the distribution tell us about whether or not your friend tends to avoid scissors?

3. Strength of evidence

Think about it: How do we use this null distribution to evaluate whether the $1/3$ model is appropriate for your friend's scissors selection process?

The key is to see whether the observed statistic (2 out of 12 ≈ 0.167) is consistent with the chance variability in the $1/3$ process. So we need to decide whether or not 0.167 is a typical value or falls in the tail of the distribution. In Figure 1.7, we see that this is a bit of a difficult call. The outcomes at 0.17 are not that far in the tail of the distribution but are not right in the middle either. What we need is a more systematic way of measuring how unusual an outcome of 0.167 is in this null distribution.

We could see how often our observed result occurs in the distribution, but that becomes difficult to compare across different distributions (see FAQ 1.2.1). A more common approach for measuring how unusual an observation is in a distribution is to determine how much of the distribution lies to the left of the value and how much to the right (think “percentiles” from when you took the SAT or ACT!). The probability of getting a result at least as extreme as the observed statistic, when the null hypothesis is true, is called the **p-value**. We can estimate the p-value from our simulated null distribution by counting how many (and what proportion) of the simulated sample proportions are as extreme, or more extreme (in the direction of the alternative hypothesis), than the observed sample proportion.

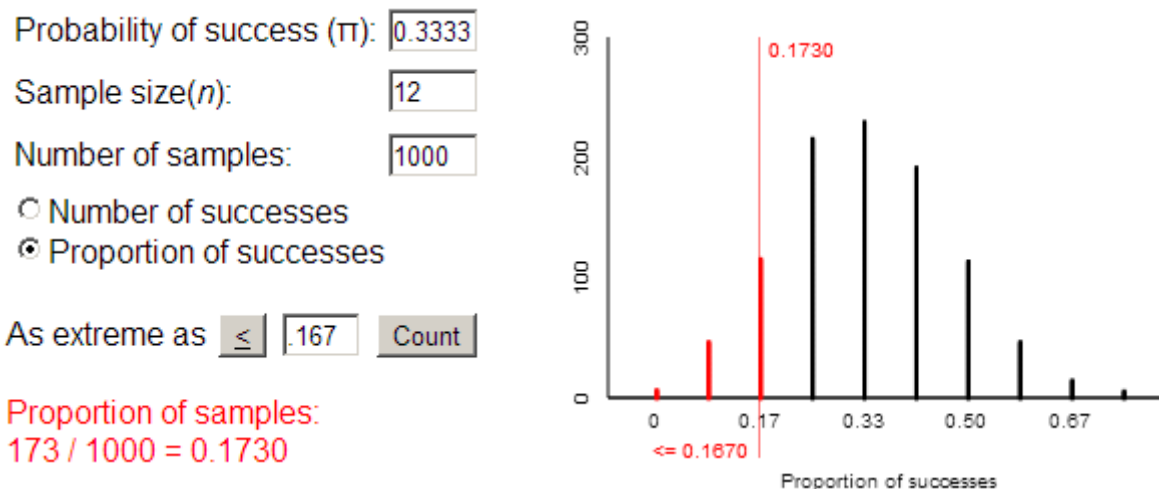
Definition: The **p-value** is the probability of obtaining a value of the statistic at least as extreme as the observed statistic when the null hypothesis is true. We can estimate the p-value by finding the proportion of the simulated statistics in the null distribution that are *at least as extreme* (in the direction of the alternative hypothesis) as the value of the statistic actually observed in the research study.

The p-value takes into account the could-have-been outcomes (assuming the null hypothesis is true) that are as extreme, or more extreme, than the one we observed. This provides a direct measure of our strength of evidence against the “by chance alone” or null model and allows for a standard, comparable value for all scientific research studies. Smaller p-values mean the

value of the observed statistic, under the null model, is more unlikely by chance alone. Hence, smaller p-values indicate stronger evidence against the null model.

Let's reconsider the distribution from Figure 1.7. In Figure 1.8, we added a red-line at the observed statistic, 0.167. The “proportion of repetitions” counts how many times 0.167 or smaller occurred under the specified chance model. In this case, we found 173 of the 1000 samples from a process with $\pi = 1/3$ resulted in a simulated sample proportion \hat{p} of 0.167 or smaller. So 0.173 is our estimate of the study's p-value.

Figure 1.8: The null distribution of simulated sample proportion of successes in 12 rounds of rock-paper-scissors for a novice that plays scissors one-third of the time in the long run with the observed proportion (0.167) or even smaller shown in red, yielding an approximate p-value of 0.1730.



Calculating a p-value from a simulation analysis is an approximation. Different simulations will have slightly different p-values based on 1000 repetitions. Performing more repetitions generally produces more accurate approximations. For our purposes, using 1000 repetitions is typically good enough to provide reasonable approximations for p-values.

Notice that the approximate p-value of 0.173 in Figure 1.8 is computed as the proportion of the 1000 simulated samples with a sample proportion of 0.167 or smaller. A p-value always computes “more extreme” in the direction of the alternative hypothesis. In this case the alternative hypothesis is “less than 1/3” and so we look at the lower tail to compute “more extreme than” for the p-value.

So how do we *interpret* this 0.173? As a probability; so we can say that in the long run, if we repeatedly generate random sets of 12 rounds of the game under identical conditions with the probability of scissors equal to 1/3, we expect to observe a sample proportion of 0.167 or smaller, by chance alone, in about 17.3% of those repetitions.

But what does 0.173 tell us about the strength of the evidence? As stated earlier, smaller p-values are stronger evidence against the null hypothesis in favor of the alternative hypothesis. Is 0.173 small enough? Although there is no hard and fast rule for determining how small is small enough to be convincing, we offer the following guidelines.

Guidelines for evaluating strength of evidence from p-values

$0.10 < \text{p-value}$	not much evidence against null hypothesis; null is plausible
$0.05 < \text{p-value} \leq 0.10$	moderate evidence against the null hypothesis
$0.01 < \text{p-value} \leq 0.05$	strong evidence against the null hypothesis
$\text{p-value} \leq 0.01$	very strong evidence against the null hypothesis

The smaller the p-value, the stronger the evidence against the null hypothesis.

Many researchers consider a p-value ≤ 0.05 to be sufficient to conclude there is convincing evidence against the null hypothesis (see FAQ 1.2.2 for some intuition on that number), but in some situations you may want stronger evidence. For now, just keep in mind that the smaller the p-value, the stronger the evidence against the null hypothesis (chance model is true) and in favor of the alternative hypothesis (typically the research conjecture).

So we would consider only 2 plays of scissors in the first 12 rounds of the game to not be much evidence against the null hypothesis your friend would play scissors $1/3$ of the time in the long run. Why? Because a p-value of 0.173 indicates that getting 2 or fewer choices of scissors in 12 plays, if the probability of scissors were really $1/3$, is *not surprising*. Hence, $1/3$ is still a plausible value for your friend's long-run probability.

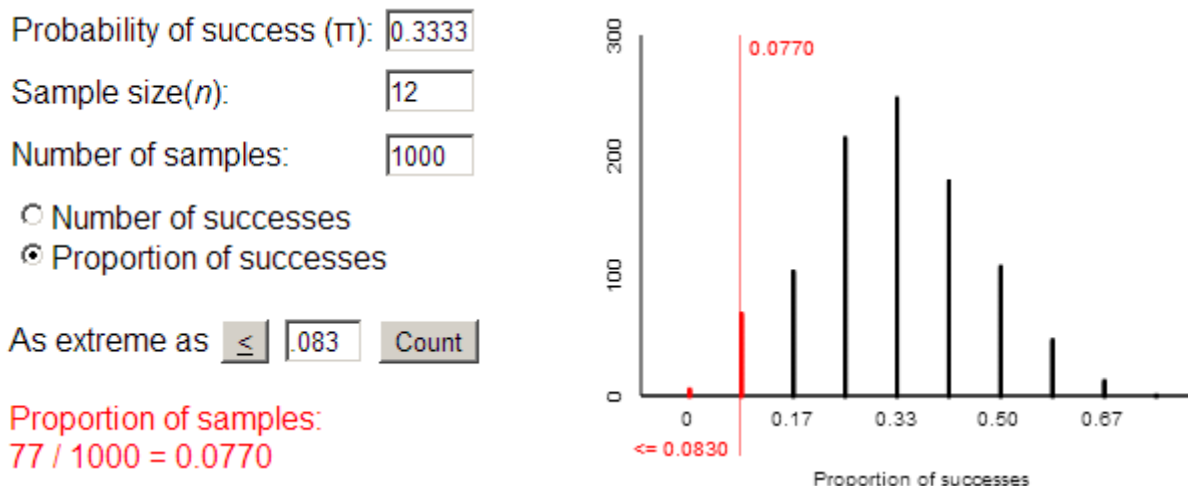
What if your friend had only played scissors once? Then our statistic becomes $\hat{p} \approx 0.083$. This is even further away from the expected one-third. So what will happen to the p-value? We can use the same null distribution but now we need to see how often we find a sample proportion as small as 0.083 or smaller. Such simulated statistics are shaded red in Figure 1.9.

Key idea: Values of the statistic that are even further from the hypothesized parameter result in a smaller p-value and stronger evidence against the null hypothesis.

Conclusions

So what can we conclude from this study? We approximated a p-value (0.173) for the first 12 plays of the game that was not small and did not provide strong evidence against the null hypothesis. Have we *proven* the null hypothesis is true? No! In fact, we will never get to prove the null hypothesis true because we had to assume it was true to do the analysis. So the results of the simulation can never “support” the null hypothesis. What we can say is that these results are not inconsistent with the type of result we would expect to see when the null hypothesis is true. In other words, the null hypothesis is one plausible (or believable) explanation for the data. If we wanted to investigate the issue more closely, we could have our friend play more games (increase the sample size). A larger sample size would give us a better chance of detecting any tendency that might be there (we'll discuss the issue of sample size more fully in section 1.4).

Figure 1.9: The null distribution of simulated sample proportion of successes in 12 rounds of rock-paper-scissors for novices that play scissors one-third of the time in the long run. If the observed proportion would only have been 0.083, we can see the p-value would have been smaller and we would therefore have stronger evidence against the null hypothesis.



The approximate p-value has decreased to 0.077. This makes sense because if the null hypothesis is true ($\pi = 1/3$), it should be more surprising to get a sample proportion further from $1/3$.

Key assumptions: Model vs. reality

Stepping back, we need to keep something in mind about the use of chance models in assessing strength of evidence. Chance models are models, and, thus, are not reality. They make key assumptions that we need to pay careful attention to, otherwise we may too quickly leap to an incorrect conclusion.

For example, with the Rock-Paper-Scissors game we simulated data from a chance model with a one-third success probability. This means each and every round has exactly the same probability of scissors. But, is that a reasonable assumption? What if your friend changes his strategy part way through? Our chance model ignores these aspects of “reality” and makes the situation much simpler than it really is.

Thus, if we could say “we have strong evidence against the chance model” are we saying that your friend subconsciously avoids scissors? Well, maybe. All we are saying is we think something else is going on other than your friend picking equally among the three gestures. When we think about Steps 4 and 5 of the Six-Step Statistical Investigation Method, we must be aware about these assumptions we’re making---specifically about the ways in which our model does not match reality.

FAQ 1.2.1 Why do we need to include “or more extreme” when computing our p-value?

Q: So Buzz pushed the correct button 15 out of 16 times, why isn’t the p-value the likelihood I would get exactly 15 heads out of 16 flips of a coin? Why do we have to also include the likelihood of getting more than 15 heads?

A: In flipping just 16 coins, the probability of getting 15 heads is very similar to the probability of getting 15 or more heads, so I can see how this would be confusing. I have an idea to help you understand this. Do you have any coins with you?

Q: I might, let me check. It looks like I have a few.

A: Good. Do you think you can toss the coin “fairly”?

Q: I don’t see why not.

A: Start flipping it. As you do this, keep track of the number of heads you get as well as the total number of flips. Also think about whether or not your results indicate whether this is an unfair tossing process.

Q: Okay, I flipped the coin 10 times and got 6 heads. Because this is close to 50-50, I don’t think I could conclude that my coin tossing is unfair.

A: Good. The probability of getting 6 heads out of 10 flips is about 0.20 whereas the probability of getting 6 or more heads (our actual p-value) is about 0.37. Either way you think of this, getting 6 heads out of 10 coin flips is not too unlikely. Let’s keep flipping.

Q: Okay, I’ve flipped it 50 times and ...

A Keep flipping.

Q: I’ve flipped it now 100 times and...

A: Keep flipping.
(1 hour later)

Q: I think my arm is going to fall off!

A: I guess you can stop now. What are your results?

Q: I flipped the coin 1000 times and got 505 heads.

A: Do you think your results show that your coin tossing is not fair?

Q: Of course not. Getting 505 heads out of 1000 flips is close enough to 50% that the result is not unexpected.

A: Do you mean you expected to get 505 heads?

Q: Well no, not exactly 505, but I did expect to get something close to 500.

