CHAPTER 4

Introducing the Logic of Inference Using Confidence Intervals

"Inference" refers to a reasoning process that begins with some information and leads to some conclusion. If you have ever taken a logic course or thought about logical reasoning, you have probably heard things like, "All mammals are warm-blooded animals; this animal is a mammal; therefore this animal is warm-blooded." That particular sentence is known as a "syllogism" and it represents a kind of inference called "deduction." Another kind of inference, **induction**, reasons from specific cases to the more general. If I observe a cat jumping from a tree and landing on its feet, and then I observe another cat, and another, and another doing the same thing, I might infer that cats generally land on their feet when jumping out of trees. Statistical inference takes this same kind of logical thinking a step further by dealing systematically with situations where we have uncertain or incomplete information. Unlike the syllogism about warm-blooded animals presented above, conclusions that we draw inductively from samples of data are never fixed or firm. We may be able to characterize our uncertainty in various ways, but we can never be 100% sure of anything when we are using statistical inference. This leads to an important idea that you should always keep in mind when reasoning from samples of data:

You cannot *prove* anything from samples or by using statistical inference.

I emphasize this point repeatedly with my students and I expect them to know the reason why (by the end of the course if not before). So, if you ever hear a journalist or a scientist or anyone else saying that statistical analysis of one or more samples of data *proves* a certain conclusion, you can be assured that he or she is mistaken, and perhaps misinformed or being intentionally misleading.

Instead of setting out to prove something, we collect data and analyze it with inferential statistics in order to build up a weight of evidence that influences our certainty about one conclusion or another. There's a great historical example of this method from the 19th-century medical researcher John Snow. Snow (1855) studied outbreaks of cholera in London. Cholera is a devastating bacterial infection of the small intestine that still claims many lives around the world. John Snow mapped cases of cholera in London and found that many cases clustered geographically near certain wells in the city where residents drew water for drinking and other needs. Using this purely graphical method, Snow was able to infer a connection between the cholera cases, the wells, and sewage contamination that led him to conclude that fecal contamination of drinking water was the primary mechanism that caused cholera. Blech! His map, his evidence, and his reasoning were not a proof—in fact, many authorities disbelieved his proposed mechanism for decades afterward—but Snow's work added substantially to the weight of evidence that eventually led to a scientific consensus about cholera infections.

Today we have many more data analysis tools than John Snow did, as well as an unrivaled ability to collect large amounts of data, and as a result we are able to more carefully quantify our ideas of certainty and confidence surrounding our data sets. In fact, while there was one predominant strategy of statistical inference used during the 20th century, known as "frequentist statistical inference," some new approaches to inference have come into play over recent years. We will begin by considering one element of the frequentist perspective in this chapter and then add to our knowledge by considering the so-called Bayesian perspective in the next chapter.

But now would be a really good point to pause and to look back over the previous chapter to make sure you have a clear understanding of the ideas around sampling distributions that I introduced there. In the previous chapter we began with a randomly generated data set that we used to represent the whole population—specifically a whole population of angles at which toast struck the ground after being dropped off a plate. It is important to now declare that this made-up example was ridiculous on many levels. First of all, I hope that there are not many scientists who spend their time worrying about toast-drop angles.

Most importantly, however, as researchers *we almost never have access to the totality of the population data*. This is a critical conceptual point. Although in some special cases it may be possible to access a whole population (think, e.g., of the population of Supreme Court justices), in most research situations it is impractical for us to reach all of the members of a population. Think about the population of humans on earth, or of asteroids in the asteroid belt, or of tablet computers deployed in the United States, or of electric utility customers in Canada. In each case we can define what we are talking about, and we can have criteria that let us know whether a particular person, or tablet, or asteroid is part of the population, *but we cannot get data on the complete population*. It is logistically impossible and/or cost-prohibitive to do so. Even in smaller populations—the

students enrolled in a college, all of the employees of a company, the complete product inventory of one store, a full list of the text messages you have ever sent—there may be good, practical reasons why we may not want to measure every single member of a population.

