Four Pillars of Inference: Strength, Size, Breadth, and Cause

The four chapters of this unit are devoted to four core concepts:

**Logic of Inference**

- **Significance**
  - How *strong* is the evidence of an effect? (Chapter 1)
  - How *strong* is the evidence that the wording of default options affects recruitment of organ donors?

- **Estimation**
  - How *large* is the effect? (Chapter 3)
  - How *much* will the neutral wording (forced choice) increase the donor pool?

- **Generalization**
  - How *broadly* do the conclusions apply? (Chapter 2)
  - Do the conclusions apply only to that one set of volunteers? To other similar groups of volunteers? To all who apply for driver’s licenses?

- **Causation**
  - Can we say what *caused* the observed difference? (Chapter 4)
  - Can we conclude that it was the *wording of the default option* that was responsible for the increased rate of donor recruitment?

The four concepts have a natural pairing. The first two, significance and estimation, deal with the *logic* of inference. They correspond to Step 4 of our six-step statistical investigation method. The last two concepts, generalization and causation, deal with the *scope* of inference. They correspond to Step 5.

In Unit 1 we present these four core concepts—*strength, size, breadth*, and *cause*—one chapter at a time, in the simplest possible statistical setting: observed values for a single binary (Yes/No) variable. This setting is mathematically simple enough to study by flipping coins or dealing cards from a shuffled deck of red and black cards. However, even though the coins and cards are simple, they can be used for inference about a variety of interesting situations. Do dolphins communicate? Can dogs detect cancer? Do red-uniformed Olympians win more often? Is the more competent-looking politician more likely to win an election?
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 trump the chance explanation and we then 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 41.8% in the “opt-in” group elected to become donors, compared with 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.

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 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.

The sample size in this example is 16. Note that the word “sample” is used as a noun (the set of observational units being studied), as an adjective, for example to mean “computed from the observed data,” as, for example, “sample statistic,” and as a verb, for example, “We need to sample 50 people.”

**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 proportion (i.e., probability) of pushing the correct button based on Doris’s whistles. This unknown long-run proportion is called a parameter.

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 of the following:

- 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 (Section 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>
<thead>
<tr>
<th>TABLE 1.1 Parallels between real study and physical simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coin flip = guess by Buzz</td>
</tr>
<tr>
<td>Heads = correct guess</td>
</tr>
<tr>
<td>Tails = wrong guess</td>
</tr>
<tr>
<td>Chance of heads = 1/2, probability of correct button when Buzz is just guessing</td>
</tr>
<tr>
<td>One repetition = one set of 16 simulated attempts by Buzz</td>
</tr>
</tbody>
</table>

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.

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.

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 100 times in Figure 1.4A (a number small enough so we can see the individual dots) and 1,000 times in Figure 1.4B—a number chosen for convenience but also 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** Dotplots showing (a) 100 repetitions and (b) 1,000 repetitions of flipping a coin 16 times and counting the number of heads.
Let’s think carefully about what the graphs in Figure 1.4 show. For these graphs, 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 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 1,000 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 just 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 our evidence is strong enough to be considered statistically significant. That is, we don’t think our study result (15 out of 16 correct) happened by chance alone, but rather, something other than “random chance” was at play.

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.
CHAPTER 1  Significance: How Strong Is the Evidence?

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 statistic 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 when 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 0.50. 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 in 16 attempts 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 proportion of pushing the correct button was not simply 0.50.

**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 his 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 redo 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 $16/28 \times 100\% \approx 57.1\%$ 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.

**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 1,000 repetitions of 28 coin flips is shown in Figure 1.6. This models 1,000 repetitions of 28 attempts with Buzz randomly pushing one of the buttons (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?**

---

**EXPLORATION**

**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 nonhuman 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 nonhuman gestures. To test this, the researchers positioned the dogs about 2.5 meters from the experimenter. Two cups were placed, one on each side of the experimenter. The experimenter would perform some sort of gesture (pointing, bowing, looking) toward one of the cups or there would be some other nonhuman 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 four-year-old mixed breed, was named Harley. Each trial involved one gesture and one pair of cups, with a total of 10 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.** 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 9 out of 10 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 experimenters mean when they bow towards the cup.

The unknown long-run proportion (i.e., probability) that Harley will choose the correct cup is called a *parameter*.

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 if Harley is not just guessing and instead understands the experimenter?
We will 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.
   a. 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?
   b. 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 Harley guessing between the two cups. Set the number of tosses to 10 and press the Draw Samples 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 Draw Samples 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 Draw Samples. The applet will now show the results for the number of heads in 1,000 different sets of 10 coin tosses. So each dot represents the number of times Harley chooses the correct cup out of 10 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 1,000 simulated sets of 10 coin flips each, would you conclude that Harley would be very unlikely to have picked the correct cup 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.

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 Table 1.2 indicating what parts of the real study correspond to the physical (coin-flipping) simulation.

<table>
<thead>
<tr>
<th>TABLE 1.2 Parallels between real study and physical simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coin flip</td>
</tr>
<tr>
<td>Heads</td>
</tr>
<tr>
<td>Tails</td>
</tr>
<tr>
<td>Chance of heads</td>
</tr>
<tr>
<td>One repetition = one set of _______ simulated attempts by Harley</td>
</tr>
</tbody>
</table>

**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, because 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 nonhuman gestures. In 10 trials, Harley chose the correct cup 6 times.
   a. Using the dotplot you obtained when you simulated 1,000 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 1,000 simulated sets of 10 coin flips each, would you conclude that Harley would be very unlikely to have picked the correct cup 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.

---

**MEASURING THE STRENGTH OF EVIDENCE**

**INTRODUCTION**

In the previous section, we discussed testing whether or not a chance model of equally choosing between two equally likely 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.

**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 (you) 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.

The underlying process here is choosing a hand gesture in each repeated play of the game. The parameter of interest is the long-run proportion that any player picks scissors. The observed statistic is the number of times scissors was chosen in the 12 plays (two) or the proportion of times that scissors was chosen ($1/6$).

What is the research conjecture about the long-run proportion?

We are told that novice players seem to pick scissors less than the other two gestures. This would imply that the long-run proportion for picking scissors is less than $1/3$. And in this study, that’s what you found. Your friend only chose scissors one-sixth of the time.

But perhaps that was just an unlucky occurrence in this study? Maybe your friend would play scissors one-third 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 three gestures and we just happened to observe fewer scissors 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 three 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.

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 $1/3$.

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 $\pi$ (pronounced “pie”), to represent a parameter that is a probability. For example, the long-run proportion that a novice player picks scissors can be represented by $\pi$. We will then use the symbol $\hat{p}$ (pronounced “p-hat”) to represent the proportion in the sample. In this example, $\hat{p} = 2/12 \approx 0.167$, 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:

\[
H_0: \pi = \frac{1}{3} \\
H_a: \pi < \frac{1}{3}
\]

where \(\pi\) represents your friend’s true probability of throwing scissors.

Notice that both of these hypothesis statements are claims about the parameter \(\pi\). When testing a single probability as we are doing in this example, we compare the actual (but unknown) value of the parameter \(\pi\) to the same numerical value in both the null and alternative hypotheses. In this case that value is \(\frac{1}{3}\) (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 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 one-sixth 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 one-third of the time in the long run. So we will use a chance model that assumes the probability of your friend choosing scissors is \(\frac{1}{3}\) 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 0.50 (as with coin tossing) but \(\frac{1}{3}\). 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 1.3.

**TABLE 1.3 Parallels between actual study and simulation**

<table>
<thead>
<tr>
<th>Observational unit</th>
<th>=</th>
<th>One round of the game</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td>=</td>
<td>Plays scissors</td>
</tr>
<tr>
<td>Failure</td>
<td>=</td>
<td>Doesn’t play scissors</td>
</tr>
<tr>
<td>Chance of success</td>
<td>(\pi = \frac{1}{3})</td>
<td>Three gestures are equally likely; chance of scissors is one out of three</td>
</tr>
<tr>
<td>One repetition</td>
<td>=</td>
<td>A simulated set of 12 rounds of the game</td>
</tr>
</tbody>
</table>

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 4 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 proportion 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 one-third 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.

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:

How do we use this null distribution to evaluate whether the one-third 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 one-third 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 either what proportion of the distribution lies at or above the value or what proportion lies at or below the value. The probability of getting a result at least as extreme in the direction of the alternative hypothesis 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 than the observed sample proportion.
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 vertical 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 1,000 samples from a process with \( \pi = \frac{1}{3} \) 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.

Calculating a p-value from a simulation analysis is an approximation. Different simulations will have slightly different p-values based on 1,000 repetitions. Performing more repetitions generally produces more accurate approximations. For our purposes, using 1,000 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 1,000 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**

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

FAQ 1.2.2  www.wiley.com/college/tintle

What p-value should make us suspicious?

So we would consider only two 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 one-third of the time in the long run. Why? Because a p-value of 0.173 indicates that getting two or fewer choices of scissors in 12 plays, if the probability of scissors was really \( \frac{1}{3} \), is not surprising. Hence, \( \frac{1}{3} \) is still a plausible value for your friend’s long-run proportion.

What if your friend had only played scissors once? Then our statistic becomes \( \hat{p} \approx 0.083 \). This is even farther away from the expected \( \frac{1}{3} \). 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 at or below the vertical line shown in Figure 1.9.

\[
\begin{align*}
\text{Probability of success (π):} & \quad 0.3333 \\
\text{Sample size (n):} & \quad 12 \\
\text{Number of samples:} & \quad 1000 \\
\circ \text{ Number of successes} & \quad 0 \\
\circ \text{ Proportion of successes} & \quad 0.083 \\
\text{As extreme as} & \quad \leq 0.083 \\
\text{Proportion of samples:} & \quad 77/1000 = 0.0770
\end{align*}
\]

**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 (\( \pi = \frac{1}{3} \)), it should be more surprising to get a sample proportion farther from \( \frac{1}{3} \).

**KEY IDEA**

Values of the statistic that are even farther 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).

Key assumptions: Model vs. reality

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

For example, with the rock-paper-scissors game we simulated data from a chance model
with a $\frac{1}{3}$ success probability. This means each and every round has exactly the same probab-
ility 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 have said “we have strong evidence against the chance model,” are we
saying that your friend subconsciously avoids scissors? Well, maybe, but maybe not. All we are
saying is we think something else is going on other than your friend picking equally among
the three gestures every round. 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.

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. Write this as a binary variable.
5. Describe the parameter of interest (in words). (Hint: The parameter is the long-run
   proportion 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
   proportion (i.e., 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 proportion that a subject

---

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.”
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.)

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 proportion, such as the probability
that a (future) subject in this study would choose tap water as most preferred, we use the
Greek letter \( \pi \) (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 \( \pi \). 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 \( \pi \).

\[ H_0: \]
\[ H_a: \]

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
was true (i.e., even if people did not really dislike the tap water and were essentially select-
ing 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 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:
   a. What does each region represent?
   b. How many spins of the spinner will you need to do in order to simulate one repetition of the experiment when there is equal preference between the four waters (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.
   a. First enter the probability of heads/probability of success value specified in the null hypothesis.
   b. Enter the appropriate sample size (number of subjects in this study).
   c. Enter 1 for the number of samples, and press Draw Samples. Report the number of “successes” in this simulated sample.
   d. 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.
   e. Leaving the “Proportion of successes” radio button selected but unchecking the “Animate” box, click on Draw Samples four more times. Do you get the same results each time?
   f. Now enter 995 for the number of samples and click on Draw Samples, bringing the number of simulated samples to 1,000. 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 \( \hat{p} = \frac{3}{27} \approx 0.1111 \). Looking at the null distribution you have simulated, is this a very unlikely result when the null hypothesis is 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 extreme than the observed value.

   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 < p-value: not much evidence against null hypothesis; null is plausible
   - 0.05 < p-value ≤ 0.10: moderate evidence against the null hypothesis
   - 0.01 < p-value ≤ 0.05: strong evidence against the null hypothesis
   - p-value ≤ 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 ≤ 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 \( \pi \)) be the probability that a subject will select 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.
   - The research conjecture is that subjects tend to select bottled water (more or less) often than tap water. (Circle your answer.)
     
     More       Less
   
   a. State the null hypothesis in words and in terms of the (newly defined) parameter \( \pi \).
   
   b. State the alternative hypothesis in words and in terms of the (new) parameter \( \pi \).
   
   c. Calculate the observed value of the relevant statistic.
   
   d. 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:
   
   e. 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 #18(f).
   
   f. 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 _____.
   
   g. Estimate the p-value from your simulation results.
   
   h. Interpret this p-value. (Hint: This is the probability of what, assuming what?)
   
   i. 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 ”$\pi < 0.25$”
- 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 proportion 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:

<table>
<thead>
<tr>
<th>p-value</th>
<th>Strength of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10 &lt; p-value</td>
<td>not much evidence against null hypothesis</td>
</tr>
<tr>
<td>0.05 &lt; p-value ≤ 0.10</td>
<td>moderate evidence against the null hypothesis</td>
</tr>
<tr>
<td>0.01 &lt; p-value ≤ 0.05</td>
<td>strong evidence against the null hypothesis</td>
</tr>
<tr>
<td>p-value ≤ 0.01</td>
<td>very strong evidence against the null hypothesis</td>
</tr>
</tbody>
</table>

### NOTATION CHECK

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

- $\pi$ 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 hypotheses); 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 provides 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 proportion of a particular outcome, \( \pi \), 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 there is another convenient way to measure how far the observed statistic is from the hypothesized value under the null hypothesis.

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.

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 dies within 30 days of the transplant.
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 postoperation” 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.

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:

\[
H_0: \pi = 0.15 \\
H_a: \pi > 0.15
\]

where \( \pi \) is the actual long-run proportion of deaths 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 1,000 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.png)

Each dot represents one set of 10 patients where \( \pi = 0.15 \).

**FIGURE 1.10** Null distribution (could-have-been simulated sample proportions) for 1,000 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 \( 1/1,000 \), as we never observed a value of 0.80 or larger by chance in these 1,000 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 1,000 values of the simulated proportions of success 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 0.60. In fact, the variability of the distribution as measured by the standard deviation is 0.113. Remember from the Preliminaries that 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 1,000 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](image)

**FIGURE 1.11** The null distribution of 1,000 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 1,000, but visually appeared to be much smaller still. Whereas 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 extreme 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: \(z\).)

\[
\text{Standardized statistic} = z = \frac{\text{statistic} - \text{mean of null distribution}}{\text{standard deviation of null distribution}}
\]

For the second study, a standardized value of the statistic would result in the following calculation because 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 is approximately 0.15 (the hypothesized probability):

\[
\text{standardized statistic} = z = \frac{0.197 - 0.15}{0.018} = 2.61
\]

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 can say 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:

\[
\text{standardized statistic} = z = \frac{0.800 - 0.15}{0.113} = 5.75
\]

This means that \(\hat{\theta} = 0.800\) is 5.75 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. For standardized statistics:

- 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 against the null hypothesis for a mound-shaped, symmetric distribution.
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 data sets. We see the stronger evidence from the small data set compared to the larger data set (5.73 > 2.61).