A: In fact, the probability of getting exactly 505 heads out of 1000 flips of a fair coin is only about 0.02. Because there are so many different outcomes possible, the probability of any one particular outcome is rather small. But we wouldn’t want to look at the probability of .02 and consider this a surprising outcome - this outcome is definitely among the typical values in the distribution. This is better conveyed by noting that the probability of getting 505 heads or more in 1000 flips is about 0.39.

Q: And because this p-value is so high, I would not conclude my coin tossing is unfair!

A: Exactly!

FAQ 1.2.2: What p-value should make us suspicious?

Q: So, that table you gave us about strength of evidence and p-values seems pretty arbitrary. Where did that come from?

A: Did you ever google “Persi Diaconis”?

Q: No, why?

A: Try it sometime. He’s unusual ... even for a statistician he’s unusual. He was one of the first people to win one of the MacArthur “genius” awards.

Q: Is that all? What else can he do?

A: Flip a coin and make it come up heads every single time.

Q: I don’t believe anyone can do that. Isn’t this whole dialog just a stat prof’s geeky way to sneak in p-values?

A: Sorry, I can’t help it. We stat pros are so transparent sometimes. That’s just the data. But humor me: How many heads in a row would Persi have to get to make you begin to think maybe it’s not just chance? Would heads on the first two tosses do it for you?

Q: Of course not. That happens 25% of the time.

A: You win a point for extra credit. What about heads on the first three tosses?

Q: You’re making me get serious about this. Three heads in a row wouldn’t happen very often, but it’s still not unusual enough to be suspicious.

A: What about four in a row? Five in a row?

Q: Now it’s really getting serious ... and suspicious. I find it hard to believe that you can get five heads in a row just by chance. Sure, it can happen, but five in a row is enough to make me think maybe there’s something else going on.

A: As it turns out the chances of getting four heads in a row is about 6% and the chances of getting five heads in a row is about 3%. This means that most people start getting suspicious when something occurs less than ~5% of the time.

Q: That sounds familiar. Didn’t you tell us that when the p-value is less than 0.05 that gives us strong evidence?

A: Exactly! So, just like getting four or five heads in a row makes us think something else is going on, getting a p-value less than 0.05 gives us strong evidence the null hypothesis may be incorrect.

Exploration 1.2: Tasting Water

People spend a lot of money on bottled water. But do they really prefer bottled water to ordinary tap water? Researchers at Longwood University (Lunsford and Dowling Fink, 2010) investigated this question by presenting people who came to a booth at a local festival with four cups of water. Three cups contained different brands of bottled water, and one cup was filled with tap water. Each **subject** (person) was asked which of the four cups of water they most preferred. Researchers kept track of how many people chose tap water in order to see whether tap water was chosen significantly less often than would be expected by random chance.

Step 1: Ask a research question.

1. What is the research question that the researchers hoped to answer?

Step 2: Design a study and collect data.

2. Identify the observational units in this study.
3. Identify the variable. Is the variable quantitative or categorical?
4. How many outcomes can the variable have?

Definition: A **binary variable** is a categorical variable with only two outcomes. Often we convert categorical variables with more than two outcomes (e.g., four brands of water) into binary variables (e.g., tap water or not). In this case we also define one outcome to be a “success” and one to be a “failure.”

5. Describe the parameter of interest (in words). (*Hint:* The parameter is the long-run probability of ...?)
6. One possibility here is that subjects have an equal preference among all four waters and so are essentially selecting one of the four cups at random. In this case what is the long-run probability that a subject in this study would select tap water?
7. Another possibility is that the subjects are less likely to prefer tap water than the bottled water brands. In this case what can you say about the long-run probability that a subject in this study would select tap water? (*Hint:* You are not to specify a particular value this time; instead indicate a *direction* from a particular value.)

Definition:

- The **null hypothesis** typically represents the “by chance alone” explanation. The chance model (or “null model”) is chosen to reflect this hypothesis.
- The **alternative hypothesis** typically represents the “there is an effect” explanation that contradicts the null hypothesis. It represents what the researchers are hoping to gather evidence to support (the research conjecture).

8. Your answers to #6 and #7 should be the null and alternative hypotheses for this study. Which is which?

The researchers found that 3 of 27 subjects selected tap water.

Step 3: Explore the data.

9. Calculate the value of the relevant statistic.

Use of symbols

We can use mathematical symbols to represent quantities and simplify our writing. Throughout the book we will emphasize written explanations, but will also show you mathematical symbols which you are free to use as a short-hand once you are comfortable with the material. The distinction between parameter and statistic is so important that we always use different symbols to refer to them.

When dealing with a parameter that is a long-run probability, such as the probability that a (future) subject in this study *would* choose tap water as most preferred, we use the Greek letter π (pronounced “pie”). But when working with a statistic that is the proportion of “successes” in a sample, such as the proportion of subjects in this study who *did* choose tap water as most preferred, we use the symbol \hat{p} (pronounced “p-hat”). Finally, we use the symbol n to represent the sample size.

10. What is the value of \hat{p} in this study?
11. What is the value of n in this study?
12. Hypotheses are always conjectures about the unknown parameter. π . You can also use H_0 and H_a as short-hand notation for the null and alternative hypotheses, respectively. A colon, ‘:’, is used to represent the word ‘is.’ Restate the null and alternative hypotheses using π .

H_0 :

H_a :


Step 4: Draw Inferences.

13. Is the sample proportion who selected tap water in this study less than the probability specified in the null hypothesis?
14. Is it *possible* that this proportion could turn out to be this small even if the null hypothesis were true (i.e., even if people did not really dislike the tap water and were essentially selecting at random from among the four cups)?

As we did with Buzz and Doris in Section 1.1, we will use simulation to investigate how surprising the observed sample result (3 of 27 selecting tap water) would be, if in fact subjects did not dislike tap water and so each had a $\frac{1}{4}$ probability of selecting tap water. (Note also that our null model assumes the same probability for all subjects.)

Think about it: Can we use a coin to represent the chance model specified by the null hypothesis like before? If not, can you suggest a different random device we could use? What needs to be different about our simulation this time?

15. Explain why we cannot use a simple coin toss to simulate the subjects' choices, as we did with the Buzz/Doris study.
16. We could do the simulation using a set of four playing cards: one black and three red. Explain how the simulation would work in this case.

17. Another option would be to use a spinner like the one shown here , like you would use when playing a child's board game. Explain how the simulation would work if you were using a spinner. In particular:

What does each region represent?

How many spins of the spinner will you need to do in order simulate one repetition of the experiment when there is equal preference (null hypothesis is true)?

18. We will now use the **One Proportion** applet to conduct this simulation analysis. Notice that the applet will show us what it would be like if we were simulating with spinners.
- First enter the **probability of success** value specified in the null hypothesis.
 - Enter the appropriate **sample size** (number of subjects in this study).
 - Enter 1 for the number of samples, and press **Draw Samples**. Report the number of “successes” in this simulated sample.
 - Now, select the radio button for “Proportion of successes.” Report the proportion of successes in this simulated sample. Use your answer to c. to verify how this value is calculated.
 - Leaving the “Proportion of successes” radio button selected, click on **Draw Samples** four more times. Do you get the same results each time?
 - Now enter 995 for the number of samples and click on **Draw Samples**, bringing the number of simulated samples to 1000. Comment on the center, variability, and shape of the resulting distribution of sample proportions.

This distribution of simulated sample proportions is called the **null distribution**, because it is based on assuming the null hypothesis to be true.

19. Recall that the observed value of the sample proportion who selected tap water in this study was $\hat{p} = 3/27 \approx 0.1111$. Looking at the null distribution you have simulated, is this a very unlikely result, if the null hypothesis were true? In other words, is this value far in the tail of the null distribution?

You might very well find that #19 is a bit of a tough call. The value 0.1111 is not far in the tail of the distribution, but it's also not near the middle of the distribution. To help make a judgment about strength of evidence in this case, we can count how many (and what proportion) of the simulated sample proportions are as extreme, or more extreme, than the observed value.

20. Use the applet to count how many (and what proportion) of the simulated sample proportions are more extreme than the observed value.

To do this, first click on the \geq inequality symbol to change it to \leq (to match the alternative hypothesis). Then enter **0.1111** (the observed sample proportion who chose tap water) in the box to the left of the Count button. Then click on the **Count** button. Record the number and proportion of simulated sample proportions that are as extreme, or more than, than the observed value.

Definition: The ***p-value*** is estimated as the proportion of simulated statistics in the null distribution that are *at least as extreme* (in the direction of the alternative hypothesis) as the value of the statistic actually observed in the research study.

How do we *evaluate* this p-value as a judgment about strength of evidence provided by the sample data against the null hypothesis? One answer is: The smaller the p-value, the stronger the evidence against the null hypothesis and in favor of the alternative hypothesis. But how small is small enough to regard as convincing? There is no definitive answer, but here are some guidelines:

Guidelines for evaluating strength of evidence from p-values

$0.10 < \text{p-value}$	not much evidence against null hypothesis; null is plausible
$0.05 < \text{p-value} \leq 0.10$	moderate evidence against the null hypothesis
$0.01 < \text{p-value} \leq 0.05$	strong evidence against the null hypothesis
$\text{p-value} \leq 0.01$	very strong evidence against the null hypothesis

The smaller the p-value, the stronger the evidence against the null hypothesis.

21. Is the approximate p-value from your simulation analysis (your answer to #20) small enough to provide much evidence against the null hypothesis that subjects prefer tap water equally to the brands of bottled water? If so, how strong is this evidence? Explain.

22. When computing p-values, “more extreme” is always measured in the direction of the alternative hypothesis. Use this fact to explain why you clicked the \leq earlier.

Step 5: Formulate conclusions.

23. Do you consider the observed sample result to be statistically significant? Recall that this means that the observed result is unlikely to have occurred by chance alone.

24. How broadly are you willing to generalize your conclusions? Would you be willing to generalize your conclusions to water drinkers beyond the subjects in this study? How broadly? Explain your reasoning.

Step 6: Look back and ahead.

25. Suggest a new research question that you might investigate next, building on what you learned in this study.

Alternate Analysis

Instead of focusing on the subjects who chose tap water, you could instead analyze the data based on the subjects who chose one of the three bottled waters. Because 3 of 27 subjects chose tap water, we know that 24 of the 27 subjects chose one of the brands of bottled water. Now let the parameter of interest (denoted by π) be the long-run probability that a subject selects one of the *bottled* water cups as most preferred.

26. Conduct a simulation analysis to assess the strength of evidence provided by the sample data.
- a. The research conjecture is that subjects tend to select bottled water (more or less) often than tap water. (Circle your answer.)
- More Less
- b. State the null hypothesis in words and in terms of the (newly defined) parameter π .
- c. State the alternative hypothesis in words and in terms of the (new) parameter π .
- d. Calculate the observed value of the relevant statistic.
- e. Before you use the **One Proportion** applet to analyze these data, indicate what values you will input:
- Probability of success:
- Sample size:
- Number of samples:
- f. Use the applet to produce the null distribution of simulated sample proportions. Comment on the center, variability, and shape of this distribution. Be sure to comment on how this null distribution differs from the null distribution in #16f.

- g. In order to approximate the p-value, you will count how many of the simulated proportions are _____ or (larger or smaller) and then divide by _____.
 - h. Estimate the p-value from your simulation results.
 - i. *Interpret* this p-value. (*Hint*: This is the probability of what, assuming what?)
 - j. *Evaluate* this p-value: How much evidence do the sample data provide against the null hypothesis?
27. Does your analysis based on the number who chose bottled water produce similar conclusions to your previous analysis based on the number who chose tap water? Explain.

You should have found that it does not matter whether you focus on the number/proportion that chose tap water or the number/proportion that chose bottled water. In other words, it does not matter which category you define to be a “success” for the *preferred water* variable. Your findings should be very similar provided that you make the appropriate adjustments in your analysis:

- using 0.75 instead of 0.25 as the null value of the parameter,
- changing the alternative hypothesis to “ $\pi > 0.75$ ” rather than “ $\pi < 0.25$ ”, and
- calculating the p-value as the proportion of samples with $\hat{p} \geq 0.8889$ rather than $\hat{p} \leq 0.1111$.

Section 1.2 Summary

The 3S Strategy for assessing statistical significance also applies to chance models other than a coin toss:

- Other probabilities of success, such as $1/3$, can be analyzed.
- The strategy can assess whether a long-run probability is *less than* a conjectured value, as well as greater than a conjectured value.

Introducing some terminology can help to clarify the 3S Strategy when used to conduct a **test of significance**:

- The **null hypothesis** is the “by chance alone” explanation.
- The **alternative hypothesis** contradicts the null hypothesis.
- The **null distribution** refers to the simulated values of the statistic generated under the assumption that the null hypothesis (“by chance alone” explanation) is true.

Strength of evidence can be assessed numerically by determining how often a simulated statistic as or more extreme than the observed value of the statistic occurs in the null distribution of simulated statistics.

- The **p-value** is estimated by determining the proportion of simulated statistic values in the null distribution that are at least as extreme as the observed value of the statistic.
- The smaller the p-value, the stronger the evidence against the null hypothesis (“by chance alone” explanation).