That's why we sample. Drawing a sample is a deliberate act of selecting elements or instances such that the collection we obtain can serve as a stand-in for the whole population. At its best, a sample is *representative* of a population, where membership in a population can be defined but the elements of the population can never be accessed or measured in their totality. As we know from our activities in the previous chapter, it is rare to ever draw a sample that perfectly mimics the population (i.e., one particular sample that has the exact same mean as the population). In this context, the goal of statistical inference is to use a sample of data to make estimates and/or inferences about a population, with some degree of certainty or confidence. (Note that in those unusual cases where we can measure every element of a population, we conduct a **census** of the population and we do not need sampling or inferential thinking at all. Whatever we measure about all Supreme Court justices is what it is, within the limits of measurement precision.)

Let's now turn to a real-world example of populations and samples to illustrate the possibilities. We will use another built-in data set that R provides called "mtcars." If you type "?mtcars" at the command line, you will learn that this is an old (1974!) data set that contains a sample of 32 different vehicles with measurements of various characteristics including fuel economy. I'll bet you are happy that I have finally stopped talking about toast! Historical point of trivia: cars from the 1974 model year were the last ones that were designed prior to the 1970s oil crisis (in most parts of the world gas prices tripled between 1974 and 1980), so the average fuel economy of these cars is shockingly low. In this chapter, we will focus on one simple research question: Do cars with automatic transmissions get better or worse mileage than cars with manual transmissions? As usual, I provide R code below to show what I am doing. In the sections below, however, the most important thing is that you should follow the conceptual arguments.

Here is the situation we are trying to address. The mtcars data set contains $n = 19$ cars with automatic transmissions and $n = 13$ cars with manual transmissions. The $n = 19$ and $n = 13$ are **independent samples** that are standing in for the whole *population* of cars (from model year 1974). The samples are independent because they were collected from two distinctive groups of cars (as opposed to one group of cars at two different points in time). What we want to do is use the sample to infer a plausible difference in mileage between the two types of transmissions among all 1974 model year cars. I use the word *plausible* intentionally, as I want to avoid the language of probability until a little later. The key thing to keep in mind is that because we are using samples, we can't possibly know *exactly* what would be true of the populations. But we can think about the evidence that we get from the sample data to see if it convinces us that a difference may exist between the respective populations (i.e., the population

of automatic cars and the population of manual cars). We will assume for the purposes of this example that the people involved in collecting the mtcars data did a good job at obtaining random samples of cars to include in the data set.

Of additional importance, we can use the statistics to give us some information about how far apart the fuel economy is between the two types of transmissions. We might have wanted to know this information to make a judgment as to which type of transmission to buy: automatics are more convenient and may be safer, but manuals might get better fuel economy and some people think they are more fun to drive. If the statistics told us that there was a big difference in fuel economy, we would have to weigh that against purchase and operating costs as well as convenience, safety, and fun. On the other hand, if the statistics told us that there was only a very small economy difference or possibly no difference between the two types of transmission, then we could make our decision based on other criteria.

Let's begin with some exploration of the mtcars data set, including a new visualization called a **box plot**. Run these commands:

```
mean( mtcars$mpg[ mtcars$am == 0 ] )     # Automatic transmissions
mean( mtcars$mpg[ mtcars$am == 1 ] )     # Manual transmissions
```

The mean miles per gallon (mpg) for the automatic transmission cars was 17.1, while the mean mpg for manual transmission cars was 24.3, a substantial difference of about 7.2 miles per gallon. Note that we have stepped up our game here with respect to R syntax. We are using the $ subsetting mechanism to access the mpg variable in the mtcars data set: that's what "mtcars$mpg" does. But we are also doing another kind of subsetting at the same time. The expressions inside the square brackets, [mtcars$am == 0] and [mtcars$am == 1] select subsets of the cases in the data set using logical expressions. The two equals signs together makes a logical test of equality, so for the first line of code we get every case in the mtcars data frame where it is true that mtcars$am is equal to 0 (0 is the code for automatic transmission; you can verify this by examining the help file you get when you type ?mtcars). For the second line of code we get every case where it is true that mtcars$am is equal to one (one is the code for manual transmission).