**Do People Use Facial Prototyping?**

A study in *Psychonomic Bulletin and Review* (Lea, Thomas, Lamkin, and 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 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?

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 **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 against the null hypothesis, 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 when 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 0.50.
   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?)

\[ \hat{p} = \]

\[ \text{standardized statistic} = \frac{\text{observed statistic} - \text{mean of null distribution}}{\text{standard deviation of null distribution}} \]

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.
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. For standardized statistics:

- 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 against the null hypothesis for a mound-shaped, symmetric distribution.

![Figure 1.13](image)

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.
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 absolute value of the standardized statistic:
      - 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:

\[
\text{standardized statistic} = \frac{\text{statistic} - \text{mean of null distribution}}{\text{standard deviation of null distribution}}
\]

A standardized statistic provides an alternative to the 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 proportion 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.

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, Mandisodoka, Goren, and 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 U.S. 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 0.50 ($\pi = 0.50$).

Alternative hypothesis: The probability this method predicts the winner is greater than 0.50 ($\pi > 0.50$).

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 $\frac{23}{32} = 0.719$, 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 1,000 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](image-url)

**FIGURE 1.14** The results of 1,000 sets of 32 coin tosses from a process with probability 0.50 of success indicating which results are at least as extreme as 23 ($\hat{\pi} \geq 0.7188$).
Only 9 of these 1,000 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 the hypothesized parameter value (the mean of that 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 nonexistent? 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 had been 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 \((0.8125 - 0.50)/0.09 = 3.47\) (meaning 0.8125 is 3.47 standard deviations above the mean of the null distribution). In short, if researchers correctly predict 26 elections, this would be 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 1,000 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 mean 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 against the null hypothesis 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 three 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 256 races (8 times as many).

![Figure 1.15](image)

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-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-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-value < 0.001; standardized statistic = 7.06). We are even more convinced that our result was not just a “fluke” outcome.

**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 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 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-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 one-sixth 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 more competent looking person being elected more frequently, it was actually the less competent looking 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 0.50.
- Alternative hypothesis: The probability this method predicts the winner is greater than 0.50.

These hypotheses can be written in symbols as

\[
H_0: \pi = 0.50 \\
H_a: \pi > 0.50
\]

where \( \pi \) = 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 hypothesis 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 0.50, we need to look at what number is 0.2188 below 0.50. Calculating 0.50 − 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 illustrates how a two-sided p-value is computed. In this case 0.7188 or greater was obtained nine times by chance, and 0.2812 or less was obtained eight times by chance. Thus, the two-sided p-value is approximately \( \frac{8 + 9}{1,000} = 0.017 \). 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 the 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 evidence against the null hypothesis. However two-sided tests are used more often in scientific practice.

Technical note: When the hypothesized probability is something other than 0.50, there are actually a couple of different ways of defining an observation as being more extreme. The One Proportion applet determines that 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.

FAQ 1.4.1  www.wiley.com/college/tintle

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

FOLLOW-UP STUDY
As a way of applying what we just learned, consider the following. The researchers inves-
tigating 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 that 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

\[
\hat{p} = \frac{189}{279} \approx 0.677, \text{ or } 67.7\% \text{ of the 279 House races.}
\]

Notice this sample percentage of 67.7\% is a bit
smaller than the 71.9\% correct predictions in 32 Senate races; however, the sample size (279
instead of 32) is lar

Think About It

Do you expect the strength of evidence for the “competent-face” method to be stron-
ger 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
0.50 as the probability under the null hypothesis.

![Figure 1.17](image-url)
Chapter 1  Significance: How Strong Is the Evidence?

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 \(z \approx 5.90\). 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, leading to much stronger evidence against the null hypothesis.

**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; and (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 \(\pi\) 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 or the standardized statistic.
   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 value of the parameter than a one-sided test.

4. If we let $\pi$ 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 to a two-sided test 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 evidence against the null hypothesis. 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 0.50 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 0.50.
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 with 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 (mean) and variability (standard deviation) 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, $\pi$, 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).
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 proportion. 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 assess strength of evidence against the null hypothesis. 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 often follow, though not always, bell-shaped curves.
2. They are centered at the null hypothesis value for $\pi$.
3. Their variability (spread, standard deviation) is influenced by the sample size.

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

![null distributions](image)

**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 ($n = 279$) and St. George’s ($n = 361$) 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 discreteness, 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, one example of a theory-based approach to statistical inference.

**Halloween Treats**

Concerned 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. They found that 148 of the trick-or-treaters took candy and the other 135 took toys (Schwartz, Chen, and Brownell, 2003). We can focus on either candy or toy to be what we call a success and we choose to focus on candy. Thus our observed sample proportion is $\frac{148}{283} = 0.523$.

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 0.50.
- **Alternative hypothesis:** The probability a trick-or-treater would choose candy is not 0.50.

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

- $H_0: \pi = 0.50$
- $H_a: \pi \neq 0.50$

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 1,000 simulated samples) shown in Figure 1.19.
We notice that the distribution is quite bell-shaped; the average value is 0.50 (the null hypothesis value for \( \pi \)) 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 proportion \((\pi)\), 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 a 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 and that is what we will use. Rules like this that must be met in order for theory-based p-values to be valid are what we will call *validity conditions*.

**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. We will consider the prediction valid 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). With the 283 trick-or-treaters in the researchers’ sample, they found that 148 of them took candy (so we have 148 successes) and 135 took toys (so we have 135 failures).
As both the number of successes and failures are greater than 10, 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 0.50, the hypothesized value of \( \pi \).
3. The simulated standard deviation of the sample proportions is 0.030, which is very close to the predicted standard deviation of \( \sqrt{0.50(1 - 0.50)/283} \approx 0.0297 \).

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, \( z \):

\[
z = \frac{0.523 - 0.50}{0.0297} = 0.77
\]

This tells us that our observed sample proportion (0.523) is 0.77 of a standard deviation above the mean of 0.50. 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 (normal) standard 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.

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 0.4670 from the simulation. We interpret this p-value and draw a conclusion from it as always: If the probability that each trick-or-treater prefers candy to toys is 0.50, 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 cutoff 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 2 of the 12 rounds played.

**THINK ABOUT IT**

Why do you think the theory-based (also known as one-proportion z-test; normal approximation) approach will not work well for these 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.

With the larger p-value, the simulation-based approach provides a bit less evidence against the null hypothesis. Because these two p-values are not all that similar 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 that 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. In the coin toss study with your class:
   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 \( \pi \) 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?

<table>
<thead>
<tr>
<th>Sample size:</th>
<th>Sample proportion:</th>
</tr>
</thead>
</table>

5. Use the One Proportion applet to test the hypotheses from #2d and #2e.
   a. 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.

<table>
<thead>
<tr>
<th>Shape:</th>
<th>Familiar?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center?</td>
<td>Why does this make sense?</td>
</tr>
<tr>
<td>Mean:</td>
<td>SD:</td>
</tr>
</tbody>
</table>

b. Approximate the p-value and summarize the strength of evidence that the sample data provide regarding the research hypothesis.

c. 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 5(a), 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 proportion used in the null hypotheses. You probably could have predicted that your null distribution was going to be somewhat bell-shaped and centered at 0.50. 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.

6. Check the box next to **Normal Approximation** in the applet. Does the region shaded in blue 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.

---

**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 proportion \( \pi \). The standard deviation can be obtained in two ways:

- First, find the standard deviation of the null distribution by simulating.
- Second, 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. The validity condition mentioned earlier says the shape will be approximately normal when the sample size is large enough where, 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 proportion \( \pi \), 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 author’s 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 proportion 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 proportion \( \pi \).
- 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 (also known as 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.

<table>
<thead>
<tr>
<th>Summary of validity conditions and applets for one proportion tests</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of data</strong></td>
</tr>
<tr>
<td>Single binary variable</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

### 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 proportion 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 3S’s are statistic, simulate, and strength of evidence.

alternative hypothesis The not by chance or there is an effect explanation, it is typically our research conjecture.

bar graph A graphical display of the distribution of a categorical variable.

binary variable Categorical variable with only two outcomes.

central limit theorem A mathematical prediction of the behavior of the null distribution when certain validity conditions are met.

chance models A real or computerized process to generate data according to a well-understood set of conditions.

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.

n A symbol used to indicate the sample size.

normally distributed How the null distribution is described when it takes the shape of a bell.

null distribution Distribution of simulated statistics that represent what could have happened in the study assuming the null hypothesis was true.

null hypothesis The by chance alone or no effect explanation; a hypothesis that can be modeled by simulation.

parameter For a random process a parameter is a long-run numerical property of the process.

pi (π) The Greek letter for p, which is pronounced “pie” and is used to represent a parameter that is a probability.

p-hat (p̂) The proportion of observational units that have a particular characteristic based on a measured variable; a statistic.

plausible A term used to indicate that the chance model is a reasonable/believable explanation for the data we observed.

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.

sample The set of observed values.

sample size The number of observational units in the sample.

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

statistic A number computed from the sample.

statistically significant Unlikely to occur just by random chance.

strength How much evidence we have against the null hypothesis.

subjects Study participants that are human.

test of significance A procedure for measuring the strength of evidence against a null hypothesis about the parameter of interest.

theory-based approach Mathematical approach which predicts the shape, center, and variability of the null distribution instead of obtaining a null distribution by simulating.

two-sided test Estimates the p-value by considering results that are at least as extreme as our observed result in either direction.

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 null distributions.

z-statistic Synonymous with standardized sample proportion, also called the standardized statistic.

CHAPTER 1
EXERCISES

SECTION 1.1

Spinning tennis racquet*

Tennis players often spin a racquet to decide who serves first. The spun racquet can land with the manufacturer’s label facing up or down. A reasonable question to investigate is whether a spun tennis racquet is equally likely to land with the label facing up or down. (If the spun racquet is equally likely to land with the label facing in either direction, we say that the spinning process is fair.) Suppose that you gather data by spinning your tennis racquet 100 times, each time recording whether it lands with the label facing up or down.
1.1.1
a. Describe the relevant long-run proportion of interest in words.
b. What statistical term is given to the long-run proportion you described in (a)?
c. What value does the chance model assert for the long-run proportion?
d. Suppose that the spun racquet lands with the label facing up 48 times out of 100. Explain, as if to a friend who has not studied statistics, why this result does not constitute strong evidence against believing that the spinning process is fair.
e. Is the result in (d) statistically significant evidence that spinning is not fair or is it plausible that the spinning process is fair?

1.1.2
a. Suppose that the spun racquet lands with the label facing up 24 times out of 100. Explain, as if to a friend who has not studied statistics, why this result does not constitute strong evidence against believing that the spinning process is fair.
b. Is the result in (a) statistically significant evidence that spinning is not fair or is it plausible that the spinning process is fair?

d. With each repetition, what would you keep track of?

e. What would be a typical value from a repetition of 1,093 coin flips? Justify your answer.

1.1.3 To conduct a simulation analysis of this racquet-spinning study for either 1.1.1 or 1.1.2, you could flip a coin _____ times and repeat that process _____ times.

A. 100, 1,000
B. 1,000, 100
C. 100, 1
D. 1, 1,000

1.1.4 Which of the following is the most important reason that a simulation analysis would repeat the coin-flipping process many times?

A. To see whether the distribution of sample proportions follows a normal, bell-shaped curve
B. To see whether the distribution of sample proportions is centered at 0.50
C. To see how much variability results in the distribution of sample proportions

1.1.5 LeBron James of the Miami Heat hit 765 of his 1354 field goal attempts in the 2012/2013 season for a shooting percentage of 56.5%. Over the lifetime of LeBron’s career, can we say he is more likely than not to make a field goal?

a. Describe the parameter of interest.
b. Is 56.5% a parameter or a statistic?
c. What value would the chance model assign to the parameter from (a)?
d. Describe how you would use coin flipping to simulate the 2012/2013 season under the assumption of the chance model.
e. What would be a typical value from a repetition of 1,354 coin flips? Justify your answer.

1.1.6 Dwyane Wade of the Miami Heat hit 569 of his 1,093 field goal attempts in the 2012/2013 season for a shooting percentage of 52.1%. Over the lifetime of Dwyane’s career, can we say that Dwyane is more likely than not to make a field goal?

a. Is the long-run proportion of Dwyane making a field goal a parameter or a statistic?
b. Is 52.1% a parameter or a statistic?
c. When simulating possible outcomes assuming the chance model, how many times would you flip a coin for one repetition of the 2012/2013 season?
d. With each repetition, what would you keep track of?
e. What would be a typical value from a repetition of 1093 coin flips? Justify your answer.

1.1.7 If Dwyane Wade of the Miami Heat hits 52 out of his first 100 field goals in the 2013/2014 season, let’s see how we might investigate if he is more likely than not to make a field goal?

a. Based on these first 100 field goals, we want to find out what Dwyane’s long-run proportion of making a field goal is. Will this value be a statistic or a parameter?
b. Is 52 out of 100 a statistic or a parameter?
c. We can use the One Proportion applet to generate 1,000 possible values of Dwyane’s last 100 field goals under the chance model that he has a long-run proportion of 50% of making a field goal. Match the aspects of the simulation in Column A to their equivalent aspects in the actual study listed in Column B.

<table>
<thead>
<tr>
<th>Column A</th>
<th>Column B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coin flip</td>
<td>Dwyane misses his field goal</td>
</tr>
<tr>
<td>Heads</td>
<td>Long-run proportion of field goals</td>
</tr>
<tr>
<td></td>
<td>Dwyane makes</td>
</tr>
<tr>
<td>Tails</td>
<td>One set of 100 field goal shots</td>
</tr>
<tr>
<td></td>
<td>by Dwyane</td>
</tr>
<tr>
<td>Chance of heads</td>
<td>Dwyane shoots a field goal</td>
</tr>
<tr>
<td>One repetition</td>
<td>Dwyane makes his field goal</td>
</tr>
</tbody>
</table>

1.1.8 One of the authors sometimes likes to play Minesweeper, and of the last 20 times she played Minesweeper, she won 12 times. That is, she won 60% of the games.

a. Based on these 20 games, we would like to learn about her long-run proportion of winning at Minesweeper. Is this value a parameter or a statistic?
b. Is 60% (12 wins in last 20 games) a parameter or a statistic?
To find out whether her performance on the last 20 games provides convincing evidence that her long-run proportion of winning at Minesweeper is higher than 50%, she used the One Proportion applet to generate 100 possible values of what her number of wins could have been if her long-run proportion of winning was equal to 50%.

c. In the table below, match the aspects of the simulation performed by the applet as listed in Column A to their equivalent aspects in the actual study as listed in Column B.

<table>
<thead>
<tr>
<th>Column A</th>
<th>Column B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coin flip</td>
<td>Author wins a game</td>
</tr>
<tr>
<td>Heads</td>
<td>Long-run proportion of games that the author wins</td>
</tr>
<tr>
<td>Tails</td>
<td>One set of 20 Minesweeper games played by author</td>
</tr>
<tr>
<td>Chance of heads</td>
<td>Author loses a game</td>
</tr>
<tr>
<td>One repetition</td>
<td>Author plays a game of Minesweeper</td>
</tr>
</tbody>
</table>

1.1.9 The dotplot generated below by the One Proportion applet assumed the long-run proportion of winning was equal to 50%. Use this dotplot to answer the following questions.