Some guidelines for evaluating **strength of evidence** based on a p-value are:

- $0.10 < \text{p-value}$ not much evidence against null hypothesis
- $0.05 < \text{p-value} \leq 0.10$ moderate evidence against the null hypothesis
- $0.01 < \text{p-value} \leq 0.05$ strong evidence against the null hypothesis
- $\text{p-value} \leq 0.01$ very strong evidence against the null hypothesis

Notation check: Here is a quick summary of the symbols we’ve introduced in this section.

- π represents the parameter when it is a probability
- \hat{p} represents the statistic when it is the observed proportion
- H_0 represents the null hypothesis
- H_a represents the alternative hypothesis
- n represents the sample size

Section 1.3: Alternative Measure of Strength of Evidence

Introduction

In the previous section, you learned the formal process of a *test of significance*: make a claim about the parameter of interest (through competing null and alternative hypothesis); gather and explore data; follow the 3S Strategy to calculate an observed statistic from the data, simulate a null distribution and measure the strength of evidence the observed statistic has against the null hypothesis; draw a conclusion about the null hypothesis. The p-value was introduced as a standard way to measure the strength of evidence against the null hypothesis. We found that the smaller the p-value, the stronger our evidence against the null hypothesis and in favor of the alternative hypothesis. However, if we find strong evidence against the null hypothesis, this does not mean we have proven the alternative hypothesis to be true. Similarly, if we don't have strong evidence against the null hypothesis, this does not mean we have proven the null hypothesis to be true. What we can conclude is whether or not the "by chance alone" explanation is reasonable. You also confirmed in the previous section that if an observed result is not as far away from the proportion under the null hypothesis, then it provides less evidence against the null hypothesis, meaning that the null hypothesis is one plausible explanation for the observed data. In this section, you will explore another method often used to measure how far away the observed statistic is from the parameter value conjectured by the null hypothesis.

The key to Step 4 of the Statistical Investigation Method, when we apply the 3S Strategy with one categorical variable, has been assuming some claim about the long-run probability of a particular outcome, π , and then seeing whether or not our observed sample proportion is consistent with that claim. One approach is to create the null distribution (the could-have-been simulated sample proportions assuming the null hypothesis to be true), and then use the p-value to measure how often we find a statistic at least as extreme as that of the actual research study. The p-value gives us a standard way of measuring whether or not our observed statistic fell in the tail of this distribution. In this section you will find that for many distributions there is another convenient way to measure how far the observed statistic is from the hypothesized value under the null.

Example 1.3: Heart Transplant Operations

In an article published in the *British Medical Journal* (2004), researchers Poloniecki, Sismanidis, Bland, and Jones reported that heart transplantations at St. George's Hospital in London had been suspended in September 2000, after a sudden spike in mortality rate. Of the last 10 heart transplants, 80% had resulted in deaths within 30 days of the transplant. Newspapers reported that this mortality rate was over five times the national average. Based on historical national data, the researchers used 15% as a reasonable value for comparison.

Think about it: What research question can we ask about these data? Identify the observational units, variable of interest, parameter, and statistic. What is the null hypothesis?

We would like to know whether the current underlying heart transplantation mortality rate at St. George's hospital exceeds the national rate. So the observational units are the individual heart transplantations (the sample size for the data above is 10) and the variable is whether or not the patient died.

When we conduct analyses with binary variables, we often call one of the outcomes a “success” and the other a “failure,” and then focus the analysis on the ‘success’ outcome. It is arbitrary which outcome is defined to be a success, but you need to make sure you do so consistently throughout the analysis. In many epidemiological studies, death is the outcome of interest and so ‘patient did not survive for 30 days post-operation’ is called a ‘success’ in this case!

Knowing the ‘success’ outcome, allows us to find the observed statistic which is the number of successes divided by the sample size. In this case, the observed statistic is 8 out of 10 or $\hat{p} = 0.80$, and the parameter is the actual long-run, current probability of a death within 30 days of a heart transplant operation at St. George’s. We don’t know the actual value of this probability; we have just observed a small sample of data from the heart transplantation operation process; if a different set of 10 people had been operated on, the statistic would mostly likely differ. We will use the symbol π to refer to this unknown probability.

The null hypothesis is that the current death rate at St. George’s hospital is no different from other hospitals, but the researchers want to know whether these data are convincing evidence that the death rate is actually higher at St. George’s. So we will state the hypotheses as follows:

Null hypothesis: Death rate at St. George’s is the same as the national rate (0.15).
Alternative hypothesis: Death rate at St. George’s is higher than the national rate.

Symbols

You can also write your null and alternative hypotheses like this

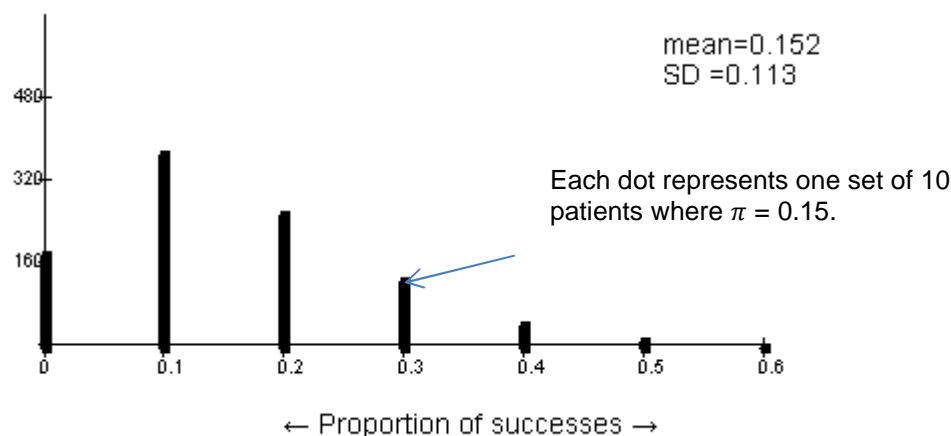
$$\begin{aligned}H_0: \pi &= 0.15 \\H_a: \pi &> 0.15\end{aligned}$$

where π is the actual long-run probability of death after a heart transplant at St. George’s.

Applying the 3S Strategy

Using the 3S Strategy from the previous section, we observed 0.80 as our statistic and now we will simulate one thousand repetitions from a process where $\pi = 0.15$ (under the null hypothesis). We did this using the One Proportion applet as shown in Figure 1.10.

Figure 1.10: Null distribution (could-have-been simulated sample proportions) for 1000 repetitions of drawing samples of 10 “patients” from a process where the probability of death is equal to 0.15. “Success” has been defined to be patient death.



Even though the sample size for this study is quite small ($n = 10$), we can see that 0.80 is not even close to the could-have-been results from the simulation. In fact, the approximate p-value is less than $1/1000$, as we never observed a value of 0.8 or larger by chance in these 1000 repetitions. This provides very strong evidence against the null hypothesis and in favor of the alternative hypothesis. Of course, it tells us nothing about *why* the death rate from heart transplantations is higher at St. George's, but it is pretty convincing that something other than random chance is at play.

Note that in Figure 1.10 the average of the 1000 proportion of successes values is 0.152, which is quite close to 0.15-- the probability of deaths if the chance model is correct. This makes sense. Also, notice that the proportions of success vary quite a lot from sample to sample, ranging from 0 to 0.6. In fact, the variability of the distribution, as measured by the standard deviation is 0.113. Remember from the Preliminaries we can think of standard deviation as the distance a typical value in the distribution is away from the mean of the distribution. In this case, 0.113 is the average distance that a simulated value of the statistic (proportion of patients who died) is from 0.152. We'll come back and use the value of the standard deviation in a moment.

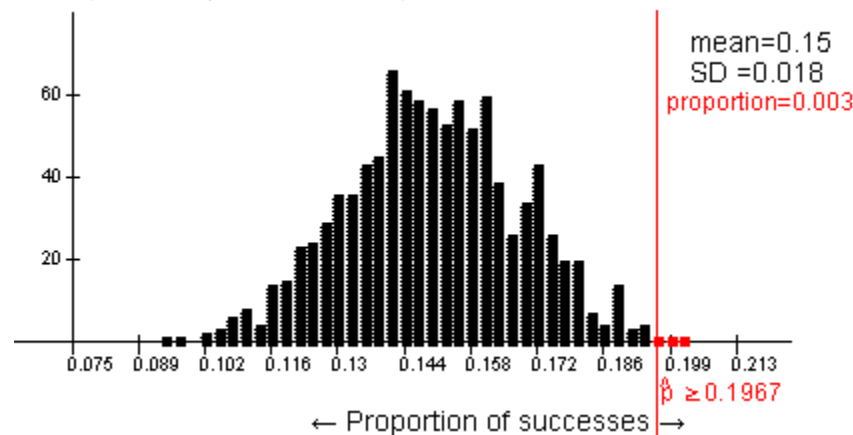
Digging deeper into the St. George's mortality data

We might still wonder whether these most recent 10 operations, which caught people's attention, are truly representative of the process as a whole. One approach is to investigate whether there were any major changes at the hospital recently (e.g., new staff, new sterilization protocols). Another is to gather more data over a longer period of time, to better represent the underlying process. So the researchers decided to examine the previous 361 heart transplantations at St. George's hospital, dating back to 1986. They found that 71 of the patients died within 30 days of the transplant.

Think about it: Now what is the value of the observed statistic? Predict how the simulated null distribution will change for this new situation. Do you think the observed statistic will still be highly statistically significant?

Figure 1.11 shows the results of 1000 repetitions of drawing samples of size $n = 361$ from a process with $\pi = 0.15$ (still assuming the null hypothesis to be true).

Figure 1.11: The null distribution of 1000 repetitions of drawing samples of 361 “patients” from a process where the probability of death is equal to 0.15



Think about it: Where does $\hat{p} = 0.197$ fall in this distribution? In the tail or among the more typical values? How does the p-value tell us whether the observed statistic is in the tail or not?

First, we note that again the p-value is small (0.003), so we still have strong evidence against the null hypothesis and in favor of the conclusion that the mortality rate at St. George's is above 0.15, but not quite as strong as in the first case. In the first case, all we could say was that the estimated p-value was less than 1 in 1000, but visually appeared to be much smaller still. While one option to more precisely quantify the p-value would be to increase the number of repetitions (say 10,000 or 100,000) we'll now look at another commonly used option.

An alternative to the p-value: Standardized value of a statistic

Using p-values is the most common way of assessing the strength of evidence by indicating the probability, under the null hypothesis, of getting a statistic as or more extreme than the one observed. Another way to measure strength of evidence is to **standardize** the observed statistic by measuring how far it is from the mean of the distribution using standard deviation units. (Common notation: z .)

Definition: To **standardize** a statistic, compute the distance of the observed statistic from the (hypothesized) mean of the null distribution and divide by the standard deviation of the null distribution.

$$\text{standardized statistic} = z = \frac{\text{statistic} - \text{mean of null distribution}}{\text{standard deviation of null distribution}}$$

For the second study, a standardized value of the statistic would result in the following calculation since Figure 1.11 tells us that the standard deviation of the null distribution is 0.018 and we know that the mean of the null distribution will be 0.15:

$$\text{standardized statistic} = z = \frac{0.197 - 0.15}{0.018} = 2.61.$$

So we would say that the observed statistic (0.197) falls 2.61 standard deviations above the mean of the distribution (0.15).

Key idea: Observations that fall more than 2 or 3 standard deviations from the mean can be considered in the tail of the distribution.

Because the observed statistic is more than 2 standard deviations above the mean, we know the statistic is in the tail of the distribution.

We can also apply this approach to the original sample of 10 patients. In that simulation, the standard deviation of the null distribution was 0.113 with a mean of 0.15. Thus, the standardized value of the statistic is computed as:

$$\text{standardized statistic} = z = \frac{0.800 - 0.15}{0.113} = 5.73$$

Thus, $\hat{p} = 0.800$ is 5.73 standard deviations above the mean of 0.15. This is another way of saying that 0.800 would be extremely unlikely to happen by chance alone if the true long-run, current mortality rate was 0.150.

There are guidelines for assessing the strength of the evidence against the null hypothesis based on the standardized value, as given next.

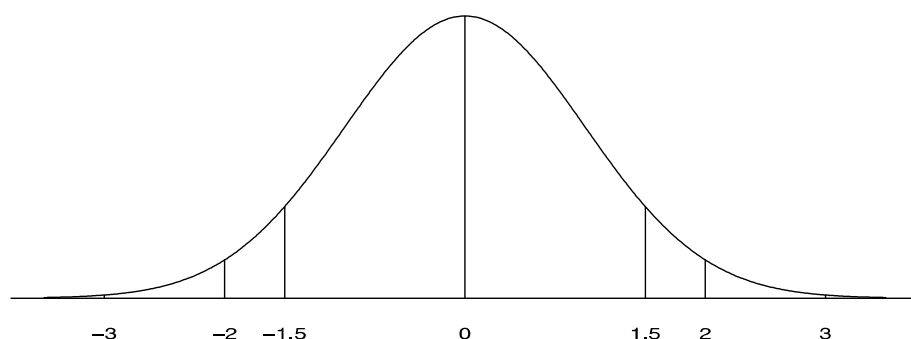
Guidelines for evaluating strength of evidence from standardized values of statistics

Standardizing gives us a quick, informal way to evaluate the strength of evidence against the null hypothesis:

between -1.5 and 1.5:	little or no evidence against the null hypothesis;
below -1.5 or above 1.5:	moderate evidence against the null hypothesis;
below -2 or above 2:	strong evidence against the null hypothesis;
below -3 or above 3:	very strong evidence against the null hypothesis.