Now if you have not yet put on your brain-enhancing headgear, you might believe that the calculated difference in means is sufficient evidence to conclude that manual transmissions are better, as the mean for manual transmission cars is more than 7 mpg higher than for automatic transmission cars. Remember the previous chapter, however, and keep in mind how often we drew samples that were a fair bit different from the population mean. Each of these two sample means is uncertain: each mean is what statisticians refer to as a **point estimate**, with the emphasis on the word "estimate." Each sample mean is an estimate of the underlying population mean, but right now we are not quite sure how good an estimate it is. One of the key goals of inferential statistics is to put some boundaries around that level of uncertainty.

We can begin to understand that uncertainty by examining the variability within each of the groups. All else being equal, a sample with high variability makes us less certain about the true population mean than a sample with low variability. So as part of our routine process of understanding a data set, let's examine the standard deviations of each of the transmission groups:

```
sd( mtcars$mpg[ mtcars$am == 0 ] )      # Automatic transmissions
sd( mtcars$mpg[ mtcars$am == 1 ] )      # Manual transmissions
```

These commands reveal that the standard deviation for automatic transmissions cars is 3.8 miles per gallon, while the standard deviation for manual transmissions is quite a bit higher at 6.1 miles per gallon. Are these standard deviation values unexpected, very large, very small, or just what the doctor ordered? We really don't know yet, and in fact it is a little tricky to judge just on the basis of seeing the two means and the two standard deviations. We might get a better feel for the comparison between these two groups with a visualization that allows us to graphically compare distributions and variability. That's where the box plot comes in:

```
boxplot(mpg ~ am, data=mtcars)        # Boxplot of mpg, grouped by am
```

This command introduces another little piece of R syntax, called "formula notation." The expression "mpg ~ am" tells R to use mpg as the dependent variable (the variable that gets plotted on the *Y*-axis) and to group the results by the contents of "am." The second piece, "data=mtcars," simply tells R where to find the data that goes with the formula. Lots of analysis commands in R use the formula notation and we will expand our knowledge of it later in the book. For now, take a look at the box plot that appears in Figure 4.1.

The box plot, sometimes also called a "box-and-whiskers plot," packs a lot of information into a small space. Figure 4.1 shows boxes and whiskers for the two groups of cars—automatic and manual—side-by-side. In each case the upper and lower boundaries of the box represent the first and third quartiles, respectively. So 25% of all the cases are above the box and 25% are below the box. The dark band in the middle of the box represents the median. You can see clearly that in the case of manual transmissions, the median is quite close to the first quartile, indicating that 25% of cases are clustering in that small region between about 21 and 23 miles per gallon. In this box plot the whiskers represent the position of the maximum and minimum values, respectively. In some other box plots these whiskers may represent the lowest or highest "extreme" value, with a few additional outliers marked beyond the whiskers. Other box plots may also notate the mean, sometimes with a dot or a plus sign.

Figure 4.1 gives us a good visual feel for the differences between the two groups. The boxes for the two groups do not overlap at all, a very intuitive and informal indication that there may be a meaningful difference between these