![Dotplot](image)

a. How many dots are in the dotplot? (Hint: You should reason it out and not count the number of dots.)

b. What does each dot represent in terms of Minesweeper games and wins?

c. At what number is the dotplot centered? Could you have anticipated that? Why or why not?

d. Based on 12 wins on the last 20 games and the above dotplot, are you convinced that the author’s long-run proportion of winning at Minesweeper is higher than 50%? Explain how you are deciding.

e. Does this analysis prove that the author’s long-run proportion of winning at Minesweeper is 50%?

f. Suppose that it so happened that when the author played the last 20 games, she was also watching her favorite TV show. Does this information change your conclusion in part (d)? If yes, how and why? If not, why not?

### Buttered toast

1.1.10 If you drop a piece of buttered toast on the floor, is it just as likely to land buttered side up as buttered side down? It sure seems like mine always lands buttered side down! Suppose that 7 of the last 10 times I dropped toast it landed buttered side down. In order to carry out a statistical analysis, the One Proportion applet was used to see if my toast fell buttered side down a majority (more than 50%) of the time. Use the dotplot generated by the applet to answer the following questions.

![Dotplot](image)

a. How many dots are in the dotplot?

b. What does each dot represent in terms of dropped toast and buttered side down?

c. At what number is the dotplot centered? Could you have determined that before running the applet simulation? Why or why not?

d. Based on 7 of the last 10 slices of toast landing buttered side down and the above dotplot, are you convinced that the long-run proportion of times my toast lands buttered side down is greater than 50%? Explain how you are deciding.

e. Does this prove that my long-run proportion of dropping toast buttered side down is 50%?

### Mark’s tennis game

1.1.11 Mark is practicing his tennis serves. He wants to be able to tell newspaper reporters the long-run proportion of getting his first serve in. Mark gets 17 of his first 20 serves in.

a. Is 17 out of 20 = 85% a statistic or a parameter?

b. Is the long-run proportion of Mark’s first serve being in a statistic or a parameter?

c. Do you think that 17 out of 20 serves in is a possible outcome if Mark is just as likely to get his first serve in as to not get his first serve in?

d. Do you think that 17 out of 20 serves in is a plausible outcome if Mark is just as likely to get his first serve in as not to get his first serve in?
**Lady tasting tea**

A famous (in statistical circles) study involves a woman who claimed to be able to tell whether tea or milk was poured first into a cup. She was presented with eight cups containing a mixture of tea and milk, and she correctly identified which had been poured first for all eight cups.

1.1.12

a. Identify the observational units and variable in this study.

b. Identify the parameter for this study. *(Hint: The long-run proportion that…)*

c. Identify the sample size. Also, identify the observed value of the statistic for this study.

d. Is it possible that the woman could get all eight correct if she were randomly guessing with each cup?

e. Is it unlikely that the woman could get all eight correct if she were randomly guessing with each cup? *(Answer for now based on your intuition, without doing any analysis.)*

1.1.13

a. Describe how you can use a coin to address the question "Is it unlikely that the woman could get all eight correct if she were randomly guessing with each cup?" Be sure to include details such as how many times you would toss the coin and why; what would “heads” and “tails” stand for; what you would record after each set of coin tosses; how many times you would repeat this process. Also, how is this repeated coin tossing going to help you address the question?

b. Now use an applet simulation to address the question “Is it unlikely that the woman could get all eight correct if she were randomly guessing with each cup?” Then explain how your answer follows from your simulation results.

c. Based on your simulation analysis, would you conclude that the woman's result produces strong evidence that she is not guessing as to whether milk or tea was poured first? Also explain the reasoning process behind your answer.

d. Is your result in (c) statistically significant or is it plausible she is just guessing?

1.1.14 Suppose that I try to discern whether tea or milk is poured first for 8 cups and make the correct identification 5 times.

a. I say: “5 out of 8 is more than half, so one must conclude that I'm doing better than random guessing.” How would you respond? Use simulation results in your response.

b. With regards to my study on 8 cups of tea and milk, does this prove that I cannot tell whether milk or tea is placed in the cup first? Why or why not?

c. Now suppose that someone gets 14 correct out of 16 cups. Conduct a new simulation analysis to assess whether this result provides strong evidence that the person actually has ability better than random guessing to distinguish which was poured first. Be sure to describe which of the applet inputs needs to change and why. Also summarize your conclusion and explain your justification.

1.1.15

a. What is the parameter of interest for this study? *(Hint: The long-run proportion that…)*

b. What are two possible explanations for Zwerg's results of choosing the correct object 37 out of 48 times?

c. Which explanation of the two suggested in part (b) do you think is the more plausible, based on Zwerg's results? Explain. *(Note: Provide an explanation based only on your intuition.)*

d. If Zwerg is just randomly choosing between the objects, what is the chance that she will choose the correct object?

1.1.16 In order to discern whether Zwerg is doing better than just guessing (50% correct), we will employ the 3S strategy:

a. **Statistic:** How many times did Zwerg pick the correct object? Out of how many attempts?

b. **Simulate:** Using an applet, simulate 1,000 repetitions of having Zwerg choose between the two objects if she is doing so randomly.

c. Based on the value of the statistic from part (a), do you think the chance model is wrong? Why or why not?

d. **Strength of evidence:** Based on your answers for (a)–(c), state your conclusions about the research question of whether the data provide convincing evidence that Zwerg can correctly follow this type of direction by an experimenter more than 50% of the time?

e. Are the results of this study statistically significant or is the chance model plausible? How are you deciding?

1.1.17 In a related study, another experimenter-given cue was to place a marker in front of the correct object. Zwerg successfully chose the object with the marker 26 out of 48 times. Does this result show that Zwerg can correctly follow this type of direction by an experimenter more than 50% of the time?

a. What are two possible explanations for Zwerg's results of choosing the correct object 26 out of 48 times?

b. Which explanation of the two suggested in part (a) do you think is the more plausible, based on Zwerg's results? Explain based on your intuition.
c. If Zwerg is just randomly choosing between the objects, what is the chance that she will choose the correct object?

1.1.18 Refer to the data in the previous question where Zwerg successfully chose the object with the marker 26 out of 48 times. In order to discern whether Zwerg is doing better than just guessing (50% correct) we will employ the 3S strategy:

a. Statistic: How many times did Zwerg pick the correct object? Out of how many attempts?

b. Simulate: Using an applet, simulate 1,000 repetitions of having Zwerg choose between the two objects if she is doing so randomly. Where is the distribution centered?

c. Based on the value of the statistic from part (a), do you think the chance model is wrong? Why or why not?

d. Strength of evidence: Based on your answers for (a)–(c), state your conclusions about the research question of whether the data provide convincing evidence that Zwerg can correctly follow this type of direction by an experimenter more than 50% of the time?

e. Are the results of this study statistically significant or is the chance model plausible? How are you deciding?

f. Compare Zwerg’s performance in this study to that mentioned in Exercise 1.1.15. Do the results of this study provide more, less, or equally convincing evidence that Zwerg can correctly follow this type of direction by an experimenter more than 50% of the time, compared to the study results in Exercise 1.1.15? Could you have anticipated this? Explain.

g. Does this prove that Zwerg is just guessing when the marker is used to indicate the correct object? Why or why not?

Janine’s volleyball game

Janine is an ambidextrous volleyball player. She is practicing “short serves” in volleyball because her team has an upcoming match against an opponent who does not pass short serves very well. Janine must land a majority of her “short serves” in bounds or her coach will not let her use her short serve in the upcoming match. When serving left-handed Janine makes 23 out of 30 serves.

1.1.19

a. What is the parameter in this scenario? (Hint: The long-run proportion that …)

b. What are two possible explanations for Janine getting 23 out of 30 left-handed short serves in bounds?

c. Which of the two explanations from part (b) do you think is more plausible? No need to perform an analysis yet; explain your answer using only your intuition.

d. If Janine is just as likely to serve in as she is to not serve in, what is the chance she will serve in?

1.1.20 In order to discern whether or not Janine can land a majority of her short serves in bounds, we will employ the 3S strategy.

a. Statistic: How many times did Janine land a short serve in bounds? Out of how many service attempts?

b. Simulate: Using an applet, simulate 1,000 repetitions of having Janine serve 30 right-handed short serves if she is just as likely to serve in as she is to not serve in. Where is the distribution centered?

c. Based on the value of the statistic from part (a), do you think the chance model is wrong? Why or why not?

d. Strength of evidence: Based on your answers for (a)–(c), state your conclusions about the research question of whether the data provide convincing evidence that Janine can get her short serve in with her left hand more than 50% of the time.

e. Are the results of this study statistically significant or is the chance model plausible? How are you deciding?

Jerry’s tennis game*

Jerry is a tennis player. He is working on a really tough first serve. While practicing his new tough serve, he gets 12 out
of 20, or 60%, of his serves in. Jerry wants to know if his long-run proportion of getting his first serve in is greater than 50%.

1.1.23 What values would you enter into the One Proportion applet to run an analysis for Jerry?

a. Probability of heads ____________

b. Number of tosses ______________

c. Number of repetitions ____________

d. Use the applet to conduct the simulation study. If Jerry’s long-run proportion of getting his new first serve in is truly 50%, is 12 out of 20 a likely value? Is it unlikely? How are you deciding?

1.1.24 Suppose Jerry continues to serve and gets 60 out of 100 serves in. Use the One Proportion applet again to test if Jerry’s long-run proportion of getting his first serve in is greater than 50%. State the values you would enter into the applet.

a. Probability of heads ____________

b. Number of tosses ______________

c. Number of repetitions ____________

d. If Jerry’s long-run proportion of getting his new first serve in is truly 50%, is 60 out of 100 a likely value? Is it unlikely? How are you deciding?

e. Why do you suppose that you found different results in part (d) to this question as compared to part (d) of the previous question even though the sample statistic was 60% in both problems?

1.1.25 If an observed statistic from a study turns out to be a likely value under the chance model, then:

a. We can say we have evidence against the chance model.

b. We can say that the chance model is plausible.

c. We can say that the chance model is true.

d. We can’t say anything about the chance model, because it isn’t real data.

FAQ

1.1.26 Read FAQ 1.1.1 that describes a random process and answer the following questions.

a. Buzz’s section on which button to push can be thought of as a random process as long as we make what two assumptions?

b. What is the parameter of interest in the Buzz and Doris story?

SECTION 1.2

1.2.1* After you conduct a coin-flipping simulation, a graph of the _______ will be centered very close to 0.50. Choose from (A)–(D).

A. Process probability

B. Sample size

C. Number of heads

D. Proportion of heads

1.2.2 The graph of a null distribution will be centered approximately on:

A. The observed proportion

B. The observed count

C. The value of the probability in the null hypothesis

D. The number of repetitions performed

1.2.3* The p-value of a test of significance is:

A. The probability, assuming the null hypothesis is true, that we would get a result at least as extreme as the one that was actually observed

B. The probability, assuming the alternative hypothesis is true, that we would get a result at least as extreme as the one that was actually observed

C. The probability the null hypothesis is true

D. The probability the alternative hypothesis is true

1.2.4 Suppose a researcher is testing to see if a basketball player can make free throws at a rate higher than the NBA average of 75%. The player is tested by shooting 10 free throws and makes 8 of them. In conducting the related test of significance we have a computer applet do an appropriate simulation, with 1,000 repetitions, and produce a null distribution. This distribution represents:

A. Repeated results if the player makes 80% of his shots in the long run

B. Repeated results if the player makes 75% of his shots in the long run

C. Repeated results if the player makes more than 75% of his shots in the long run

D. Repeated results if the player makes more than 80% of his shots in the long run

1.2.5* The simulation (flipping coins or using the applet) done to develop the distribution we use to find our p-values assumes which hypothesis is true?

A. Null hypothesis

B. Alternative hypothesis

C. Both hypotheses

D. Neither hypothesis

1.2.6 When using the coin-flipping chance model, the most important reason you repeat a simulation of the study many times is:

A. To see whether the null distribution follows a symmetric, bell-shaped curve

B. To see whether the null distribution is centered at 0.50
C. To see whether your coin is really fair
D. To see how much variability there is in the null distribution

1.2.7* When we get a p-value that is very large, we may conclude that:
A. The null hypothesis has been proven to be true.
B. There is strong evidence for the alternative hypothesis.
C. The null hypothesis is plausible.
D. The alternative hypothesis has been proven to be false.

1.2.8 When we get a p-value that is very small, we may conclude that:
A. The null hypothesis has been proven to be true.
B. There is strong evidence for the alternative hypothesis.
C. The null hypothesis is plausible.
D. The alternative hypothesis has been proven to be false.

1.2.9* Suppose you are testing the hypotheses $H_0: \pi = 0.25$ and $H_a: \pi < 0.25$ and the observed statistic, $\hat{\pi}$, is equal to 0.30 with a sample size of 100.

a. If you are using a proportion as your statistic, where do you expect your null distribution to be centered?
b. If you are using a count as your statistic, where do you expect your null distribution to be centered?

1.2.10 Explain the meaning of each of the following symbols.
a. $H_0$
b. $H_a$
c. $\hat{\pi}$
d. $\pi$
e. $n$

1.2.11* What is the difference between $\hat{\pi}$ and the p-value?

Sarah the chimpanzee

1.2.12 A chimpanzee named Sarah was the subject in a study of whether chimpanzees can solve problems. Sarah was shown 30-second videos of a human actor struggling with one of several problems (for example, not able to reach bananas hanging from the ceiling). Then Sarah was shown two photographs, one that depicted a solution to the problem (like stepping onto a box) and one that did not match that scenario. Researchers watched Sarah select one of the photos, and they kept track of whether Sarah chose the correct photo depicting a solution to the problem. Sarah chose the correct photo in seven of eight scenarios that she was presented.

We want to run a test of significance to determine whether Sarah understands how to solve problems and will thus pick the photo of the correct solution more often than what would be done by random chance.

a. Describe the parameter of interest in the context of this study and assign a symbol to denote it.
b. What is the observed value of the statistic in this case?
c. Write out the null and alternative hypotheses for this first in words and then in symbols.
d. We conducted a test of significance using the One Proportion applet and got the following null distribution for the “number of heads.” (Note: this null distribution uses only 100 simulated samples and not the usual 1,000 or more.) Based on the null distribution, what is the p-value for the test? How did you calculate it?

<table>
<thead>
<tr>
<th>Probability of heads:</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of tosses:</td>
<td>8</td>
</tr>
<tr>
<td>Number of repetitions:</td>
<td>100</td>
</tr>
<tr>
<td>Animate</td>
<td></td>
</tr>
<tr>
<td>Toss Coins</td>
<td></td>
</tr>
</tbody>
</table>

Total = 100

![Null Distribution Graph]

e. Based on your p-value, do you have strong evidence that Sarah understands how to solve problems similar to those she was presented? How are you deciding? Justify.
f. Complete this sentence describing what the p-value means: If Sarah doesn’t understand how to solve problems and is just guessing at which picture to select, the probability she would …
g. What does a single dot in the null distribution represent in terms of Sarah and the photos?