Figure 1.12 illustrates the basis for using a standardized statistic to assess strength of evidence.

Figure 1.12: Extreme values of the bell-shaped curve for a standardized statistic



Notice that the farther from zero the standardized statistic is, the stronger the evidence against the null hypothesis. Like the p-value, we can directly compare the standardized values across datasets. We see the stronger evidence from the small dataset compared to the larger dataset ($5.73 > 2.61$).

Exploration 1.3: Do People Use Facial Prototyping?

A study in *Psychonomic Bulletin and Review* (Lea, Thomas, Lamkin, & Bell, 2007) presented evidence that “people use facial prototypes when they encounter different names.” Participants were given two faces and asked to identify which one was Tim and which one was Bob. The researchers wrote that their participants “overwhelmingly agreed” on which face belonged to Tim and which face belonged to Bob, but did not provide the exact results of their study.

Step 1: Ask a research question. We will gather data from your class to investigate the research question of whether students have a tendency to associate certain facial features with a name.

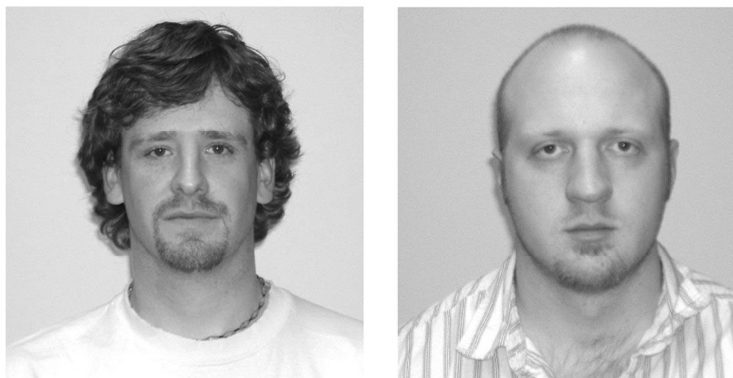
Step 2: Design a study and collect data. Each student in your class will be shown the same two pictures of men’s faces used in the research study. You will be asked to assign the name Bob to one photo and the name Tim to the other. Each student will then submit the name that he or she assigned to the picture on the left. Then the name that the researchers identify with the face on the left will be revealed.

1. Identify the observational units in this study.
2. Identify the variable. Is the variable categorical or quantitative?

The parameter of interest here is the long-run probability that a student in your class would assign the same name to the face on the left.

3. State the null and alternative hypotheses to be tested when the data are collected. Express these both in words and symbols. (*Hint:* Think about the parameter and the research question of interest here.)

Consider these two photos:



4. Do you think the face on the left is Bob or Tim? Collect the responses (data) for all the students in your class.

Step 3: Explore the data.

5. How many students put Tim as the name on the left? How many students participated in this study (sample size)? What proportion put Tim's name on the left?

Number (Tim on left):

Sample size:

Proportion (Tim on left):

When we conduct analyses with binary variables, we often call one of the outcomes a “success” and the other a “failure,” and then focus the analysis on the ‘success’ outcome. It is arbitrary which outcome is defined to be a success, but you need to make sure you do so consistently throughout the analysis. In this case we’ll call “Tim on left” a success because that’s what previous studies have found to be a popular choice.

Step 4: Draw inferences. You will use the **One Proportion** applet to investigate how surprising the observed class statistic would be, if students were just randomly selecting which name to put with which face.

6. Before you use the applet, indicate what you will enter for the following values:
 - a. Probability of success:
 - b. Sample size:
 - c. Number of repetitions:

7. Conduct this simulation analysis. Make sure the “Proportion of heads” button is selected in the applet and not the “Number of heads.”
 - a. Indicate how to calculate the approximate p-value (count the number of simulated statistics that equal ____ or _____).
 - b. Report the approximate p-value.
 - c. Use the p-value to evaluate the strength of evidence provided by the sample data against the null hypothesis, in favor of the alternative that students really do tend to assign the name Tim (as the researchers predicted) to the face on the left.

The p-value is the most common way to evaluate strength of evidence, but now we will explore a common alternative way to evaluate strength of evidence. The goal of any measure of strength of evidence is to use a number to assess whether the observed statistic falls in the tail of the null distribution (and is therefore surprising if the null hypothesis is true) or among the typical values we see when the null hypothesis is true.

8. Check the **Summary Stats** box in the applet.
 - a. Report the mean (average) value of the simulated statistics.
 - b. Explain why it makes sense that this mean is close to 0.5.
 - c. Report the standard deviation (SD) of the simulated statistics.
 - d. Report (again) the observed class value of the statistic. (What proportion of students in your class put Tim’s name on the left?)

$$\hat{p} =$$
 - e. Calculate how many standard deviations the observed class value of the statistic is from the hypothesized mean of the null distribution, 0.5. In other words, subtract the 0.5 from the observed value, and then divide by the standard deviation. In still other words, calculate:
 (observed statistic (\hat{p}) – 0.5) / SD of null distribution.

Your calculation in #8e is called standardizing the statistic. It is telling us how far above the mean the observed statistic is in terms of the ‘how many standard deviations.’

Definition: To **standardize** a statistic, compute the distance of the statistic from the (hypothesized) mean of the null distribution and divide by the standard deviation of the null distribution.

$$\text{standardized statistic} = z = \frac{\text{statistic} - \text{mean of null distribution}}{\text{standard deviation of null distribution}}$$

Once you calculate this value, you interpret it as “how many standard deviations the observed statistic falls from the hypothesized parameter value.”

The next question is how to evaluate strength of evidence against the null hypothesis based on a standardized value. Here are some guidelines:

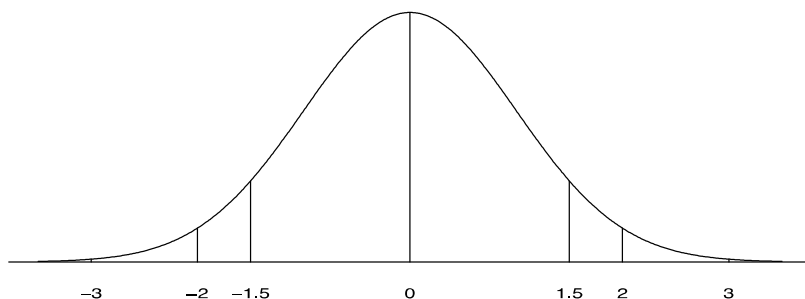
Guidelines for evaluating strength of evidence from standardized values of statistics

Standardizing gives us a quick, informal way to evaluate the strength of evidence against the null hypothesis:

between -1.5 and 1.5: **little or no** evidence against the null hypothesis;
 below -1.5 or above 1.5: **moderate** evidence against the null hypothesis;
 below -2 or above 2: **strong** evidence against the null hypothesis;
 below -3 or above 3: **very strong** evidence against the null hypothesis.

The diagram in Figure 1.13 illustrates the basis for using a standardized statistic to assess strength of evidence.

Figure 1.13: Extreme values of the bell-shaped curve for a standardized statistic



The figure can be summarized by the following key idea.

Key idea: Observations that fall more than 2 or 3 standard deviations from the mean can be considered in the tail of the distribution.

Step 5: Formulate conclusions.

9. Let's examine the strength of evidence against the null.
 - a. Based on the value of the standardized statistic, z , in #8e and the guidelines shown above, how much evidence do the class data provide against the null hypothesis?

- b. How closely does your evaluation of strength of evidence based on the standardized statistic compare to the strength of evidence based on the p-value in #7c?

Now, let's step back a bit further and think about the scope of inference. We have found that in most classes, the observed data provide strong evidence that students do better than random guessing which face is Tim's and which is Bob's. In that case, do you think that most students at your school would agree on which face is Tim's? Do you think this means that most people can agree on which face belongs to Tim? Furthermore, does this mean that all people do ascribe to the same facial prototyping?

Step 6: Look back and ahead.

10. Based on the limitations of this study, suggest a new research question that you would investigate next.

Extensions

11. In #5 you recorded the proportion of students in your class who put Tim's name with the photo on the left. Imagine that the proportion was actually larger than that (e.g., if your class was 60%, imagine it was 70%).

- a. How would this have affected the p-value:

Larger Same Smaller

- b. How would this have affected the standardized value:

Larger Same Smaller

- c. How would this have affected the strength of evidence against the null hypothesis:

Stronger Same Weaker

12. Suppose that less than half of the students in your class had put Tim's name on the left, so your class result was in the opposite direction of the research conjecture and the alternative hypothesis.

- a. What can you say about the standardized value of the statistic in this case? Explain. (*Hint:* You cannot give a value for the standardized statistic, but you can say something specific about its value.)

- b. What can you say about the strength of evidence against the null hypothesis and in favor of the alternative hypothesis in this case?

Section 1.3 Summary

In addition to the p-value, a second way to evaluate strength of evidence numerically is to calculate a **standardized statistic**:

- Standardized statistic: $\frac{\text{statistic} - \text{mean of null distribution}}{\text{standard deviation of null distribution}}$

A standardized statistic provides an alternative to p-value for measuring how far an observed statistics falls in the tail of the null distribution. More specifically:

- A standardized statistic indicates how many standard deviations the observed value of the statistic is above or below the hypothesized process probability.
- Larger values of the standardized statistic (in absolute value) indicate stronger evidence against the null model.
- Values of a standardized statistic greater than 2 or less than -2 indicate strong evidence against the null model; values greater than 3 or less than -3 indicate very strong evidence against the null model.

Section 1.4: What Impacts Strength of Evidence?

Introduction

When we are conducting a test of significance for a single proportion we assume the null hypothesis is true (or that the long-run probability equals some number) and then determine how unlikely it would be to get a sample proportion that is as far away (or farther) from the probability assumed in the null hypothesis. The p-value and standardized scores are measures of how unlikely this is. Small p-values and large standardized scores (in absolute value) give us strong evidence against the null.

In this section we will explore some of the factors that affect the strength of evidence. You should have already seen that as the sample proportion moves farther away from the probability in the null hypothesis, we get more evidence against the null. We will review this factor and explore two more. We will see how sample size and what are called “two-sided tests” affect strength of evidence.

Example 1.4: Predicting Elections from Faces?

Do voters make judgments about a political candidate based on his/her facial appearance? Can you correctly predict the outcome of an election, more often than not, simply by choosing the candidate whose face is judged to be more competent-looking? Researchers investigated this question in a study published in *Science* (Todorov, Mandisodka, Goren, & Hall, 2005). Participants were shown pictures of two candidates and asked who has the more competent-looking face. Researchers then predicted the winner to be the candidate whose face was judged to look more competent by most of the participants. In particular the researchers predicted the outcomes of the 32 US Senate races in 2004.

Think about it: What are the observational units? What is the variable measured? Is the variable categorical or quantitative? What is the null hypothesis?

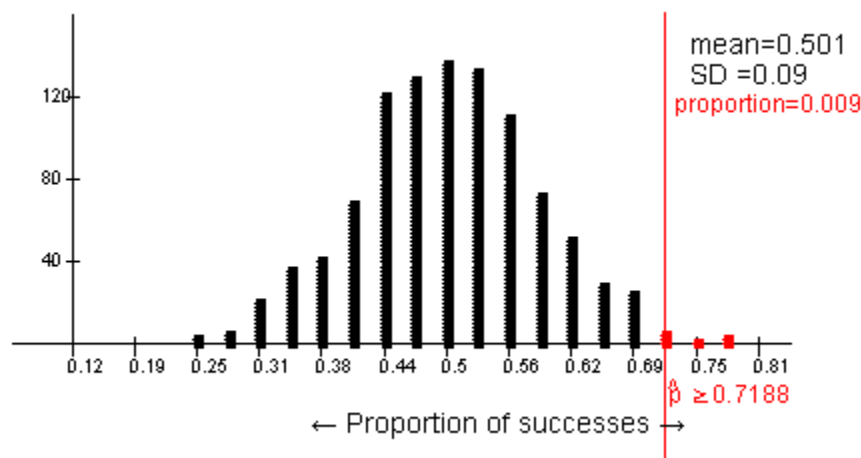
The observational units are the 32 Senate races and the variable is whether or not this method correctly predicted the winner – a categorical variable. Because we are looking for evidence that this method works better than guessing we will state:

Null hypothesis: The probability this method predicts the winner between the two candidates equals 0.5 ($\pi = 0.5$)

Alternative hypothesis: The probability this method predicts the winner is greater than 0.5 ($\pi > 0.5$).

The researchers found the competent face method of predicting election outcomes to be successful in 23 of the 32 Senate races. Thus the observed statistic, or observed proportion of correct predictions is $23/32 \approx 0.719$, or 71.9%. We can use simulation to investigate whether this provides us with strong enough evidence against the null to conclude the competent face method is better than randomly choosing the winner. Figure 1.14 displays the results of 1000 simulated sets of 32 races, assuming that the “competent face” method is no better than flipping a coin, using the One Proportion applet.