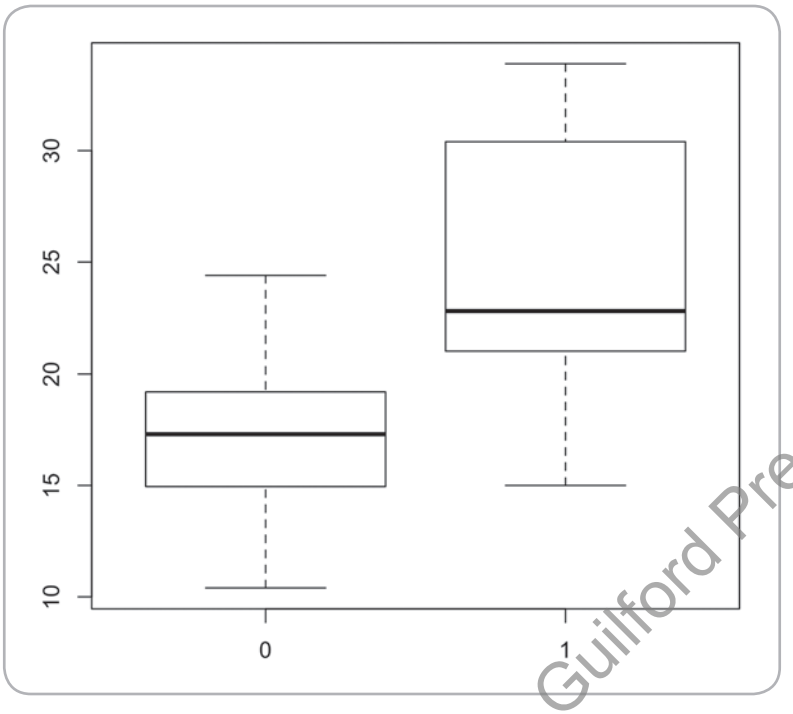

FIGURE 4.1. Box plot of mpg by transmission type from the mtcars data set.

two groups. In comparing the heights of the boxes, we also see a reflection of what we learned before from the standard deviations: the automatic transmission cars are considerably less variable than the manual transmission cars. Finally, the whiskers show that the very lowest value for manual transmissions falls at the first quartile for the automatic transmissions, further reaffirming the differences between these groups.

But we are still cautious, because we know that samples can fluctuate all over the place, and we can't be certain that the differences between these two groups can be trusted. So now let's do something clever: we can use the resampling and replication techniques we developed in the previous chapter to create a simulation.

EXPLORING THE VARIABILITY OF SAMPLE MEANS WITH REPETITIOUS SAMPLING

This simulation will show, in an informal way, the amount of uncertainty involved in these two samples. Let's try to visualize those boundaries using some of the tricks we learned in the previous chapter. First, we will have a little fun by sampling from our samples:

```
mean( sample(mtcars$mpg[ mtcars$am == 0 ],size=19,replace=TRUE) )
mean( sample(mtcars$mpg[ mtcars$am == 1 ],size=13,replace=TRUE) )
```

These functions should be familiar to you now: We are drawing a sample of $n = 19$ from the automatic transmission group, with replacement. Likewise, we are drawing a sample of $n = 13$ from the manual transmission group. I got 16.8 mpg for automatic and 28.4 mpg for manual, but your mileage will vary (ha ha!) because of the randomness in the sample() command. It may seem kind of goofy to "resample" from a sample, but bear with me: we are building up toward creating a histogram that will give us a graphical feel for the uncertainty that arises from sampling. Each of the sample mean differences will be close to, but probably not exactly equal to, the mean difference that we observed between the two original samples. Now, let's calculate the difference between those two means, which is, after all, what we are most interested in:

```
mean(sample(mtcars$mpg[mtcars$am == 0],size=19,replace=TRUE)) –
    mean(sample( mtcars$mpg[mtcars$am == 1],size=13,replace=TRUE) )
```

If you are typing that code, put the whole thing on one line and make sure to type the minus sign in between the two pieces. Keep in mind that we are usually looking at *negative* numbers here, because we are expecting that manual transmission mileage will on average be higher than automatic transmission mileage. Now let's take that same command and replicate it one hundred times:

```
meanDiffs <- replicate(100, mean( sample(mtcars$mpg[ mtcars$am == 0 ],
    size=19,replace=TRUE) ) - mean( sample(mtcars$mpg[ mtcars$am ==
    1 ], size=13,replace=TRUE) ))
```

Now plot a histogram of that sampling distribution of mean differences:

```
hist(meanDiffs)
```

The first statement above uses replicate() to run the same chunk of code 100 times, each time getting one sample mean from the automatic group and one sample mean from the manual group and subtracting the latter from the former. We store the list of 100 sample mean differences in a new vector called meanDiff and then request a histogram of it. The result appears in Figure 4.2.

Let's make sense out of this histogram. The code says that whenever we draw a sample of automatic transmission cars, we calculate a mean for it. Then we draw a sample of manual transmission data and calculate a mean for it. Then we subtract the manual mean from the automatic mean to create a mean difference between two samples. Every time we do this we end up with a slightly different result because of the randomness of random sampling (as accomplished by the sample() command). We append each mean difference onto a vector

of mean differences. As the histogram shows, in a lot of cases the difference between the two means is right around –7, as you would expect from looking at the two original samples. But we can also see that on occasion manual transmission samples are better by as much as 13 miles per gallon (the left end of the *X*-axis), while in other cases manual transmissions are only better by 1 mile per gallon (the right end of the *X*-axis). Note that if you run this code yourself, your results may be slightly different because of the inherent randomness involved in sampling.