Hope the dog*

1.2.13 Suppose you are testing to see if your dog, Hope, understands pointing towards an object. You put Hope through 20 trials and 12 times (or 60%) she goes to the correct object when given a choice between two objects. You then conduct a test of significance and generate the following 100 simulations using an applet.
1.2.15 Suppose you determine that in order to evaluate some data you need to conduct a simulation analysis.

a. In your own words explain how to conduct a simulation using a six-sided die where the sample size is 20 and the null hypothesis probability is $1/6$. Be specific.

b. Referring to part (a), explain how you would conduct the simulation using a regular deck of playing cards. Be specific.

c. How would your answer to (a) change if the sample size was 30?

d. How would your answer to (b) change if the sample size was 30?

e. How would your answer to (a) change if the null hypothesis probability is $2/3$?

f. How would your answer to (b) change if the null hypothesis probability is $2/3$?

Love, first

A recent study (Ackerman, Griskevicius, and Li, 2011) examined expressions of commitment between two partners in a committed romantic relationship. One aspect of the study involved 47 heterosexual couples who are part of an online pool of people willing to participate in surveys. These 47 couples were asked about which person was the first to say “I love you.” For 7 of those couples, the two people disagreed about the answer to this question. But both people agreed for the other 40 couples, so those 40 responses were included in the analysis. Previous studies have suggested that males tend to say “I love you” first.

1.2.16

a. Identify the observational units and variable in this study. Also classify the variable as categorical or quantitative.

b. State the appropriate null and alternative hypotheses (in words) for testing whether males are more likely to say “I love you” first.

We can express these hypotheses with symbols as:

<table>
<thead>
<tr>
<th>Null: $\pi = 0.50$</th>
<th>Alternative: $\pi &gt; 0.50$</th>
</tr>
</thead>
</table>

c. Describe (in words) what the symbol $\pi$ stands for here. It turned out that for 28 of the 40 couples in the sample (after the 7 couples who could not agree were excluded), the man said “I love you” before the woman did.

d. Determine the sample proportion (written as a decimal) of couples for whom the man was the first to say “I love you.” What symbol do we use to denote this proportion?

e. Describe how you could use a coin-flipping model to find the p-value for these data.

f. Use an applet to conduct a simulation analysis to assess the strength of evidence against the null hypothesis provided by the sample data. Report the values that you enter into the applet and describe how you approximate the p-value. Also report the value of this (approximate) p-value.

1.2.14 Suppose two researchers are conducting studies about animal behavior, and they both have the same null hypothesis and the same alternative hypothesis as each other. Also, suppose that Researcher A’s study results produce a p-value of 0.034, whereas Researcher B’s study results produce a p-value of 0.055. Which researcher has found stronger evidence against the null hypothesis and for the alternative hypothesis? Explain how you know.
g. Write a sentence or two interpreting this p-value: the probability of ____, assuming _____.

h. Summarize your conclusion from this p-value.

### 1.2.17

The researchers also interviewed 96 university students that had been or were currently involved in a romantic heterosexual relationship where at least one person said “I love you.” The students were asked, “Think about your last or current romantic relationship in which someone confessed their love. In this relationship, who admitted love first?”

a. Identify the observational units and variable in this study. Also classify the variable as categorical or quantitative.

b. State the appropriate null and alternative hypotheses (in words) for testing whether rhesus monkeys can interpret the head jerk better than random chance.

c. Determine the sample proportion (written as a decimal) of couples for whom the man was the first to say “I love you.” What symbol do we use to denote this proportion?

d. Describe how you could use a coin-flipping model to find the p-value for these data.

e. Use an applet to conduct a simulation analysis with at least 1,000 repetitions. Based on the null distribution provided by the applet, explain how you are deciding how you would decide whether the observed data provide convincing evidence that rhesus monkeys can read human gestures better than random chance. Summarize the conclusion that you would draw about the research question of whether rhesus monkeys have some ability to understand gestures made by humans.

f. Write a sentence or two interpreting this p-value: the probability of ____, assuming _____.

g. Summarize your conclusion from this p-value.

### Rhesus monkeys*

A recent article (Hauser, Glynn, and Wood, 2007) described a study that investigated whether rhesus monkeys have some ability to understand gestures made by humans. In one part of the study, the experimenter approached individual rhesus monkeys and placed two boxes an equal distance from the monkey. The experimenter then placed food in one of the boxes, making sure that the monkey could tell that one of the boxes received food without revealing which one. Finally, the researcher pointed and looked toward the box that contained the food. This process was repeated for a total of 40 rhesus monkeys. It turned out that 31 of the monkeys approached the box that the human had gestured toward, and 9 approached the other box. The purpose is to investigate whether rhesus monkeys can interpret human gestures better than random chance.

a. Identify the observational units and variable in this study. Also classify the variable as categorical or quantitative.

b. Describe in words the parameter of interest for this study and assign a symbol to it.

c. Determine the sample proportion of monkeys who picked the box towards which the human had gestured. Is this value a parameter or a statistic? What is the symbol you should use to denote this proportion?

d. State the appropriate null and alternative hypotheses in the context of this study, first in words and then in symbols.

e. Describe how you could use a coin to conduct a simulation analysis of this study and its result. Give sufficient detail that someone else could implement this simulation analysis based on your description. Be sure to indicate how you would decide whether the observed data provide convincing evidence that rhesus monkeys can read human gestures better than random chance. Summarize the conclusion that you would draw about the research question of whether rhesus monkeys have some ability to understand gestures made by humans.

### 1.2.19

In one part of the study, the experimenter approached individual rhesus monkeys and placed two boxes an equal distance from the monkey. The experimenter then placed food in one of the boxes, making sure that the monkey could tell that one of the boxes received food without revealing which one. Finally, the researcher pointed and looked toward the box that contained the food. This process was repeated for a total of 40 rhesus monkeys. It turned out that 31 of the monkeys approached the box that the human had gestured toward, and 9 approached the other box. The purpose is to investigate whether rhesus monkeys can interpret human gestures better than random chance.

a. Identify the observational units and variable in this study. Also classify the variable as categorical or quantitative.

b. Describe in words the parameter of interest for this study and assign a symbol to it.

c. Determine the sample proportion of monkeys who picked the box towards which the human had gestured. Is this value a parameter or a statistic? What is the symbol you should use to denote this proportion?

d. State the appropriate null and alternative hypotheses in the context of this study, first in words and then in symbols.

e. Describe how you could use a coin to conduct a simulation analysis of this study and its result. Give sufficient detail that someone else could implement this
simulation analysis based on your description. Be sure to indicate how you would decide whether the observed data provide convincing evidence that rhesus monkeys can interpret human gestures better than random chance.

e. Use an applet to conduct a simulation analysis with at least 1,000 repetitions. Based on the null distribution produced by the applet, explain how you are deciding whether the observed data provide convincing evidence that rhesus monkeys can read human gestures better than random chance. Summarize the conclusion that you would draw about the research question of whether rhesus monkeys have some ability to understand gestures made by humans.

Minesweeper*

1.2.20 Recall that one of the authors liked to play the game Minesweeper (Exercise 1.1.8) and she won 12 of the last 20 games she played. Now that you know how to calculate p-values, use an applet to find the p-value to investigate whether her results provide convincing evidence that her long-run proportion of winning at Minesweeper is higher than 0.50. Report the p-value and use the p-value to state a conclusion about the strength of evidence.

Spider Solitaire

1.2.21 While the author in the previous question likes to play Minesweeper, another author likes to play Spider Solitaire on the computer. In his last 40 games, he won 24 of them. Use an applet to find the p-value to investigate whether his results provide convincing evidence that his long-run proportion of winning at Spider Solitaire is higher than 0.50. Report the p-value and use the p-value to state a conclusion about the strength of evidence.

Spinning a coin*

1.2.22 It has been stated that spinning a coin on a table will result in it landing heads side up fewer than 50% of the time in the long run. One of the authors tested this by spinning a penny 50 times on a table and it landed heads side up 21 times. A test of significance was then conducted with the following hypotheses.

\[ H_0: \pi = 0.50 \quad \quad H_1: \pi < 0.50 \]

a. Describe what the symbol \( \pi \) stands for in this context.

b. Use an applet to conduct a simulation with at least 1,000 repetitions. What is your p-value? Based on your p-value, is there strong evidence that the probability the spun coin will land heads up is less than 0.50?

c. Suppose you focused on the proportion of times the coin landed tails up instead of heads up. How would your hypotheses be different? What would you do differently to calculate your p-value?

Flipping a coin

1.2.23 According to Stanford mathematics and statistics professor Persi Diaconis, the probability a flipped coin that starts out heads up will also land heads up is 0.51. Suppose you want to test this. More specifically, you want to test to determine if the probability that a coin that starts out heads up will also land heads up is more than 0.50. Suppose you flip a coin (that starts out heads up) 100 times and find that it lands heads up 53 of those times.

a. If \( \pi \) stands for the probability a coin that starts heads up will also land heads up, write out the hypotheses for this study in symbols.

b. Use an applet to conduct a simulation with at least 1,000 repetitions. What is your p-value? Based on your p-value, is there strong evidence that the probability that a coin that starts heads up will also land heads up is more than 0.50?

c. Even though the coin in the study landed heads up more than 50% of the time, you should not have found strong evidence that it will land heads up more than 50% of the time in the long run. Does this mean that Diaconis is wrong about his assertion that it will land heads up 51% of the time in the long run? Explain.

d. A newspaper article described the 0.51 probability by saying, “This means that if a coin is flipped with its heads side facing up, it will land the same way 51 out of 100 times.” What is wrong with this statement?

Free throws*

1.2.24 Suppose your friend says he can shoot free throws as well as someone in the NBA and you don't think he is that good. You know that the NBA average for shooting free throws is 75% and decide to test your friend. You have him shoot 20 free throws and he makes 12 of them. Based on this, do you have strong evidence that he is worse than the average NBA player? Answer this by conducting a simulation and reporting your p-value.

1.2.25 Suppose you retest your friend from the previous question to see if he is a worse free throw shooter than the NBA average of 75%. He shoots 20 more free throws and again makes 12 of them. You combine the data from the two tests together so he made 24 of his 40 attempts. Based on this, do you have strong evidence that he is worse than the average NBA player? Answer this by conducting a simulation and reporting your p-value.

FAQ

1.2.26 Read FAQ 1.2.1 that describes why we need to include "or more extreme" when computing a p-value and, in your own words, describe why we don't just compute the p-value based on the probability of a single outcome, but also include outcomes that are more extreme in that probability.
1.2.27 Read FAQ 1.2.2 that explores how small a p-value needs to be in order for us to have strong evidence against the null hypothesis and answer the following questions.

a. What does Persi Diaconis claim to be able to do?
b. How many heads does Diaconis have to flip in a row in order for the probability of such an outcome to drop below 0.05? Is that the number you would pick for a point where you would start to become convinced that he could flip a coin and get heads every time?

SECTION 1.3

1.3.1* Which standardized statistic (standardized sample proportion) gives you the strongest evidence against the null hypothesis?
A. $z = 1$
B. $z = 0$
C. $z = -3$
D. $z = 1.80$

1.3.2 Consider the output given below that was obtained using the One Proportion applet. Use information from the output to find the standardized statistic for a sample proportion value of 0.45.

1.3.3* Consider the two null distributions (A and B) given below, both for proportion of successes. For which null distribution will the standardized statistic for a sample proportion value of 0.60 be farther from 0? How are you deciding?

1.3.4 Suppose that your hypotheses are $H_0: \pi = 0.25$ and $H_1: \pi < 0.25$. In the context of these hypotheses, which of the following standardized statistics would provide the strongest evidence against the null hypothesis and for the alternative hypothesis? Why?
A. $z = -1$
B. $z = 0$
C. $z = 3$
D. $z = -1.80$

1.3.5* Identify these statements as either true or false.

a. A p-value can be negative.
b. A standardized statistic can be negative.
c. We run tests of significance to determine whether $\hat{p}$ is larger or smaller than some value.
d. As a p-value gets smaller, its corresponding standardized statistic gets closer to zero.

1.3.6 Suppose that a standardized statistic (standardized sample proportion) for a study is calculated to be 2.45. Which of the following is the most appropriate interpretation of this standardized statistic?
A. The observed value of the sample proportion is 2.45 SDs away from the hypothesized parameter value.
B. The observed value of the sample proportion is 2.45 SDs above the hypothesized parameter value.
C. The observed value of the sample proportion is 2.45 times the hypothesized parameter value.
D. The study results are statistically significant.
**Rock-paper-scissors**

Refer to Example 1.2, which explored players’ choices in the game rock-paper-scissors.

1.3.7 Suppose that you play the game with three different friends separately with the following results:

   Friend A chose scissors 100 times out of 400 games.
   Friend B chose scissors 20 times out of 120 games.
   Friend C chose scissors 65 times out of 300 games.

   Suppose that for each friend you want to test whether the long-run proportion that the friend will pick scissors is less than 1/3.

   a. Determine the appropriate standardized statistic for each friend’s results. (Hint: You will need to get the standard deviation of the simulated statistics from the null distribution produced by an applet.)

   b. Based on the standardized statistics, which friend’s data provides the strongest evidence that the long-run proportion that the friend will choose scissors is less than 1/3? Which friend’s data provides the least strong evidence?

1.3.8 Suppose that you play the game, again, with two other friends separately with the following results:

   Friend D chose rock 200 times out of 400 games.
   Friend E chose rock 20 times out of 40 games.

   Suppose that for each friend you want to test whether the long-run proportion that the friend will pick rock is more than 1/3.

   a. Even though both of the friends played rock the same proportion of times, one of the two friends’ data provide more evidence the null hypothesis is wrong. Which friend’s data do you think provides more evidence against the null hypothesis? Why?

   b. Based on your answer to (a), which friend’s data yields a smaller p-value? Why?

   c. Based on your answer to (a), which friend’s data yields a larger standardized statistic? Why?

   d. Based on your answer to (c), which friend’s null distribution has a smaller standard deviation? Why?

1.3.9 Suppose that you play the game, again, with two other friends separately with the following results:

   Friend F chose rock 15 times out of 40 games.
   Friend G chose rock 30 times out of 40 games.

   Suppose that for each friend you want to test whether the long-run proportion that the friend will pick rock is more than 1/3. In which friend’s case (F or G) will you find a larger (more extreme) standardized statistic? Explain your reasoning, without actually calculating the standardized statistics.

**Minesweeper**

1.3.10 Recall that one of the authors likes to play the game Minesweeper (Exercise 1.1.8) and in the last 20 games she played she won 12. Use an applet to conduct appropriate simulations in order to calculate the standardized statistic for these data and investigate whether her results provide convincing evidence that her long-run proportion of winning at Minesweeper is higher than 50%? Report the calculated standardized statistic and use it to state a conclusion about the strength of evidence.