Figure 1.14: The results of 1000 sets of 32 coin tosses from a process with probability 0.5 of success indicating which results are at least as extreme as 23 ($\hat{p} \geq 0.7188$)



Only 9 of these 1000 simulated sets of 32 races show 23 or more correct predictions (or a proportion of $23/32 = 0.7188$ or more), so the approximate p-value is 0.009. This p-value is small enough to provide strong evidence against the null hypothesis, in favor of concluding that the “competent face” method makes the correct prediction more than half the time in the long run. Alternatively, we can compute the standardized value of the statistic as

$$\text{standardized statistic} = \frac{0.7188 - 0.500}{0.09} = 2.43.$$

Thus, the observed statistic is 2.43 standard deviations away from hypothesized parameter value (the mean of the null distribution) specified by the chance model, and so this (being larger than 2), again, confirms the observation is in the tail of the null distribution, so there is strong evidence that the chance model is wrong.

What impacts strength of evidence?

In the previous two sections we’ve looked at two measures of the strength of evidence: p-value and standardized statistic, however we’ve not yet formally looked at what factors *impact* the strength of evidence. In other words, why is the strength of evidence (measured by p-value or standardized statistic) sometimes strong and sometimes weak or non-existent? We will look at three factors that impact the strength of evidence: the difference between the observed statistic (\hat{p}) and null hypothesis parameter value, the sample size, and whether we do a one or two-sided test.

1. Difference between statistic and null hypothesis parameter value

Think about it: What if instead of 23 correct predictions out of 32, the researchers were able to correctly predict 26 elections? Or, what if they only correctly predicted 20 elections? How would the number of correct predictions in the sample impact our strength of evidence against the null

hypothesis?

Intuitively, the more extreme the observed statistic, the more evidence there is against the null hypothesis. But, let's be a bit more precise.

If the researchers correctly predicted 26 elections, that is a success rate of $\hat{p} = 26/32 = 0.8125$, which is farther away from what would occur, in the long-run, if they were just guessing (0.50). Back in Figure 1.14 you can see that a value of 0.81 or larger never occurs just by chance, approximating a p-value < 0.001 . Similarly, the standardized statistic would be 3.47 (meaning 0.8125 is 3.47 standard deviations above the mean). In short, if researchers correctly predict 26 elections, this is extremely strong evidence against the null hypothesis because the observed statistic is farther out in the tail of the null distribution.

On the other hand, if the researchers correctly predicted only 20 elections, the success rate drops to $\hat{p} = 20/32 = 0.625$. In this case, the observed statistic (0.625) is something that is fairly likely to happen just by chance if the researchers were guessing who would win the election. The p-value increases to 0.115 (there were 115 out of 1000 times that 62.5% or more correct predictions occurred by chance), and the standardized statistic is closer to zero $[(0.625 - 0.50)/0.09=1.39]$, suggesting that if the researchers correctly predicted 20 of the 32 elections, there is little evidence that the researchers' method performs better (in the long run) than guessing.

Key idea: The farther away the observed statistic is from the average value of the null distribution, the more evidence there is against the null hypothesis.

2. Sample size

Think about it: Do you think that increasing the sample size will increase the strength of evidence, decrease the strength of evidence, or have no impact on the strength of evidence against the null hypothesis (assuming that both the value of the observed statistic and the chance model do not change)?

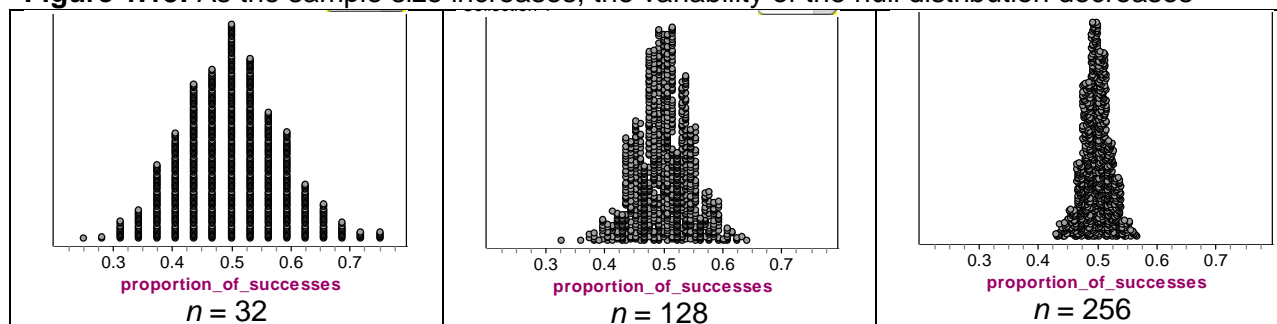
Intuitively, it seems reasonable to think that as we increase the sample size, the strength of evidence will increase. Observing Doris and Buzz longer, playing rock paper scissors longer and looking at St. George's heart transplant patients outside of the initial set of 10 patients all intuitively suggest we'll have more knowledge about the truth. Let's dig a bit deeper into this intuition.

In Exploration P.3, we looked at how probability is the long-run proportion of times an outcome occurs. The key there is long-run. We realize that in the short-term, we expect more chance variability than in the long-term. For example, long-term investment strategies are typically good (variability is reduced). Whereas if you flip a coin just 3 times and get heads each time, you aren't convinced the coin isn't fair because with such a small sample size, almost "anything can happen."

In terms of statistical inference, our sample size (number of observational units) is what dictates how precise a measure of the parameter we have. In Figure 1.15, we can see how the null

distribution changes as we increase the sample size from 32 senate races to 128 races (4 times as many) or 512 races (16 times as many).

Figure 1.15: As the sample size increases, the variability of the null distribution decreases



In each case the distribution is centered at 0.50, the null hypothesis value. What, then, is changing? What's changing is the variability of the distribution. With 32 races the variability of the distribution, as measured by the standard deviation is 0.088, with 128 races, it is 0.043, and with 256 races it is only 0.031. You can see this visually by noting that the distributions are getting squeezed closer and closer to the null hypothesis value. Stated yet another way, there is less sample to sample variability in the sample proportions as the sample size gets bigger.

What does this decrease in variability mean in terms of statistical significance? Remember from our observed data that in 71.9% of elections the competent face method predicted the winning candidate. This is strong evidence ($p\text{-value} = 0.009$; standardized statistic = 2.43) that the competent face method does better than guessing. But, what if the sample size was 128 elections and the competent face method correctly predicted 71.9% of the races? Now the $p\text{-value}$ would be < 0.001 and the standardized statistic would be $z = 5.09$ —the strength of evidence against the chance model of just guessing. The strength of evidence against the null hypothesis increases even more if the competent face method can get 71.9% correct in 256 elections ($p < 0.001$; standardized statistic = 7.06).

Key idea: As the sample size increases (and the value of the observed sample statistic stays the same) the strength of evidence against the null hypothesis increases.

Two more quick points: (1) If you are trying to pass a true/false test (let's say get a 60% or higher) but know NOTHING about what is going to be on the test, would you rather have more questions or fewer questions on the test? You, the student, would rather have fewer questions. If there was only one question on the test you'd have a 50% chance of passing! The teacher would rather have more questions on the test because the more questions they have on the test the more likely your outcome on the test will be close to 50% (just guessing!) and the less likely you would be to "get lucky" and pass the test by just guessing; (2) Importantly, we can't automatically assume that if we have Doris and Buzz do more trials or have your friend play more rounds of rock-paper-scissors or collect more data at St. George's Hospital that the strength of evidence will increase (smaller $p\text{-value}$, larger standardized statistic). Why not? Because when we collect more data our observed statistic will almost always change as well. If we have Doris and Buzz do more trials, they won't always get exactly 93.75% correct, your friend may not always throw scissors in exactly $1/6^{\text{th}}$ of the rounds played, and the other heart transplant patients may not have 80% mortality after 30 days.

3. One sided versus two-sided tests

What if the researchers were wrong, and instead of the competent person being elected more frequently, it was actually the less competent person who was more likely to win the election?

Currently, as we've stated our null and alternative hypotheses, we haven't allowed for this possibility. The null hypothesis says that the competent face method predicts the winner 50% of the time, the alternative hypothesis says greater than 50%, but less than 50% doesn't appear.

Null hypothesis: The probability this method predicts the winner equals 0.5.

Alternative hypothesis: The probability this method predicts the winner is greater than 0.5.

These hypotheses can be written in symbols as:

$$H_0: \pi = 0.5$$

$$H_a: \pi > 0.5, \text{ where } \pi = \text{the probability this method predicts the correct winner}$$

This type of alternative hypothesis is called "one-sided" because it only looks at one of the two possible ways that the null hypothesis could be wrong. In this case, it only considers that the null hypothesis could be wrong if the probability is more than 0.50. Many researchers consider this way of formulating the alternative hypothesis to be too narrow and too biased towards assuming the researchers are correct ahead of time. Instead, a more objective approach is to conduct a two-sided test, which can be formulated as follows.

Null hypothesis: The probability this method predicts the winner equals 0.50.

Alternative hypothesis: The probability this method predicts the winner is not 0.50.

These hypotheses can be written in symbols as:

$$H_0: \pi = 0.50$$

$$H_a: \pi \neq 0.50$$

In this case, the alternative hypothesis states that the probability the competent face method predicts the winner is not 0.50—might be more, might be less. This change to the alternative hypothesis, however, has ramifications on the rest of the analysis. Recall that p-values are computed as the probability under the null of obtaining a value that is equal to or more extreme than your observed statistic, where more extreme goes in the direction of the alternative hypothesis (greater than **or** less than). In the case of a **two-sided test**, more extreme must go in both directions. The way this is operationalized is that the p-value is computed by finding out how frequently the observed statistic or more extreme occurred in one tail of the distribution, and adding that to the corresponding probability of being at least as extreme in the other direction - in the other tail of the null distribution.

For example, how often does 0.7188 or more occur by chance? Because 0.7188 is 0.2188 *above* 0.5, we need to look at what number is 0.2188 *below* 0.5. Calculating $0.5 - 0.2188 = 0.2812$, we need to also look at the proportion of outcomes at or below 0.2812. (See Figure 1.16.)

Figure 1.16: When determining a two-sided p-value we can look at outcomes in both tails for the null distribution that are the same distance above or below the mean

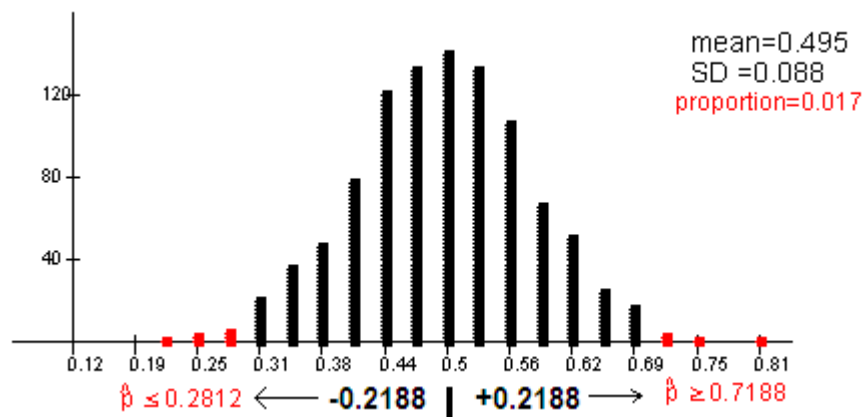


Figure 1.16 illustrates how a two-sided p-value is computed. In this case 0.7188 or greater was obtained 9 times by chance, and 0.2812 or less was obtained 8 times by chance. Thus, the two-sided p-value is approximately $(8 + 9 = 17)/1000 = 0.017$. Two-sided tests always increase the p-value (approximately doubling it from a one-sided test), and thus, using two-sided tests always decreases the strength of evidence. Two-sided tests are more conservative and are used most of time in scientific practice. However, because of their objectivity they require even stronger results before the observed statistic will be statistically significant.

Key idea: Because the p-value for a two-sided test is about twice as large as that for a one-sided test, they provide less strength of evidence. However two-sided tests are used more often in scientific practice.

Technical note: When the hypothesized probability is something other than 0.5, there are actually a couple of different ways of defining an observation as being more extreme. The One Proportion applet determines an outcome is more extreme if it has a smaller individual probability than that of the observed outcome. In other words, if an observation is more unlikely, then it is considered more extreme. See FAQ 1.4.1 for more on one-sided vs. two-sided p-values.

Follow-up Study

As a way of applying what we just learned, consider the following. The researchers investigating the competent face method, also predicted the outcomes of 279 races for the U.S. House of Representatives in 2004, looking for whether the probability the “competent face” method predicts the correct winner (π) is different from 0.50. In these 279 races, the method correctly predicted the winner in 189 of the races, which is a proportion of $\hat{p} = 189/279 \approx 0.677$, or 67.7% of the 279 House races. Notice this sample percentage of 67.7% is bit smaller than the 71.9% correct predictions in 32 Senate races, however the sample size (279 instead of 32) is larger. We are also now considering a two-sided alternative hypothesis.