You can think of this informally as what might have happened if we had replicated our study of transmissions and fuel economy 100 times. Based solely on the simulation represented by this histogram, we might feel comfortable saying this: manual transmissions may provide better fuel economy than automatic transmissions with a difference of about 7 miles per gallon, but that could on rare occasions be as little as 1 mile per gallon or as much as 13 miles per gallon. The width of the span between –1 and –13 is one very concrete representation of our uncertainty. Further, one might say that we have a certain amount of *confidence* that we do have a real difference between the two kinds of transmissions. I love that word "confidence." We are using it

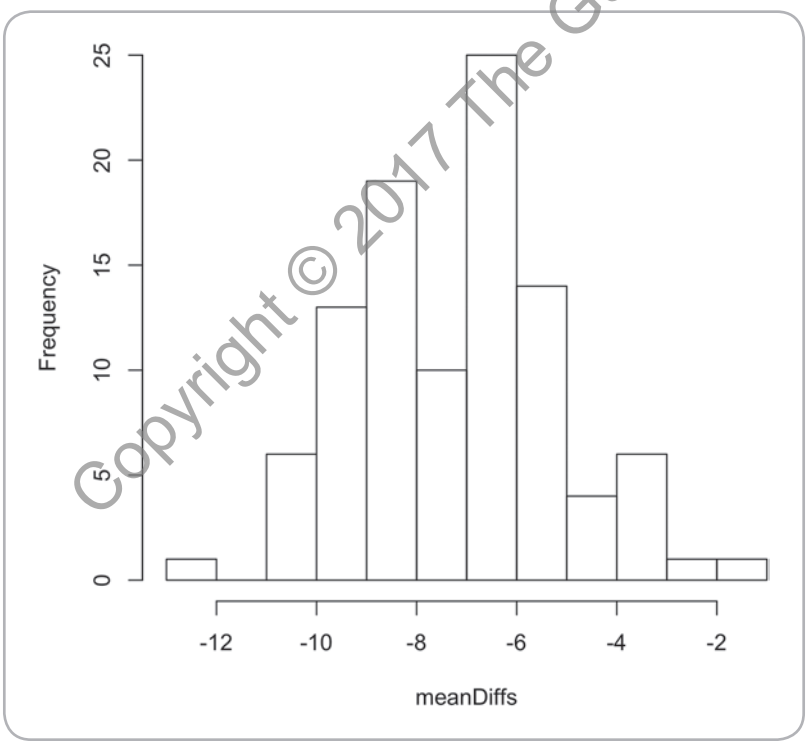

FIGURE 4.2. Histogram of mean differences between automatic transmission cars and manual transmission cars.

to signify that there is a span of different possibilities and although we know roughly how wide that span is, we don't know where exactly the truth lies within that span.

If we really wanted to follow the methods described in Chapter 3, we could go one step further by calculating quantiles for our distribution of mean differences. The following command provides the values for the 0.025 and 0.975 quantiles: this divides up the distribution of simulated sampling means into 95% in the center and 5% in the tails.

```
quantile(meanDiffs, c(0.025, 0.975))
```

For my simulated distribution of 100 mean differences, I got –10.8 on the low end and –3.1 on the high end. So we could now update our previous informal statement of uncertainty to become slightly more specific: manual transmissions may provide better fuel economy than automatic transmissions with a difference of about 7 mpg, but for 95% of the simulated mean differences that could be as much as 10.8 mpg or as little as 3.1 mpg. The width of this span, about plus or minus 4 mpg, is a representation of our uncertainty, showing what might happen in about 95 out of 100 trials if we repeatedly sampled fuel economy data for cars with the two types of transmissions.