**Doris and Buzz**

Refer to Example 1.1 involving two dolphins, Buzz and Doris, where the researcher Dr. Bastian was investigating whether dolphins can communicate.

1.3.11 Recall that in the first study, out of 16 attempts, Buzz pushed the correct button 15 times. Calculate the standardized statistic for these data to investigate whether Buzz’s results provide convincing evidence that his long-run proportion of pushing the correct button is higher than 0.50. Report the calculated standardized statistic and use it to state a conclusion about the strength of evidence.

1.3.12 Recall that in the second study mentioned (where they later found out that the fish-delivering equipment had been malfunctioning), out of 28 attempts, Buzz pushed the correct button 16 times. Calculate the standardized statistic for these data to investigate whether Buzz’s results provide convincing evidence that his long-run proportion of pushing the correct button is higher than 0.50. Report the calculated standardized statistic and use it to state a conclusion about the strength of evidence.

**Right or left**

1.3.13 Most people are right-handed, and even the right eye is dominant for most people. Molecular biologists have suggested that late-stage human embryos tend to turn their heads to the right. In a study reported in *Nature* (2003), German bio-psychologist Onur Güntürkün conjectured that this tendency to turn to the right manifests itself in other ways as well, so he studied kissing couples to see which side they tended to lean their heads while kissing. He and his researchers observed kissing couples in public places such as airports, train stations, beaches, and parks. They were careful not to include couples who were holding objects such as luggage that might have affected which direction they turned. For each kissing couple observed, the researchers noted whether the couple leaned their heads to the right or to the left. They observed 124 couples, age 13–70 years. Suppose that we want to use the data from this study to investigate whether kissing couples tend to lean their heads right more often than would happen by random chance.
a. Define the parameter of interest in the context of the study and assign a symbol to it.
b. State the null hypothesis and the alternative hypothesis using the symbol defined in part (a).
c. Of the 124 kissing couples, 80 were observed to lean their heads right. What is the observed proportion of kissing couples who leaned their heads to the right? What symbol should you use to represent this value?
d. Determine the standardized statistic from the data. *(Hint: You will need to get the standard deviation of the simulated statistics from the null distribution.)*
e. Interpret the standardized statistic in the context of the study. *(Hint: You need to talk about the value of your observed statistic in terms of standard deviations assuming ______ is true.)*
f. Based on the standardized statistic, state the conclusion that you would draw about the null and alternative hypotheses.

**1.3.14** Suppose that instead of $H_0: \pi = 0.50$ like it was in the previous exercise, our null hypothesis was $H_0: \pi = 0.60$.

a. In the context of this null hypothesis, determine the standardized statistic from the data where 80 of 124 kissing couples leaned their heads right. *(Hint: You will need to get the standard deviation of the simulated statistics from the null distribution.)*
b. How, if at all, does the standardized statistic calculated here differ from that when $H_0: \pi = 0.50$? Explain why this makes sense.

**Love, first**

**1.3.15** A previous exercise (1.2.16) introduced you to a study of 40 heterosexual couples. In 28 of the 40 couples the male said “I love you” first. The researchers were interested in learning whether these data provided evidence that significantly more than 50% of couples the male says “I love you” first.

a. State the null hypothesis and the alternative hypothesis in the context of the study.
b. Determine the standardized statistic from the data. *(Hint: You will need to get the standard deviation of the simulated statistics from the null distribution.)*
c. Interpret the standardized statistic in the context of the study. *(Hint: You need to talk about the value of your observed statistic in terms of standard deviations assuming ______ is true.)*
d. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether males are more likely to say “I love you” first.

**Rhesus monkeys**

Revisit Exercise 1.2.18 about the study on Rhesus monkeys. When given a choice between two boxes, 30 out of 40 monkeys approached the box that the human had gestured toward, and 10 approached the other box. The purpose is to investigate whether rhesus monkeys can interpret human gestures better than random chance.

**1.3.16** For this study:

a. State the null hypothesis and the alternative hypothesis in the context of the study.
b. Determine the standardized statistic from the data. *(Hint: You will need to get the standard deviation of the simulated statistics from the null distribution in an applet.)*
c. Interpret the standardized statistic in the context of the study. *(Hint: You need to talk about the value of your observed statistic in terms of standard deviations assuming ______ is true.)*
d. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether rhesus monkeys have some ability to understand gestures made by humans.

**Tasting tea**

Revisit Exercise 1.1.12 about the study on a lady tasting tea. When presented with eight cups containing a mixture of milk and tea, she correctly identified whether tea or milk was poured first for all eight cups. Is she doing better than if she were just guessing?

**1.3.17** For this study:

a. Define the parameter of interest in the context of the study and assign a symbol to it.
b. State the null hypothesis and the alternative hypothesis using the symbol defined in part (a).
c. What is the observed proportion of times the lady correctly identified what was poured first into the cup? What symbol should you use to represent this value?
d. Suppose that you were to generate the null distribution of the sample proportion of correct answers, that is, the distribution of possible values of sample proportion of correct identifications if the lady always guesses. Where would you anticipate this distribution would center? Also, do you anticipate the SD of the null distribution to be negative, positive, or 0? Why?
e. Use an applet to generate the null distribution of sample proportion of correct identifications and use it to determine the standardized statistic.
f. Interpret the standardized statistic in the context of the study. *(Hint: You need to talk about the value of your observed statistic in terms of standard deviations assuming ______ is true.)*
g. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether the lady does better than randomly guess.
Zwerg

Refer to a previous exercise, 1.1.15, involving Zwerg the sea lioness.

1.3.18  In one study, Zwerg, when given two choices, successfully chose the object pointed at by the experimenter 37 times out of 48 trials. Does this result show that Zwerg can correctly follow this type of direction (“pointing at”) by an experimenter more than 50% of the time?

a. Define the parameter of interest in the context of the study and assign a symbol to it.

b. State the null hypothesis and the alternative hypothesis using the symbol defined in part (a).

c. What is the observed proportion of times Zwerg picked the correct object? What symbol should you use to represent this value?

d. Suppose that you were to generate the null distribution of the sample proportion of correct answers, that is, the distribution of possible values of sample proportion of correct identifications if Zwerg always guesses. Where would you anticipate this distribution would center? Also, do you anticipate the SD of the null distribution to be negative, positive, or 0? Why?

e. Use an applet to generate the null distribution of sample proportion of correct answers and use it to determine the standardized statistic.

f. Interpret the standardized statistic in the context of the study. (Hint: You need to talk about the value of your observed statistic in terms of standard deviations assuming ______ is true.)

g. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether Zwerg does better than randomly guess when the cue is that the experimenter places a marker in front of the correct object.

1.3.19  In a related study, Zwerg, when given two choices, successfully chose the object indicated by a “marker” 26 out of 48 times. Does this result show that Zwerg can correctly follow this type of direction (“placing a marker”) by an experimenter more than 50% of the time?

a. Define the parameter of interest in the context of the study and assign a symbol to it.

b. State the null hypothesis and the alternative hypothesis using the symbol defined in part (a).

c. What is the observed proportion of times Zwerg picked the correct object? What symbol should you use to represent this value?

d. Suppose that you were to generate the null distribution of the sample proportion of correct answers, that is, the distribution of possible values of sample proportion of correct identifications if Zwerg always guesses. Where would you anticipate this distribution would center? Also, do you anticipate the SD of the null distribution to be negative, positive, or 0? Why?

e. Use an applet to generate the null distribution of sample proportion of correct answers and use it to determine the standardized statistic.

f. Interpret the standardized statistic in the context of the study. (Hint: You need to talk about the value of your observed statistic in terms of standard deviations assuming ______ is true.)

g. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether Zwerg does better than randomly guess when the cue is that the experimenter places a marker in front of the correct object.

Helper or hinderer*

Use the following information to answer the next two questions. An investigation reported in the November 2007 issue of Nature (Hamlin, Wynn, and Bloom) aimed at assessing whether infants take into account an individual’s actions towards others in evaluating that individual as appealing or aversive, perhaps laying the foundation for social interaction. In one component of the study, each of sixteen 10-month-old healthy infants (in New England) were shown a “climber” character (a piece of wood with “gooble” eyes glued onto it) that could not make it up a hill in two tries. Then they were shown two scenarios for the climber’s next try, one where the climber was pushed to the top of the hill by another character (“helper”) and one where the climber was pushed back down the hill by another character (“hinderer”). Each infant was alternately shown these two scenarios many times. Then the child was presented with both pieces of wood (the “helper” and the “hinderer”) and asked to pick one to play with. In the study 14 out of the 16 babies picked the “helper” toy. Does this provide evidence that such 10-month-old infants have a genuine preference for the helper toy?

1.3.20  a. Define the parameter of interest in the context of the study and assign a symbol to it.

b. State the null hypothesis and the alternative hypothesis using the symbol defined in part (a).

c. What is the observed proportion of times the infants picked the helper toy? What symbol should you use to represent this value?

d. Use an applet to generate the null distribution of the proportion of “successes.” Determine the standardized statistic for the observed sample proportion of “successes.” Show your work.

e. Interpret the standardized statistic in the context of the study. (Hint: You need to talk about the value of your
observed statistic in terms of standard deviations assuming _____ is true.)

f. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether such 10-month-old infants have a genuine preference for the helper toy.

1.3.21
a. Based on the value of the standardized statistic obtained in the previous question, do you anticipate the p-value to be small? Why or why not?
b. Use an appropriate applet to find and report the p-value. Also, interpret the p-value. (Hint: The p-value is the probability of … assuming. …)
c. Based on the p-value, state the conclusion that you would draw about the research question of whether such 10-month-old infants have a genuine preference for the helper toy.
d. Do the p-value and standardized statistic provide similar strength of evidence? Do they both lead you to the same conclusion? Did you anticipate this? Why or why not?

Choosing numbers

Use the following information to answer the next four questions. One of the authors read somewhere that it’s been conjectured that when people are asked to choose a number from the choices 1, 2, 3, and 4, they tend to choose “3” more than would be expected by random chance.

To investigate this, she collected data in her class. Here is the table of responses from her students:

<table>
<thead>
<tr>
<th>Chose 1</th>
<th>Chose 2</th>
<th>Chose 3</th>
<th>Chose 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4</td>
<td>14</td>
<td>5</td>
</tr>
</tbody>
</table>

1.3.22
a. Define the parameter of interest in the context of the study and assign a symbol to it.
b. State the null hypothesis and the alternative hypothesis using the symbol defined in part (a).
c. What is the observed proportion of times students chose the number 3? What symbol should you use to represent this value?
d. Use an applet to generate the null distribution of the proportion of “successes.” Report the mean and SD of this null distribution.
e. Determine the standardized statistic for the observed sample proportion of “successes.”
f. Interpret the standardized statistic in the context of the study. (Hint: You need to talk about the value of your observed statistic in terms of standard deviations assuming _____ is true.)
g. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether students tend to have a genuine preference for the number 3 when given the choices 1, 2, 3, and 4.

1.3.23
a. Based on the value of the standardized statistic obtained in the previous question, do you anticipate the p-value to be small? Why or why not?
b. Use an appropriate applet to find and report the p-value. Also, interpret the p-value. (Hint: The p-value is the probability of … assuming. …)
c. Based on the p-value, state the conclusion that you would draw about the research question of whether when people are asked to choose from the numbers 1, 2, 3, and 4, they tend to choose “3” more than would be expected by random chance.
d. Do the p-value and standardized statistic provide similar strength of evidence? Do they both lead you to the same conclusion? Did you anticipate this? Why or why not?

1.3.24 Suppose that you wanted to investigate whether people tend to pick a “big” number (3 or 4) rather than a “small” number (1 or 2).
a. In this context, define the parameter of interest and assign a symbol to it.
b. State the null hypothesis and the alternative hypothesis using the symbol defined in part (a).
c. Recall that, of 33 students, 19 picked a “big” number (that is, 3 or 4). What is the observed proportion of times students picked a “big” number? What symbol should you use to represent this value?
d. Use an applet to generate the null distribution of the proportion of “successes.” Report the mean and SD of this null distribution.
e. Determine the standardized statistic for the observed sample proportion of “successes.”
f. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether people tend to pick a “big” number.

g. Based on the standardized statistic, state the conclusion that you would draw about the research question of whether students tend to have a genuine preference for the number 3 when given the choices 1, 2, 3, and 4.


are asked to pick a number from the choices 1, 2, 3, and 4, they tend to pick a “big” number.

d. Do the p-value and standardized statistic provide similar strength of evidence? Do they both lead you to the same conclusion? Did you anticipate this? Why or why not?

SECTION 1.4

Racquet spinning*

Use the following information to answer the following three questions. Reconsider the racquet-spinning Exercise 1.1.1 from earlier. Researchers wanted to investigate whether a spun tennis racquet is equally likely to land with the label facing up or down.

1.4.1 Does this racquet-spinning study call for a one-sided or a two-sided alternative?
A. One-sided, because there is only one variable: how the label lands
B. Two-sided, because there are two possible outcomes: up or down
C. One-sided, because the researchers want to know whether the label is more likely to land face up
D. Two-sided, because the researchers want to know whether the spinning process is fair or biased in either direction

1.4.2 Which of the following will always be true about the standardized statistic for the racquet-spinning study?
A. The standardized statistic increases as the sample proportion that land “up” increases.
B. The standardized statistic decreases as the sample proportion that land “up” increases.
C. The standardized statistic increases as the sample proportion that land “up” gets farther from 0.50.
D. The standardized statistic decreases as the sample proportion that land “up” gets farther from 0.50.

1.4.3 Which of the following will always be true about the p-value for the racquet-spinning study?
A. The p-value increases as the sample proportion that land “up” increases.
B. The p-value decreases as the sample proportion that land “up” increases.
C. The p-value increases as the sample proportion that land “up” gets farther from 0.50.
D. The p-value decreases as the sample proportion that land “up” gets farther from 0.50.

Chess-boxing

Use the following information to answer the following three questions. You have heard that in sports like boxing there might be some competitive advantage to those wearing red uniforms. You want to test this with your new favorite sport of chess-boxing. You randomly assign blue and red uniforms to contestants in 20 matches and find that those wearing red won 14 times (or 70%). You conduct a test of significance using simulation and get the following null distribution. (Note this null distribution uses only 100 simulated samples and not the usual 1,000 or more.)

- Probability of success ($\pi$): 0.5
- Sample size ($n$): 20
- Number of samples: 100

Draw Samples

Total = 100

1.4.4 One-sided or two?

a. Suppose you want to see if competitors wearing red win more than 50% of the matches in the long run, so you test $H_0: \pi = 0.50$ versus $H_a: \pi > 0.50$. What is your p-value based on the above null distribution?
b. Suppose you now want to see if competitors wearing either red or blue have an advantage, so you test $H_0: \pi = 0.50$ versus $H_a: \pi \neq 0.50$. What is your p-value now based on the above null distribution?