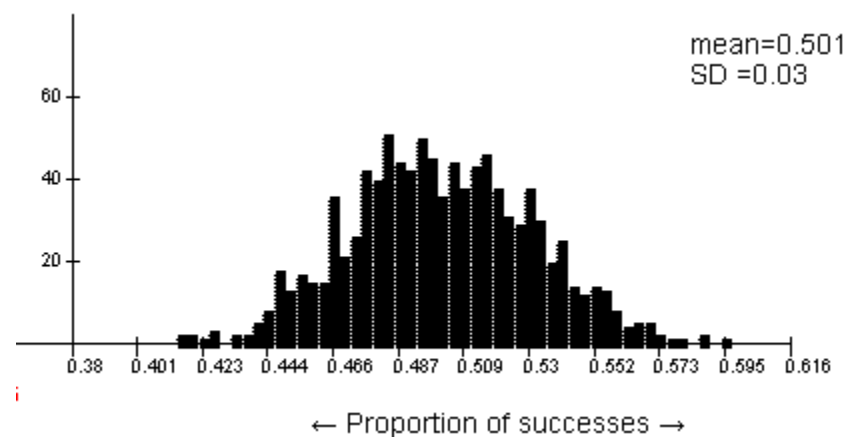
Think about it: Do you expect the strength of evidence for the “competent face” method to be stronger for the House results, weaker for the House results, or essentially the same for the House results as compared to the Senate results? Why?

Let’s take the three factors separately:

1. **Distance of the observed statistic to the null hypothesis value.** In the new study the observed statistic is 0.677, whereas in the original study it was 0.719—this is a small change, closer to the hypothesized value of the parameter, which will slightly decrease the strength of evidence against the null hypothesis.
2. **Sample size.** The sample size is almost 10 times as large (279 vs. 32) in the new study which will have a large impact on the strength of evidence against the null hypothesis. This will increase the strength of evidence quite a bit, because the observed statistic didn't change much.
3. **One- or two-sided test.** When we looked at the Senate races, we obtained a p-value of 0.009 with a one-sided test and 0.017 with a two-sided test. Because we are asked to use a two-sided test here that means we will obtain a p-value about twice as large as what we would obtain with a one-sided test.

So, what's the p-value? Figure 1.17 shows the null distribution for a sample size of 279 with 0.5 as the probability under the null hypothesis.

Figure 1.17: A null distribution for the Senate sample of 279 races



Notice that a proportion of 0.677 or larger does not occur just by chance in the null distribution. Neither do any values of 0.323 or smaller. (We need to look down there ($0.677 - 0.500 \approx 0.177$, $0.500 - 0.177 \approx 0.323$) because this is a two-sided test.) Thus the p-value is approximately zero and the evidence against the null hypothesis in the House of Representatives sample is stronger than in the Senate sample. This is mainly due to the increased sample size and the fact that the observed sample statistic didn't change too much.

Looking at the standardized statistics, for the Senate races, we find $z = (0.719 - 0.500)/0.088 \approx 2.49$ and for the House of Representatives we find $z = (0.677 - 0.500)/0.03 \approx 5.90$. We see that the numerator is a bit smaller in the second case, but the denominator (the standard deviation of the null distribution) is much smaller, providing much stronger evidence against the null hypothesis.

FAQ 1.4.1: When and why do we use two-sided alternative hypotheses?

Q: Why would we use a two-sided alternative?

A: Two-sided alternative hypotheses are common in scientific practice when researchers don't make a prior assumption about the direction of the outcome in the study. This can happen if the researchers don't have a prior suspicion about the direction, or if they are interested in finding out whether the parameter (in this case, the long-run probability, π) differs from the value hypothesized in the null hypothesis. In contrast, to use a one-sided test, one needs to have prior suspicion that the outcome will be in a certain direction.

Q: Why don't we look at the direction of the outcome in the data collected and state our alternative hypothesis to match that?

A: Because that would be considered cheating. Remember that hypotheses are always statements about the parameter(s) of interest, and are stated before we begin to explore the data. We use the data to test our hypotheses, not to form them.

Q: So, should we always do two-sided tests? Are there any drawbacks to using two-sided tests?

A: The reality is that most scientists do use two-sided tests the majority of the time. There are, however, downsides to using two-sided tests. The first is related to the size of the p-value. When we compute a p-value for a two-sided test, the p-value is approximately twice the size it would be had we computed it for the one-sided test. This means that two-sided tests always yield larger p-values than one-sided tests, making it harder to find strong evidence in support of your research conjecture. So, in some cases scientists will do one-sided tests if they are focused on only one type of result (and have prior evidence that that is the direction of the effect or that's the only direction of interest). For example, in some of the examples we've looked at in previous chapters (e.g., Doris and Buzz), we studied whether the dolphins could do better than just guessing, it makes less sense to study whether the dolphins did worse than just guessing (this would mean that they were purposely pushing the wrong button!). In these cases a one-sided test makes the most sense.

Q: And the second drawback?

A: The second drawback to two-sided tests is that the two-sided p-value is the strength of evidence for a difference, but does not tell us much about *how* (bigger or smaller?) the parameter (in this case, π) differs from the value hypothesized in the null hypothesis.

Exploration 1.4: Competitive Advantage to Uniform Colors?

In this exploration, we are going to explore three factors that influence the strength of evidence in a test of significance: (1) The difference between the observed sample statistic and the value of the parameter used in the null hypothesis; (2) The sample size; (3) One-sided tests versus two-sided tests. To do this we will look at a study conducted by Hill and Barton (*Nature*, 2005) to investigate whether Olympic athletes in certain uniform colors have an advantage over their competitors. They noticed that competitors in the combat sports of boxing, tae kwon do, Greco-Roman wrestling, and freestyle wrestling are randomly assigned red or blue uniforms. For each match in the 2004 Olympics, they recorded the uniform color of the winner.

Think about it: What are the observational units? What is the variable measured? Is the variable categorical or quantitative?

The observational units in this study are the matches, and the variable is whether the match was won by someone wearing red or someone wearing blue – a categorical variable. Let's suppose that going into this study, the researchers wanted to see whether the color red had an advantage over blue. In other words, competitors that wear red uniforms will win a majority of the time.

1. State the null and the alternative hypotheses in words.
2. We will let π represent the probability that a competitor wearing a red uniform wins. Using this, restate the hypotheses using symbols.
3. Researchers Hill and Barton used data collected on the results of 457 matches and found that the competitor wearing red won 248 times, whereas the competitor wearing blue won 209 times. We will carry out a simulation to assess whether or not the observed data provide evidence in support of the research conjecture. This simulation will employ the **3S Strategy**: Determine the **statistic**, **simulate** could-have-been outcomes of the statistic under the null model, and assess the **strength** of evidence against the null model by estimating the p-value.
 - a. What is the statistic we will use? Calculate the observed value of the statistic in this study.
 - b. Describe how you could use a coin to develop a null distribution to test our hypothesis.
 - c. Use the **One Proportion** applet to test our hypothesis. Based on your simulation, find the p-value and write a conclusion. Also write down the mean and standard deviation from your null distribution when the proportion of successes is used for the variable on the horizontal axis. You will need this later.

One sided vs. two-sided tests

One factor that influences strength of evidence is whether we conduct a two-sided or a one-sided test. Up until now we have only done one-sided tests. In a one-sided test the alternative hypothesis is either $>$ or $<$. In a two-sided test, the alternative hypothesis is \neq . So, why would you do a two-sided test and what are the implications?

Suppose the researchers did not necessarily think that red would win more often, but they also didn't necessarily think that blue would win more often. They were just interested in whether one color would win more often than the other. A two sided alternative hypothesis (red wins at a rate other than 50% of the time (or at a rate not equal to 50%)) allows the researchers to be less sure of the anticipated location of the parameter than a one-sided test.

4. If we let π equal the probability that a competitor wearing a red uniform wins, state the hypotheses for this study in symbols using a two-sided alternative.
5. Return to the **One Proportion** applet to approximate the p-value for our original overall proportion of red winning 248 times out of 457 matches, but now select the "Two-sided" check box to find the "two-sided p-value."
 - a. Describe how the portion of the null distribution that is shaded red is different than our first test done in #3c.
 - b. Describe how the p-value is different from the p-value that was obtained in our original test done in #3c.
 - c. To find the two-sided p-value, the applet is looking to see how often 0.543 or larger occurs and how often the comparable value on the other side of the null distribution (or smaller) occurs. To find this value, first compute how far 0.543 is from the center of the null distribution, and then go that same distance to the left (less than) the center of the null distribution. What is the comparable value?
 - d. Complete the following sentence: The two-sided p-value of _____ is the probability of obtaining _____ or larger plus the probability of obtaining _____ or smaller if the _____ is true.
 - e. You should have seen that when the alternative hypothesis is *two-sided*, the p-value is computed by looking at how extreme the observed data is in *both* tails on the null distribution. This makes the p-value about twice as large. Because of this, explain how switching from a one-sided test to a two-sided influences the strength of evidence against the null.

Key idea: Because the p-value for a two-sided test is about twice as large as that for a one-sided test, they provide less strength of evidence. However two-sided tests are used more often in scientific practice.

Difference between statistic and null hypothesis parameter value

6. A second factor that influences the strength of evidence against the null is how far apart the observed sample statistic and the value of the parameter specified under the null hypothesis are. For this study the null value was 0.5 and the observed sample statistic was about 0.543 (or 54.3% of the competitors wearing red won their matches). Suppose a larger proportion of competitors wearing red won their matches. If fact, suppose 57% of the 457 matches were won by a competitor wearing red.
 - a. Go back to the **One Proportion** applet and approximate the (one-sided) p-value for this situation where again we are testing to see whether the overall probability of winning is more than 0.5.
 - b. Is your p-value larger or smaller than your original one? Explain why this makes sense.
 - c. Write a sentence explaining the relationship between the distance between the observed sample statistic and the value of the parameter specified under the null hypothesis to the strength of evidence against the null hypothesis.

Key idea: The farther away the observed statistic is from the average value of the null distribution, the more evidence there is against the null hypothesis.

Sample size

7. The third factor we will look at that influences strength of evidence against the null hypothesis is the sample size. As we said earlier, the data for this study came from four combat sports in the 2004 Olympics. One of those sports was boxing. The researchers found that out of the 272 boxing matches, 150 of them were won by competitors wearing red. This proportion of $150/272 \approx 0.551$ is similar to the overall proportion of times the competitor wearing red won. Let's see what the smaller sample size does to the strength of evidence. Use the **One Proportion** applet to test the same hypotheses as we originally did, but just the boxing matches as our sample.
 - a. Compare the null distribution you generate in this case to that generated in #3. In particular, how do the center and variability compare?

- b. What is your new p-value? Is it larger or smaller than your original p-value from #3c? Explain why this makes sense.
- c. Write a sentence explaining the relationship between sample size and the strength of evidence against the null hypothesis.

Key idea: As the sample size increases (and the value of the observed sample statistic stays the same) the strength of evidence against the null hypothesis increases.

Section 1.4 Summary

Three factors impact the strength of evidence provided by sample data against a null hypothesis:

- **Two-sided alternatives/tests** are used when the researcher does not have a prior suspicion about the direction that the parameter value is from the hypothesized value.
 - A two-sided test produces a larger p-value than a one-sided test based on the same sample data.
 - Two-sided tests therefore require a higher “standard of proof” to provide convincing evidence against the null hypothesis.
 - The p-value for a two-sided test is generally twice as large as the p-value would have been from a one-sided test on the same sample data.
- The farther away the observed statistic is from the hypothesized value of the parameter, π , in the direction of the alternative hypothesis, the stronger the evidence against the null hypothesis.
- A larger sample size generally produces stronger evidence against the null hypothesis if the observed value of the statistic does not change (and if the observed result is in the direction of the alternative hypothesis).