OUR FIRST INFERENTIAL TEST: THE CONFIDENCE INTERVAL

Now we are ready to perform our very first official inferential test:

```
t.test(mtcars$mpg[mtcars$am==0] ,mtcars$mpg[mtcars$am==1])
```

That command produces the following output:

```
        Welch Two Sample t-test
data: mtcars$mpg[mtcars$am == 0] and mtcars$mpg[mtcars$am == 1]
t = -3.7671, df = 18.332, p-value = 0.001374
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-11.280194 -3.209684
sample estimates:
mean of x mean of y
17.14737 24.39231
```

The t.test() function above invokes the single most popular and basic form of inferential test in the world, called the "Student's *t*-Test." If you go all the way back to the introduction to this book, you will remember that "Student"

was William Sealy Gosset (1876–1937), the Guinness Brewery scientist. Gosset developed this "independent groups" *t*-test in order to generalize to a population of mean differences using sample data from two independent groups of observations. The output on page 60 is designated the "Welch Two Sample *t*-test" because the 20th-century statistician Bernard Lewis Welch (1911–1989) developed an adaptation of Gosset's original work that made the test more capable with different kinds of "unruly" data (more specifically, situations where

Formulas for the Confidence Interval

I promised to postpone a deeper discussion of the meaning of *t* until the next chapter, but now is the right moment to show you the formulas for the confidence interval for the difference between two independent means:

$$\text{Confidence interval: Lower bound} = \left(\bar{x}_1 - \bar{x}_2\right) - t^{\star}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{Upper bound} = \left(\bar{x}_1 - \bar{x}_2\right) + t^{\star}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

You'll notice that the top and bottom equations only have one difference between them: the top equation has a minus sign between the first and second part and the bottom equation has a plus sign. The first half of each equation, a subtraction between two "*x*-bars," is simply the observed difference in sample means. In our mtcars example, that difference was 17.14 – 24.39 = –7.2. The second part of the equation calculates the width of the confidence interval, in the top case subtracting it from the mean difference and in the bottom case adding it.

The width of the confidence interval starts with t^*—this is a so-called critical value from the *t*-distribution. I won't lay the details on you until the next chapter, but this critical value is calculated based on the sample sizes of the two samples. The important thing to note is that the critical value of *t* will differ based on both sample size and the selected confidence level. We have used a "95% confidence interval" throughout this chapter, but it is also possible to use 99% or on occasion other values as well.

All of the stuff under the square root symbol is a combination of the variability information from each of the samples: technically a quantity called the **standard error.** Sounds complicated, but it is really nothing more than the standard deviation of the sampling distribution of means (or in this case, mean differences). In each case we square the standard deviation to get the variance and then divide the variance by the sample size. Once we have added together the two pieces, we square root the result to get the standard error.

the two groups have different levels of variability; Welch, 1947). We are going to postpone a detailed consideration of what "*t*" actually means until the next chapter. So for now, the key piece of output to examine above is the **95 percent confidence interval.** The *t*-test procedure has used these two samples to calculate a confidence interval ranging from a mean difference of –11.3 miles per gallon to –3.2 miles per gallon. That range should seem familiar: it is darned close to what our little resampling simulation produced! In fact, there is some conceptual similarity to what we did in our informal simulation and the meaning of a confidence interval.

In our simulation we sampled from the existing sample, because we had no way of sampling new data from the population. But statisticians have figured out what would happen if we *could* have sampled new data from the population. Specifically, if we *reran our whole study of transmissions and fuel economy many times*—sampling from the population and taking means of both a new group of automatic transmission cars and a new group of manual transmission cars—and each time we constructed a new confidence interval, in 95% of those replications the confidence interval *would contain the true population mean difference*. In the previous sentence the phrase "would contain" signifies that the true population mean difference would fall somewhere in between the low boundary of the confidence interval and the high boundary. Based on this definition it is really, extremely, super important to note that *this particular confidence interval* (the one that came out of our *t*-test above) does not necessarily contain the true population value of the mean difference. Likewise, the 95% is *not* a statement about the probability that *this particular confidence interval* is correct. Instead, the 95% is a long-run prediction about what would happen if we replicated the study—sampling again and again from the populations—and in each case calculated new confidence intervals.