1.4.5 Suppose you are testing the hypothesis $H_0: \pi = 0.50$ versus $H_a: \pi > 0.50$. You get a sample proportion of 0.54 and find that your p-value is 0.08. Now suppose you redid your study with each of the following changes. Will your new p-value be larger or smaller than the 0.08 your first obtained?
a. You increase the sample size and still find a sample proportion of 0.54.
b. Keeping the sample size the same, you take a new sample and find a sample proportion of 0.55.
c. With your original sample, you decided to test a two-sided alternative instead of $H_a: \pi > 0.50$.

1.4.6 Identify these statements as either true or false.
a. Using a simulation-based test, the p-value for a two-sided test will be about twice as large as the corresponding p-value for a one-sided test.
b. We run tests of significance to determine if \( \hat{p} \) is larger, smaller, or different than some value.

**Minesweeper**

**1.4.7** Look back to Exercise 1.1.8. Recall that in the last 20 games of Minesweeper she played, the author won 12.

a. What if she had played 20 games and won 18? Would that provide stronger, weaker, or evidence of similar strength compared to 12 wins out of 20, to conclude that her long-run proportion of winning at Minesweeper is higher than 50%? Explain how you are deciding.

b. What if she had played 100 games and won 60? Would that provide stronger, weaker, or evidence of similar strength compared to 12 wins out of 20, to conclude that her long-run proportion of winning at Minesweeper is higher than 50%? Explain how you are deciding.

c. What if she had played 30 games and won 12? Would that provide evidence that her long-run proportion of winning at Minesweeper is higher than 50%? Explain how you are deciding.

**Divine Providence?**

Use the following information to answer the next five questions. Dr. John Arbuthnot (1667–1735) was physician to England’s Queen Anne. He was also one of the first scientists to publish a use of p-values and to apply the logic of statistical inference.1 His goal was to prove the existence of God, and he took as evidence the fact that for 82 years in a row male christenings had outnumbered female christenings. His argument: The p-value is so extremely tiny that the evidence against the null hypothesis to be strong or weak? Explain your reasoning.

Based solely on the observed value of \( \hat{p} \) and the null value for \( \pi \), would you expect the evidence against the null hypothesis to be strong or weak? Explain your reasoning.

**1.4.10** Suppose (against all historical evidence) Arbuthnot had been an early feminist, equally willing to accept an excess of female births as evidence of “Divine Providence.” In that case, his p-value would have been ______ (half, twice) as large.

**1.4.11** Summary. For the analysis based on the data above:

a. The distance between \( \hat{p} \) and \( \pi \) is ______ (tiny, small, large, huge).

b. The sample size is ______ (tiny, small, large, huge).

c. The alternative hypothesis is ______-sided (one, two).

**Buzz and Doris**

Use the following information to answer the next three questions. Refer to Example 1.1. Recall that Buzz pushed the correct lever 15 out of 16 times when given the choice between two levers to push.

**1.4.12** Which of the three influences (distance, sample size, one- or two-sided) would change? Which would stay the same?

**1.4.13** In the actual experiment, Buzz got 15 right out of 16. Suppose the numbers had been exactly double that: 30 right out of 32.

**1.4.14** In the actual experiment, Buzz got 15 right out of 16. Suppose he had only guessed right 14 times out of 16.

**1.4.15** Just for the sake of this exercise, imagine that the investigators wanted their statistical test to include the possibility that Buzz was a very sadistic dolphin, so malicious that he would willingly give up his own chance at a fish just to deprive Doris.

**Harley**

Use the following information to answer the next two questions. Refer to Exploration 1.1. Researchers investigated

---

whether Harley the dog could correctly choose between two objects when a researcher bowed towards one of the cups.

1.4.16  Harley got 9 right out of 10. Suppose instead the numbers had been 18 right out of 20.

a. Which of the three influences on strength of evidence (distance, sample size, one- or two-sided) would change? Which would stay the same?

b. Overall, the evidence against the null hypothesis would be ______ (stronger, weaker).

1.4.17  Suppose that instead of the actual result (9 right out of 10) Harley had only been right 8 times out of 10.

a. Which of the three influences on strength (distance, sample size, one- or two-sided) would change? Which would stay the same?

b. Overall, the evidence against the null hypothesis would be ______ (stronger, weaker).

Krieger*

Can domestic dogs understand human body cues such as bowing, pointing, or glancing? The experimenter presented a body cue toward one of two objects and recorded whether or not the dog being tested correctly chose the object indicated. A four-year-old male pit bull named Krieger participated in this study. He chose the correct object 6 out of 10 times when the experimenter turned and looked towards the correct object.

1.4.18  

a. Identify the parameter of interest for this study and assign a symbol to it.

b. If Krieger is just randomly choosing between the two objects, what is the chance that he will choose the correct object?

c. State your null hypothesis and an appropriate one-sided alternative hypothesis.

1.4.19  

a. Statistic: How many times did Krieger choose the correct object? Out of how many attempts? Thus, what proportion of the time did Krieger choose the correct object?

b. Simulate: Using an applet, simulate 1,000 repetitions of having the dog choose between the two objects if he is doing so randomly. Report the mean and standard deviation.

c. Based on the study's result, what is the p-value for this test?

d. Strength of evidence: What are your conclusions based on the p-value?

e. How would the evidence change if Krieger got 8 out of 10 correct?

1.4.20  

a. Suppose that you decide to use a two-sided alternative hypothesis. What additional values for the parameter will now be part of the alternative?

b. Conjecture how, if at all, the p-value for the two-sided alternative will change compared to that reported in Exercise 1.4.19, part (c) for the one-sided alternative hypothesis. Explain your choice.

Increase  Stay the same  Decrease

c. Use an applet to find and report the p-value for the two-sided alternative. Did the p-value behave as you had conjectured in (b)?

d. Complete the following: A p-value corresponding to a one-sided alternative hypothesis provides ______ (weaker/the same/stronger) evidence against the null hypothesis, compared to a p-value calculated from the same data, but corresponding to a two-sided alternative hypothesis.

1.4.21  Suppose that we repeated the same study with Krieger, and this time he chose the correct object 12 out of 20 times.

a. Conjecture how, if at all, the p-value would change from that reported in Exercise 1.4.19, part (c) (one-sided alternative).

Increase  Stay the same  Decrease

c. Use an applet, find and report the p-value for the two-sided alternative. Did it behave the way you conjectured in (a)?

d. Complete the following: When the sample proportion of successes stays the same, as the sample size gets larger, then the strength of evidence against the null hypothesis gets ______ (weaker/the same/stronger).

1.4.22  In another part of the study, instead of looking at the object, the experimenter kept her eyes on the dog and leaned toward the object. For this part of the study, Krieger got 9 right out of 10.

a. Identify the parameter of interest for this part of the study and assign a symbol to it.

b. If Krieger is just randomly choosing between the two objects, what is the chance that he will choose the correct object?

c. State your null hypothesis and an appropriate one-sided alternative hypothesis.

d. Conjecture how, if at all, the p-value for these data will change compared to that from the data in Exercises 1.4.18 and 1.4.19 (where Krieger got 6 out of 10 correct). Explain your choice.

Increase  Stay the same  Decrease
1.4.23  For the “leaning” version of the study from the previous question:

a. Statistic: How many times did Krieger choose the correct object? Out of how many attempts? Thus, what proportion of the time did Krieger choose the correct object?

b. Simulate: Using an applet, simulate 1,000 repetitions of having the dog choose between the two objects if he is doing so randomly. Report the null and standard deviation.

c. Based on the study’s result, what is the p-value for this test?

d. Approximately what proportion of the 10 attempts would Krieger have needed to get correct in order to yield a p-value of approximately 0.05?

1.4.24

a. Based on the study’s result, what is the standardized statistic for this test?

b. Strength of evidence: What are your conclusions based on the p-value you found in part (d) from the previous exercise? Are the conclusions the same if you base them off the standardized-statistic you found in (a)?

c. Revisit your conjecture in Exercise 1.4.22, part (d). Did the p-value behave the way you had conjectured?

The sign test

So far, the outcome has always been binary—Yes/No, Right/Wrong, Heads/Tails, etc. What if outcomes are quantitative, like heights or percentages? Although there are specialized methods for such data that you will learn in a later chapter, you can also use the methods and logic you have already learned for situations of a very different sort: (1) outcomes are quantitative, (2) you want to compare two conditions A and B, and (3) your data come in pairs, one A and one B in each pair. To apply the coin toss model, you simply ask for each pair, “Is the A value bigger than the B value?” The resulting test is called the “sign test” because the difference \((A - B)\) is either plus or minus. Here’s a summary table:

<table>
<thead>
<tr>
<th>Coin toss</th>
<th>Heads</th>
<th>P(Heads)</th>
<th>Null hypothesis</th>
<th>Statistic</th>
</tr>
</thead>
</table>
| Buzz's guess| Right | \( \pi = P \text{ (Right)} \) | \( \pi = 0.50 \) | \( \hat{p} \)
| Each pair   | \( A > B \) | \( \pi = P \text{ (A > B)} \) | \( \pi = 0.50 \) | \( \hat{p} \)

Divine providence

1.4.25* Refer to Exercises 1.4.8 to 1.4.12. Dr. Arbuthnot’s actual analysis was different from the analysis you saw earlier. Instead of using each individual birth as a coin toss, Arbuthnot used a sign test with each of the 82 years as a coin toss, and a year with more male births counted as a “success.”

a. Complete the following table of comparisons:

<table>
<thead>
<tr>
<th>Analysis method</th>
<th>Sample size n</th>
<th>Null value ( \pi_0 )</th>
<th>Value of ( \hat{p} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: 1.4.8 – 1.4.12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B: 1.4.25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. For each method of analysis, rate the strength of evidence against the null hypothesis, as one of: inconclusive, weak but suggestive, moderately strong, strong, or overwhelming.

Healthy lungs

1.4.26 Researchers wanted to test the hypothesis that living in the country is better for your lungs than living in a city. To eliminate the possible variation due to genetic differences, they located seven pairs of identical twins with one member of each twin living in the country, the other in a city. For each person, they measured the percentage of inhaled tracer particles remaining in the lungs after one hour: the higher the percentage, the less healthy the lungs. They found that for six of the seven twin pairs the one living in the country had healthier lungs.

a. Is the alternative hypothesis one-sided or two-sided?

b. Based on the sample size and distance between the null value and the observed proportion, estimate the strength of evidence: inconclusive, weak but suggestive, moderately strong, strong, or overwhelming.

c. Here are probabilities for the number of heads in seven tosses of a fair coin:

<table>
<thead>
<tr>
<th># Heads</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.0078</td>
<td>0.0547</td>
<td>0.1641</td>
<td>0.2734</td>
<td>0.2734</td>
<td>0.1641</td>
<td>0.0547</td>
<td>0.0078</td>
</tr>
</tbody>
</table>

Compute the p-value and state your conclusion.

Bee stings

1.4.27* Scientists gathered data to test the research hypothesis that bees are more likely to sting a target that has already been stung by other bees. On eight separate occasions, they offered a pair of targets to a hive of angry bees; one target in each pair had been previously stung, the other pristine. On six of the eight occasions, the target that had been previously stung accumulated more new stingers.

a. Is the alternative hypothesis one-sided or two-sided?

b. Based on the sample size and distance between the null value and the observed proportion, estimate the strength of evidence: inconclusive, weak but suggestive, moderately strong, strong, or overwhelming.

c. Here are probabilities for the number of heads in eight tosses of a fair coin:

<table>
<thead>
<tr>
<th># Heads</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>0.0039</td>
<td>0.0313</td>
<td>0.1094</td>
<td>0.2188</td>
<td>0.2734</td>
<td>0.2188</td>
<td>0.1094</td>
<td>0.0313</td>
<td>0.0039</td>
</tr>
</tbody>
</table>

Compute the p-value and state your conclusion.


**Presidential stature**

In a race for U.S. president, is the taller candidate more likely to win?

**1.4.28** In the first election of the 20th century, Theodore Roosevelt (178 cm) defeated Alton B. Parker (175 cm). There have been 27 additional elections since then, for a total of 28. Of these, 25 elections had only two major party candidates with one taller than the other. In 19 of the 25 elections, the taller candidate won.

**a.** Let \( \pi = P(\text{taller wins}) \). State the research hypothesis in words and in symbols.

**b.** State the null and alternative hypotheses in words and symbols.

**c.** Compute the appropriate p-value using an applet.

**d.** If you take the p-value at face value, what do you conclude?

**e.** Are there reasons not to take the p-value at face value? Is yes, list them.

**1.4.29** In this exercise, as in the one before, we eliminate elections with more than two major party candidates as well as elections with two candidates of the same height. In addition, we eliminate the two elections in which George Washington was unopposed and five elections with missing data. Consider four different data sets:

**A.** Elections from 1960 (Kennedy) to the present: \( n = 14 \), \( \hat{p} = 8/14 = 0.5714 \).

**B.** Elections from Theodore Roosevelt (1904) to the present: \( n = 25 \), \( \hat{p} = 19/25 = 0.76 \).

**C.** Elections from John Adams (1796) through William McKinley (1900): \( n = 16 \), \( \hat{p} = 5/16 = 0.3125 \).

**D.** Elections from John Adams (1796) to the present: \( n = 41 \), \( \hat{p} = 24/41 = 0.5854 \).

**a.** The four p-values, from smallest to largest, are 0.007, 0.174, 0.395, 0.961. Match each data set (A–D) with its p-value.

**b.** What do you conclude about the hypothesis that taller candidates are more likely to win?

**Three features that affect p-values: An abstract summary**

**1.4.30** The graph below shows six different curves labeled A–F. Each curve shows the relationship between the p-value \( (y\text{-axis}) \) and the distance \( \hat{p} - \pi \) \( (x\text{-axis}) \) for testing the null hypothesis \( \pi = 0.50 \).

Match each curve A–F with one of the descriptions:

- Sample size is \( n = 25 \), alternative hypothesis is right-sided: \( \pi > \frac{1}{2} \).
- Sample size is \( n = 225 \), alternative hypothesis is right-sided: \( \pi > \frac{1}{2} \).
- Sample size is \( n = 25 \), alternative hypothesis is left-sided: \( \pi < \frac{1}{2} \).
- Sample size is \( n = 225 \), alternative hypothesis is left-sided: \( \pi < \frac{1}{2} \).
- Sample size is \( n = 25 \), alternative hypothesis is two-sided: \( \pi \neq \frac{1}{2} \).
- Sample size is \( n = 225 \), alternative hypothesis is two-sided: \( \pi \neq \frac{1}{2} \).

Comment: This is not an easy exercise, but don’t let the surface mess of the picture put you off. The purpose of the exercise is to challenge you to think hard about very important issues in a new way and at the same time to challenge you to become more skillful at connecting graphs with ideas. We hope you won’t give up and that you’ll feel that what you learn from the time you spend is worth it. Give it a shot, see how far you can get, and if you get stuck, skip to the five exercises that come after the graphs, do those, and then come back to this one.
1.4.31 Notice that there are three basic shapes for the curves: increasing, decreasing, and “up-down.” Tell the letters for each shape:
a. Increasing:
b. Decreasing:
c. Up-down:

1.4.32 Notice that the distance \( \hat{p} - \pi \) (x-axis) can tell you percent correct. Recall Example 1.1. For Buzz and Doris:
a. A distance of 0.50 means Buzz is right _____% of the time.
b. A distance of 0.00 means Buzz is right _____% of the time.
c. A distance of −0.50 means Buzz is right _____% of the time.