Section 1.5: Inference on a single proportion: Theory-based approach

Introduction

The focus of this chapter has been on Step 4: Drawing Inferences. We have learned that we can draw inferences from data by comparing our observed statistic to a conjecture or claim about a long-run probability. In order to assess the strength of evidence for the claim we have

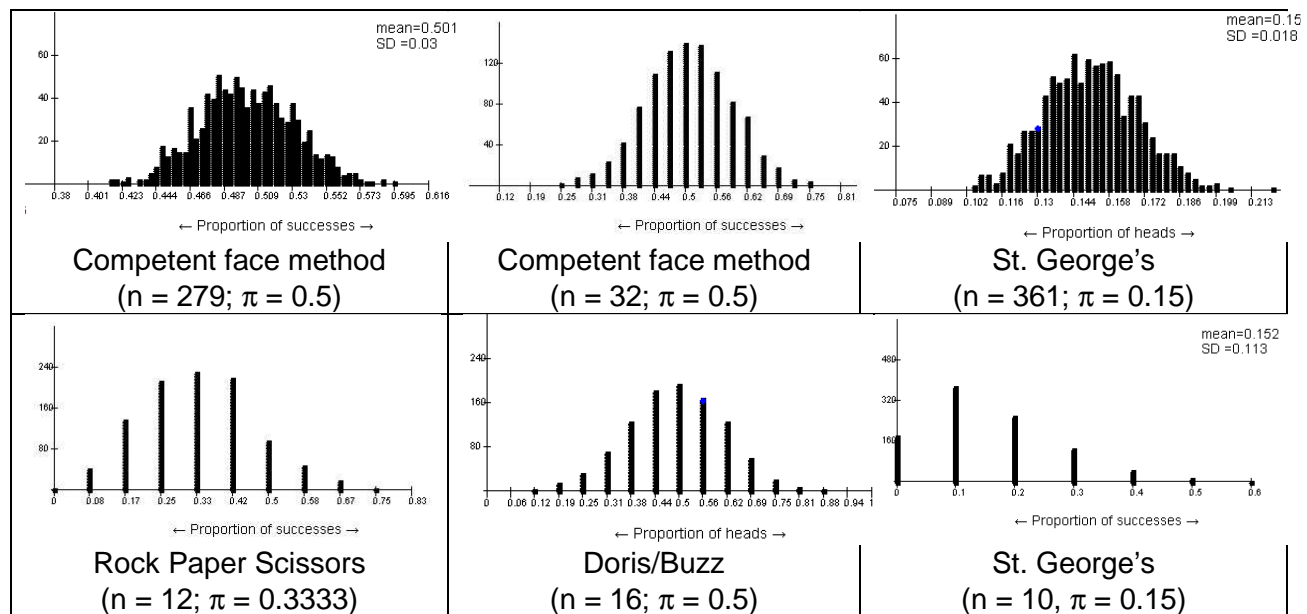
always simulated the null distribution. However, simulation takes a lot of computer power, and historically that wasn't always possible. In this section we will see how, in many cases, we can predict what will happen when you simulate, thus avoiding the need to conduct a simulation, but still providing the ability to generate either a p-value or standardized statistic to assess strength of evidence. As we'll find out in this and later sections, in addition to being necessary historically, this method (we'll call it a theory-based approach) also gives insights into and advantages in standardization (Section 1.3) and confidence intervals (the theme of Chapter 3).

At the heart of this method is the fact that the null distributions of sample proportions that we have been simulating often exhibit a common and familiar shape. In particular:

- (1) They are often, though not always, bell-shaped curves;
- (2) They are centered at the null hypothesis value for π ; and
- (3) Their variability (spread, standard deviation) is determined by the sample size.

Figure 1.18 shows a few examples of null distributions from this chapter.

Figure 1.18: Null distributions for various studies we've explored so far. A bell-shaped distribution, called a normal distribution, can be used to nicely approximate some of them, but will not work well for others.



The competent face method and St. George's ($n = 279$) simulations show bell-shaped curves. However, the Rock/Paper/Scissors study and the first sample of St. George's heart patients ($n = 10$) are not bell-shaped curves, in part because the distribution is not symmetric and in part because there are so few values that the sample proportion can be. For the Doris/Buzz dolphin study, even though the distribution of sample proportions is quite symmetric, an approximation based on a bell-shaped curve will not be very good because of the small number of possible values (i.e., because of the extreme discrete-ness, the gaps between the values) for the sample proportion in those simulations.

As we will learn in this section, in many, but not all, cases we can predict when the simulated distribution is bell-shaped (or **normally distributed**), where it will be centered, and how variable it will be. All of these predictions can be used to generate p-values and standardized statistics without simulating, in an approach called a One proportion z test., which is one example of Theory-based approach to statistical inference.

Example 1.5: Halloween Treats

Stemming from concern over the nation's obesity epidemic, researchers investigated whether children might be as tempted by toys as by candy for Halloween treats. Test households in five Connecticut neighborhoods offered children two plates: one with lollipops or fruit candy and one containing small, inexpensive Halloween toys, like plastic bugs that glow in the dark. The researchers observed the selections of 283 trick-or-treaters between the ages of 3 and 14 (Schwartz, Chen, and Brownell, 2003).

To investigate whether children show a preference for either the candy or the toys, we test the following hypotheses.

Null hypothesis: The probability a trick-or-treater would choose candy is 0.5.

Alternative hypothesis: The probability a trick-or-treater would choose candy is not 0.5.

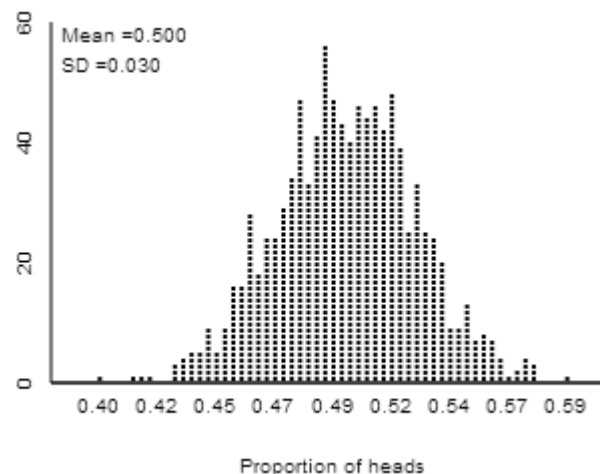
Note that our null model assumes that the probability of choosing candy (π) is the same for all children. In symbols, these hypotheses translate to

$$H_0: \pi = 0.5$$

$$H_a: \pi \neq 0.5$$

The researchers collect the reactions of 283 children for the study. With a sample size of 283, under our null hypothesis, we simulated the null distribution (using 10,000 simulated samples) shown in Figure 1.19.

Figure 1.19: A null distribution representing 10,000 simulated samples of 283 children where the probability that an individual child would choose candy is 0.5



We notice that the distribution is quite bell-shaped, the average value is 0.5 (the null hypothesis value for π) and the standard deviation is 0.030. The theory-based approach could have predicted this would happen!

Theory-based approach (One proportion z test)

In the early 1900s, and even earlier, computers weren't available to do simulations, and as people didn't want to sit around and flip coins all day long, they focused their attention on mathematical and probabilistic rules and theories that could predict what would happen if someone did simulate.

They proved the following key result (often called the **Central Limit Theorem**):

Central Limit Theorem: If the sample size (n) is large enough, the distribution of sample proportions will be bell-shaped (or normal), centered at the long-run probability (π), with a standard deviation of $\sqrt{\pi(1 - \pi)/n}$.

One bit of ambiguity in the statement is how large is large enough for the sample size? As it turns out, the larger the sample size is, the better the prediction of bell-shaped behavior in the null distribution is, but there is not a sample size where all of the sudden the prediction is good. However, some people have used the convention that you should have at least 10 successes and at least 10 failures in the sample.

Validity conditions: The normal approximation can be thought of as a prediction of what would occur if simulation was done. Many times this prediction is valid, but not always, only when the validity condition (at least 10 successes and at least 10 failures) is met.

Let's see how this prediction compares to what actually happened in the simulation (back in Figure 1.19). Because the sample size is 283 (with 148 and 135 the two categories) the prediction should work. Looking at Figure 1.19 we see that:

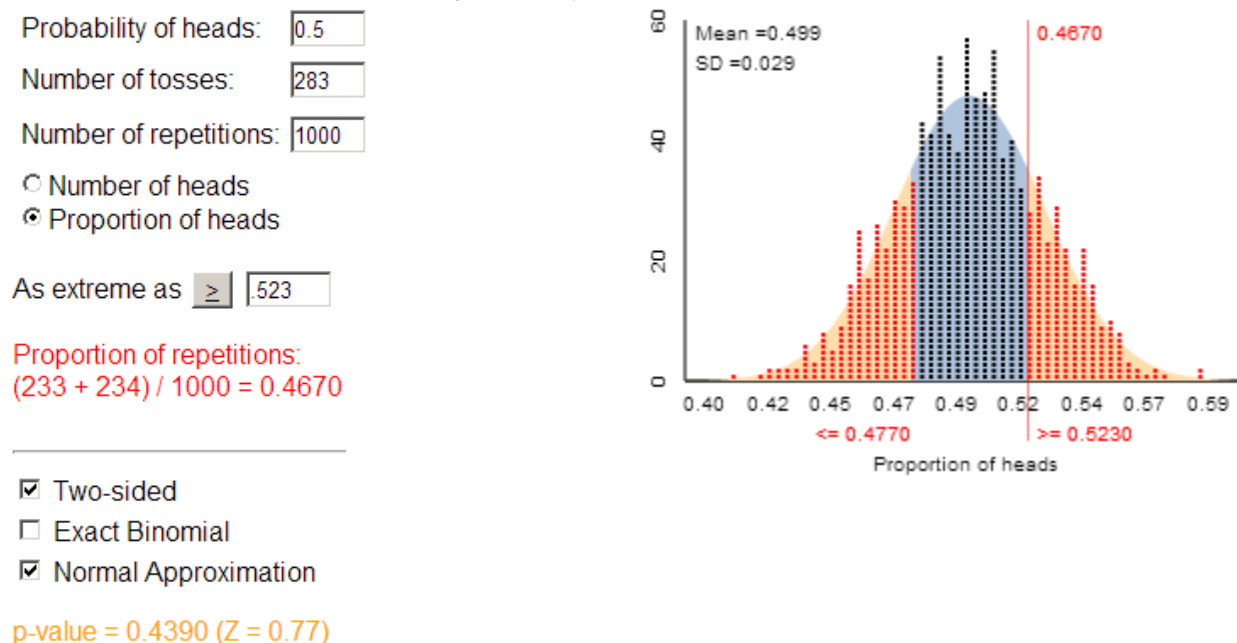
1. The simulated distribution certainly looks bell-shaped.
2. The simulated distribution is centered at 0.5, the hypothesized value of π .
3. The simulated standard deviation of the sample proportions is 0.030, which is very close to the predicted standard deviation of $\sqrt{0.5(1 - 0.5)/283} = 0.0297$.

One advantage of using this result is we can determine the standard deviation of the sample proportion without having to conduct the simulation. Knowing the standard deviation allows us to calculate the standardized statistic, z :

$$z = \frac{0.523 - 0.5}{0.0297} = 0.77.$$

This tells us that our observed sample proportion is 0.77 standard deviations above the mean of 0.5. Because this is less than one standard deviation above the mean, it gives us little evidence against the null hypothesis.

To convert this standardized score to a two-sided p-value, we could look at the area under the (standard) normal curve to the right of 0.77 and to the left of -0.77. In the One Proportion applet, checking the box for **Normal Approximation** will overlay the theoretical normal curve and give you the p-value from the normal distribution. Figure 1.20 shows the results of this applet for the Halloween treat study where in the sample of 283 trick-or-treaters, 148 (52.3%) chose candy and 135 (47.7%) chose toys.

Figure 1.20: Simulated and Theory-Based p-values for Halloween study

The normal distribution does a nice job of predicting the behavior of the null distribution of sample proportions in this case. The p-value from the theory-based test is 0.4390, compared to .4670 from the simulation. We interpret this p-value, and draw a conclusion from it, as always: If the probability each trick-or-treater prefers candy to toys is 0.5, then there's about a 44% chance that a random sample of 283 trick-or-treaters would have found 52.3% or more, or 47.7% or fewer, choosing the candy. Because this is not a small p-value, we do not have substantial evidence to suggest that trick-or-treaters prefer either type of treat. This was actually good news to the researchers, to learn that both types of treats are viable and we could probably distribute less candy at Halloween (at least in this Connecticut neighborhood).

The p-value of 0.4390 from the theory-based approach (One proportion z test) is very similar to that obtained from the simulation (see Figure 1.20) because the prediction of the shape, center, and variability of the null distribution is very good. [Note: We can improve the normal approximation by employing a "continuity correction" that would use a value just below 0.523 and just above 0.477 to include more area right at those cut-off values.]

A situation where a theory-based approach doesn't work

In Section 1.2, we looked at the Rock/Paper/Scissors game where novice players threw scissors two of the 12 rounds played.

Think about it: Why do you think the Theory-based (One proportion z-test; Normal approximation) approach will not work well for this data?

In this case, the theory-based approach is not expected to work well because the sample size (12) is small. Recall that the validity conditions for the theory-based approach state the need for at least 10 successes and 10 failures, we have 2 scissors and 10 not scissors so this condition is not met.

But, let's see what happens when we use the theory-based approach anyway. The theory based approach predicts that:

1. The null distribution will be bell-shaped and approximately normal
2. The null distribution will be centered at 0.333
3. The standard deviation of the null distribution will be $\sqrt{\frac{0.333(1-0.333)}{12}} \approx 0.136$.

Figure 1.21 shows a picture of the simulated null distribution with the theory-based normal distribution overlaid.

Figure 1.21 shows that the approximate p-value from the simulation (0.1980) is not all that similar to the theory-based p-value (0.1108). Although both distributions (theoretical and simulation) are centered at nearly 0.3333, with similar standard deviations (0.134 vs. 0.136), the distribution is not “normal” in that it is not “filled in.” In other words, the null distribution is too discrete (too much empty space between the possible observations) for it to be well-modeled by a normal distribution.