This is definitely a major brain stretcher, so here's a scenario to help you think about it. You know how in soccer, the goal posts are fixed but the player kicks the ball differently each time? Now completely reverse that idea in your mind: pretend that the player does the same kick every time, but you get the job of moving the goal posts to different locations. In fact, let's say you get 100 tries at moving the goal posts around. A 95% confidence interval indicates that 95 out of 100 times, you moved the goal posts to a spot where the mystery kick went right through. The player always does the same kick: that is the unseen and mysterious population mean value that we can never exactly know. Each of the 100 times that you move the goal posts represents a new experiment and a new sample where the two posts are the two edges of the confidence interval calculated from that sample. You can create a nifty animation in R that helps to demonstrate this idea with the following code:

```
install.packages("animation")
library(animation)
conf.int(level=0.95)
```

The animation creates 50 different confidence intervals around a population mean of 0 by repeatedly sampling from the normal distribution. The dotted line in the middle of the graph shows where the population mean lies. Each constructed confidence interval looks like a tall capital letter "I." The circle in the middle of the "I" shows where the mean of each sample falls. In a few cases, the confidence interval will not overlap the population mean: these are marked in red. Most of the time these "goal posts" overlap the population mean, but in about 5% of the cases they do not. If for any reason this code did not work for you in R, try to search online for "confidence interval simulation" and you will find many animations you can view in a browser.

Now back to the mtcars transmission data: remember that when we say "95% confidence interval," we are referring to the proportion of constructed confidence intervals that would likely contain the true population value. So if we ran our transmission and fuel economy study 100 times, in about 95 of those replications the samples of transmission data would lead to the calculation of a confidence interval that overlapped the true mean difference in mpg. As well, about five of those 100 replications would give us a confidence interval that was either too high or too low—both ends of the confidence interval would either be above or below the population mean (just like the red-colored intervals in the animation described above). And for the typical situation where we only get the chance to do one study and draw one sample, we will never know if our particular confidence interval is one of the 95 or one of the five.

From the *t*-test, the span of –11.3 up to –3.2 is what statisticians call an **interval estimate** of the population value. The fact that it is a range of values and not just a single value helps to represent uncertainty: we do not know and can never know exactly what the population value is. The width of the confidence interval is important! A wide interval would suggest that there is quite a large span where the population mean difference may lie: in such cases we have high uncertainty about the population value. A narrow interval would signify that we have a pretty sharp idea of where the population mean difference is located: low uncertainty. We would want to keep that uncertainty in mind as we think about which kind of automobile transmission to choose, based on our preferences about fuel economy as well as other criteria such as cost.

Finally, I want to emphasize again the importance of the idea that the observed confidence interval does not tell us where the true population mean difference lies. The population mean difference *does not* lie at the center point of the confidence interval. The confidence interval we calculate from our sample of data may, in fact, not even contain the actual population mean difference. In keeping with our consideration of inferential thinking, the confidence interval adds to the weight of evidence about our beliefs. The span of –11.3 up to –3.2 strengthens the weight of evidence that the population difference in fuel economy between automatic and manual transmissions is a negative number somewhere in the region of –7.2 mpg plus or minus about 4 mpg. The confidence

interval does not *prove* that there is a difference in fuel economy between the two types of transmissions, but it does suggest that possibility, and it gives us a sense of the uncertainty of that conclusion (the plus or minus 4 mpg represents the uncertainty).

CONCLUSION

We began this chapter by constructing our own informal simulation model of a sampling distribution of mean differences by extrapolating from two samples of data about the fuel economy of 1974 cars. This was only an informal model because we did not have access to the complete populations of car data, but the simulation did give us the chance to put the idea of the uncertainty into a graphical context.

We then conducted a *t*-test, which calculated a confidence interval based upon the two samples of car data. The confidence interval suggested that manual transmissions from the 1974 model year might be more efficient than automatic transmissions, by somewhere in the neighborhood of 7 mpg. The interval estimate of the population mean difference ranged from –11.3 up to –3.2, a span of about 8 mpg with the observed mean difference between the samples –7.2, right in the center of that range. The width of the confidence interval, that is, the plus or minus 4 from the center point of –7.2, was an indication of our uncertainty. If the interval had been wider we would have been less certain. If the interval had been narrower, we would have been more certain.