1.4.33 Try to relate the shape of the curve to the alternate hypothesis.
a. Suppose Buzz really wants to earn his fish and tries to guess right. Your alternative hypothesis is that \( \pi > 0.50 \). The better Buzz does, the _____ (smaller, larger) the p-value. The graph should be ________ (increasing, decreasing, up-down). The graphs that match this alternative are _______.
b. Suppose you are studying sadism in dolphins and your research hypothesis is that Buzz is able to guess right but he wants only to deprive Doris of her fish. Your alternative hypothesis is that \( \pi < 0.50 \). The better Buzz does, the _____ (smaller, larger) the p-value. The graph should be ________ (increasing, decreasing, up-down). The graphs that match this alternative are _______.
c. You are unsure about whether Buzz is fish-greedy or sadistic. Your more objective alternative hypothesis is that \( \pi \neq 0.50 \). The graphs that match this alternative are ________ (increasing, decreasing, up-down). Their letters are _______.
d. Complete the following summary table. For “Strongest evidence,” choose from among \( \hat{p} = 0 \), \( \hat{p} = 1 \), and \( \hat{p} = 0 \) or 1. For “Shape of curve,” choose from increasing, decreasing, and up-down.

<table>
<thead>
<tr>
<th>Alternative hypothesis</th>
<th>Strongest evidence</th>
<th>Shape of curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi &gt; 0.50 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \pi &lt; 0.50 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \pi \neq 0.50 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1.4.34 The six graphs are for two different sample sizes: one small (25) and one large (225). Your challenge in this exercise is to figure out how to tell the difference between a small-sample curve and a large-sample curve. Exercise 1.4.33 illustrates that the form of the alternative hypothesis determines the shape of the curve.
a. Consider two sample sizes with the same distance: If Buzz is right 30 times out of 32, the p-value will be ________ than if Buzz is right 15 times out of 16. (Choose from: higher, lower, can’t tell—it depends on the alternative).
b. More generally, for any given value of the distance \( \hat{p} - \pi \) (x-axis), the p-value is smaller when the sample size \( n \) is ________ (larger, smaller, can’t tell—it depends on the alternative).
c. This means that compared to the graph for \( n = 225 \), the graph for \( n = 25 \) _________ (always lies above, always lies below, neither: it depends on the alternative hypothesis.).
d. “Above/below” is not a useful comparison. Instead, use the steepness of the curve. Compared to the graph for \( n = 225 \), the graph for \( n = 25 \) is _______ (steeper, less steep).

1.4.35 Look at the figure long enough to follow each curve with your eye and recognize its shape and steepness. Keep these two facts in mind—three shapes, two curves of each shape—and try to relate them to the three main influences on p-values: distance, sample size, and one- versus two-sided alternatives. Then fill in the following table indicating which letter (A–F) corresponds to combination of shape descriptions.

<table>
<thead>
<tr>
<th>Increasing</th>
<th>Decreasing</th>
<th>Up-Down</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steeper</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flatter</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Now go back to Exercise 1.4.30 and use your table to do the matching.

FAQ

1.4.36 Read FAQ 1.4.1 about two-sided alternative hypotheses and answer the following questions.
a. Why don’t we first take a sample and use the observed sample proportion to help us decide which direction the alternative hypothesis should go? For example if we get a sample proportion greater than 0.50 then we will test to see if the parameter is greater than 0.50.
b. What is a downside of doing a two-sided test?

SECTION 1.5

1.5.1* Which long-run proportion of success, \( \pi \), gives the largest standard deviation of the null distribution when the sample size is 10?
A. 0.05     B. 0.25
C. 0.50     D. 0.90

1.5.2 Would your answer change in Exercise 1.5.1 if the sample size were 100?

1.5.3* Which sample size, \( n \), gives the smallest standard deviation of the null distribution where the long-run proportion, \( \pi \), is 0.25?
A. 30     B. 40
C. 50     D. 60
1.5.4 What is calculated using the formula $\sqrt{\pi(1-\pi)/n}$?

1.5.5* What is calculated using the formula (statistic − mean of null distr.)/(SD of null distr.)?

1.5.6 Suppose you are using theory-based techniques (e.g., a one-proportion z-test) to determine p-values. How will a two-sided p-value compare to a one-sided p-value (assuming the one-sided p-value is less than 0.50)?
A. The two-sided p-value will be about the same as the one-sided.
B. The two-sided p-value will be close to twice as large as the one-sided.
C. The two-sided p-value will be exactly twice as large as the one-sided.
D. The two-sided p-value will be half as much as the one-sided.

1.5.7* What is a z-statistic?

1.5.8 Suppose you ride to school with a friend and often arrive at a certain stop light when it is red. One day she states, “It seems like this light is green only 10% of the time when we get here.” You think it is more often than 10% and want to test this. You keep track of the color (green/not green) the next 20 times you go to school and find that 4 times (4/20 = 20%) the light is green when you arrive. You wish to see if your sample provides strong evidence that the true proportion of times the light is green is greater than 10%. In other words, you are testing the hypotheses

\[ H_0: \pi = 0.10 \text{ versus } H_a: \pi > 0.10 \]

where $\pi$ = the long-run proportion of times the light is green.

Two different approaches were taken in order to yield a p-value and both are shown in the applet output.

- **Option 1.** A simulation-based test was done and found a p-value of 0.148, showing weak evidence against the null.
- **Option 2.** A one-proportion z-test was conducted and found a p-value of 0.068, yielding moderate evidence against the null.

a. Which test gives a more valid p-value?
b. Give one or two sentences why you picked the option you did.

1.5.9* According to statistician Persi Diaconis, the probability of a penny landing heads when it is spun on its edge is only about 0.20. Suppose you doubt this claim and think that it should be more than 0.20. To test this, you spin a penny 12 times and it lands heads side up 5 times. You put this information in the One Proportion applet and determine a simulation-based p-value of 0.0770, but the one-proportion z-test p-value is 0.0303.

a. Write down the hypotheses for this study.
b. Which p-value would you say is the most valid? Explain.
1.5.9  Do you have strong evidence that a spun penny will land heads more that 20% of the time in the long run?

1.5.10  According to researchers, a coin flip may not have a 50% chance of landing heads and a 50% chance of landing tails. In fact, they believe that a coin is more likely to land the same way it started. So if it starts out heads up, it is more likely to land heads up. Suppose someone tests this hypothesis with 1,000 flips of a coin where it starts out heads up each time.

a. Describe what the symbol π stands for in this context.

b. State your null and alternative hypotheses.

c. Suppose 52% of the sample of 1,000 flips landed heads facing up. Verify the validity conditions that allow us to use a theory-based test.


e. A theory-based test reports a p-value of 0.1030. State your conclusion in terms of strength of evidence and what that means in the context of the study.

Heart transplant operations*

For the next two questions recall Example 1.3, which describes how in September 2000 the last 10 heart transplant operations at St. George’s Hospital had resulted in 8 deaths. Based on this, concerns arose and heart transplantations were suspended at this hospital. To investigate whether the data provide evidence that the long-run proportion of dying within 30 days of a heart transplant at St. George’s is higher than 15%, one can use the One Proportion applet to produce the null distribution of sample proportion of deaths as given below:

EXERCISE 1.5.9

a. According to the dotplot produced by the applet, what is the mean of the null distribution? Explain how you could have anticipated this.

b. According to the dotplot produced by the applet, is the overall shape of the null distribution bell-shaped? That is, can the null distribution be described as being a normal distribution? Explain how you could have anticipated this.
1.5.12 Recall that, to explore more deeply, the researchers went on to look at data from the most recent 361 heart transplantations at St. George's and found that 71 had resulted in deaths within 30 days of the heart transplant procedure. Once again, one can use the One Proportion applet to produce the null distribution of sample proportion of deaths assuming the long-run proportion of death is 0.15:

![Dotplot of null distribution]

Mean = 0.15
SD = 0.019

a. According to the dotplot produced by the applet, what is the mean of the null distribution? Explain how you could have anticipated this.

b. According to the dotplot produced by the applet, what is the SD of the null distribution? Explain how you could have anticipated this.

c. According to the dotplot produced by the applet, is the overall shape of the null distribution bell-shaped? That is, can the null distribution be described as being a normal distribution?

d. Explain why the normal approximation works better in the context of the null distribution in Exercise 1.5.12 compared to the null distribution in Exercise 1.5.11.

Psychic abilities

Use the following information to answer the next two questions. Statistician Jessica Utts has conducted extensive analysis of studies that have investigated psychic functioning. One type of study involves having one person (called the “sender”) concentrate on an image while a person in another room (the “receiver”) tries to determine which image is being “sent.” The receiver is given four images to choose from, one of which is the actual image that the sender is concentrating on. This is a technique called Ganzfeld.

1.5.13
a. Describe what the symbol π stands for in this context.

b. State your null and alternative hypotheses.

c. If the subjects in these studies have no psychic ability, approximately what proportion will identify the correct image? Is this the null hypothesis or the alternative hypothesis?

1.5.14 In the same research article, Utts (1995) also cites research from Morris et al. (1995) where, out of 97 Ganzfeld sessions, the receiver could correctly identify the image sent 32 times. You would like to know whether these data provide evidence that in Ganzfeld studies receivers will do better than just guess when picking which image is being sent.

a. Find the sample proportion of “hits” in Morris et al.’s study. How does this compare to the sample proportion of “hits” in Bem and Honorton’s study: larger, smaller, or about the same?

b. Conjecture whether the p-value for Morris et al.’s study results will be larger, smaller, or about the same compared to that of Bem and Honorton’s. Explain your reasoning.

c. Use simulation to find and report a p-value. Based on this p-value, summarize your conclusion in the context of this study and explain your reasoning for having arrived at this conclusion.

d. Now carry out the same analysis, only instead of simulating the null, use the theory-based (normal approximation) approach for that null distribution. How does the p-value you found using the theory-based approach compare to the one you found using simulation? Does this surprise you? Why or why not?

e. Why do the p-values from the simulation versus theory-based approaches in the Bern and Honorton study not differ as much from each other as do the p-values from the simulation versus theory-based approaches as here?

I love you, first*

1.5.15 Recall Exercise 1.2.16, in which researchers asked 40 heterosexual couples which person said “I love you first.”
They found that the man said “I love you” in 28 of the 40 couples before the woman did. Since previous research indicated that the man tends to say “I love you” before the woman, we tested to see if the probability a man would say “I love you” before the woman is more than 50%.

a. State the appropriate null and alternative hypotheses (in symbols) for testing whether males are more likely to say “I love you” first.

b. Using an appropriate applet, find the p-value using a theory-based test (one-proportion z-test; normal approximation).

c. Summarize your conclusion from your p-value.

d. What is the z-statistic for the test? What does this number represent in the context of this study?

e. Without using an applet, determine the value of the theory-based (normal approximation) p-value if we used a two-sided alternative hypothesis.

Rhesus monkeys

1.5.16 Recall Exercise 1.2.19, where researchers investigated whether rhesus monkeys have some ability to understand gestures made by humans. In one part of the study, they found that 31 of 40 monkeys correctly went to one of two boxes that researchers pointed towards.

a. State the appropriate null and alternative hypotheses in the context of this study, first in words and then in symbols.

b. Using an appropriate applet, find the p-value using a theory-based test (one-proportion z-test; normal approximation).

c. Summarize your conclusion from your p-value.

d. What is the z-statistic for the test? What does this number represent in the context of this study?

e. Without using an applet, determine the value of the theory-based (normal approximation) p-value if we used a two-sided alternative hypothesis.

Rock-paper-scissors*

Use the following information to answer the next two questions. In Example 1.2, we investigated some results of the game rock-paper-scissors. In the article referenced in the example, the researchers had 119 people play rock-paper-scissors against a computer. They found 66 players (55.5%) started with rock, 39 (32.8%) started with paper, and 14 (11.8%) started with scissors. We want to see if players start with scissors with a long-term probability that is different than 1/3.

1.5.17

a. State the appropriate null and alternative hypotheses in the context of this study, first in words and then in symbols.

b. Using an appropriate applet, find the p-value using a theory-based test (one-proportion z-test; normal approximation).

c. Summarize your conclusion from your p-value.

1.5.18 Based on the statistics from the previous question where people were playing rock-paper-scissors, we want to see if players start with rock with a probability that is different than 1/3.

a. State the appropriate null and alternative hypotheses in the context of this study, first in words and then in symbols.

b. Using an appropriate applet, find the p-value using a theory-based test (one-proportion z-test; normal approximation).

c. Summarize your conclusion from your p-value.

Facial prototyping

1.5.19 Reconsider Exploration 1.3, where we explored whether the long-run proportion that students in a class would correctly assign the names Bob and Tim to the pictures. We tested to see if this long-run proportion is more than 0.50.

a. State the appropriate null and alternative hypotheses in the context of this study.

b. In a class of one of the authors 19 out of 30 students correctly identified that the face on the left belonged to Tim. Using an appropriate applet, find the p-value using a one-proportion z-test (theory-based/normal approximation).

c. Summarize your conclusion from your p-value.

d. What is the z-statistic for the test? What does this number represent in the context of this study?

Predicting elections from faces*

1.5.20 In Example 1.4, we explored whether voters make decisions based on a candidate’s facial appearance. The researchers described how they had people determine which of two candidates for the U.S. House of Representatives in 2004 had the more competent face. They then looked to see if the one with the more competent face won the election. We will let \( \pi \) = the probability this method predicts the winner.

a. State the appropriate null and alternative hypotheses in the context of this study.

b. This method predicted the winner in 189 of the 279 House races that year. Based on these results, what are the values of \( \hat{p} \) and \( n \)?

c. Using an appropriate applet, find the p-value using a theory-based test (one-proportion z-test; normal approximation).

d. Summarize your conclusion from your p-value.
Uniform color

Use the following information to answer the next two questions. Reconsider Exploration 1.4, where we looked to see if an athlete’s uniform color affects whether or not they win or lose. The athletes were randomly assigned red or blue uniforms in matches of four combat sports. We want to see if one color has an advantage over the other. If we focus on the red uniform, we can ask if the probability the competitor in the red uniform will win is different than 0.50.

1.5.21 The researchers found that of the 457 matches, the competitor in red won 248 times.
   a. State the appropriate null and alternative hypotheses in the context of this study.
   b. Using an appropriate applet, find the p-value using a theory-based test (one-proportion z-test; normal approximation).
   c. Summarize your conclusion from your p-value.
   d. What is the z-statistic for the test? What does this number represent in the context of this study?

1.5.22 The researchers found that of the 272 boxing matches, the competitor in red won 150 times.
   a. State the appropriate null and alternative hypotheses in the context of this study.
   b. Using an appropriate applet, find the p-value using a theory-based test (one-proportion z-test; normal approximation).
   c. Summarize your conclusion from your p-value.
   d. What is the z-statistic for the test? What does this number represent in the context of this study?