Figure 1.21: The null distribution for the Rock Paper Scissors example

Probability of success (π):

Sample size(n):

Number of samples:

☐ Number of successes

☒ Proportion of successes

As extreme as

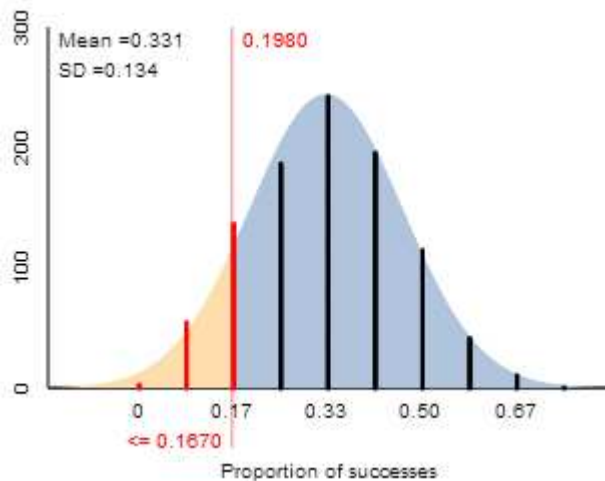
Proportion of samples:
198 / 1000 = 0.1980

☐ Two-sided

☐ Exact Binomial

☒ Normal Approximation

p-value = 0.1108 (Z = 1.22)



With the slightly larger p-value, the simulation-based approach provides a bit less evidence against the null hypothesis. Although there is not a large difference between these two p-values, in this case we would have a bit more faith in the p-value obtained from the simulation. In summary, the theory-based approach should not have been used in this case because the sample size was too small, and, thus, the end result is that the theory suggested there was more evidence against the null than there was.

Exploration 1.5: Calling Heads or Tails

When asked to call the outcome of a coin toss, are people equally likely to choose heads or tails? Let's investigate this question by collecting some data from you and your classmates.

1. What would you call: heads or tails?

Before we even collect data from your classmates, let's think about what we want to test here. Conventional wisdom says that people tend to pick heads more often than tails, so that's the research hypothesis we'll investigate.

2.
 - a. What are the observational units in this study?
 - b. What is the variable that is recorded?
 - c. Describe the parameter of interest in words. (Use the symbol π to represent this parameter.)
 - d. If people do *not* have a tendency to pick heads more often than tails (or tails more often than heads) what would you expect the numerical value of the parameter to be? Is this the null hypothesis or the alternative hypothesis?
 - e. If people *do* have a tendency to pick heads more often than tails, what can you say about the numerical value of the parameter? Is this the null hypothesis or the alternative hypothesis?
3. Including yourself and your classmates, how many people participated in this study? How many picked heads? Calculate the sample proportion that picked heads.
4. To have a larger sample size to analyze, combine your class results with the results from one of the author's classes, in which 54 of 83 students picked heads. Now what are the sample size and the sample proportion that picked heads?

Sample size:

Sample proportion:

5. Use the **One Proportion** applet to test the hypotheses from #2d and #2e.
- Describe the shape of the null distribution of sample proportions. Does this shape look familiar? Where is the null distribution centered? Does this make sense? Check the **Summary Stats** box, and report the mean and standard deviation as reported by the applet.

Shape:	Familiar?
Center?	Why does this make sense?
Mean:	SD:
 - Approximate the p-value, and summarize the strength of evidence that the sample data provide regarding the research hypothesis.
 - Determine the standardized statistic, z , and summarize the strength of evidence. Confirm that the strength of evidence obtained using the standardized statistic is similar to that obtained using the p-value.

Theory-based approach (One proportion z test)

In question 5a, you probably described the shape of the null distribution using words such as bell-shaped, symmetric, or maybe even normal. You have seen many null distributions in this chapter that have had this same basic shape. You should have also noticed that the null distributions have all been centered at the hypothesized value of the long-run probability used in the null hypotheses. You probably could have predicted that your null distribution was going to be somewhat bell-shaped and centered at 0.5. You probably would have a harder time predicting your null distribution's variability (standard deviation), but this too can be predicted in advance, as we will see shortly.

We can use mathematical models known as normal distributions (bell-shaped curves) to approximate many of the null distributions we have generated so far in this text. When rules and theories are used to predict what the value of the standardized statistic and p-value would be if someone did simulate, we call the approach a **theory-based approach**. The normal distribution provides a second way, in addition to simulation, to approximate a p-value.

- Check the box next to **Normal Approximation** in the applet. Does the shaded in blue region seem to be a good description (model) of what we actually got in the simulation?

Validity Conditions for Theory-Based Approach

The normal approximation to the null distribution is valid whenever the sample size is reasonably large. One convention is to consider the sample size to be large enough whenever there are at least 10 observations in each category.

7. According to this convention, is the sample size large enough in this study to use the normal approximation and theory-based inference? Justify your answer.

Validity conditions: The normal approximation can be thought of as a prediction of what would occur if simulation was done. Many times this prediction is valid, but not always, only when the validity condition (at least 10 successes and at least 10 failures) is met.

Formulas

The normal approximation will also give you values of the standardized statistic and p-value based on its mathematical predictions. As you learned in Section 1.3, the standardized score is calculated as:

$$z = \frac{\text{statistic} - \text{mean of null distribution}}{\text{standard deviation of null distribution}}$$

The mean of the null distribution is the hypothesized value of the long-run probability (π). The standard deviation can be obtained in two ways:

- (1) Find the standard deviation of the null distribution by simulating
- (2) Predict the value of the standard deviation by plugging into this formula: $\sqrt{\pi(1 - \pi)/n}$

8. Use the formula to determine the (theoretical; predicted) standard deviation of the sample proportion. Then compare this to the SD from your simulated sample proportions, as recorded in #5a. Are they similar?

The predicted value of the standard deviation (using the formula) will be very close to the simulated standard deviation of the null distribution when the sample size is large enough. The Validity Condition mentioned earlier, says that a “large enough” sample size means at least 10 successes and at least 10 failures. This mathematical prediction is often called the “Central Limit Theorem.”

Central Limit Theorem: If the sample size (n) is large enough, the distribution of sample proportions will be bell-shaped (or normal), centered at the long-run probability (π), with a standard deviation of $\sqrt{\pi(1 - \pi)/n}$.

9. Use the predicted value of the standard deviation from (8) to calculate the standardized statistic (z) by hand, and confirm that your answer is very close to what you found in #5c when using simulation.

In the applet, see that the predicted value of the standardized statistic, z , is given immediately below the button for “Normal approximation” in parentheses and should match your answer to (9)

10. The theory-based (Normal approximation) p-value is also now displayed. Compare this p-value to the one you got from simulation (#5b). Are they similar?
11. Why are the standard deviation (#8), standardized statistic (#9) and p-value (#10) similar when using the theory-based (one proportion z-test; normal approximation) to what you got in your simulation? When would they be different?

Follow-up Analysis #1

In his book *Statistics You Can't Trust*, Steve Campbell claims that people pick heads 70% of the time when they're asked to predict the outcome of a coin toss.

12. Use the theory-based approach to test Campbell's claim based on the sample data used above (your class combined with authors class) using a two-sided alternative. Report the null and alternative hypothesis, standardized statistic, and p-value. Summarize your conclusion, and explain the reasoning process by which it follows from your analysis.

Follow-up Analysis #2

In a small class of eight students, seven students picked heads when given the choice between heads and tails.

13. Use simulation to generate a two-sided p-value evaluating the strength of evidence that the long-run probability of students picking heads is different than 50% based on this small class's data alone.
14. Why can't you use the normal approximation in this case?
15. Use the normal approximation anyway. Compare and comment on the p-values obtained from the two methods.

Section 1.5 Summary

Most of the null distributions for a sample proportion that we have seen follow a common and familiar shape. This bell-shaped curve, known mathematically as a **normal distribution**, allows us to anticipate what a null distribution will look like, and to determine an approximate p-value, without bothering to conduct a simulation analysis.

All normal distributions have a bell-shaped curve, but they can differ with regard to center and variability. When working with the null distribution of a sample proportion:

- The center is described by the mean, which for a null distribution equals the hypothesized value of the long-run probability π .
- The variability is described by the standard deviation, which is determined primarily by the sample size.
 - The larger the sample size, the smaller the variability in sample proportions.
 - The standard deviation of the sample proportion equals $\sqrt{\pi(1 - \pi)/n}$.

The **theory-based approach** (one-proportion z-test; normal approximation) standardizes the statistic based on the observed value of the statistic (sample proportion) and these theoretical mean and standard deviation values.

- The p-value is calculated by software as the area under the normal curve in the appropriate direction (as specified by the alternative hypothesis) from the standardized statistic.
- The p-value is interpreted and evaluated just as with the simulation-based method.

This theory-based method works well whenever the sample size is large enough for the normal curve to provide a good approximation to the null distribution.

- We consider the theory-based method to be valid when there are at least 10 observations of each category (“success” and “failure”) in the sample.
- When this validity condition is not met, and even when it is met, an alternative is simply to use the simulation-based method.

The names, conditions and applets used for the simulation and theory-based tests are shown in the following table.

Figure 1.22 Summary of validity conditions and applets for one proportion tests

Type of data	Name of test	Applet	Validity conditions
Single binary variable	Simulation test for a single proportion	One proportion	---

	Theory-based test for a single proportion (1 Proportion z test)	One proportion	At least 10 successes and at least 10 failures
--	--	----------------	---

Chapter 1 Summary

This chapter has focused on Step 4 of the Statistical Investigation Method: Assessing the statistical significance of an observed result against some claim about the long-run probability of successes. This reasoning will stay the same as we consider other statistics (e.g., sample mean, difference in sample proportions) and other parameters (e.g., a difference in long-run probabilities) in other chapters. The basic approach is the 3S Strategy: Decide on an appropriate statistic, simulate values of that statistic under the assumption that the null hypothesis is true, and then assess the strength of evidence against the null hypothesis and in favor of the alternative hypothesis. Keep in mind that the null and alternative hypotheses are competing claims about the parameter value. We expect the observed sample statistic to differ from the parameter by chance. The goal is to see whether the difference observed in the study is larger than can be reasonably expected by chance alone. This unusualness (deciding whether the observed result in the tail of the null distribution) can be measured through the standardized statistic (e.g., z-statistic), which measures the distance between the observed statistic and the hypothesized parameter value in terms of number of standard deviations away, and/or the p-value, which measures how often a statistic at least as extreme would occur when the null hypothesis is true.

In the case of one categorical variable, when the sample size is large (e.g., at least 10 successes and at least 10 failures in the sample), then we can approximate the p-value using the normal distribution. This theory-based approach is referred to as a one-proportion z-test

You also considered factors that affect the strength of evidence against the null hypothesis:

- How far the observed sample statistic is from the hypothesized parameter value
- Sample size
- One-sided vs. two-sided alternatives.

The relationships you learned in this chapter will also apply in other scenarios as well.

Chapter 1 Glossary

3S Strategy

A framework for evaluating the strength of evidence against the chance model (null hypothesis). The 3 S's are: Statistic, Simulate, and Strength of Evidence.. 1-9, 1-20, 1-23

alternative hypothesis

The *not by chance* or *there is an effect* explanation, it is our research conjecture. 1-26, 1-37

bar graph

A graphical display of the distribution of a categorical variable 1-11

binary variable

Categorical variable with only two outcomes..... 1-25, 1-36

Central Limit Theorem

A mathematical prediction of the standard deviation of the null distribution when certain validity conditions are met..... 1-76

chance models

A real or computerized process to generate data according to a well-understood set of conditions1-5, 1-23

model

A mathematical or probabilistic conceptualization meant to closely match reality, but always making assumptions about the reality which may or may not be true:1-5

n

A symbol used to indicate the sample size 1-26

normally distributed

how the null distribution is described when it takes the shape of a bell 1-74

null distribution

Distribution of simulated statistics that represent what could have happened in the study assuming the null hypothesis was true 1-28, 1-40

null hypothesis

The *by chance alone* or *no effect* explanation; A hypothesis that can be modeled by simulation. 1-26, 1-37

one-proportion z-test

name for the theory-based approach with one proportion..... 1-75

parameter

For a random process a parameter is a long-run numerical property of the process **1-4**

p-hat (*p*)

The proportion or percentage of observational units that have a particular characteristic based on a measured variable. A statistic..... 1-26

plausible

A term used to indicate that the chance model is a reasonable/believable explanation for the data we observed1-8, 1-19

p-value

The probability of obtaining a value of the statistic at least as extreme as the observed statistic when the null hypothesis is true..... 1-29, 1-41

sample

The set of observed values.....1-4, 1-15

sample size

The number of observational units in the sample1-4, 1-15

standardize

To standardize an observation, compute the distance of the observation from the mean and divide by the standard deviation of the distribution..... 1-50, 1-55

statistic

A number computed from the sample.....1-4, 1-15

statistical significance

Results unlikely to have arisen by chance alone1-1

statistically significant

Unlikely to occur just by random chance.....1-8, 1-19

strength

How much evidence we have against the null hypothesis1-1

subjects

Study participants that are human..... 1-36

test of significance

A procedure for measuring the strength of evidence against a null hypothesis about the parameter of interest..... 1-24

Theory-based approach

mathematical approach which predicts the shape, center and variability of the null distribution instead of obtaining a null distribution by simulating..... 1-75

two-sided test

Estimates the p-value by considering results that are at least as extreme as our observed result *in either direction*..... 1-64

Validity conditions

Check to see that certain conditions are met that render the theory-based approach valid. Often these conditions deal with sample size and shape and variability of distributions. 1-76

z-statistic

z-statistic is synonymous with standardized sample proportion, also called the standardized statistic. 1-77