Throughout this process, we firmly held in mind the idea that the meaning of a 95% confidence interval is that in 95 out of 100 study replications, we would be likely to find that whatever confidence interval we constructed in a given study did contain the actual population value—in this case a mean difference in fuel economy between two types of cars. Similar to the concepts we explored in the previous chapter, this is a statement about probabilities over the long run and not a statement about the particular confidence interval we constructed from this specific data set. This particular confidence interval may or may not contain the true population value: we will never know for sure.

EXERCISES

1. In your own words, write a definition of a 95% confidence interval.
2. Answer the following true/false questions about confidence intervals: (a) The center of the confidence interval is the population value; (b) The confidence interval always contains the population value; (c) A wider confidence interval is better, because it signals more certainty; (d) When we say "95% confidence interval," what we mean is that we are 95% certain that this particular confidence interval contains the population value.
3. Run the code shown in this chapter that created the animation of confidence intervals:

   ```
   install.packages("animation")
   library(animation)
   conf.int(level=0.95)
   ```

   Once the animation has finished running, comment on your results. Pay particular attention to the number of times that the confidence interval did not contain the population mean value (0). You may have gotten a different answer from other people who completed this exercise. Explain why this is so in your own words.
4. Some doctors conducted clinical trials on each of two new pain relievers. In the first trial, Drug A was compared to a placebo. In the second trial, Drug B was also compared to a placebo. In both trials, patients rated their pain relief such that a more negative number, such as –10, signified better pain relief than a less negative number, such as –5. As you may have already guessed, a rating of 0 meant that the patient's pain did not change, and a positive rating meant that pain actually increased after taking the drug (yikes!). After running the trials, the doctors calculated confidence intervals. Drug A had a confidence interval from –10 to –2 (these are mean differences from the placebo condition). Drug B had a confidence interval from –4 to +2 (again, mean differences from the placebo condition). Which drug is better at providing pain relief and why? Which drug gives us more certainty about the result and how do you know?
5. Assume the same conditions as for the previous question, but consider two new drugs, X and Y. When comparing Drug X to placebo, the confidence interval was –15 to +5. When comparing Drug Y to placebo, the confidence interval was –7 to –3. Which drug is better at providing pain relief and why? Which drug gives us more certainty about the result and how do you know?
6. Use the set.seed() command with the value of 5 to control randomization, and then calculate a confidence interval using the rnorm() command to generate two samples, like this:

   ```
   set.seed(5)
   t.test(rnorm(20,mean=100,sd=10),rnorm(20,mean=100,sd=10))
   ```

   The set.seed() function controls the sequencing of random numbers in R to help with the reproducibility of code that contains random elements. Review and interpret

the confidence interval output from that t.test() command. Keep in mind that the two rnorm() commands that generated the data were identical and therefore each lead to the creation of a sample representing a population with a mean of 100. Explain in your own words why the resulting confidence interval is or is not surprising.

7. The built-in PlantGrowth data set contains three different groups, each representing a different plant food diet (you may need to type data(PlantGrowth) to activate it). The group labeled "ctrl" is the control group, while the other two groups are each a different type of experimental treatment. Run the summary() command on PlantGrowth and explain the output. Create a histogram of the ctrl group. As a hint about R syntax, here is one way that you can access the ctrl group data:

```
PlantGrowth$weight[PlantGrowth$group==”ctrl”]
```

Also create histograms of the trt1 and trt2 groups. What can you say about the differences in the groups by looking at the histograms?

8. Create a boxplot of the plant growth data, using the model "weight ~ group." What can you say about the differences in the groups by looking at the boxplots for the different groups?

9. Run a *t*-test to compare the means of ctrl and trt1 in the PlantGrowth data. Report and interpret the confidence interval. Make sure to include a carefully worded statement about what the confidence interval implies with respect to the population mean difference between the ctrl and trt1 groups.

10. Run a *t*-test to compare the means of ctrl and trt2 in the PlantGrowth data. Report and interpret the confidence interval.

Guilford Publications
370 Seventh Avenue, Suite 1200
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com