Yahtzee*

Use the following information to answer the next two questions. One of the authors used to have an electronic Yahtzee game that he played frequently. The game would “roll” five virtual six-sided dice. But were the dice fair?

1.5.23 It seemed to the author that sixes showed up more than what they should if the dice were fair and he wanted to test this. He had the machine roll 500 dice and obtained 92 sixes. But were the dice fair?
   a. Describe what the parameter is in the context of this problem.
   b. State the appropriate null and alternative hypotheses in the context of this study.
   c. Using an appropriate applet, find the p-value using a theory-based test (one-proportion z-test; normal approximation).
   d. Summarize your conclusion from your p-value.

1.5.24 It also seemed to the author that ones showed up less often than what they should if the dice were fair and he wanted to test this. In his 500 dice rolls, 72 resulted in ones.
   a. Describe what the parameter is in the context of this problem.
   b. State the appropriate null and alternative hypotheses in the context of this study.
   c. Using an appropriate applet, find the p-value using a theory-based test (one-proportion z-test; normal approximation).
   d. Summarize your conclusion from your p-value.

Mario and Luigi

1.5.25 Suppose two brothers named Mario and Luigi like to compete by playing a certain video game. Mario thinks he is better at this game than Luigi and sets out to prove it by keeping track of who wins. After playing the game 30 times, Mario won 18 of them (or 60%). Mario then declares that this proves he is obviously the better player. Luigi, who just finished Chapter 1 in his statistics class, realizes that perhaps he and his brother are evenly matched and just by chance Mario won 60% of the last 30 games. Luigi is going to test this by running a test of significance.
   a. Describe what the parameter is that Luigi should be testing.
   b. State the appropriate null and alternative hypotheses in the context of this study. 
   c. Using an appropriate applet, find the p-value using a theory-based test (one-proportion z-test; normal approximation).
   d. Summarize your conclusion from your p-value.
   e. The p-value you found in part (c) should have been greater than 0.05. Suppose the two brothers continue to compete and Mario continues to win 60% of the games. How many games will they have to play until he gets a p-value less than 0.05 after retesting?

END OF CHAPTER

1.CE.1* A p-value is calculated assuming that which hypothesis is true: null or alternative?

1.CE.2 The standard deviation of a null distribution is calculated using which value: the probability value in the null hypothesis or the observed value of the sample proportion?

Fantasy golf*

1.CE.3 A statistics professor is in a fantasy golf league with four friends. Each week one of the five people in the league is the winner of that week’s tournament. During the 2010 season, this particular professor was the winner in 7 of the first 12 weeks of the season. Does this constitute strong evidence that his probability of winning in one week was larger than would be expected if the five competitors were equally likely to win?
CHAPTER 1  Significance: How Strong Is the Evidence?

a. State the null hypothesis and the alternative hypothesis in the context of the study.

b. Describe how you could use playing cards to conduct a tactile simulation analysis of this study and its result. Give sufficient detail that someone else could implement this simulation analysis based on your description. Be sure to indicate how you would decide whether the observed data provide convincing evidence that the statistics professor’s probability of winning in one week was larger than would be expected if the five competitors were equally likely to win?

c. Use an applet to conduct a simulation analysis with at least 1,000 repetitions. Based on the null distribution produced by the applet, explain how you are deciding whether the observed data provide convincing evidence that the statistics professor’s probability of winning in one week was larger than would be expected if the five competitors were equally likely to win.

d. In answering (c), you should have reported a p-value. Interpret this p-value in the context of the study. (Hint: The p-value is the probability of what, assuming what?)

e. Explain why it is not advisable to use a one-proportion z-test (theory-based approach) to find a p-value for the data from this study.

Mythbusters

1.CE.4 On the television show Mythbusters, the hosts Jamie and Adam wanted to investigate which side buttered toast prefers to land on when it falls through the air. To replicate a piece of toast falling through the air, they set up a specially designed rig on the roof of the Mythbusters’ headquarters. In 48 attempts, 19 pieces of toast fell buttered side down. Do these data provide evidence that buttered toast falling through air has a preference for any specific side (buttered or not buttered) when landing? Write a report describing all six steps of the statistical investigation method as they apply to this question and these data.

Cheating*

Use this information to answer the following two questions. The August 19, 2013, issue of Sports Illustrated included a brief article titled “Why Men Cheat,” in which the author L. Jon Wertheim argued that Major League Baseball (MLB) players from the United States are less likely than those from poorer countries to take performance-enhancing drugs (PEDs). The article mentioned that 57.3% of all professional baseball players between 2005 and 2013 were from the U.S., and 206 of the 595 (34.6%) players suspended for PED use in these years were from the U.S.

1.CE.5

a. One question is whether 0.346, the observed proportion of PED-suspended players from the U.S., is (statistically) significantly less than 0.573, the proportion of all players from the U.S. Before you attempt to answer this question, explain what the question means (i.e., what statistical significance means) in this context. Write as if to a friend who is a baseball fan but has not studied statistics.

b. Perform a simulation analysis to investigate the question in part (a). Write a paragraph summarizing the results and conclusion that you draw from your simulation analysis, again as if to a friend who is a baseball fan but has not studied statistics.

c. Calculate the value of the standardized statistic (z-value) for comparing 34.6% to 57.3% in this context. Interpret what this value means and summarize the conclusion that you can draw from it.

1.CE.6 Additional information supplied in the Sports Illustrated article was that 34.6% of all professional baseball players are from Latin-American countries, but 368 of the 595 (61.8%) players suspended for PED use between 2005 and 2013 are from Latin-American countries. Answer parts (a) to (c) of the previous exercise with respect to the question of whether 61.8% is (statistically) significantly greater than 34.6% in this study.

Bob or Tim?

Use the following information to answer the next two questions. Gary Judkins is a teacher in New Zealand who asked his students the question about facial prototyping from Exploration 1.3. Mr. Judkins asked this question of 135 New Zealand students and found that 105 of the students associated the name Tim with the face on the left.

1.CE.7

a. Express the appropriate null and alternative hypotheses, in symbols and in words, for testing whether New Zealand students have a tendency to associate the name Tim with the face on the left.

b. Conduct a simulation analysis to test these hypotheses with data from Mr. Judkins’s students. Summarize your conclusion.

c. Is it appropriate to use a theory-based approach with these data? Justify your answer.

d. Implement a theory-based approach to test the hypotheses in part (a). Report the values of the test statistic and p-value. Summarize your conclusion.

1.CE.8 Reconsider the previous exercise. Suppose that you were to define the parameter π, not as the probability that a New Zealand student would associate the name Tim with the face on the left, but as the probability that a New Zealand student would associate the name Bob with the face on the left. Re-answer parts (a) to (d) of the previous exercise with this different set-up. Comment on how (if at all) your calculations and conclusions change.
Coin flipping*

1.CE.9 Suppose that Jose and Roberto both collect data to investigate whether people tend to call “heads” more often than “tails” when they are asked to call the result of a coin flip. If Jose uses a larger sample size than Roberto, is Jose guaranteed to have a smaller p-value than Roberto? Explain why or why not.

1.CE.10 Suppose that Sasha and Jayla both collect data to investigate whether people tend to call “heads” more often than “tails” when they are asked to call the result of a coin flip. If Sasha has a smaller p-value than Jayla, which of the following would you conclude? (There may be more than one correct conclusion.)

a. Sasha has stronger evidence than Jayla that people tend to call “heads” more often than “tails.”

b. Jayla has stronger evidence than Sasha that people tend to call “heads” more often than “tails.”

c. Both Sasha and Jayla have convincing evidence that people tend to call “heads” more often than “tails.”

d. Sasha must have used a larger sample size than Jayla.

e. Sasha must have used a smaller sample size than Jayla.

Family structure

1.CE.11 Josephine wants to investigate whether more than one-third of families with two children have one child of each sex, so she collects data for 50 two-child families. Describe how she could use a die to perform a physical (by hand, without using a computer) simulation to approximate the null distribution.

1.CE.12 Alfonso wants to investigate whether fewer than one-fourth of families with two children have two girls, so he collects data for 25 two-child families. Describe how he could use a die to perform a physical (by hand, without using a computer) simulation to approximate the null distribution.

Athletic performance*

1.CE.13 Rick is a basketball player who wants to investigate whether he successfully makes more than 90% of his free throws by shooting with an underhand style.

a. Define (in words) the relevant parameter of interest for Rick.

b. State (in words and in symbols) the appropriate hypotheses to be tested by Rick.

1.CE.14 Lorena is a golfer who wants to investigate whether she successfully makes more than 60% of her 10-foot putts.

a. Define (in words) the relevant parameter of interest for Lorena.

b. State (in words and in symbols) the appropriate hypotheses to be tested by Lorena.

Odd numbers

Use the following information to answer the following two questions. Many studies have investigated the question of whether people tend to think of an odd number when they are asked to think of a single-digit number (0 through 9). Combining results from several studies, Kubovy and Psotka (1976) used a sample size of 1,770 people, of whom 741 thought of an even number and 1,029 thought of an odd number. Consider investigating whether these data provide strong evidence that people have a tendency to think of an odd number rather than an even number in this situation.

1.CE.15

a. Describe the relevant parameter of interest.

b. State the appropriate hypotheses to be tested.

c. Calculate the observed value of the appropriate statistic.

d. Check whether a theory-based approach is appropriate for these data and hypotheses.

e. Calculate the standardized value of the statistic.

f. Determine an approximate p-value.

g. Summarize your conclusion.

1.CE.16 Refer to the previous exercise. The researchers also found that 503 of the 1,770 people thought of the number 7. Do these data provide strong evidence that people have a tendency to select the number 7 disproportionately often? Investigate this question by answering parts (a) to (g) of the previous exercise.

Racket spinning*

1.CE.17 One of the authors once spun his tennis racquet 100 times and found that it landed with the label up for 46 of those 100 spins (0.46). For testing whether the probability (that a spun tennis racquet lands with the label up) differs from 0.50, he calculated the p-value to be 0.484.

a. Summarize the conclusion that you would draw from this study.

b. Explain the reasoning process that leads to your conclusion, as if to someone with no knowledge of statistics.

1.CE.18 Reconsider the previous exercise. Notice that three different decimal numbers appear: 0.46, 0.484, and 0.50. Identify which is the null-hypothesized probability, which is the observed value of the statistic, and which is the p-value.

Your own random process

1.CE.19 Think of a random process of interest to you for which you would be interested in assessing evidence regarding a claim about the value of the long-run proportion associated with the random process.

a. Describe the random process.

b. Describe the parameter of interest.

c. State the null and alternative hypotheses that you would be interested in investigating.

d. Describe how you could collect data to conduct this statistical investigation.
INVESTIGATION: TIRE STORY FALLS FLAT

PART I

A legendary story on college campuses concerns two students who miss a chemistry exam because of excessive partying but blame their absence on a flat tire. The professor allows them to take a make-up exam, and sends them to separate rooms to take it. The first question, worth 5 points, is quite easy. The second question, worth 95 points, asks: Which tire was it?

STEP 1: Ask a research question. Do students pick which tire went flat in equal proportions? It has been conjectured that when students are asked this question and forced to give an answer (left front, left rear, right front, or right rear) off the top of their head, they tend to answer “right front” more than would be expected by random chance.

STEP 2: Design a study and collect data. To test this conjecture about the right front tire, a recent class of 28 students was asked, if they were in this situation, which tire would they say had gone flat. We obtained the following results:

<table>
<thead>
<tr>
<th>Tire</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left front</td>
<td>6</td>
</tr>
<tr>
<td>Left rear</td>
<td>4</td>
</tr>
<tr>
<td>Right front</td>
<td>14</td>
</tr>
<tr>
<td>Right rear</td>
<td>4</td>
</tr>
</tbody>
</table>

1. What are the observational units?
2. What is the variable that is measured/recorded on each observational unit?
3. Describe the parameter of interest in words. (You can use the symbol \( \pi \) to represent this parameter.)
4. State the appropriate null and alternative hypotheses to be tested.

STEP 3: Explore the data

5. What percentage of the students picked the right front tire? Is this more than you would expect if students randomly pick one of the four tires?
6. Is it possible that we could observe 14 students from this class of 28 students even if all of the students were just selecting randomly among the four tires?

STEP 4: Draw inferences. Let’s use our 3S strategy to help us investigate how much evidence the sample data provide to support our conjecture that there is something special about the right front tire.

Statistic

7. What is the statistic that you can use to summarize the data collected in the study and what symbol is associated with this statistic?

Simulate

8. Use the One Proportion applet to simulate 1,000 repetitions of this study, assuming that every student in class selects randomly (equally) among the four tires. Report what values you input into the applet.
   Probability of success (\( \pi \)) ___________
   Sample size (\( n \)) _______________
   Number of samples ___________

9. Using the proportion of successes for the values on the horizontal axis, what is the center of your null distribution? Does it make sense that this is the center? Explain.

Strength of evidence

10. Let’s examine the strength of evidence with the three ways we used in Chapter 1.
a. Determine the p-value from your simulation analysis. Also interpret what this p-value represents (i.e., the probability of what, assuming what?).

b. Determine the standardized statistic from your simulation analysis. Also interpret what this standardized statistic represents.

c. Use a theory-based test (or normal approximation) to determine the p-value for this test. Are the validity conditions met for this test? Why or why not?

11. Is there strong evidence against the null hypothesis for all three of the methods used in Question 11? Explain.

**STEP 5: Formulate conclusions.**

12. Summarize the conclusion that you draw from this study and your simulation analysis. Also explain the reasoning process behind your conclusion.

13. Now, let's step back a bit and think about the scope of our inference. What are the wider implications? Do you think that your conclusion holds true for people in general? (These are extremely important questions that we'll discuss more when we talk about the scope of inference in Chapter 2.)

**STEP 5: Look back and ahead.**

14. If you were to repeat this study, what improvements might you make? What further research might you propose related to this topic in the future?

**PART II**

Now suppose another class conducts the same study with exactly half as many students and the proportional breakdown in the four categories is identical to the class of 28. In other words, 7 out of 14 students answered “right front.”

15. Before you analyze the data, would you expect to find stronger evidence for the research conjecture (that people pick the right front tire more than one-fourth of the time), weaker evidence, or the same strength of evidence? Explain your thinking.

16. Conduct a simulation analysis to produce a simulated p-value. How does it compare to the p-value from the study of the class with 28 students? Is this what you expected? Explain.

**PART III**

Suppose we didn't have a preconceived notion that the right front tire would be chosen more often, but just wanted to find out if it was chosen at a rate that was different than 1/4. Let's use our original data where 14 out of 28 chose the right front tire to test this.

17. Write the null and alternative hypotheses for this new question.

18. Use a theory-based test to find the p-value. How does this compare with the p-value you obtained back in Question 11c.

19. Do you have strong evidence that the long-run proportion of times a student will choose the right front tire is different than 1/4?

---

**Research Article**  [www.wiley.com/college/tintle](www.wiley.com/college/tintle)

Infants prefer to harm those who are different  Read “Not like me = Bad: Infants prefer those who harm dissimilar others” by Hamlin, Mahajan, Liberman, and Wynn (*Psychological Science*, 2013, 24:589)